ROMA TRE
UNIVERSITÀ DEGLI STUDI

Department of Sciences

Doctoral program in "BIOMEDICAL SCIENCES AND TECHNOLOGIES"

XXXI° Cycle

"Study of the ordered domains of Huntingtin"

Candidate
Valentina Brandi

Tutor
Prof. Fabio Polticelli

Coordinator
Prof. Paolo Visca

ACADEMIC YEAR 2017/2018

Contents

Abstract

Huntington's disease (HD) is a rare neurodegenerative and autosomal dominant disorder. The neuropathology of HD consists in the dysfunction and death of specific neurons within the brain. The most affected are the neurons of the striatum and the neurons that project from the cortex to the striatum. The symptoms vary between individuals but are usually characterized by a triad of motor, cognitive, and psychiatric symptoms. From its earliest stages, HD dramatically affects patients quality of life and prevents their functional abilities. Indeed, motor symptoms such as the choreiform movements and gait disturbances emerge early in the course of the disease while motor impairments occur in later stages. In many cases patients die because of fatal aspiration pneumonia. HD is caused by a mutation in the IT15 gene, located on chromosome 4. The result is the production of a mutant htt (mhtt) with an abnormally long polyglutamine (polyQ) repeat in the N-terminal that leads to htt aggregation into fibrils and a variety of other aggregates. Htt is a huge protein of 348 kDa and 3144 amino acids. The polyQ stretch, which begins at the 18th amino acid, is followed by a proline-rich domain (PRD). Downstream of the polyQ there are several HEAT repeats, sequences of 40 amino acids that consist of two antiparallel α-helices linked by a short loop. The HEAT repeats are packed together to form a flexible rod (denoted α-rod) and can act as a scaffold for diverse protein complexes and mediate inter- and intra-molecular interactions. Htt undergoes proteolysis at numerous sites, mostly found in disordered regions. It is subjected to various post-translational modifications, such as phosphorylation, acetylation, palmitoylation, ubiquitynation, and sumoylation. The protein is present predominantly in the cytoplasm of neurons and is enriched in compartments containing vesicle-associated proteins. Htt controls the transport of organelles, in axons and dendrites within neurons, in both the anterograde and retrograde directions. It is located not only in neurons but also at the spindle poles, mitotic spindles, and astral microtubules in dividing cells, where it regulates the cortical localization of proteins essential for mitotic spindle positioning. Further, htt mediates the transport of proteins required for ciliogenesis. It binds several proteins involved in endocytosis and that support membrane invagination and the assembly of the endocytosis vescicles coating mediated by clathrin. Htt may directly regulate autophagy and the transcription by the interaction with transcription factors and transcriptional regulators, and may also act as a transcriptional cofactor of itself. Different studies have shown that htt is essential for embryonic development and for the formation of the nervous system, for embryogenesis and development. Some studies indicate that htt is a regulator of tissue maintenance and that has a prosurvival role. Other experimental studies have highlighted a possible involvement of htt in the neuroprotective role mediated by the 17β-estradiol (E2), a sex steroid hormone that has a neurotrophic and neuroprotective role in several brain regions. E2 protective effects depend on the over-expression of neuroglobin (Ngb), a brain globin promoted by ERβ (estrogen receptor β). E2 stimulation induces the over-expression of both htt and Ngb, promoting the formation of a htt/Ngb complex, by direct interaction or mediated by other interactors, at the mitochondrial compartment. However, how Ngb

1

translocation to mitochondria occurs is completely unknown. Given a possible role of htt in vesicular trafficking, htt may be a putative Ngb carrier. Recently the structure of full-length human htt in a complex with HAP40 (Htt-associated protein 40) has been determined by cryo-electron microscopy but it lacks atomic details of several regions. Due to the lack of structural data, when this work began, the aim of this thesis has been that of obtaining information on the structure-function relationships of htt through a structural bioinformatics approach.

Through the use of several methods that predict order and disorder within a given protein, five htt regions predicted to fold into ordered domains have been detected. Structural models of htt ordered domains have been proposed for the first time, leading to the identification of a previously undetected HEAT repeats region in the third ordered domain (hunt3) and providing a putative function for four out of the five domains. Hunt3 has shown structural similarity with a Karyopherin, a protein involved in nuclear import by interacting with GTP-binding proteins. Surprisingly, the analysis of the htt interactome has revealed that it interacts with several GTP binding proteins. Molecular docking simulations have confirmed that Hunt3 can be considered the putative interaction site between htt and G proteins. Among htt interactors, the G protein $G\alpha_o1$ (encoded by the GNAO1 gene) is particularly interesting. In fact, different heterozygous mutations in the GNAO1 gene are the cause of a severe neurodevelopmental disorder, the epileptic encephalopathy type 17, which is characterized by symptoms similar to those observed in HD.

In addition, the analyses of htt and Ngb interactome revealed that $G\alpha_o1$ interacts both with htt and with Ngb. Molecular docking simulations indicated that a ternary complex between htt, $G\alpha_o1$ and Ngb is plausible, leading to the hypothesis that G proteins mediate the interaction between htt and Ngb, and possibly the transport of the latter protein to mitochondria.

The cryo-electron microscopy structure of htt evidences that the putative interaction site of hunt3 with the G proteins is occupied by an α-helix connected to the opposite side of the domain through disordered loop regions and stabilized by electrostatic interactions. Moreover, this α-helix contains two phosphorylation sites, whose function is not known, and a signature sequence shared with several G proteins. Molecular docking simulations have shown that when the α-helix is removed from the structure, G proteins can bind to htt by replacing the α-helix.

To validate the hypothesis that the α-helix could act as a molecular switch for the binding to the G proteins when phosphorylation occurs, 0.5 μs molecular dynamics simulations of the hunt3 domain in different phosphorylation states have been performed.

Preliminary results suggest that the phosphorylation of Ser1345 increases the flexibility of the entire hunt3 domain with respect to the wild type. An opposite effect is observed on the double phosphorylated domain. Thus, at variance to what has been hypothesized, only the phosphorylation on Ser might act as a positive regulator of htt binding to G proteins, while double phosphorylation would revert this effect.

However, it must be taken into account that replicas of all systems are currently in progress and that molecular dynamics results have been obtained on an isolated domain of htt, and it cannot be excluded that this region displays a different behaviour in the context of the entire protein. Therefore, the outcome of this work is currently being validated experimentally by recombinant expression of hunt3 domain in *E. coli* and proteomic analysis of its interactome.

Another line of research of this work has been the study of htt sequences and structures of less complex model organisms. The htt sequences of *C. intestinalis* and *B. floridae* have been analysed, identifying ordered domains and performing structure prediction. All the models have shown structural similarity with the human ones. Finally, an htt-like protein, not previously identified, has been detected for the first time in *C. elegans*.

Riassunto

La malattia di Huntington (HD) è una malattia neurodegenerativa a carattere autosomico dominante. La neuropatologia consiste nella morte dei neuroni dello striato e dei neuroni che proiettano dalla corteccia allo striato. La malattia determina un progressivo cambiamento della personalità, disfunzioni motorie e demenza, e può svilupparsi in adulti e bambini. Fin dalle prime fasi la malattia compromette la qualità di vita dei pazienti impedendo le loro capacità funzionali. Infatti, all'inizio è associata a progressivi disturbi emotivi, psichiatrici e cognitivi mentre negli ultimi stadi è caratterizzata da sintomi motori, demenza progressiva o una graduale perdita dei processi mentali coinvolti nella comprensione, nel ragionamento, nel giudizio e nella memoria. La maggior parte dei pazienti alla fine soccombe a causa di polmonite da aspirazione per difficoltà nella deglutizione.

La malattia è causata da una mutazione nel gene IT15 localizzato nel braccio corto del cromosoma 4 che codifica per una proteina chiamata huntingtina (htt). La mutazione consiste nell'espansione della ripetizione del codone CAG, all'interno del gene che codifica per l'htt. Il codone CAG codifica per l'amminoacido glutammina, perciò l'espansione comporta un lungo tratto anomalo di glutammine responsabile dell'aggregazione dell'htt in fibrille. La proteina ha un peso molecolare di 348 kDa e 3144 amminoacidi. La regione N-terminale include il tratto di poliglutammine che è seguito da un dominio ricco di proline (PRD). A valle del tratto di poliglutammine ci sono diverse ripetizioni chiamate HEAT, sequenze di 40 amminoacidi costituite da due α-eliche unite da un corto loop e disposte a formare un bastoncello flessibile denominato α-rod. I domini HEAT possono agire come impalcatura per la formazione di complessi proteici e mediare interazioni intra- e inter-molecolari. L'htt è soggetta a proteolisi in numerosi siti, la maggior parte dei quali si trovano nelle regioni disordinate. Inoltre è soggetta a varie modificazioni post-traduzionali, tra cui fosforilazione, acetilazione, palmitoilazione, ubiquitinazione e sumoilazione. L'htt si trova principalmente nel citoplasma dei neuroni e nei compartimenti contenenti proteine associate a vescicole. È infatti coinvolta nel controllo del trasporto di organelli, sia retrogrado che anterogrado, negli assoni e dendriti. L'htt è localizzata non solo nei neuroni ma anche ai poli del fuso, nel fuso mitotico e nei microtubuli astrali delle cellule in divisione dove regola la localizzazione corticale di proteine essenziali per il posizionamento del fuso mitotico. L'htt media anche il trasporto di proteine, necessarie per la ciliogenesi, alla matrice pericentriolare,

L'htt agisce da impalcatura per l'interazione di proteine coinvolte nell'endocitosi, regolando le prime fasi di rivestimento e invaginazione della membrana. Può regolare direttamente l'autofagia e la trascrizione attraverso l'interazione con attivatori trascrizionali e repressori, e può anche agire da fattore trascrizionale di se stessa. Differenti studi hanno mostrato che l'htt è essenziale per lo sviluppo

embrionale, per la formazione del sistema nervoso per l'embriogenesi e lo sviluppo. Alcuni studi indicano che è anche un regolatore dell'omeostasi tissutale e che ha un ruolo antiapoptotico. Altri studi hanno evidenziato un possibile coinvolgimento dell'htt nel ruolo neuroprotettivo mediato dal 17-β estradiolo (E2). Gli effetti protettivi di quest'ultimo dipendono dalla sovraespressione della neuroglobina (Ngb), una globina del sistema nervoso regolata dal recettore β degli estrogeni. La stimolazione di E2 induce la sovraespressione sia dell'htt che della Ngb, promuovendo la formazione del complesso htt/Ngb nei mitocondri, dove la Ngb lega il citrocromo c, impedendone il rilascio nel citosol e la conseguente attivazione dell'apoptosi. Tuttavia, come avvenga la traslocazione della Ngb ai mitocondri non è ancora noto. Dato un possibile ruolo dell'htt nel traffico vescicolare, questa potrebbe essere un buon candidato al trasporto della Ngb ai mitocondri.

L'elevato peso molecolare e le grandi dimensioni dell'htt impediscono la cristallizzazione e studi di diffrazione a raggi X. Solo recentemente la struttura è stata determinata in complesso con la proteina HAP40 attraverso crio-microscopia elettronica, anche se ci sono diverse regioni che mancano di dettagli atomici. A causa dell'assenza di dati strutturali sull'htt quando questo lavoro è iniziato, il principale scopo è stato quello di ottenere maggiori informazioni sulla struttura e funzione di questa proteina, utilizzando un approccio di tipo bioinformatico.

Attraverso l'uso di diversi metodi che predicono l'ordine e il disordine strutturale di una proteina, sono state identificate 5 regioni dell'htt predette formare domini ordinati.

Per la prima volta sono stati proposti modelli strutturali di questi domini, portando all'identificazione di una nuova regione contenente ripetizioni HEAT nel terzo dominio ordinato (hunt3). Una presunta funzione è stata predetta soltanto per 4 domini dell'htt. Hunt3 ha mostrato somiglianza strutturale con una carioferina, coinvolta nell'importo nucleare di proteine cargo attraverso l'interazione con proteine che legano il GTP (proteine G). Sorprendentemente, l'analisi degli interattori dell'htt ha rivelato che quest'ultima interagisce con proteine G. Simulazioni di docking molecolare hanno mostrato che hunt3 può essere il possibile sito di interazione tra l'htt e le proteine G. Tra gli interattori dell'htt, la proteina $G\alpha_o1$ (codificata dal gene GNAO1), sembra essere particolarmente interessante. Infatti differenti mutazioni in eterozigosi nel gene GNAO1 causano una patologia caratterizzata da sintomi simili a quelli dell'HD. Da un confronto dell'interattoma dell'htt e della Ngb è emerso che la proteina $G\alpha_o1$ interagisce sia con l'htt che con la Ngb. I risultati delle simulazioni di docking molecolare indicano che la formazione di un complesso ternario tra htt, $G\alpha_o1$ e Ngb è plausibile, facendo ipotizzare che le proteine G possano mediare l'interazione tra htt e Ngb, e probabilmente il trasporto di quest'ultima ai mitocondri.

La struttura dell'htt evidenzia che in hunt3, il sito di interazione predetto con le proteine G è occupato da un'α-elica, connessa al lato opposto del dominio da due loop disordinati e stabilizzata da

interazioni elettrostatiche. Inoltre l'α-elica contiene due siti di fosforilazione, la cui funzione non è ancora chiara, e alcuni residui che sono conservati in diverse proteine G. Per validare l'ipotesi che l'α-elica possa agire da interruttore molecolare per l'interazione con le proteine G come conseguenza della fosforilazione, sono state eseguite simulazioni di dinamica molecolare (ciascuna di 0.5 μs) di hunt3 in differenti stati di fosforilazione. Risultati preliminari indicano che la fosforilazione del residuo S1345 aumenterebbe la flessibilità dell'intero dominio hunt3, mentre un effetto inverso si osserva quando il dominio è fosforilato su entrambi i siti. Contrariamente a quanto era stato ipotizzato, solo la fosforilazione sul residuo S1345 agirebbe da regolatore positivo del legame tra l'htt e le proteine G. Tuttavia bisogna considerare che le repliche delle simulazioni di tutti i sistemi analizzati sono attualmente in corso e che le simulazioni sono state eseguite unicamente su un dominio isolato della proteina. Non si può quindi escludere che tale dominio possa avere un comportamento diverso nel contesto dell'intera proteina. Pertanto il risultato di questo lavoro è attualmente in corso di validazione attraverso l'espressione ricombinante del dominio hunt3 in *E. coli* ed analisi proteomica dell'interattoma.

Un'altra linea di ricerca ha riguardato lo studio della sequenza dell'htt in alcuni organismi modello quali *C. intestinalis* e *B. floridae*. Infine una proteina simile all'htt è stata identificata per la prima volta in *C. elegans*.

Chapter 1: Introduction

1.1 Huntington Disease

The first description of Huntington's disease (HD) as an epidemic dancing mania dates back to the 14th century. Later Paracelsus assigned the name *chorea* to this movement disorder, denoting its cause in the central nervous system (CNS). In the 17th century HD was named "that disorder" or "San Vitus" dance. It was believed that people with chorea, involuntary muscle twitches typical of HD, were possessed by the devil (Zuccato, Valenza, and Cattaneo, 2010). In 1872 an American doctor, George Huntington published in the "Medical and Surgical Reporter of Philadelphia" journal the first paper on HD, called "*On Chorea*", in which he described in detail the disease. Only in the 20th century researchers determined the caudate nucleus as the central target of brain cell death in HD (Zuccato *et al.*, 2010).

HD is a rare neurodegenerative disorder with a prevalence of 5–10 individuals per 100.000 in the Caucasian population. The symptoms vary between individuals but are usually characterized by a triad of motor, cognitive, and psychiatric symptoms (Saudou and Humbert, 2016). From its earliest stages, HD dramatically affects patients quality of life and prevents their functional abilities. Indeed, motor symptoms such as the choreiform movements and gait disturbances emerge early in the course of the disease while motor impairments occur in later stages. The deficits consist of cognitive slowing and decreases in both attention and mental flexibility (Saudou and Humbert, 2016). The neuropathology of HD consists in the dysfunction and death of specific neurons within the brain. The most affected are the neurons of the striatum and the neurons that project from the cortex to the striatum. Moreover, patients show metabolic and immune disturbances, skeletal-muscle wasting, weight loss, cardiac failure, testicular atrophy, and osteoporosis. The mean age of onset is 35 to 44 years and the median survival time is 15 to 18 years after onset. In many cases patients die because of fatal aspiration pneumonia (Saudou and Humbert, 2016).

The gene responsible for the disease and its associated mutation were identified in 1993 by an international team of 58 scientists called the "Huntington Disease Collaborative Research Group". They found that HD was linked to IT15 gene, located on chromosome 4, and was caused by an autosomal dominantly inherited expansion of CAG trinucleotide repeat on the first exon of the protein called huntingtin (htt) (HD Collaborative research group, 1993). The result is the production of a mutant htt (mhtt) with an abnormally long polyglutamine (polyQ) repeat that leads to htt aggregation

into fibrils and a variety of other aggregates (Poirier *et al.*, 2002; Legleiter *et al.*, 2010; McColgan and Tabrizi, 2018) . In healthy individual the CAG sequence is repeated 9 to 35 times, while CAG expansion exceeding 35 repeats results in HD (Saudou and Humbert, 2016).

1.2 Huntingtin

Htt is a huge protein of 348 kDa and 3144 amino acids. The protein is well conserved from flies to mammals, with the highest identity being found between mammals. Instead the polyQ tail within its exon 1 is not conserved. This suggests that the polyQ tail of htt may have a role in the precise modulation of its functions as a result of recent evolutionary achievements (Tartari *et al.*, 2008). The N-terminal region of the protein includes the polyQ stretch, which begins at the 18th amino acid and is followed by a proline-rich domain (PRD) found only in mammals, suggesting recent evolution of htt protein (Tartari *et al.*, 2008; Saudou and Humbert, 2016). The 17 residues at the N-terminal end (Nt17) are composed of an amphipathic α-helix that contains a nuclear export signal (NES) (Atwal *et al.*, 2007) and undergoes post-translation modifications that alter htt clearance and its subcellular localization (Thompson *et al.*, 2009; Steffan *et al.*, 2004; Atwal *et al.*, 2007; Maiuri *et al.*, 2013). The PRD is made up of a rigid helix with a kinked or straight conformation, which may be significant for stabilizing the structure of the polyQ stretch (Saudou and Humbert, 2016).

Although the polyQ length is the cause of htt aggregation and toxicity, the Nt17 influences aggregation, aggregates stability, other important biochemical properties of the protein, and finally its role in pathogenesis (Arndt *et al.*, 2015). Indeed, Nt17 has been involved in the first phase of htt intermolecular interaction that leads to the formation of small, α-helix rich oligomers (Arndt *et al.*, 2015). Instead the polyP domain interacts with vesicle trafficking proteins (HIP1, SH3GL3, and dynamin) that are seized into htt inclusion bodies, causing neuronal dysfunction (Arndt *et al.*, 2015). Downstream of the polyQ, between positions 69 and 3144 there are several HEAT repeats. They are sequences of 40 amino acids, initially found in **H**tt, **E**longation factor 3, protein phosphatase 2**A**, and **T**OR1 and consist of two antiparallel α-helices linked by a short loop. The heat repeats are packed together to form a flexible rod (denoted α-rod). (Takano and Gusella, 2002; Tartari *et al.*, 2008; Warby *et al.*, 2008; Palidwor *et al.*, 2009). The HEAT repeat domains can act as a scaffold for diverse protein complexes and mediate inter- and intra-molecular interactions. The central region of htt (507–1,230) can bind to the N-terminal (1–506) and C-terminal (2,721–3,144) domains of htt itself (Palidwor *et al.*, 2009). Further, the middle domain can self-associate to form homodimers of htt. In addition, N-terminal parts of htt (1–416 and 1–586) bind to different C-terminal regions of htt (1,725–2,800 and 2,416–3,144, respectively), and these intramolecular interactions are disrupted upon

proteolysis (Ochaba *et al.*, 2014; El-Daher *et al.*, 2015). This suggests that htt can adopt various three-dimensional conformations and involve other protein complexes, as htt interacts with a number of different partners. Indeed, purified htt can adopt up to 100 structurally distinguishable conformations (Seong *et al.*, 2009).

1.3 Huntingtin through the evolution

The comparison of htt homologs can help to understand if there is a functional conservation of domains in the evolution of mammals. In the vertebrate subphylum htt homologs are highly conserved (>80% sequence identity) (Cattaneo *et al.*, 2005). Instead the only entirely known amino acid sequence among invertebrates is the one from *Drosophila melanogaster*, in the protostome branch. It is characterized by an additional region and five conserved regions (20–50% sequence identity with the human protein) distributed throughout the length of the protein. These regions may represent a remnant of the ancestral htt at the origin of the protostome-deuterostoma branches (**Figure 1**).
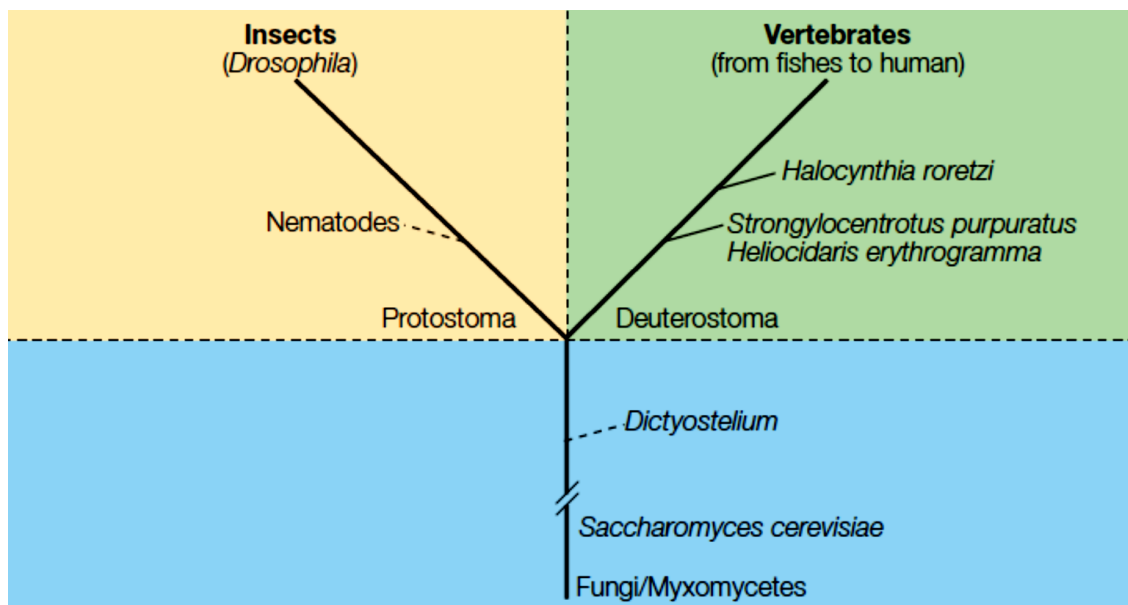
Figure 1. Huntingtin through the evolution. Htt homologs in the vertebrate subphylum are highly conserved, whereas *Drosophila melanogaster* htt is characterized by an additional region and by five 20–50% conserved regions. These may represent a remnant of the ancestral htt at the origin of the protostoma–deuterostoma branches (yellow–green). Htt seems to be present in *Dictyostelium discoideum* but not in the evolutionary older *Saccharomyces cerevisiae* (fungi, light blue) or in previously diverged plants, which suggests that htt is an evolutionarily old gene that has been lost in some animals (for example, *Caenorhabditis elegans*) (Cattaneo *et al.*, 2005).

Htt is present in an lower complexity deuterostome, the tunicate *Halocynthia roretzi* (sea pineapple), and in the echinoderm *Heliocidaris herithrogramma* (sea urchin), but *Dictyostelium discoideum*, an amoeba, was the first organism known to carry the gene in a form that is slightly different from the

human version (**Figure 2**) (Cattaneo *et al.*, 2005). Also, the amoeba's gene differs from the human one because it doesn't possess CAG triplets (**Figure 2**) (Palidwor *et al.*, 2009). Miguel Andrade and colleagues have predicted the presence of htt in the nematode *Caenorhabdtis elegans*, but not in *Saccharomyces cerevisiae*, thus confirming that HTT is, evolutionarily, an old gene (Palidwor *et al.*, 2009).

HTT starts to acquire CAG triplets in a category of basal deuterostomes called echinoderms (such as the sea urchin *Strongylocentrotus purpuratus*), where two CAG triplets in the initial part of the gene have been identified (Tartari *et al.*, 2008) (**Figure 2**). Notwithstanding the presence of a primitive nervous system in sea urchins, the gene is present mainly in non-neural tissues, indicating that early on in evolution the gene with its two CAG triplets did not have a crucial role in the nervous system. Research on the triplets in protostomes is not completed yet, but it is clear that they are uncommon (Zuccato and Cattaneo, 2016). The analyses of the DNA sequences in the HTT deuterostomes revealed, as for sea urchins, that two CAG triplets occur in the sequence of the amphioxus, of the Cephalochordata family (Zuccato and Cattaneo, 2016). However, in amphioxus, the nucleotide sequence around the triplet pair is similar to that in vertebrates, including humans, and the protein encoded by the gene is largely confined to neural tissues. This may have contributed to the formation of the primitive brain. In fact, the CAG triplets tend to increase in organisms with a more developed nervous systems, until they reach their maximum extension in humans (Zuccato and Cattaneo, 2016).
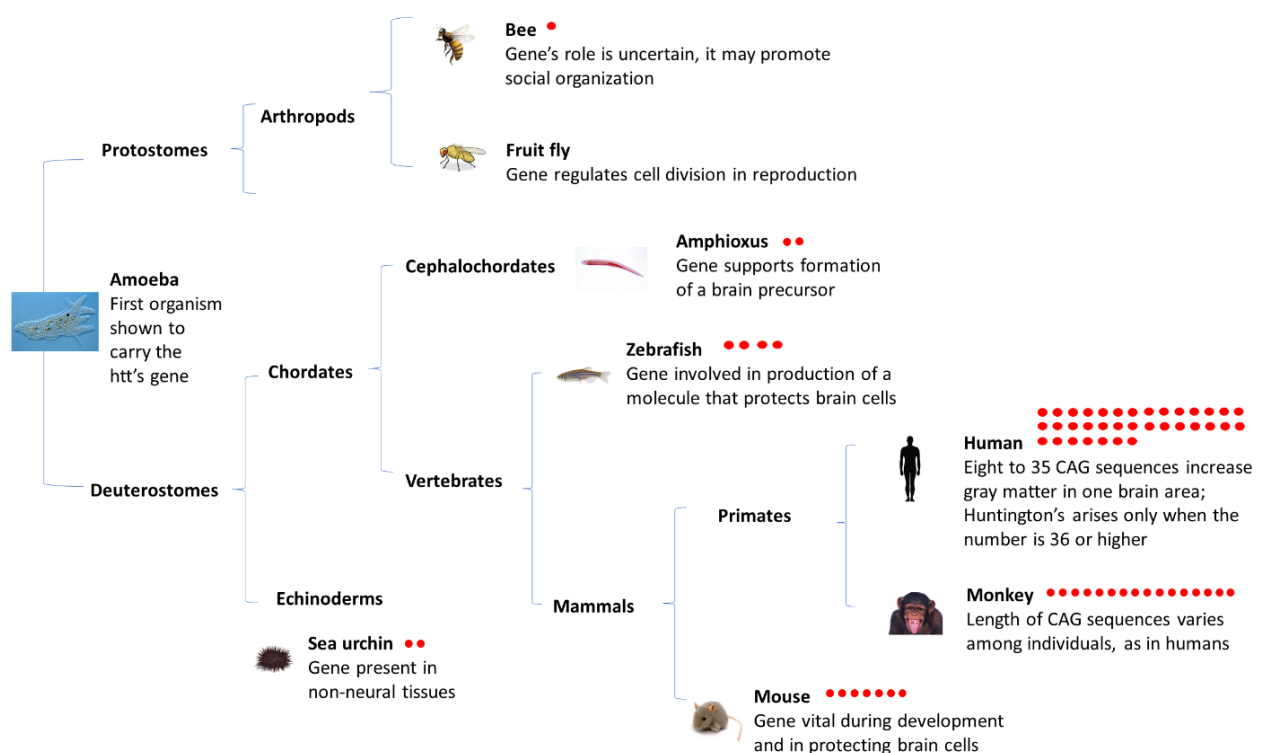
Figure 2. Evolution of HTT gene and CAG triplets.
Dictyostelium discoideum amoebas were the most ancient organisms to carry the HTT gene, but without CAG triplets. The gene regulates a number of vital cellular processes, including the transition of *Dictyostelium* to its multicellular stage. The amoeba preceded the division of the tree of life into its two branches more than 550 million years ago: the protostomes, which include insects, crustaceans, mollusks, and the deuterostomes, which led to the first vertebrates, the fishes, birds, amphibians, reptiles, mammals, primates and modern human beings. Only the deuterostomes went on to accumulate CAG triplets at the same place in which, in the human gene, occurs the mutation that causes HD. HTT gene starts to accumulate CAG triplets in the echinoderms, where two triplets have been found. Also, two triplets occur in the sequence of amphioxus. CAG triplets begin to lengthen appreciably in organisms with more sophisticated nervous systems. The number of red dots indicates the number of CAG triplets (modified from Zuccato and Cattaneo, 2016).

1.4 Huntingtin proteolysis

Htt undergoes proteolysis at numerous sites by a variety of proteases. These proteolytic sites are located in PEST domains, rich in proline (**P**), glutamic acid (**E**) or aspartic acid (**D**), serine (**S**), and threonine (**T**) amino acids, mostly found in disordered regions (Warby *et al.*, 2008). Several caspases, calpain, cathepsins, and the metalloproteinase MMP10 have been reported to cleave htt (Gafni and Ellerby, 2002; Hermel *et al.*, 2004; Kim *et al.*, 2001; Lunkes *et al.*, 2002; Miller *et al.*, 2011; Ratovitski *et al.*, 2009; Tebbenkamp *et al.*, 2012; Kim *et al.*, 2006; Goldberg *et al.*, 1996). Both the wild type and mutated htt are good substrates for cleavage in vitro (Goldberg *et al.*, 1996). Moreover, in HD an increase of the activity of these proteases occurs in brains of patients. This proteolysis leads to the generation and accumulation of small N-terminal fragments containing the polyQ stretch, which are toxic (Benn *et al.*, 2005; Graham *et al.*, 2006; Saudou *et al.*, 1998).

It has been reported that proteolysis of wild-type htt may turn off some of its functions (El-Daher *et al.*, 2015). Indeed, the proteolysis of wild-type full-length htt causes the toxicity of the protein. The C-terminal htt fragment is toxic in the cytoplasmic compartment by inactivating dynamin 1 at ER membranes, leading to ER stress and cell death (El-Daher *et al.*, 2015). Proteolysis may occur in different disease conditions, in which massive cell death and/or caspases are activated or may be associated to developmental apoptosis. Thus, proteolysis of full-length htt could transform the neuroprotective function of wild-type htt into a toxic one (Saudou and Humbert, 2016).

1.5 Post translational modifications

Htt is subjected to various post-translational modifications, such as phosphorylation, acetylation, palmitoylation, ubiquitynation, and sumoylation (Saudou and Humbert, 2016).

The function of several modifications is not known, but some of these can have a therapeutic significance since they regulate mutant htt toxicity (Saudou and Humbert, 2016). Wild-type htt is less acetylated than the mutated protein, in which acetylation mediates its clearance

by the autophagic lysosomal pathway (Jeong *et al.*, 2009). Phosphorylation of mutant htt at Ser434 or Ser536 reduces htt proteolysis by caspases 3 and calpain and causes a reduction of polyQ-htt toxicity (Luo *et al.*, 2005; Schilling *et al.*, 2006). However, if these mechanisms are crucial for the regulation of htt functions is unknown. Moreover, phosphorylation at Ser13 and Ser16 favours clearance of both wild-type and mutant htt and decreases its toxicity (Thompson *et al.*, 2009). Phosphorylation at Ser13 and Ser16 may be essential for htt nuclear localization, because the N-terminal domain sequence L4F11S16 (leucine 4, phenylalanine 11, serine 16) functions as a nuclear export signal (NES, as mentioned before) and a phospho-mimicking mutant at Ser16 causes the block of the CRM1-dependent nuclear export of htt, increasing its nuclear localization (Maiuri *et al.*, 2013). Phosphorylation/dephosphorylation at Ser421 and at Ser1181/Ser1201 modulates microtubules-dependent intracellular transport of organelles. Htt phosphorylation at Ser1181/Ser1201 raises both anterograde and retrograde transport, while phosphorylation at Ser421 promotes anterograde transport (Ben M'Barek *et al.*, 2013; Colin *et al.*, 2008; Humbert *et al.*, 2002).

Htt interacts with two palmitoyl-acyl transferases (PATs), HIP14 and HIP14L (DHHC17 and DHHC13, respectively) (Yanai *et al.*, 2006). Htt is itself palmitoylated at Cys214, and the expansion of the polyQ tract of htt decreases the interaction between mhtt and HIP14 and consequently a noteworthy reduction in palmitoylation (Yanai *et al.*, 2006; Huang *et al.*, 2011). Moreover, htt acts as a modulator of HIP14 activity such that in HD condition, HIP14 is less active. Thus the palmitoylation of others HIP14 substrates decreases, leading to neuronal toxicity (Sanders and Hayden, 2015).

Regarding sumoylation, this modification modulates htt subcellular localization, activity, and stability (Steffan *et al.*, 2004).

Htt is ubiquitinated at the N-terminal lysines K6, K9, and K15 and targeted to the proteasome (Kalchman *et al.*, 1996; Anderson *et al.*, 1997). Ubiquitination controls the stability, function, and intracellular localization of htt, thus contributing to maintain htt homeostasis in the cell. Htt mutation impairs this process, causing proteosomal dysfunction and accumulation of htt fragments that become toxic to neurons (Zuccato *et al.*, 2010).

1.6 Huntingtin functions

Htt has been involved in several cellular events during development and adulthood, mostly in the nervous system (Saudou and Humbert, 2016).

1.6.1. Vesicles transport

Htt is present predominantly in the cytoplasm of neurons and is enriched in compartments containing vesicle-associated proteins (Velier *et al.*, 1998). Htt controls the transport of organelles, in axons and

dendrites within neurons, in both the anterograde and retrograde directions, by interacting either directly with dynein or indirectly through the Huntingtin-associated protein 1 (HAP1) (Engelender, 1997; Li *et al.*, 1998) (**Figure 3**). The latter functions as an adaptor protein, interacting with the p150^{Glued} subunit of dynactin and the kinesin 1 member KIF5C (Caviston *et al.*, 2007; Engelender, 1997; Li *et al.*, 1998; McGuire *et al.*, 2006; Twelvetrees *et al.*, 2010).
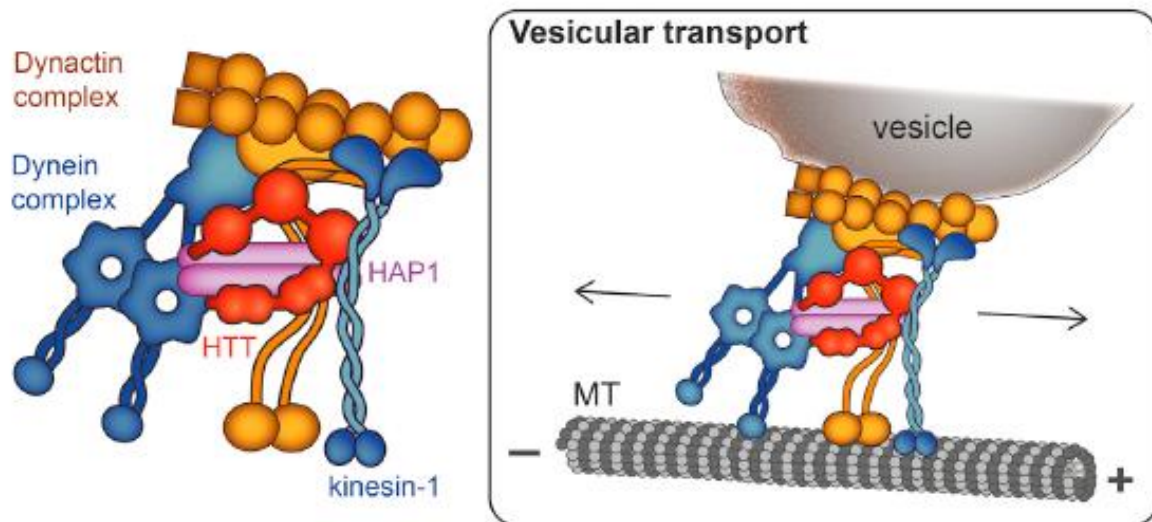
Figure 3. Htt controls the transport of organelles, both anterograde and retrograde, by the direct and indirect interaction with dynein and dynactin, respectively (Saudou and Humbert, 2016).

The transport processes mediated by htt include precursor vesicles, vesicles that contain the v-SNARE VAMP7 protein, autophagosomes, endosomes and lysosomes, brain-derived neurotrophic factor (BDNF)-containing vesicles, APP (amyloid precursor protein)- positive vesicles, and GABA-receptor-containing vesicles (Colin *et al.*, 2008; Wong and Holzbaur, 2014; Caviston *et al.*, 2011; Liot *et al.*, 2013; Gauthier *et al.*, 2004; Her and Goldstein, 2008; Twelvetrees *et al.*, 2010).

Further, the interaction between htt, HAP1, dynein and kinesin is the main mechanism in trafficking of the mitochondria in neurons and mhtt is able to alter mitochondrial homeostasis and dynamics (Jodeiri Farshbaf and Ghaedi, 2017; Caviston and Holzabaur, 2009). Htt may improve the speed of the axonal transport of vesicles, by the binding and retention of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) on vesicles, that supplies energy for the transport (Zala *et al.*, 2013).

As already mentioned in section 1.5, htt can coordinate the directionality of transport through its phosphorylation at Ser421 (Colin *et al.*, 2008) mediated by the kinases Akt/PKB and the serum- and glucocorticoid-induced kinase SGK (Humbert *et al.*, 2002; Rangone *et al.*, 2004). Instead, the de-phosphorylation at Ser421 occurs upon calcineurin activation (Pardo *et al.*, 2006).

1.6.2 Cell Division

Htt is located not only in neurons but also at the spindle poles, mitotic spindles, and astral microtubules in dividing cells (Elias *et al.*, 2014; Godin *et al.*, 2010; Gutekunst *et al.*, 1995). During mitosis, htt is headed to the spindle poles through dynein and induces the accumulation of NUMA (nuclear mitotic apparatus) and LGN (G protein regulator leucine-glycine-asparagine repeat) proteins (Elias *et al.*, 2014; Godin *et al.*, 2010) (**Figure 4**).
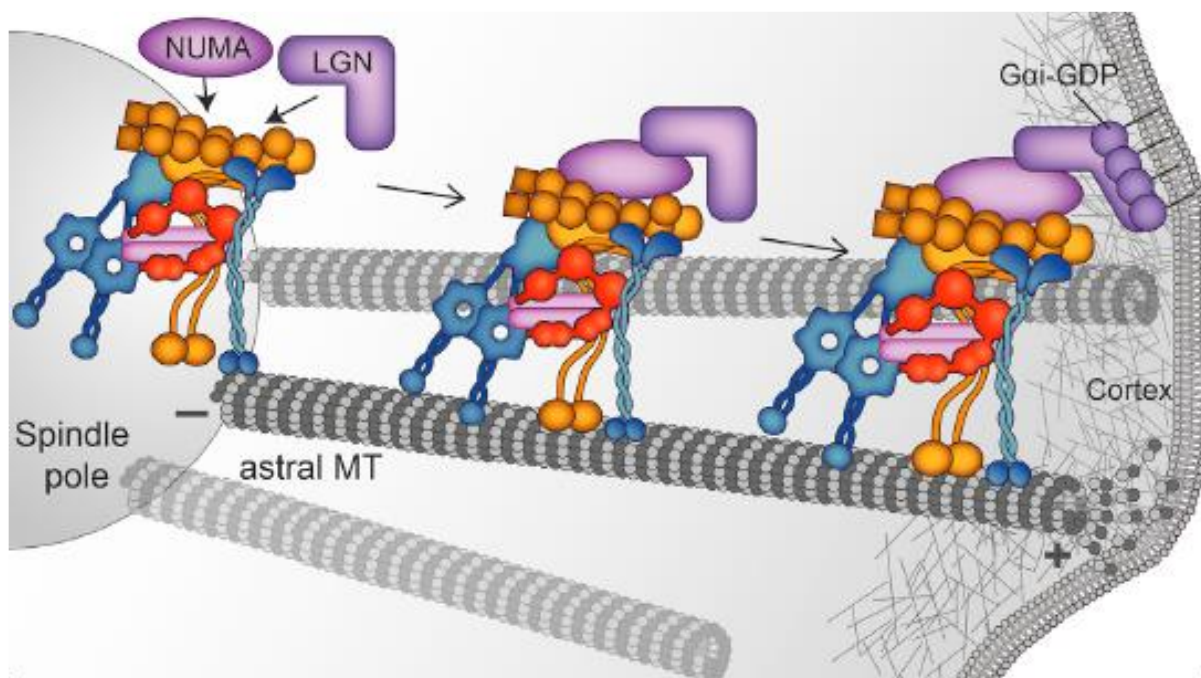
Figure 4. Htt is important for spindle pole assembly and regulates also the kinesin 1-dependent trafficking of dynein/dynactin/NUMA/LGN to the cell cortex (Saudou and Humbert, 2016).

Moreover htt regulates the cortical localization of dynein, dynactin, NUMA, and LGN, essential for mitotic spindle positioning (Elias *et al.*, 2014).

1.6.3 Ciliogenesis

Htt has been detected at the base of the cilia in primary cilia in neurons, photoreceptor cilia, and cilia in multiciliated cells (Haremaki *et al.*, 2015; Karam *et al.*, 2015; Keryer *et al.*, 2011). Htt interacts with HAP1 and pericentriolar material 1 protein (PCM1) at the centrosome in interphasic cells in a microtubule-dependent manner (Keryer *et al.*, 2011). The transport of PCM1 to the pericentriolar material, mediated by htt, is crucial for ciliogenesis (**Figure 5**). Indeed, the absence of htt in mouse cells threatens the retrograde trafficking of PCM1 and prevents the formation of the primary cilium.

The phosphorylation of the NES motif present on the N-terminal domain could mediate the localization of htt in the axomeme (Maiuri *et al.*, 2013).
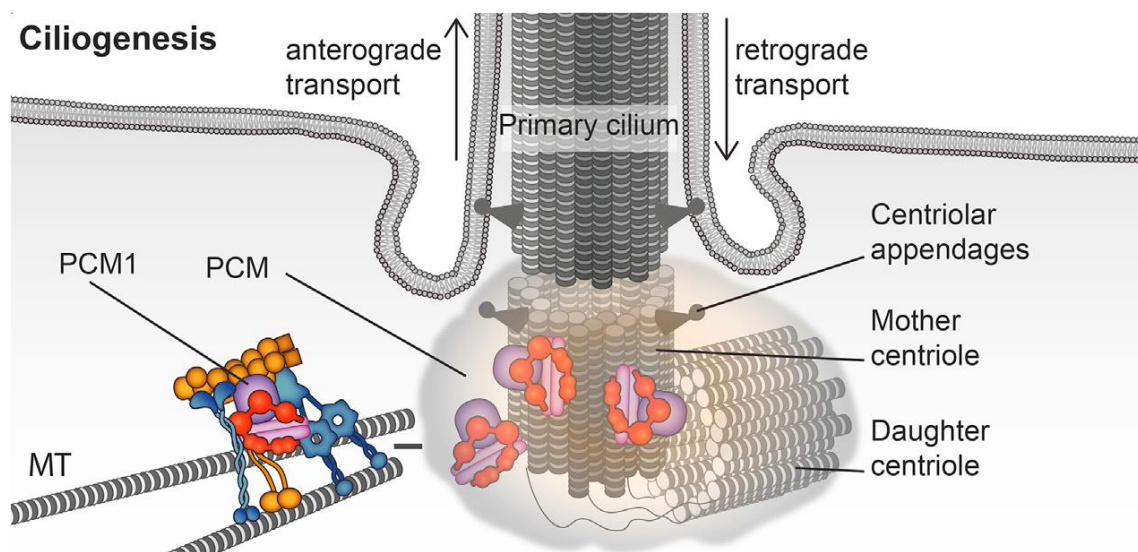
Figure 5. Htt mediates the dynein/dynactin/HAP1-dependent transport of proteins to the pericentriolar material (PCM), including PCM1 protein that is required for ciliogenesis (Saudou and Humbert, 2016).

1.6.4 Endocytosis, Vesicle Recycling, and Endosomal Trafficking

Htt binds several proteins involved in endocytosis, including htt-interacting protein 1 (HIP1) and HIP1R (HIP12), that support membrane invagination and the assembly of the coating endocytosis mediated by clathrin (Engqvist-Goldstein *et al.*, 2001; Legendre-Guillemin *et al.*, 2002; Waelter *et al.*, 2001). Htt interacts with dynamin 1 through both its N- and C-terminal domains, acting as a scaffold and regulating the first step of endocytosis though membrane coating and invagination (El-Daher *et al.*, 2015; Kaltenbach *et al.*, 2007; Moreira Sousa *et al.*, 2013). Htt may participate in a complex containing SH3GL3/endophilin-A3, endophilin-B1, amphiphysin, and dynamin, acting during both endocytosis and vesicle recycling (Sittler *et al.*, 1998; Modregger *et al.*, 2003). Moreover, htt interacts with and activates Rab11, a GTPase that regulates intracellular trafficking, coordinating apical translocation of PAR3-PAR6-aPKC complex, which regulates apical polarity in several tissues, including the mammary epithelium (Li *et al.*, 2008; Elias *et al.*, 2015). Another key regulator of endocytosis is HAP40 (htt-associated protein 40). A complex between htt and HAP40 can be recruited onto early endosomes by directly interacting with Rab5, reducing endosome motility and regulating the switch from microtubules to actin filaments (Pal *et al.*, 2006). In addition, htt interaction with the Rab8/optineurin/myosinVI complex (Hattula and Peränen, 2000; Faber *et al.*, 1998; Sahlender *et al.*, 2005) and the HAP1/dynactin complex (Engelender *et al*., 1997; Gauthier *et*

al., 2004; Li *et al.*, 1998) may mediate transition of cargo between microtubules and actin cytoskeletons.

1.6.5 Autophagy

Htt may directly regulate autophagy by several systems. As mentioned before, htt acts as scaffold for the formation of the dynein/dynactin/HAP1 complex, regulating the retrograde transport of autophagosomes along axons (Wong and Holzbaur, 2014). Mhtt disturbs the trafficking from Golgi to lysosomal compartments by delocalizing the optineurin/Rab8 complex, which, in turn, impairs the lysosomal function (del Toro *et al.*, 2009). Also, htt sequence contains 11 LIR (LC3-interacting repeats) motifs found in Atg8 family-interacting proteins, p62 and optineurin, which play a critical role in autophagy because ensures the targeting of autophagy receptors to LC3 anchored in the phagophore membrane (Ochaba *et al.*, 2014). The interaction between htt and p62 may facilitate the recognition of ubiquitinylated proteins at K63 by p62 and the loading of the cargo into the autophagosomes (Ochaba *et al.*, 2014). When htt binds to ULK1 (Unc-51-like kinase 1, also known as ATG1), the latter is set free from its interaction with the protein kinase mTOR (mammalian **T**arget **O**f **R**apamycin), causing authophagy. Furthermore, the C-terminal fragments, produced as a result of the proteolysis, may dysregulate autophagy (Ochaba *et al.*, 2014).

1.6.6 Transcription

Htt is a cytoplasmic protein, but it is also present in the nucleus. The polyQ motif is generally found in transcription factors, acting as transcriptional regulating domains and mediating the binding between transcription factors and transcriptional regulators. Among htt interactors there are numerous transcription factors, including the cAMP-response element (CREB)-binding protein (CBP), NeuroD, the specificity protein-1 (SP1), the nuclear factor-kB (NF-kB), and the tumor suppressor protein 53 (p53) (Steffan *et al.*, 2000; Marcora *et al.*, 2003; Li *et al.*, 2002; Takano and Gusella, 2002; Dunah *et al.*, 2002). In addition, htt interacts with transcriptional activators and repressors: the Gln-Ala repeat transcriptional activator CA150, the co-activator TAFII130, the nuclear co-repressor (NCOR), the repressor element-1 transcription factor/neuron restrictive silencer factor (REST/NRSF), and the transcriptional corepressor C-terminal-binding protein (CtBP) (Kegel *et al.*, 2002; Holbert *et al.*, 2001; Dunah *et al.*, 2002; Yohrling IV *et al.*, 2003; Zuccato *et al.*, 2003). Lastly, htt binds nuclear receptors, including LXRa, peroxisome proliferator-activated receptor-γ (PPARγ), vitamin D receptor (VDR), and thyroid hormone receptor a1 (TRa1) (Futter *et al.*, 2009). Through these interactions htt can modulate genes transcription and, consequently, the processes regulated by gene transcription. Upon the interaction with p53, htt influences the transcription of p53 target genes that are involved in cell-cycle control, apoptosis, cellular stress responses, and DNA repair (Bae *et al.*,

2005). Furthermore, htt activates the transcription of genes containing a repressor sequence named RE1, also known as the neuron-restrictive silencer element (NRSE) (Zuccato *et al.*, 2003). The NRSE sequence is recognized by the transcriptional regulator REST, that acts as a transcriptional silencer. The promoter II of the BDNF (brain-derived neurotrophic factor) gene contains an NRSE element. Wild-type htt induces BDNF transcription by confining the available REST/NRSF in the cytoplasm, thereby preventing the formation of the co-repressor complex at the RE1/NRSE nuclear site (Zuccato *et al.*, 2001).

In the nucleus, htt binds to both SP1 and TAFII130, linking the DNA-bound SP1 and core components of the basal transcription machinery, thus acting as a scaffold for the transcriptional complex (Dunah *et al.*, 2002). Htt may also act as a transcriptional cofactor of itself (Benn *et al.*, 2008).

Wild-type htt also regulates chromatin remodelling, interacting with Ezh2 (enhancer of zeste 2) and Suz12 (Suppressor of zeste 12), two components of the polycomb repressive complex 2 (PRC2) endowed with histone methyltransferase activity (Seong *et al.*, 2009).

Htt may bind transcription factors and regulators and mediate their transport to the nucleus. Through HAP1, htt binds NeuroD (Neurogenic differentiation factor), possibly facilitating its nuclear translocation activation through the phosphorylation mediated by the mixed-lineage kinase 2 (MLK2) (Marcora *et al.*, 2003).

1.6.7 Embryonic Development

Different studies have shown that htt is essential for embryonic development. The inactivation of the gene by targeting exon 1 or 5 is lethal in mice at embryonic day 7.5 (Nasir *et al.*, 1995; Duyao *et al.*, 1995; Zeitlin *et al.*, 1995). Thus, htt is crucial in the early stage of embryonic development, before the emergence of the nervous system. The absence of htt may cause an altered nutritive function of the extraembryonic tissues. Indeed extra-embryonic tissues lacking htt display defects in iron transport (Dragatsis *et al.*, 1998).

Htt was found to be important for the formation of the nervous system. Indeed, mice expressing less than 50% of the normal level of the protein display defects in the formation of the precursor of the epiblast and malformations of the cortex and striatum, and die shortly after birth (White *et al.*, 1997). In addition, analysis of chimeric embryos, in which some cells are depleted of htt, revealed that the protein is essential for the differentiation of neuroblasts in the striatum, cortex, and thalamus (Reiner *et al.*, 2001). Thus, htt is essential for embryogenesis and development. The molecular mechanisms

underlying the two processes include htt regulation of the dynein complex during mitosis and, consequently, the mitotic spindle orientation of the dividing cortical progenitors (Godin *et al.*, 2010).

1.6.8 Tissue Maintenance and Cell Morphology

Some studies indicate that htt is a regulator of tissue maintenance, such as the mammary epithelium, that consists of two major layers of cells, the myoepithelial basal cells and the luminal cells surrounding the lumen. Htt is located more in luminal cells than in basal cells and increases as differentiation progresses (Elias *et al.*, 2014, 2015). Htt is also important for the establishment of apical polarity and the specific depletion of htt from luminal cells alters ductal morphogenesis and lumen formation (Elias *et al.*, 2014). Htt seems to be essential for homotypic interactions between neuroepithelial cells; it allows neurulation and rosette formation by modulating metalloprotease ADAM10 activity and N-cadherin cleavage (Lo Sardo *et al.*, 2012). Htt may also be involved in histogenesis and organogenesis by regulating metabolism. The overexpression of wild-type htt causes an increase of the body weight and increased weight of heart, liver, kidneys, lungs, and spleen in mice (Pouladi *et al.*, 2010; Van Raamsdonk *et al.*, 2005).

1.6.9 Cell Survival

Several studies show that wild-type htt has a prosurvival role. The first in vitro evidence of an antiapoptotic role of htt revealed that the overexpression of wild-type htt in brain-derived cells protects them from toxic stimuli (Rigamonti *et al.*, 2000, 2001). On the contrary, deficiency of htt make cells more sensitive to cell death (Zhang *et al.*, 2006). These in vitro studies were confirmed by in vivo observations in mouse, demonstrating that neuroprotection is obtained by overexpression of htt against excitotoxicity, ischemic injuries and cell death, when htt levels are decreased (Leavitt *et al.*, 2006; Nasir *et al.*, 1995; Zhang *et al.*, 2003; Dragatsis *et al.*, 1998, 2000; O'Kusky *et al.*, 1999). The first efforts to understand how wild-type htt promotes survival focused on the apoptotic machinery. It was found that htt blocks the activation of caspase-3 and -9 (Rigamonti *et al.*, 2001; Zhang *et al.*, 2003).

Another way by which htt supports cell survival is depicted in **Figure 6**. Since the striatum does not produce BDNF, it depends almost exclusively on the BDNF released by cortico-striatal afferences (Baquet, 2004). Htt favours the transcription of BDNF (Zuccato *et al.*, 2001) and facilitates the axonal transport and delivery of vesicles containing BDNF to the cortico-striatal synapse (Gauthier *et al.*, 2004). After BDNF arrival at the cortico-striatal synapse, the activation and the endocytosis of TrkB (Neurotrophic tyrosine kinase receptor type 2) receptors occurs, these receptors are transported retrogradely to striatal cell bodies, where they activate survival signaling (Liot *et al.*, 2013).

This retrograde transport of TrkB is mediated by the htt-dynein intermediate chain (IC)-1B complex (Liot *et al.*, 2013). Thus, htt allows both the anterograde transfer of BDNF to cortico-striatal synapses and the retrograde transport of the BDNF-TrkB endosomes along striatal dendrites (Saudou and Humbert, 2016).
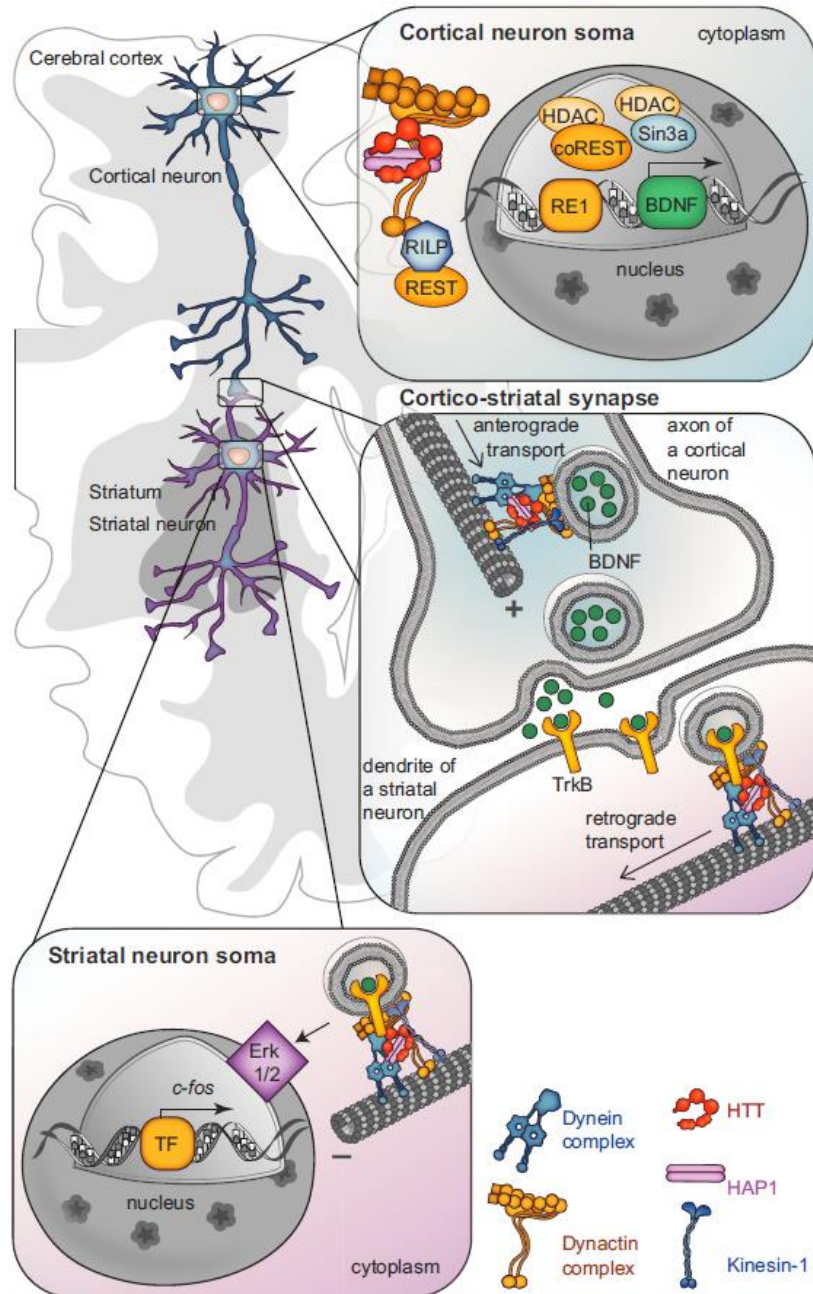

Figure 6. Htt favours the transcription of BDNF and promotes the axonal transport and delivery of vesicles containing BDNF to the cortico-striatal synapse (Saudou and Humbert, 2016).

1.7 Events causing neuronal death

As soon as mhtt is cleaved by proteases, it causes the formation of N-terminal fragments containing the abnormal polyQ stretch. These fragments are transported to the nucleus, where they cause neuronal cell death by interfering with transcription (Ross and Tabrizi, 2011; Saudou *et al.*, 1998; Valor, 2014). In addition, mhtt forms aggregates that at first were considered toxic because they have effects on cellular functions and cause cell death (Davies *et al.*, 1997). Nevertheless, it has been reported that mhtt inclusions reduce the level of the toxic soluble protein (Miller *et al.*, 2010; Saudou *et al.*, 1998). The degree of toxicity is inversely related to the length of the polyQ-containing fragments and the mutation implicates a gain of function in the nucleus (Ross and Tabrizi, 2011).

Also the C-terminal fragments produced by proteolysis are toxic, since they dysregulate the activity of dynamin 1 (El-Daher *et al.*, 2015). The situation is complicated by other effects caused by mutant htt, including proteasome and autophagy inhibition, mitochondrial abnormalities and related metabolic impairments, alteration in endocytosis and microtubule-based transport, synaptic activity deficiencies, impairment in calcium signalling, excitotoxicity caused by increased glutamate release, and dysfunction of several neurotransmitter receptors involved in signalling (glutamate receptor, adenosine A2, cannabinoid, and GLT-1 receptors) (Ross and Tabrizi, 2011). However, this information has been deduced from the overexpression of mhtt, but whether these processes are regulated by wild type htt is not known.

1.8 Huntingtin structure in complex with HAP40

Until a few months ago no data on htt structure were available. The only structural information concerned the N-terminal domain (1-17 amino acids) (Michalek *et al.*, 2013). In 2018, Guo, Huang and their coworkers have determined, through cryo-electron microscopy, the structure of full-length human htt in a complex with HAP40 to an overall resolution of 4 Å (Guo *et al.*, 2018) (**Figure 7**).

For both htt and HAP40, all secondary-structure elements resolved in the model corresponded to α-helices. For htt, 72% of the helices were arranged in HEAT or other tandem repeats. On the other hand, most of the regions not resolved in the map were predicted to be unstructured, such as the N-terminal region (residues 1–90), for which no density was observed even at very low thresholds, indicating the flexibility of this region. The data obtained by Guo and co-workers (2018) show that htt consist of three domains: a smaller bridge domain that connects the N- and C-terminal domains, all containing multiple HEAT repeats (**Figure 8**).
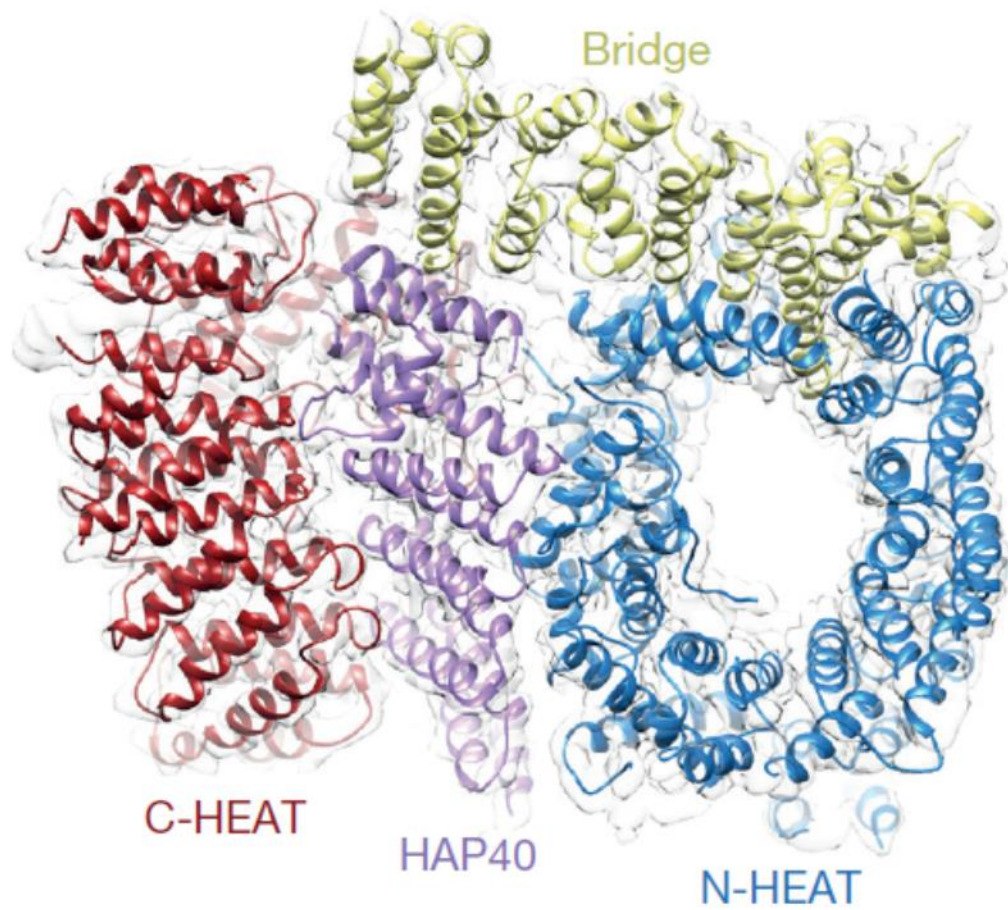
Figure 7. Schematic representation of the three-dimensional structure of Htt-Hap40 complex (Guo *et al.*, 2018).
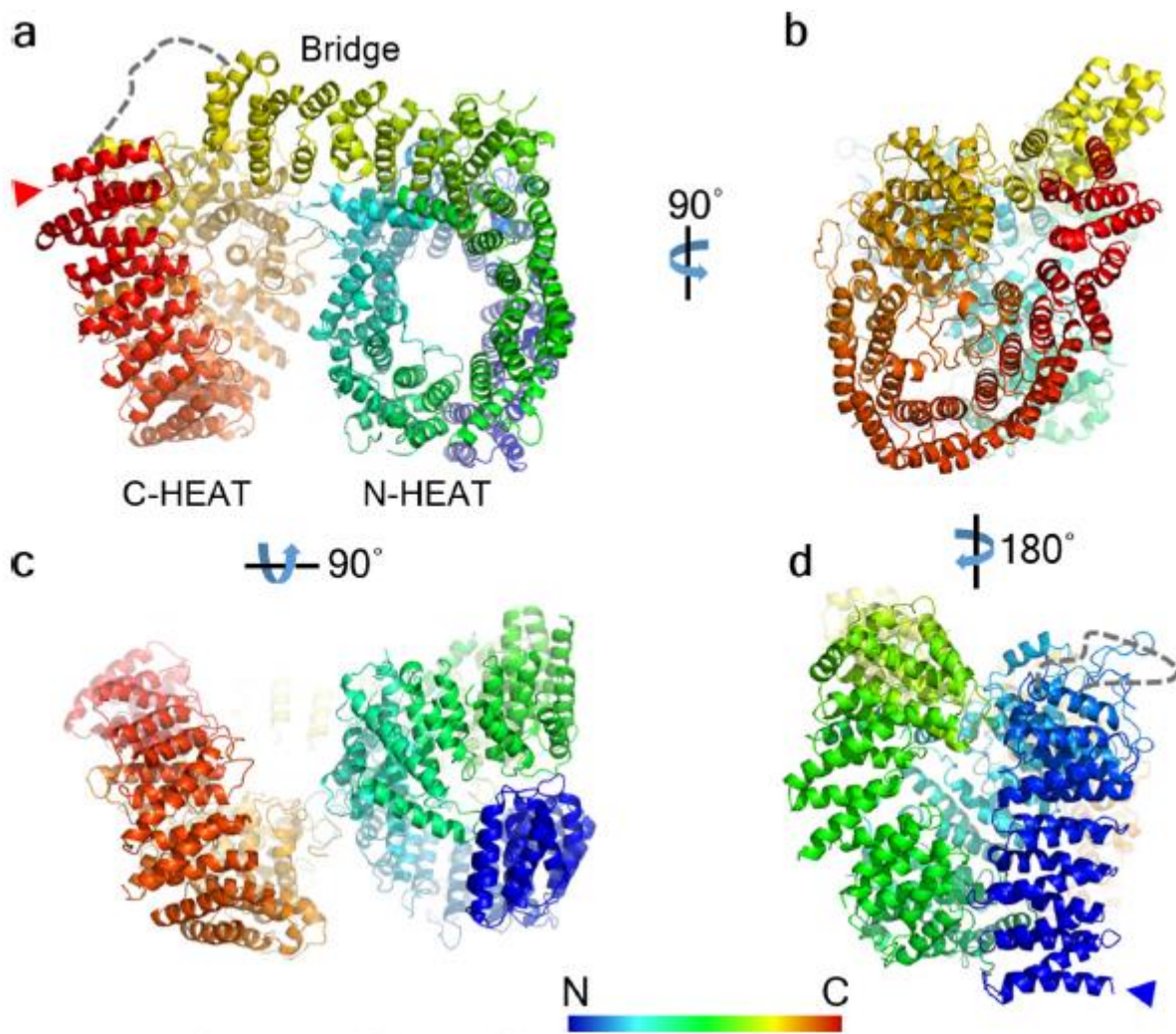
Figure 8. Different views of htt atomic model. Dashed lines indicate unresolved regions (Guo *et al.*, 2018).

The htt N- terminal domain has a typical α-solenoid shape and is made up of 21 HEAT repeats, arranged as a one-and-a-half-turn right-handed superhelix (**Figure 9a**) (Guo *et al.*, 2018).

Within htt N-terminal domain, two membrane-binding regions have been detected: 1-17 residues that can form an amphipathic helix (Arndt *et al.*, 2015) and the second region at residues 168–366, which contains a functionally important palmitoylation site at Cys208 (corresponding to Cys214 of htt sequence). The residues 160–275 of the N-terminal domain form a positively charged portion at the second putative membrane-binding region (Guo *et al.*, 2018). Residues 400–674 constitute a large disordered insertion (**Figure 9a**), consistent with previous computational predictions and results obtained in this thesis work (see chapter 2). Moreover, multiple proteolytic cleavage sites have been mapped to this region (Saudou and Humbert, 2016; Wellington *et al.*, 1998). However, the continuous packing of HEAT repeats makes the release of N-terminal fragments upon proteolysis not easy, in line with previously reported data (Li *et al.*, 2006; El-Daher *et al.*, 2015). Further, this region includes multiple phosphorylation sites that may modulate protein–protein interactions and proteolytic

accessibility (Saudou and Humbert, 2016; Luo *et al.*, 2005; Ratovitski *et al.*, 2017). Most other known post-translational modifications are located in unstructured regions, not resolved in the map, including protease cleavage sites that release N-terminal fragments (Saudou and Humbert, 2016; Wellington *et al.*, 1998; Ratovitski *et al.*, 2009).

The C-terminal domain (residues 2,092–3,098) (**Figure 9b**) is characterized by 12 HEAT repeats that form an elliptical ring with the two terminal HEAT repeats interacting to close the ring.

The repeats are interrupted by two insertions. The first insertion (residues 2,121–2,456) consists of 12 helical regions that separate HEAT repeats 1 and 2 of the C-terminal domain. Instead, insertion 2 (residues 2,510–2,663) is mostly unstructured and does not prevent the interaction between HEAT repeats 2 and 3 of the C-terminal domain.
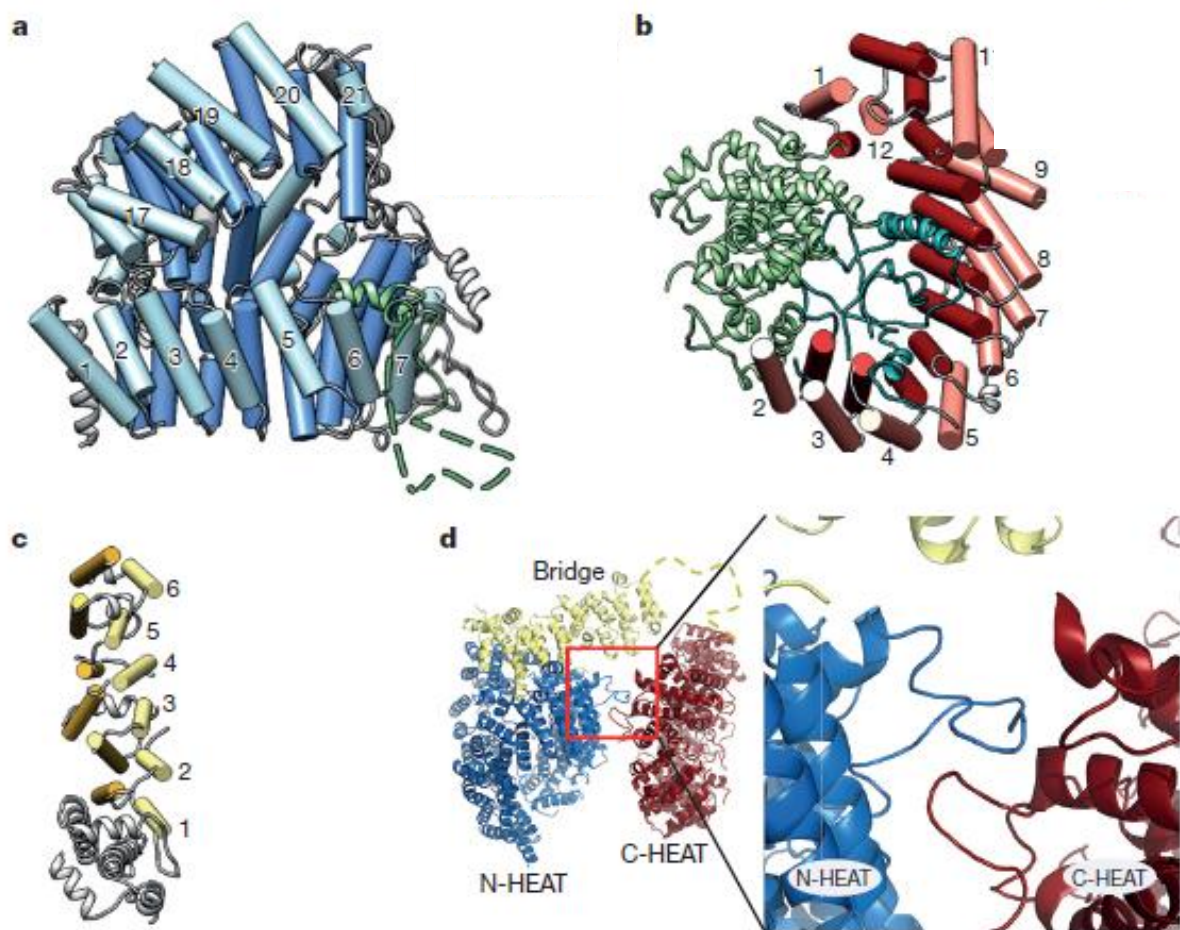
Figure 9. Structure of htt domains. (**a**) The N-terminal domain. The insertion, between HEAT repeats 6 and 7 (corresponding to residues 400-674) is shown in green, with the unresolved structure as a dashed line; (**b**) the C-terminal domain, with the insertions between HEAT repeats shown in green (between HEAT repeats 1 and 2) and greenish-blue (between HEAT repeats 2 and 3); (**c**) the bridge domain; (**d**) a view of htt showing the interaction region between loops of the N- and C-terminal domains (Guo *et al.*, 2018).

The insertions are arranged in the concave surface of the C-terminal domain, likely protecting the region from protein-protein interactions. On the contrary, the convex and concave surfaces of N-terminal domain are accessible in the structure and could thus act as cargo-binding sites. The bridge domain (residues 1,685–2,091, **Figure 9c**) connects the N and C-terminal domains. It is composed of six tandem α-helical repeats, flanked by five non-repeat helices and a flexible C-terminus unresolved. Other than this bridge domain, there is only a weak interaction between the loops of the N- and C-terminal domains (**Figure 9d**) (Guo *et al.*, 2018).

1.9 Huntingtin, Neuroglobin and their neuroprotective role

Experimental studies have highlighted a possible involvement of htt in the neuroprotective role mediated by17β-estradiol (E2) (Cardinale *et al.*, 2018; Nuzzo *et al.*, 2016; De Marinis *et al.*, 2013, 2011; Nuzzo *et al.*, 2017). E2, a sex steroid hormone, has a neurotrophic and neuroprotective role in several brain regions (Arevalo *et al.*, 2015; Mcewen, 2014). Indeed, E2 is a potent trophic factor that controls development, differentiation plasticity, cell survival, cell migration, neuronal somatic and dendritic growth, and synapse formation. In particular E2 shows neuroprotection against β-amyloid toxicity and oxidative stress-induced neuronal death (Fiocchetti *et al.*, 2012).

E2 neuroprotection is regulated by the two estrogen receptors subtypes (ERα and ERβ), members of nuclear receptors family, present in several male and female brain regions including hypothalamus, hippocampus, striatum and cortex (Mitra *et al.*, 2003).

E2 protective effects depend on the over-expression of neuroglobin (Ngb), a brain globin, promoted by ERβ (De Marinis *et al.*, 2013, 2011).

Upon E2 stimulation Ngb localizes mainly into mitochondria, where it binds to cytochrome *c*. An oxidative insult with H_2O_2 increases the E2-mediated association between Ngb and cytochrome *c*, thus reducing the release of cytochrome *c* into the cytosol. This event causes a decrease of caspase-3 activation and, consequently, of apoptosis activation (De Marinis *et al.*, 2013). Experimental studies have shown that htt is also an E2-inducible protein under oxidative stress conditions. E2 stimulation induces the over-expression of both htt and Ngb, promoting the formation of a htt/Ngb complex, by direct interaction or mediated by other interactors, at the mitochondrial compartment (Nuzzo *et al.*, 2017).

However, how Ngb translocation to mitochondria occurs is completely unknown. Given a possible role of htt in vesicular trafficking, htt may be a putative Ngb carrier (Nuzzo *et al.*, 2017; Cardinale *et al.*, 2018).

1.10 Aim of the work

Huntington's disease is a rare neurodegenerative disorder elicited by an increase of CAG repeats within the coding region of a protein called huntingtin. The mutation causes the presence of an elongated stretch of glutamine residues near the N-terminus of the protein. Several studies revealed a critical role for this protein in a number of cellular activities including transcriptional regulation, vesicular shuttling, embryonic development, regulation of apoptosis and autophagy. Its huge size, together with the presence of disordered regions, still doesn't allow crystallization and X-ray diffraction studies. Furthermore, homology modeling techniques are not applicable to htt due to the lack of homologous proteins with known structure. The situation is complicated by the putative existence of htt variants, as result of proteolysis events. Unfortunately, no treatments are currentlty available that can alter the course of Huntington's disease.

Given the relevance of htt in the pathogenesis of Huntington's disease, the determination of the three-dimensional structure of this protein can provide valuable information for understanding its functions and the mechanisms by which the expansion of the polyQ tract in mutated huntingtin leads to neurodegeneration.

In light of the unavailability of structural data on htt when the present work started, the main aim has been that to use a bioinformatics approach to gain information concerning the structure-function relationships of htt, to analyze the htt interactome and to study htt homologs of less complex model organisms. Molecular modelling techniques have been complemented by molecular docking and molecular dynamics techniques, used to dissect the function of a HEAT repeats domain, identified for the first time in the course of this work, and predicted to be the interaction site between htt and G proteins. A related aim has been also that of examining a possible involvement of htt in the neuroprotective role of Ngb.

1.11 References

Anderson,K.P. *et al.* (1997) Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, **277**, 1990–1993.

Arevalo,M.A. *et al.* (2015) The neuroprotective actions of oestradiol and oestrogen receptors. *Nat. Rev. Neurosci.*, **16**, 17–29.

Arndt,J.R. *et al.* (2015) The emerging role of the first 17 amino acids of huntingtin in Huntington's disease. *Biomol. Concepts*, **6**, 33–46.

Atwal,R.S. *et al.* (2007) Huntingtin has a membrane association signal that can modulate huntingtin aggregation, nuclear entry and toxicity. *Hum. Mol. Genet.*, **16**, 2600–2615.

Bae,B. Il *et al.* (2005) p53 mediates cellular dysfunction and behavioral abnormalities in Huntington's disease. *Neuron*, **47**, 29–41.

Baquet,Z.C. (2004) Early Striatal Dendrite Deficits followed by Neuron Loss with Advanced Age in the Absence of Anterograde Cortical Brain-Derived Neurotrophic Factor. *J. Neurosci.*, **24**, 4250–4258.

Benn,C.L. *et al.* (2005) Contribution of nuclear and extranuclear polyQ to neurological phenotypes in mouse models of Huntington's disease. *Hum. Mol. Genet.*, **14**, 3065–3078.

Benn,C.L. *et al.* (2008) Huntingtin Modulates Transcription, Occupies Gene Promoters In Vivo, and Binds Directly to DNA in a Polyglutamine-Dependent Manner. *J. Neurosci.*, **28**, 10720–10733.

Cardinale,A. *et al.* (2018) Localization of neuroglobin in the brain of R6/2 mouse model of Huntington's disease. *Neurol. Sci.*, **39**, 275–285.

Cattaneo,E. *et al.* (2005) Normal huntingtin function: An alternative approach to Huntington's disease. *Nat. Rev. Neurosci.*, **6**, 919–930.

Caviston,J.P. *et al.* (2011) Huntingtin coordinates the dynein-mediated dynamic positioning of endosomes and lysosomes. *Mol. Biol. Cell*, **22**, 478–492.

Caviston,J.P. *et al.* (2007) Huntingtin facilitates dynein/dynactin-mediated vesicle transport. *Proc. Natl. Acad. Sci.*, **104**, 10045–10050.

Caviston,J.P. and Holzabaur,E.L.F. (2009) Huntingtin as an essential integrator of intracellular vesicular Trafficking. *Trends Cell Biol*, **19**, 147–155.

Colin,E. *et al.* (2008) Huntingtin phosphorylation acts as a molecular switch for anterograde/retrograde transport in neurons. *EMBO J.*, **27**, 2124–2134.

Davies,S.W. *et al.* (1997) Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell*, **90**, 537–548.

De Marinis,E. *et al.* (2011) 17β-Estradiol - A new modulator of neuroglobin levels in neurons: Role in neuroprotection against H 2O 2-induced toxicity. *NeuroSignals*, **18**, 223–235.

De Marinis,E. *et al.* (2013) Neuroglobin upregulation induced by 17β-estradiol sequesters cytocrome c in the mitochondria preventing H2O2-induced apoptosis of neuroblastoma cells. *Cell Death Dis.*, **4**, e508.

Dragatsis,I. *et al.* (2000) Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. *Nat. Genet.*, **26**, 300–306.

Dragatsis,I. *et al.* (1998) Mouse mutant embryos lacking huntingtin are rescued from lethality by wild-type extraembryonic tissues. *Development*, **125**, 1529–1539.

Dunah,A.W. *et al.* (2002) Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. *Science (80-.).*, **296**, 2238–2243.

Duyao,M. *et al.* (1995) Inactivation of the Mouse Huntington's Disease Gene Homolog Hdh. *Science (80-.).*, **269**, 407–411.

El-Daher,M.-T. *et al.* (2015) Huntingtin proteolysis releases non-polyQ fragments that cause toxicity through dynamin 1 dysregulation. *EMBO J.*, **34**, 2255–2271.

Elias,S. *et al.* (2015) Huntingtin Is Required for Epithelial Polarity through RAB11A-Mediated Apical Trafficking of PAR3-aPKC. *PLoS Biol.*, **13**, 1–27.

Elias,S. *et al.* (2014) Huntingtin regulates mammary stem cell division and differentiation. *Stem Cell Reports*, **2**, 491–506.

Engelender,S. (1997) Huntingtin-associated protein 1 (HAP1) interacts with the p150Glued subunit of dynactin. *Hum. Mol. Genet.*, **6**, 2205–2212.

Engqvist-Goldstein,Å.E.Y. *et al.* (2001) The actin-binding protein Hip1R associates with clathrin during early stages of endocytosis and promotes clathrin assembly in vitro. *J. Cell Biol.*, **154**, 1209–1223.

Faber,P.W. *et al.* (1998) Huntingtin interacts with a family of WW domain proteins. *Hum. Mol. Genet.*, **7**, 1463–1474.

Fiocchetti,M. *et al.* (2012) Neuroprotective effects of 17β-estradiol rely on estrogen receptor membrane initiated signals. *Front. Physiol.*, **3**.

Futter,M. *et al.* (2009) Wild-type but not mutant huntingtin modulates the transcriptional activity of liver X receptors. *J. Med. Genet.*, **46**, 438–446.

Gafni,J. and Ellerby,L.M. (2002) Calpain Activation in Huntington's Disease. *J. Neurosci.*, **22**, 4842–4849.

Gauthier,L.R. *et al.* (2004) Huntingtin controls neurotrophic support and survival of neurons by enhancing BDNF vesicular transport along microtubules. *Cell*, **118**, 127–138.

Godin,J.D. *et al.* (2010) Huntingtin Is Required for Mitotic Spindle Orientation and Mammalian Neurogenesis. *Neuron*, **67**, 392–406.

Goldberg,Y. *et al.* (1996) Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. *Nat Genet.*, **14**, 353–6.

Graham,R.K. *et al.* (2006) Cleavage at the Caspase-6 Site Is Required for Neuronal Dysfunction and Degeneration Due to Mutant Huntingtin. *Cell*, **125**, 1179–1191.

Guo,Q. *et al.* (2018) The cryo-electron microscopy structure of huntingtin. *Nature*, **555**, 117–120.

Gutekunst,C. a *et al.* (1995) Identification and localization of huntingtin in brain and human lymphoblastoid cell lines with anti-fusion protein antibodies. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 8710–4.

Haremaki,T. *et al.* (2015) Huntingtin is required for ciliogenesis and neurogenesis during early

Xenopus development. *Dev. Biol.*, **408**, 305–315.

Hattula,K. and Peränen,J. (2000) FIP-2, a coiled-coil protein, links Huntingtin to Rab8 and modulates, cellular morphogenesis. *Curr. Biol.*, **10**, 1603–1606.

HD Collaborative research group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.

Her,L.S. and Goldstein,L.S.B. (2008) Enhanced Sensitivity of Striatal Neurons to Axonal Transport Defects Induced by Mutant Huntingtin. *J. Neurosci.*, **28**, 13662–13672.

Hermel,E. *et al.* (2004) Specific caspase interactions and amplification are involved in selective neuronal vulnerability in Huntington's disease. *Cell Death Differ.*, **11**, 424–438.

Holbert,S. *et al.* (2001) The Gln-Ala repeat transcriptional activator CA150 interacts with huntingtin: Neuropathologic and genetic evidence for a role in Huntington's disease pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 1811–1816.

Huang,K. *et al.* (2011) Wild-type HTT modulates the enzymatic activity of the neuronal palmitoyl transferase HIP14. *Hum. Mol. Genet.*, **20**, 3356–3365.

Humbert,S. *et al.* (2002) The IGF-1/Akt pathway is neuroprotective in Huntington's disease and involves huntingtin phosphorylation by Akt. *Dev. Cell*, **2**, 831–837.

Jeong,H. *et al.* (2009) Acetylation Targets Mutant Huntingtin to Autophagosomes for Degradation. *Cell*, **137**, 60–72.

Jodeiri Farshbaf,M. and Ghaedi,K. (2017) Huntington's Disease and Mitochondria. *Neurotox. Res.*, **32**, 518–529.

Kalchman,M.A. *et al.* (1996) Huntingtin Is Ubiquitinated and Interacts with a Specific Ubiquitin-conjugating Enzyme. *J. Biol. Chem.*, **271**, 19385–19394.

Kaltenbach,L.S. *et al.* (2007) Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet.*, **3**, 689–708.

Karam,A. *et al.* (2015) A novel function of Huntingtin in the cilium and retinal ciliopathy in Huntington's disease mice. *Neurobiol. Dis.*, **80**, 15–28.

Kegel,K.B. *et al.* (2002) Huntingtin is present in the nucleus, interacts with the transcriptional corepressor C-terminal binding protein, and represses transcription. *J. Biol. Chem.*, **277**, 7466–7476.

Keryer,G. *et al.* (2011) Ciliogenesis is regulated by a huntingtin-HAP1-PCM1 pathway and is altered in Huntington disease. *J. Clin. Invest.*, **121**, 4372–4382.

Kim,Y.J. *et al.* (2001) Caspase 3-cleaved N-terminal fragments of wild-type and mutant huntingtin are present in normal and Huntington's disease brains, associate with membranes, and undergo calpain-dependent proteolysis. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 12784–9.

Kim,Y.J. *et al.* (2006) Lysosomal proteases are involved in generation of N-terminal huntingtin fragments. *Neurobiol. Dis.*, **22**, 346–356.

Leavitt,B.R. *et al.* (2006) Wild-type huntingtin protects neurons from excitotoxicity. *J. Neurochem.*, **96**, 1121–1129.

Legendre-Guillemin,V. *et al.* (2002) HIP1 and HIP12 display differential binding to F-actin, AP2, and clathrin. Identification of a novel interaction with clathrin light chain. *J. Biol. Chem.*, **277**,

19897–19904.

Legleiter,J. *et al.* (2010) Mutant huntingtin fragments form oligomers in a polyglutamine length-dependent manner in Vitro and in Vivo. *J. Biol. Chem.*, **285**, 14777–14790.

Li,S.-H. *et al.* (2002) Interaction of Huntington disease protein with transcriptional activator Sp1. *Mol. Cell. Biol.*, **22**, 1277–87.

Li,S.H. *et al.* (1998) Interaction of huntingtin-associated protein with dynactin P150Glued. *J. Neurosci.*, **18**, 1261–1269.

Li,W. *et al.* (2006) Expression and characterization of full-length human huntingtin, an elongated HEAT repeat protein. *J. Biol. Chem.*, **281**, 15916–15922.

Li,X. *et al.* (2008) A function of huntingtin in guanine nucleotide exchange on Rab11. *Neuroreport*, **19**, 1643–1647.

Liot,G. *et al.* (2013) Mutant Huntingtin Alters Retrograde Transport of TrkB Receptors in Striatal Dendrites. *J. Neurosci.*, **33**, 6298–6309.

Lunkes,A. *et al.* (2002) Proteases acting on mutant huntingtin generate cleaved products that differentially build up cytoplasmic and nuclear inclusions. *Mol. Cell*, **10**, 259–269.

Luo,S. *et al.* (2005) Cdk5 phosphorylation of huntingtin reduces its cleavage by caspases: Implications for mutant huntingtin toxicity. *J. Cell Biol.*, **169**, 647–656.

Ben M'Barek,K. *et al.* (2013) Huntingtin Mediates Anxiety/Depression-Related Behaviors and Hippocampal Neurogenesis. *J. Neurosci.*, **33**, 8608–8620.

Maiuri,T. *et al.* (2013) The huntingtin N17 domain is a multifunctional CRM1 and ran-dependent nuclear and cilial export signal. *Hum. Mol. Genet.*, **22**, 1383–1394.

Marcora,E. *et al.* (2003) Stimulation of NeuroD activity by huntingtin and huntingtin-associated proteins HAP1 and MLK2. *Proc. Natl. Acad. Sci.*, **100**, 9578–9583.

McColgan,P. and Tabrizi,S.J. (2018) Huntington's disease: a clinical review. *Eur. J. Neurol.*, **25**, 24–34.

Mcewen,B.S. (2014) Sex, stress and the brain: Interactive actions of hormones on the developing and adult brain. *Climacteric*, **17**, 18–25.

McGuire,J.R. *et al.* (2006) Interaction of Huntingtin-associated protein-1 with kinesin light chain: Implications in intracellular trafficking in neurons. *J. Biol. Chem.*, **281**, 3552–3559.

Michalek,M. *et al.* (2013) Structure and topology of the huntingtin 1-17 membrane anchor by a combined solution and solid-state nmr approach. *Biophys. J.*, **105**, 699–710.

Miller,J. *et al.* (2011) NIH Public Access. **30**, 10541–10550.

Miller,J. *et al.* (2010) Quantitative Relationships between Huntingtin Levels, Polyglutamine Length, Inclusion Body Formation, and Neuronal Death Provide Novel Insight into Huntington's Disease Molecular Pathogenesis. *J. Neurosci.*, **30**, 10541–10550.

Mitra,S.W. *et al.* (2003) Immunolocalization of estrogen receptor β in the mouse brain: Comparison with estrogen receptor α. *Endocrinology*, **144**, 2055–2067.

Modregger,J. *et al.* (2003) Characterization of endophilin B1b, a brain-specific membrane-associated lysophosphatidic acid acyl transferase with properties distinct from endophilin A1. *J. Biol. Chem.*, **278**, 4160–4167.

Moreira Sousa,C. *et al.* (2013) The Huntington disease protein accelerates breast tumour development and metastasis through ErbB2/HER2 signalling. *EMBO Mol. Med.*, **5**, 309–325.

Nasir,J. *et al.* (1995) Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell*, **81**, 811–823.

Nuzzo,M.T. *et al.* (2016) 17β-Estradiol modulates huntingtin levels in rat tissues and in human neuroblastoma cell line. *Neurosci. Res.*, **103**, 59–63.

Nuzzo,M.T. *et al.* (2017) Huntingtin polyQ Mutation Impairs the 17β-Estradiol/Neuroglobin Pathway Devoted to Neuron Survival. *Mol. Neurobiol.*, **54**, 6634–6646.

O'Kusky,J.R. *et al.* (1999) Neuronal degeneration in the basal ganglia and loss of pallido-subthalamic synapses in mice with targeted disruption of the Huntington's disease gene. *Brain Res.*, **818**, 468–479.

Ochaba,J. *et al.* (2014) Potential function for the Huntingtin protein as a scaffold for selective autophagy. *Proc. Natl. Acad. Sci.*, **111**, 16889–16894.

Pal,A. *et al.* (2006) Huntingtin-HAP40 complex is a novel Rab5 effector that regulates early endosome motility and is up-regulated in Huntington's disease. *J. Cell Biol.*, **172**, 605–618.

Palidwor,G.A. *et al.* (2009) Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput. Biol.*, **5**.

Pardo,R. *et al.* (2006) Inhibition of Calcineurin by FK506 Protects against Polyglutamine-Huntingtin Toxicity through an Increase of Huntingtin Phosphorylation at S421. *J. Neurosci.*, **26**, 1635–1645.

Poirier,M.A. *et al.* (2002) Huntingtin spheroids and protofibrils as precursors in polyglutamine fibrilization. *J. Biol. Chem.*, **277**, 41032–41037.

Pouladi,M.A. *et al.* (2010) Full-length huntingtin levels modulate body weight by influencing insulin-like growth factor 1 expression. *Hum. Mol. Genet.*, **19**, 1528–1538.

Van Raamsdonk,J.M. *et al.* (2005) Loss of wild-type huntingtin influences motor dysfunction and survival in the YAC128 mouse model of Huntington disease. *Hum. Mol. Genet.*, **14**, 1379–1392.

Rangone,H. *et al.* (2004) The serum- and glucocorticoid-induced kinase SGK inhibits mutant huntingtin-induced toxicity by phosphorylating serine 421 of huntingtin. *Eur. J. Neurosci.*, **19**, 273–279.

Ratovitski,T. *et al.* (2009) Mutant huntingtin N-terminal fragments of specific size mediate aggregation and toxicity in neuronal cells. *J. Biol. Chem.*, **284**, 10855–10867.

Ratovitski,T. *et al.* (2017) Post-Translational Modifications (PTMs), Identified on Endogenous Huntingtin, Cluster within Proteolytic Domains between HEAT Repeats.

Reiner,A. *et al.* (2001) Neurons lacking huntingtin differentially colonize brain and survive in chimeric mice. *J. Neurosci.*, **21**, 7608–19.

Rigamonti,D. *et al.* (2001) Huntingtin's Neuroprotective Activity Occurs via Inhibition of Procaspase-9 Processing. *J. Biol. Chem.*, **276**, 14545–14548.

Rigamonti,D. *et al.* (2000) Wild-type huntingtin protects from apoptosis upstream of caspase-3. *J. Neurosci.*, **20**, 3705–13.

Ross,C.A. and Tabrizi,S.J. (2011) Huntington ' s disease : from molecular pathogenesis to clinical

treatment. *Lancet Neurol.*, **10**, 83–98.

Sahlender,D.A. *et al.* (2005) Optineurin links myosin VI to the Golgi complex and is involved in Golgi organization and exocytosis. *J. Cell Biol.*, **169**, 285–295.

Sanders,S.S. and Hayden,M.R. (2015) Aberrant palmitoylation in Huntington disease. *Biochem. Soc. Trans.*, **43**, 205–210.

Lo Sardo,V. *et al.* (2012) An evolutionary recent neuroepithelial cell adhesion function of huntingtin implicates ADAM10-Ncadherin. *Nat. Neurosci.*, **15**, 713–721.

Saudou,F. *et al.* (1998) Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. *Cell*, **95**, 55–56.

Saudou,F. and Humbert,S. (2016) The Biology of Huntingtin. *Neuron*, **89**, 910–926.

Schilling,B. *et al.* (2006) Huntingtin phosphorylation sites mapped by mass spectrometry: Modulation of cleavage and toxicity. *J. Biol. Chem.*, **281**, 23686–23697.

Seong,I.S. *et al.* (2009) Huntingtin facilitates polycomb repressive complex 2. *Hum. Mol. Genet.*, **19**, 573–583.

Sittler,A. *et al.* (1998) SH3GL3 associates with the huntingtin exon-1 protein and promotes the formation of polygln-containing protein\raggregates. *Mol. Cell*, **2**, 427–436.

Steffan,J.S. *et al.* (2004) SUMO Modification of Huntingtin and Huntington's Disease Pathology. *Science (80-.).*, **304**, 100–104.

Steffan,J.S. *et al.* (2000) The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. *Proc Natl Acad Sci U S A*, **97**, 6763–6768.

Takano,H. and Gusella,J.F. (2002) The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor. *BMC Neurosci.*, **3**, 1–13.

Tartari,M. *et al.* (2008) Phylogenetic comparison of huntingtin homologues reveals the appearance of a primitive polyQ in sea urchin. *Mol. Biol. Evol.*, **25**, 330–338.

Tebbenkamp,A.T.N. *et al.* (2012) Analysis of Proteolytic Processes and Enzymatic Activities in the Generation of Huntingtin N-Terminal Fragments in an HEK293 Cell Model. *PLoS One*, **7**, 1–10.

Thompson,L.M. *et al.* (2009) IKK phosphorylates Huntingtin and targets it for degradation by the proteasome and lysosome. *J. Cell Biol.*, **187**, 1083–1099.

del Toro,D. *et al.* (2009) Mutant Huntingtin Impairs Post-Golgi Trafficking to Lysosomes by Delocalizing Optineurin/Rab8 Complex from the Golgi Apparatus. *Mol. Biol. Cell*, **20**, 1478–1492.

Twelvetrees,A.E. *et al.* (2010) Delivery of GABAARs to Synapses Is Mediated by HAP1-KIF5 and Disrupted by Mutant Huntingtin. *Neuron*, **65**, 53–65.

Valor,L.M. (2014) Transcription, Epigenetics and Ameliorative Strategies in Huntington's Disease: a Genome-Wide Perspective. *Mol. Neurobiol.*, **51**, 406–423.

Velier,J. *et al.* (1998) Wild-type and mutant huntingtins function in vesicle trafficking in the secretory and endocytic pathways. *Exp. Neurol.*, **152**, 34–40.

Waelter,S. *et al.* (2001) The huntingtin interacting protein HIP1 is a clathrin and alpha-adaptin-

binding protein involved in receptor-mediated endocytosis. *Hum. Mol. Genet.*, **10**, 1807–1817.

Warby,S.C. *et al.* (2008) Activated caspase-6 and caspase-6-cleaved fragments of huntingtin specifically colocalize in the nucleus. *Hum. Mol. Genet.*, **17**, 2390–2404.

Wellington,C.L. *et al.* (1998) Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *J. Biol. Chem.*, **273**, 9158–9167.

White,J.K. *et al.* (1997) Huntington is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nat. Genet.*, **17**, 404–410.

Wong,Y.C. and Holzbaur,E.L.F. (2014) The Regulation of Autophagosome Dynamics by Huntingtin and HAP1 Is Disrupted by Expression of Mutant Huntingtin, Leading to Defective Cargo Degradation. *J. Neurosci.*, **34**, 1293–1305.

Yanai,A. *et al.* (2006) Palmitoylation of huntingtin by HIP14is essential for its trafficking and function. *Nat. Neurosci.*, **9**, 824–831.

Yohrling IV,G.J. *et al.* (2003) Mutant huntingtin increases nuclear corepressor function and enhances ligand-dependent nuclear hormone receptor activation. *Mol. Cell. Neurosci.*, **23**, 28–38.

Zala,D. *et al.* (2013) Vesicular glycolysis provides on-board energy for fast axonal transport. *Cell*, **152**, 479–491.

Zeitlin,S. *et al.* (1995) Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. *Nat. Genet.*, **11**, 155–163.

Zhang,Y. *et al.* (2006) Huntingtin inhibits caspase-3 activation. *EMBO J.*, **25**, 5896–5906.

Zhang,Y. *et al.* (2003) Start Cast Breakouts Preventative Prediction Using Multi-Way PCA Technology. *IFAC Proc. Vol.*, **36**, 101–106.

Zuccato,C. *et al.* (2003) Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat. Genet.*, **35**, 76–83.

Zuccato,C. *et al.* (2001) Loss of Huntingtin-Mediated BDNF Gene Transcription in Huntington ' s Disease. **293**, 493–498.

Zuccato,C. *et al.* (2010) Molecular Mechanisms and Potential Therapeutical Targets in Huntington ' s Disease. *Physiol Rev*, **90**, 905–981.

Zuccato,C. and Cattaneo,E. (2016) The Huntington's Paradox. *Sci. Am.*, **315**, 56–61.

Chapter 2: "A comprehensive *in silico* analysis of huntingtin and its interactome"

In the paper entitled "A comprehensive *in silico* analysis of huntingtin and its interactome" five htt regions predicted to fold into ordered domains have been detected. *Ab initio* molecular modelling of each ordered domain led to the identification of a previously undetected HEAT repeats region in the third ordered domain. Structural models of htt ordered domains have been proposed for the first time, providing a putative function for four out of the five domains. The third ordered domain, named hunt3, has shown structural similarity with a Karyopherin, a protein involved in nuclear import by interacting with GTP-binding proteins. Interestingly, the analysis of htt interactome has revealed that it interacts with several GTP-binding proteins. Molecular docking simulations have confirmed that the hunt3 domain is the putative interaction site between htt and G proteins. Among htt interactors, the G protein $G\alpha_o1$ (encoded by the GNAO1 gene) is particularly interesting. In fact, different heterozygous mutations in the GNAO1 gene are the cause of a severe neurodevelopmental disorder, the epileptic encephalopathy type 17, which is characterized by symptoms similar to those observed in HD. Interestingly, some of these mutations are located in the putative $G\alpha_o1$-Hunt3 interface. These results lead to the hypothesis that mutations of the $G\alpha_o1$ subunit impair the interaction with htt, thus providing a possible explanation for the similar symptoms observed in HD and epileptic encephalopathy.
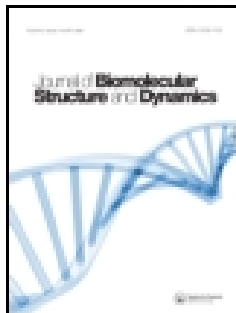
A comprehensive in silico analysis of huntingtin and its interactome

Valentina Brandi, Valentina Di Lella, Maria Marino, Paolo Ascenzi & Fabio Polticelli

   View supplementary material

   Accepted author version posted online: 18 Sep 2017.
   Published online: 28 Sep 2017.

   Submit your article to this journal

   Article views: 225

   View related articles

   View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

A comprehensive *in silico* analysis of huntingtin and its interactome

Valentina Brandi[a], Valentina Di Lella[a], Maria Marino[a], Paolo Ascenzi[a,b] and Fabio Polticelli[a,c]*

[a]Department of Sciences, Roma Tre University, Viale Guglielmo Marconi 446, Roma I–00146, Italy; [b]Interdepartmental Laboratory for Electron Microscopy, Roma Tre University, Roma I–00146, Italy; [c]National Institute of Nuclear Physics, Roma Tre University, Roma Tre Section, Roma I–00146, Italy

A polyglutamine expansion of the *N*-terminal region of huntingtin (Htt) causes Huntington's disease, a severe neurodegenerative disorder. Htt huge multidomain structure, the presence of disordered regions, and the lack of sequence homologs of known structure, so far prevented structural studies of Htt, making the study of its structure-function relationships very difficult. In this work, the presence and location of five Htt ordered domains (named from Hunt1 to Hunt5) has been detected and the structure of these domains has been predicted for the first time using a combined threading/*ab initio* modeling approach. This work has led to the identification of a previously undetected HEAT repeats region in the Hunt3 domain. Furthermore, a putative function has been assigned to four out of the five domains. Hunt1 and Hunt5, displaying structural similarity with the regulatory subunit A of protein phosphatase 2A, are predicted to play a role in regulating the phosphorylation status of cellular proteins. Hunt2 and Hunt3 are predicted to be homologs of two yeast importins and to mediate vescicles transport and protein trafficking. Finally, a comprehensive analysis of the Htt interactome has been carried out and is discussed to provide a global picture of the Htt's structure–function relationships.

Keywords: huntingtin; *in silico* analysis; huntingtin domains; huntingtin interacting proteins; molecular modeling; molecular docking

1. Introduction

Huntington's disease (HD) is a rare neurodegenerative disorder elicited by an increase of CAG repeats within the coding region of a protein called huntingtin (Htt). The mutation entails an elongated stretch of glutamine residues near the *N*-terminus of the protein. The broad expression and the complex amino acid sequence of Htt render difficult the search for Htt function(s), but several studies revealed a critical role for this protein in a number of cellular activities including transcriptional regulation, vesicular shuttling, embryonic development, regulation of apoptosis and autophagy (Nuzzo et al., 2016; Nuzzo & Marino, 2016; Zuccato, Valenza, & Cattaneo, 2010). Intriguingly, the Htt mutation could impair the ability of normal Htt to execute these functions (i.e., loss of function) (Cattaneo, Zuccato, & Tartari, 2005; Zuccato & Cattaneo, 2007) or could even promote new functions including protein aggregation (i.e. gain of function; Arndt, Chaibva, & Legleiter, 2015; Nuzzo et al., 2016; Ratovitski et al., 2012; Zuccato et al., 2010). Both loss and gain of function of mutated Htt lead to the impairment of neuron survival and functioning that culminate in neuro-degeneration, even though the exact mechanisms are still unclear.

Htt is a 3144 amino acids protein; its high molecular weight together with the presence of disordered regions hinders crystallization and X-ray diffraction studies. Further, no Htt homologs with known structure are available, making the study of its structure–function relationships very difficult (Arndt et al., 2015; Ratovitski et al., 2012; Zuccato et al., 2010). The *N*-terminal region of Htt includes the polyQ region, starting at the 18th residue and followed by a proline-rich domain (PRD) (Saudou & Humbert, 2016). The structure of the *N*-terminal region of wild type Htt, containing 17 Gln residues (17Q), has been determined showing that the first *N*-terminal 17 amino acids form an α-helical structure, but the 17Q stretch is flexible and can adopt several conformations (α-helix, random coil, and extended loop) (Kim, Chelliah, Kim, Otwinowski, & Bezprozvanny, 2009). This region contains a nuclear export signal (NES) and undergoes post-translational modifications (Atwal et al., 2007; Maiuri, Woloshansky, Xia, & Truant, 2013; Steffan et al., 2004; Thompson et al., 2009). The PRD is made up of a polyproline helix, a rigid structure that may be significant for maintaining the structure of the polyQ stretch and have an effect on the tendency of mutant Htt to aggregate. The remaining 97% of the

---

protein is less well characterized (Saudou & Humbert, 2016). The protein is composed of several HEAT repeats, each formed by short antiparallel α-helices, usually involved in protein–protein interactions (Palidwor et al., 2009). Several bioinformatics analyses of Htt sequence identified a number of HEAT repeats varying between 16 and 36, grouped into three to five large α-rod domains separated by disordered regions (Palidwor et al., 2009; Takano & Gusella, 2002; Tartari et al., 2008; Warby et al., 2008). Intramolecular interactions occur between different Htt regions: a central part of Htt (507–1230) can form homodimers or can bind to the *N*-terminal (1–416) and *C*-terminal (2721–3144) domains of Htt itself; two *N*-terminal portions of Htt (1–416; 1–586) can bind to different *C*-terminal parts (1725–2800; 2416–3144, respectively). Nevertheless, these interactions are broken up after proteolysis (El-Daher et al., 2015; Ochaba et al., 2014). These data suggest diverse three-dimensional conformations of Htt, depending on its intra-molecular interactions (Saudou & Humbert, 2016). Furthermore, purified Htt can assume up to 100 structurally distinct conformations (Seong et al., 2010). Htt undergoes multiple post-translational modifications, such as phosphorylation, acetylation, palmitoylation, ubiquitylation, and sumoylation. Their role has been mainly investigated in the mutant Htt and their implication in the regulation of the normal functions of the protein is still not known (Saudou & Humbert, 2016).

In the attempt to obtain structural data on Htt that could help to clarify its numerous physio-pathological functions, in this work a comprehensive *in silico* analysis of wild-type Htt has been carried out. First, using a consensus among a variety of protein disorder prediction tools, five Htt's ordered domains have been detected. The corresponding sequence regions have been subjected to protein structure prediction using a combined threading/*ab initio* modeling strategy and a putative function has been assigned to four out of the five domains. Finally, the Htt interactome has been analyzed *in silico* as well to provide a global picture of Htt structure-function relationships.

2. Methods

2.1. *Prediction of Htt ordered domains*

Prediction of Htt ordered regions has been carried out using a consensus between the following methods.

Foldindex   (https://fold.weizmann.ac.il/fldbin/findex) predicts if a given protein sequence is intrinsically unfolded using the algorithm of Uversky and co-workers (Uversky, Gillespie, & Fink, 2000). This algorithm is based on average hydrophobicity and net charge and provides a single 'foldability' score (Prilusky et al., 2005).

Globplot (https://globplot.embl.de/) detects disordered regions within protein sequences. It measures and displays the propensity of a sequence region to be ordered or disordered (Linding et al., 2003).

PredictProtein (https://www.predictprotein.org) is a meta-service for sequence analysis that predicts structural and functional features of proteins. For most prediction methods, the prediction quality is estimated by a reliability score (Yachdav et al., 2014).

ANCHOR (https://anchor.enzim.hu/) is a web service that predicts protein binding regions that are disordered in isolation, but can become ordered upon binding of a molecular partner (Dosztányi, Mészáros, & Simon, 2009). ANCHOR deduces the basic biophysical properties of disordered binding regions using estimated energy calculations (Mészáros, Simon, & Dosztányi, 2009). ANCHOR's output is a score, which represents the probability of each residue of the input sequence to be in a disordered binding region. Regions with a score higher than .5 are predicted as disordered binding regions (Dosztányi et al., 2009).

InterPro database (https://www.ebi.ac.uk/interpro/) classifies sequences into protein families and predicts the presence and function of important domains and sites, making use of diagnostic models known as signatures (Hunter et al., 2012; Mitchell et al., 2014).

SMART (https://smart.embl-heidelberg.de/) is a web resource for the identification of protein domains and the analysis of their architectures; it is Nuzzo et al. (2016); Uniprot (The UniProt Consortium, 2014), ENSEMBL (Flicek et al., 2014), and STRING (Franceschini et al., 2013). The SMART database integrates manually curated hidden Markov models for many domains (Krogh, Brown, Mian, Sjolander, & Haussler, 1994) and offers a variety of analysis and visualization tools (Letunic, Doerks, & Bork, 2014).

HMMER (https://www.ebi.ac.uk/Tools/hmmer/) is a suite for sequence similarity searches using profile hidden Markov models (HMMs) (Finn et al., 2015) focussed primarily on UniProtKB (The UniProt Consortium, 2015). The query sequence is scanned against the Pfam profile HMM library using hmmscan, to detect any Pfam family (Finn et al., 2014) and analyzed for the presence of disordered regions, using IUPred (Dosztányi, Csizmok, Tompa, & Simon, 2005a; Dosztányi, Csizmok, Tompa, & Simon, 2005b), signal peptides and transmembrane regions, using Phobius (Käll, Krogh, & Sonnhammer, 2004), and coiled-coil regions (Lupas, Van Dyke, & Stock, 1991).

MobiDB (https://mobidb.bio.unipd.it/), a database of intrinsically disordered and mobile proteins (Potenza, Di Domenico, Walsh, & Tosatto, 2015), provides a complete picture of different types of protein disorder covering all Uniprot sequences. MobiDB includes 10 predictors: three ESpritz flavors (Walsh, Martin, Di Domenico, & Tosatto, 2012), two IUPred flavors (Dosztányi et al., 2005b), two DisEMBL flavors (Linding, Russell, Neduva, &

Gibson, 2003b), GlobPlot, VSL2b (Peng, Radivojac, Vucetic, Dunker, & Obradovic, 2006), and JRONN (Yang, Thomson, McNeil, & Esnouf, 2005).

PONDR (https://www.pondr.com) is a neural network predictor of natural protein disordered regions. It includes different predictors for short, medium, and long disordered regions. For Htt has been used PONDR VL-XT, which integrates the VL1 predictor (for internal regions), the N-terminus predictor (XN), and the C-terminus predictor (XC). The XN and XC predictors are called XT because trained using X-ray crystallographic data (Romero et al., 2001).

2.2. Protein structure prediction

The structural models of the ordered domains of Htt have been built using I-TASSER and Robetta. For comparative structure modeling of the open conformation of the Hunt3 domain (see below) the program MODELLER has been used.

I-TASSER (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) is an integrated platform to perform automated protein structure and function prediction. It generates three-dimensional models of the query protein using a threading approach to identifying suitable modeling templates for the query sequence (Roy, Kucukural, & Zhang, 2010). The quality of the template alignments is judged accounting for the statistical significance of the best threading alignment, the Z-score (Roy et al., 2010). Models are built using a combination of threading and *ab initio* techniques for the template-aligned and template-unaligned regions, respectively. The function of the query protein is predicted by structurally matching the models obtained against the protein structures present in the PDB (Roy et al., 2010) and ranking the functional analogs through a combination of TM-score, rmsd, sequence identity, and coverage of the structure alignment (Roy et al., 2010). The reliability of the structural models can be assessed by the C-score parameter provided by the server. Supplementary materials Table S1 reports the C-score values obtained for the models of the predicted ordered domains of Htt.

The automated server Robetta (https://robetta.bakerlab.org/; Kim, Chivian, & Baker, 2004) implements the Rosetta *ab initio* protein structure prediction protocol (Rohl, Strauss, Misura, & Baker, 2004). In detail, the query sequence is first divided into putative domains using the Ginzu method (Chivian et al., 2003). Ginzu identifies regions of the query sequence that are homologous to known experimental structures and predicts the putative domains composing the query protein (Kaufmann, Lemmon, DeLuca, Sheehan, & Meiler, 2010). The corresponding structural models are built using comparative modeling, if a confident match to a protein of known structure was found, or else the *de*

novo Rosetta fragment insertion method (Kaufmann et al., 2010). The side chains of the final model are then repacked using a Monte Carlo algorithm in combination with a backbone-dependent side chain rotamers library (Dunbrack & Cohen, 1997). Robetta server has a sequence length limit for single domains structure prediction of 200 amino acids for *ab initio* modeling and 650 amino acids for comparative modeling. Therefore, it was not possible to obtain a structural model of the Hunt3 domain of Htt (1118 amino acid residues) using Robetta.

MODELLER is a program for comparative structure modeling (Fiser, Do, & Sali, 2000; Šali & Blundell, 1993) which, starting from the alignment of the query protein sequence with the template structures and the atomic coordinates of the templates, builds a model containing all non-hydrogen atoms (Webb & Sali, 2014). In this work, MODELLER has been used only to model the Htt region Hunt3 in the same conformation of the structural homolog Kap121p, bound to a GTP-binding protein (see Results and discussion Section 3.2).

Docking simulations between the model of the Htt Hunt3 region and its putative partner Rab11a were performed using the protein–protein docking server ZDOCK (https://zdock.umassmed.edu; version 3.0.2; Pierce et al., 2014). ZDOCK implements a Fast Fourier Transform algorithm and a scoring system based on a combination of shape complementarity, electrostatics and statistical potential terms (Pierce et al., 2014). The 2000 complexes predicted by ZDOCK were re-ranked using ZRANK (Pierce & Weng, 2007) that uses a more detailed potential including electrostatics, van der Waals, and desolvation terms (Pierce & Weng, 2007).

2.3. Analysis of the Htt interactome

Predicted and experimentally determined Htt interactors were collected using the following tools.

PrePPI (available at https://bhapp.c2b2.columbia.edu/PrePPI) is a protein–protein interaction (PPI) database which combines predicted and experimentally confirmed PPIs (Zhang, Petrey, Garzón, Deng, & Honig, 2013). Predicted interactions are assigned likelihood ratios (LRs) based on structural, functional, evolutionary, and expression information. Similarly, LRs are assigned also to the experimentally determined PPIs, collected from the MIPS, DIP, IntAct, MINT, HPRD, and BioGRID databases. A final probability is assigned to each interaction by combining the LRs of the predicted and the experimentally confirmed interactions (Zhang et al., 2013).

BioGRID (available at https://www.thebiogrid.org) is an open access database collecting and annotating experimentally determined interactions for most model organisms (Chatr-Aryamontri et al., 2013). BioGRID is manually curated scanning and rating the literature available for each interaction (Hirschman et al., 2012).

IntAct (available at https://www.ebi.ac.uk/intact/) is a manually curated database containing molecular interactions derived from the literature and from data repositories. Each record reports also the experimental conditions in which the interaction was observed. Each interaction is reviewed and scored according to the methods of detection of the interaction and the interaction type (Kerrien et al., 2012).

HIPPIE (available at https://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/) is a database of experimentally derived human PPIs, which integrates and scores multiple data-sets. HIPPIE normalized scoring scheme rates the reliability of a given PPI evidence by combining information on the experimental techniques used, the number of studies evidencing the PPI, and its reproducibility. A score in the range 0 and 1 is calculated for each PPI as a weighted sum of three different sub-scores. Higher scores are assigned to methods with higher reliability (e.g., X-ray crystallography), lower scores are assigned to methods which do not provide evidence of a direct interaction such as co-localization (Schaefer et al., 2012).

3. Results and discussion

3.1. Identification of Htt-ordered domains

Due to the lack of structural data and templates with significant sequence similarity with Htt, the approach used in this work has been initially to identify Htt regions predicted to fold into ordered domains. Starting from the Htt amino acid sequence, several order/disorder prediction methods were used. The results of this analysis are schematically reported in Figure 1 and in detailed form in the Supplementary materials Table S2. As shown in Figure 1, the various methods display a high degree of consensus on the presence of five putative ordered domains (named from now on from Hunt1 to Hunt5) whose location along Htt's sequence is shown in Figure 1. This result is consistent with data obtained by cross-linking mass spectrometry analysis which also indicate the presence of five ordered domains in Htt (Vijayvargia et al., 2016).

3.2. Prediction of the structural models of Htt's ordered domains

3.2.1. Hunt1

Hunt1 structural models obtained with I-TASSER and Robetta display the topology typical of HEAT repeats (Figure 2(a) and (b)). They both have a concave shape, more evident in the Robetta model (Figure 2(b)). The overall similarity of the two models, given the different approach characterizing the two methods, reinforces the reliability of the prediction. Among the structural

homologs identified by I-TASSER (Supplementary materials Table S3) is present the Protein phosphatase 2A (PP2A) A subunit (PDB code: 1B3U). PP2A is the most important cellular serine/threonine phosphatase, which is involved in many essential aspects of cellular function and is characterized by a HEAT-like repeat structure (Xu et al., 2006). The superimposition between the structural model of Hunt1 (obtained with I-TASSER) and the three-dimensional structure of PP2A A-subunit (Figure 2(c)) shows that Hunt1 well matches with half of PP2A A-subunit.

3.2.2. Hunt2

The Hunt2 model obtained with I-TASSER is characterized by an α-helical structure, which displays an internal torsion determining an opposite orientation of the *N-* and *C*-termini (Figure 3(a)). The Robetta model (Figure 3(b)) consists of a α-helical fold as well, characterized by a 'horseshoe' shape. The best structural homolog of Hunt2 model is the yeast karyopherin α (importin-α) (PDB code: 1EE5; Supplementary materials Table S4). The list of homologs contains also other importins/karyopherins (Supplementary materials Table S4). The superimposition of Hunt2 structural model with the importin-α structure indicates that the two match very well, even though the Hunt2 structural model displays a longer tail at the *C*-terminus and two internal loops not present in the homolog's structure (Figure 3(c)). Yeast importin Karyopherin α is an adaptor molecule that binds cargo proteins containing nuclear localization signals (NLSs) (Conti & Kuriyan, 2000). The NLSs lack a specific consensus sequence and fall into two classes: monopartite and bipartite NLSs, containing one or two stretches of basic amino acids, respectively. These NLSs are both recognized by the same receptor, made up by a heterodimer of α and β subunits (Fontes, Teh, & Kobe, 2000). The NLS binding site is located on importin-α while importin-β is responsible for binding to the nucleopore complex and translocation of the cargo through the pore (Fontes et al., 2000). Most of the residues involved in the interactions of NLSs with importin-α (PDB code: 1EJL) are conserved or conservatively substituted in the Hunt2 model (Table 1), thus reinforcing the hypothesis that Hunt2 can be one of the Htt domains involved in protein trafficking.

3.2.3. Hunt3

Among the predicted ordered domains of Htt, Hunt3 is the largest one (1118 amino acids). Thus, for the known Robetta sequence length limitations (see Section 2.2), a structural model of this domain was obtained only using I-TASSER. The model, shown in Figure 4(a), displays the α-helical structure typical of HEAT repeats. The Hunt3
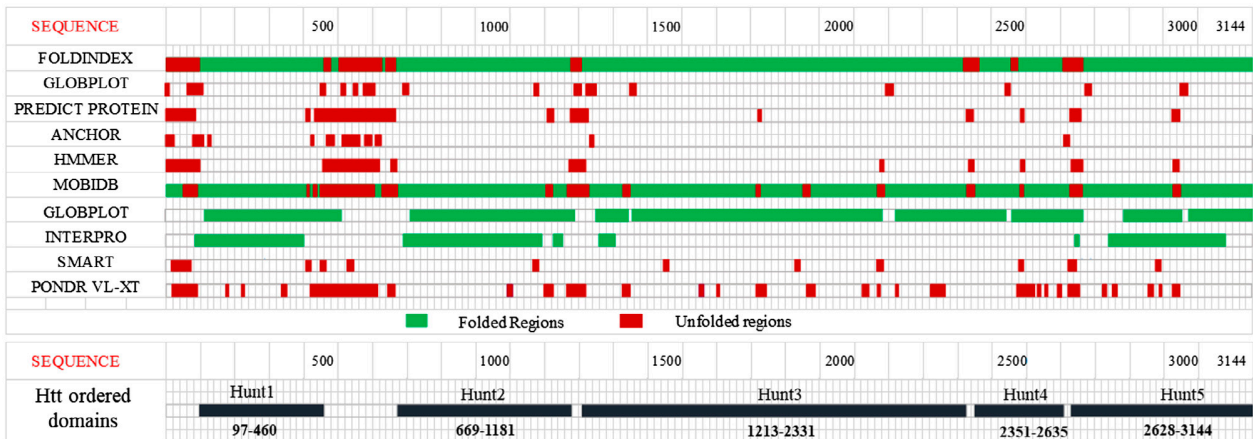
Figure 1. Predicted ordered (green) and disordered (red) regions of Htt (see Methods Section for details). The Htt domains predicted as ordered by consensus among the different methods are displayed in the bottom part of the figure as black bars.

model displays structural similarity with Kap121p (PDB code: 3W3T), a yeast karyopherin that mediates the nuclear import of several cargoes (Kobayashi & Matsuura, 2013) (Supplementary materials Table S5). As shown in Figure 4(b), the Hunt3 structural model matches very well with the Kap121p three-dimensional structure.

Since Kap121p is a Ran-GTP-binding protein and changes its conformation upon binding its partner, the structure of Kap121p bound to Ran-GTP (PDB code: 3W3Z) was used as a template to obtain a model of Hunt3 in the same conformation (Figure 4(c)). A structural comparison between Hunt3 model and the template used is depicted in Figure 4(d). Karyopherin Kap121p (also known as Pse1p) is responsible for the nuclear import of several cargoes (e.g., cell cycle regulators, transcription factors, and ribosomal proteins) by binding to nonconventional NLSs rich in lysine residues (Kobayashi & Matsuura, 2013). The direction of transport, as for other karyopherins, is the result of specific interactions with cargo proteins and the GTPase Ran. Kap121p, forms a right-handed super-helical structure and binds RanGTP to the inner surface of its *N*-terminal half through a conformational change involving the opening of the super-helix (Kobayashi & Matsuura, 2013). A comparison between the residues of Kap121p involved in RanGTP binding and the horthologous residues present in the Hunt3 structural model is shown in Table 2 and Figure 5(a). Although not all of the residues in the Kap121p interaction region are conserved in Hunt3, literature data indicate that Htt interacts with Rab11a and several other Rab GTPases (Elias, McGuire, Yu, & Humbert, 2015; Hattula & Peränen, 2000; Pal, Severin, Lommer, Shevchenko, & Zerial, 2006), which are members of the Ras superfamily of monomeric G proteins. In order to probe the hypothesis that Hunt3 is the Htt

domain involved in Rab11a binding, the 'open' model of Hunt3 was used to perform 'blind' molecular docking simulations with Rab11a three-dimensional structure (PDB code: 1OIW) (Figure 5(b)). Interestingly, in the best-ranked ZDOCK complex (Figure 5(b)), Rab11a is located in the same region where Kap121p binds to RanGTP (Figure 5(c)). The Hunt3-Rab11a complex obtained by molecular docking (Figure 5(b)) supports the hypothesis that Hunt3 could be the putative Rab-binding domain of Htt. It must be noted that Rab11a displays several differences with respect to RanGTP both at the sequence and structure levels (Supplementary materials Figures S1 and S2). Many of the residues involved in the binding of RanGTP to Kap121p are not conserved in Rab11a (Table 3). However, the interaction between Hunt3 and Rab11a could occur, at least in part, through different residues with respect to those involved in the binding of Kap121p to RanGTP.

These results highlight the possibility that Hunt3 is the Htt domain that recognizes GTP-binding proteins. No data concerning the interaction of Htt with GTP-binding proteins other than Rab11 are available in the literature. However, a number of Htt-interacting GTP-binding proteins have been identified in this work (see Section 3.3 and Supplementary materials Table S7). Among these, particularly intriguing is GNAO1, the G_o subunit α of a Guanine nucleotide-binding protein. Interestingly, heterozygous mutations in the GNAO1 gene (Kehrl et al., 2014; Marcé-Grau et al., 2016; Nakamura et al., 2013; Talvik et al., 2015) are the cause of a severe neurodevelopmental disorder, featuring early infantile seizures, profound cognitive dysfunction and, occasionally, movement disorders (early infantile epileptic encephalopathy-17) (Marcé-Grau et al., 2016), the latter symptoms being similar to those observed in HD.

(a)

(b)

(c)
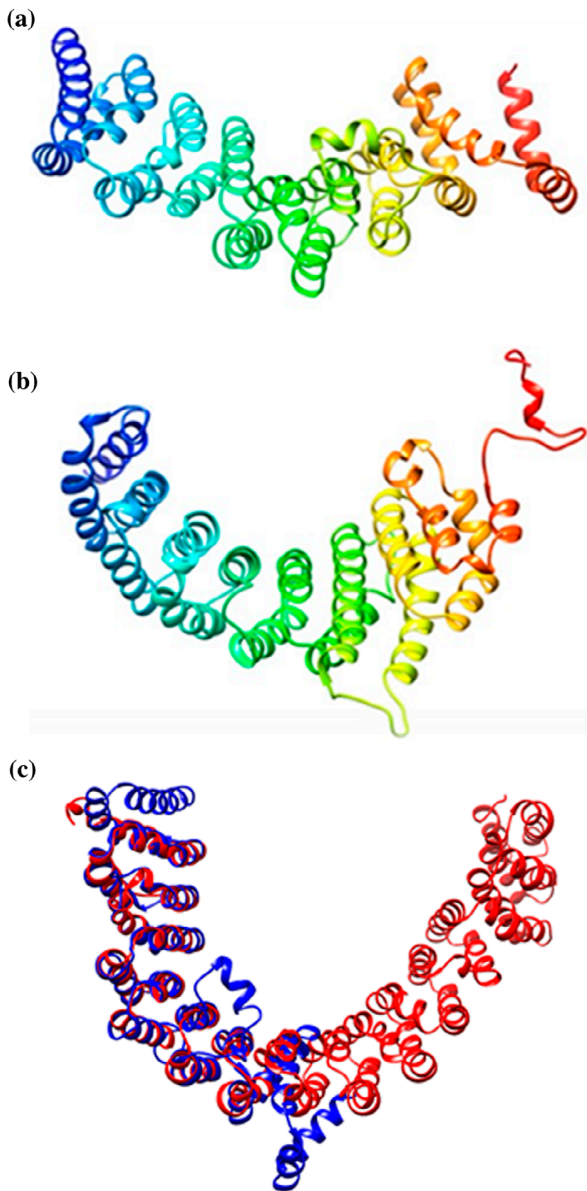
Figure 2.   Schematic representation of Hunt1 models obtained with I-TASSER (a) and Robetta (b), and superimposition between the Hunt1 model obtained with I-TASSER (blue) and PP2A (Protein phosphatase 2A) A subunit (red; PDB code: 1B3U) (c).

In order to probe the hypothesis that GNAO1 could interact with the Hunt3 domain of Htt, a highly reliable homology model of human GNAO1 has been built using the three-dimensional structure of the closely related mouse protein as a template (PDB code 3C7K; 98% sequence identity with the human counterpart; Supplementary materials Figure S3). The model, shown in Supplementary materials Figure S4, has then been used to perform docking simulations against the Hunt3 structural model using ZDOCK. Interestingly, in the best-ranked

(a)

(b)

(c)
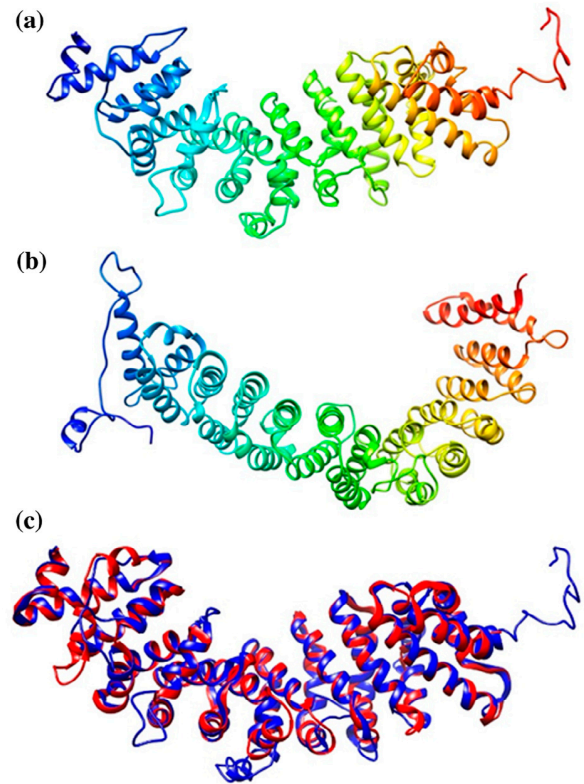
Figure 3.   Structural models of Hunt2, obtained with I-TASSER (a) and Robetta (b), and structural comparison between Hunt2 model obtained with I-TASSER (blue) and the yeast karyopherin α (importin-α; red; PDB code: 1EE5) (c).

Table 1.   Comparison between the residues of importin α involved in the binding of NLSs and orthologous residues of Hunt2. Residues conserved or conservatively substituted in the Hunt2 model are highlighted in red.

Importin-α	Hunt2
Asn146	Thr108
Asn188	Asn166
Asn35	Glu208
Asn319	Asn322
Asn361	Thr368
Asn403	Asn429
Glu180	Lys158
Glu266	Asp254
Glu354	Glu361
Glu396	Asp416
Asp192	–
Asp270	Val260
Asp325	Asp329

ZDOCK complex (Figure 6(a)), the Gα$_o$ subunit is located in the same region of Hunt3 in which Rab11a has been predicted to bind (Figure 5(c)).

The GNAO1 mutations causing epileptic encephalopathy-17, listed in Supplementary materials Table S6, have been mapped onto the predicted
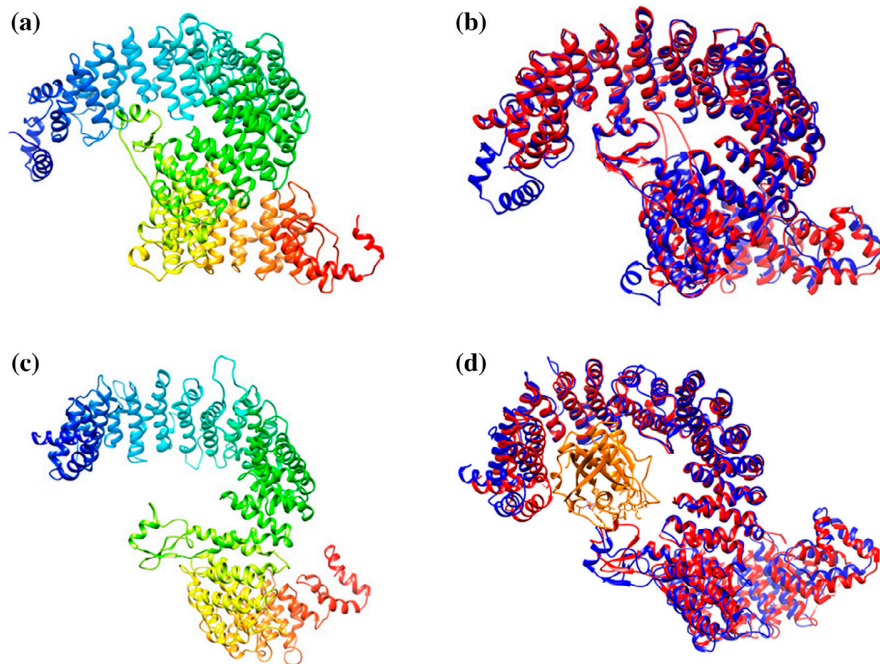
Figure 4.   Structural model of Hunt3 obtained with I-TASSER (a) and superimposition between this model (blue) and the karyopherin Kap121p(red; PDB code: 3W3T) (b). The Karyopherin Kap121p-Ran-GTP complex (PDB code: 3W3Z) was used as a template to obtain a model of Hunt3 in the same conformation (c). Superimposition between the Hunt3 model (blue) and Kap121p-RanGTP complex, used as template (red; PDB code: 3W3Z); RanGTP structure is colored in orange (d).

Hunt3-Gα$_o$ complex (Figure 6(b)). Interestingly, some GNAO1 mutations are located in the putative Gα$_o$-Hunt3 interface. In particular, the mutation Leu199Pro affects two areas (187–196; 236–241; Figure 6(c)) near the interface. Thus, Gα$_o$ subunit mutations could impair the interaction with Htt, thus providing an explanation for the similar symptoms observed in HD and epileptic encephalopathy and a link between these two pathologies.

Another interesting result of this work is the identification of Hunt3 as an additional region of Htt formed by HEAT repeats. Palidwor et al. (2009) identified two *N*-terminal domains of repeats (H1 from residue 114 to 413 and H2 from residue 672 to 969) and suggested the existence of a *C*-terminal domain of repeats (H3 from residue 2667 to 2938). Recently, Saudou and Humbert (Saudou & Humbert, 2016) have hypothesized the presence of HEAT repeats in the central regions of Htt (residues 1289–1710 and 2175–2475). The present work confirms the likely presence of HEAT repeats in these regions, but also between residues 1710 and 2175, part of the Hunt3 domain (residues 1213–2331; Figure 1).

3.2.4.   *Hunt4*

For the Hunt4 domain no well-structured models were obtained either with I-TASSER or with Rosetta. This prevented to hypothesize any function for this domain coherent with literature data on Htt functions. This negative result could be due to a number of different causes as the Hunt4 domain could become structured only upon intra- or intermolecular interactions with other Htt molecules or with a yet unknown molecular partner, which would lead to failure of the modeling approaches. Indeed, PONDR, at variance with the other methods, predicts that approximately half of the fourth huntingtin domain (Hunt4) could be disordered (Figure 1). This is the likely reason why both Rosetta and I-TASSER failed to provide a well-structured model of this domain.

3.2.5.   *Hunt5*

Both the structural models of Hunt5 obtained with I-TASSER and Robetta display the α-helical structure typical of HEAT repeats (Figure 7(a) and (b)). Further, their overall fold is very similar, reinforcing the reliability of the prediction. Among the best structural homologs of the Hunt5 structural model obtained with I-TASSER there is the PP2A A-subunit (Supplementary materials Table S7), as already found for the Hunt1 domain. Figure 7(c) displays the superimposition between the Hunt5 structural model obtained with I-TASSER and the three-dimensional structure of the PP2A A-subunit (PDB code: 1B3U). At variance with what has been observed for Hunt1, the Hunt5 model covers the entire structure of the PP2A A-subunit (Figure 7(c)).

Table 2. Comparison between the residues of Kap121p involved in the binding of RanGTP and the orthologous residues of the Hunt3 structural model. Residues conserved or conservatively substituted are highlighted in red.

Kap121p	Hunt3
Glu644	Gly687
Asn648	Leu691
Gln651	Asp694
Tyr652	Tyr695
Glu244	Asp240
Glu248	Gln244
Glu287	Val292
Glu340	Ser363
Gln350	Gln373
Lys121	Thr108
Thr283	Phe288
Glu251	Gln247
His347	Gln370
Asp346	Ile369
Arg54	Met377
Leu357	Leu380
Glu31	Leu29
Asn24	Ttr21
Arg27	Pro24
Pro22	Ser19
Lys73	Ala71
His120	Lys107
Arg161	Thr154
Leu66	Leu64
Thr62	Ala60

The structure prediction here presented indicates that two of the putative ordered domains of Htt (Hunt1 and Hunt5) display a significant structural similarity with the subunit A of PP2A. This is one of the major cellular serine/threonine phosphatases, which is involved in the regulation of a number of fundamental cellular activities (Sangodkar et al., 2016). The association of a catalytic subunit (PP2AC) with a structural subunit (PR65/A) and variable regulatory B subunits generates several structurally and functionally diverse hetero-trimeric forms of PP2A. Literature data (Palidwor et al., 2009) indicate that many Htt functions depend on a modular structure in which different domains can associate in a transient mode depending on effectors and post-translation modifications. Therefore, Hunt5, whose structural model matches with the entire structure of PP2A subunit A, could interact with regulatory and catalytic subunits of this PP2A to form a holo-enzyme with phosphatase activity. Indeed, in the present work, the PP2A catalytic subunit PP2AC is identified as one of the Htt interactors (see Section 3.3 and Supplementary materials Table S8). Of note, Htt may directly modify the phosphorylation state of other proteins. In this regard it's interesting to note that, in the presence of mutant Htt, a significant decreased level of Tau phosphatases (PP1, PP2A, and PP2B) is detected within cells, promoting Tau hyperphosphorylation (p-Tau) (Blum et al., 2014).
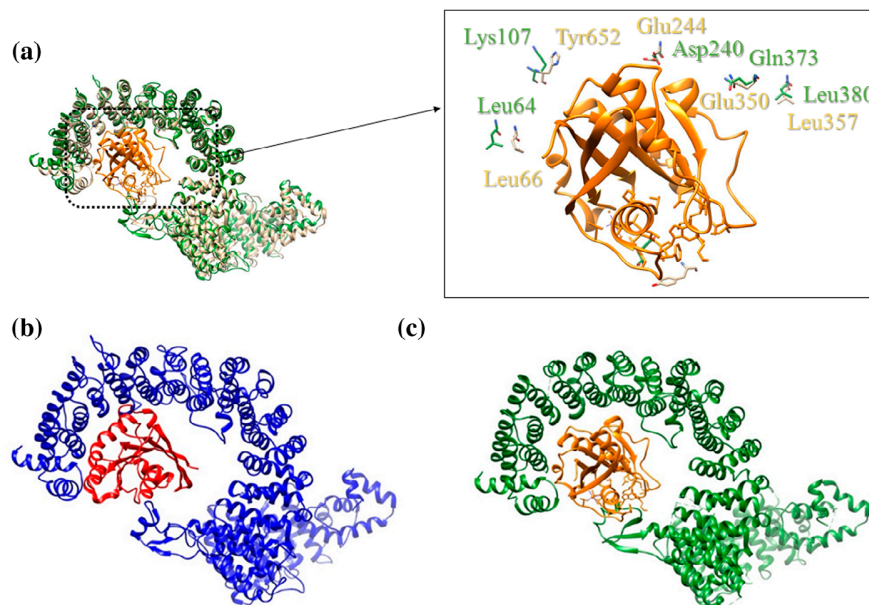
Figure 5.   Superimposition between the Hunt3 structural model obtained with Modeller (in green) and Kap121p (in beige) in complex with RanGTP (in orange) (a). The right panel shows a detail of the recognition region highlighting Hunt3 residues conserved or conservatively substituted with respect to Kap121p. Comparison between the best ranked Hunt3-Rab11a ZDock complex (b) and the three-dimensional structure of the Kap121p-RanGTP complex (c). Kap121p is depicted in green, RanGTP in orange, Hunt3 in blue and Rab11a (PDB code: 1OIW) in red.

Table 3. Residues involved in the binding of RanGTP to Kap121p and the orthologous residues of Rab11a. Residues highlighted in red are conserved or conservatively substituted.

RanGTP	Rab11a
Lys71	Arg72
Lys38	Glu39
Tyr39	Ser40
Gly20	Gly21
Arg140	Ala143
Lys142	Lys145
Lys141	Glu144
Lys159	Ala163
Tyr79	Tyr80
Asp77	Ser78
Val47	Phe48
Asn143	Asn146
Glu113	Asp114
Ile81	Arg82

3.3. Prediction of the Htt interactome

Htt is involved in many different cellular functions in association with other proteins (Saudou & Humbert, 2016). Thus, the analysis of the Htt interactome may be an important step also in view of the structural information on Htt domains here reported.

Here, four different databases that combine predicted and experimentally determined protein–protein interactions (PPIs), have been interrogated to obtain a complete and reliable list of Htt interactors. A total of 979 predicted Htt-interacting proteins have been identified (Supplementary material Table S8). Among these Htt interactors, 575 have been predicted by PrePPI. To limit

the analysis to the most reliable Htt interactors, a consensus has been built between the different databases, reducing the number of interactors to the 80 proteins found by all the methods (Figure 8 and Table 4). In agreement with Htt involvement in many and diverse cellular processes, these interactors belong to many functional classes. The most relevant interactors in relation to the biological roles hypothesized for each Htt domain, are discussed below.

Eleven out of the 80 Htt interactors are proteins involved in gene expression, chromatin remodeling, and DNA stability and repair. Among these, CHD3, a component of the histone deacetylase NuRD complex, is one of the importin α-specific cargoes (Hügel et al., 2014). CBP (CREB-binding protein) another protein of the same class, is a histone acetylase, and its sequestration mediated by mutant Htt has been reported to play a significant role in the pathogenesis of HD (Choi et al., 2012). CBP can also acetylate proteins not directly involved in transcription, such as Rch1 (human importin-α) and importin-α7 (Bannister, Miska, Görlich, & Kouzarides, 2000).

Another big group of interactors (23 proteins) includes proteins involved in RNA transcription and processing, including many transcription factors. This is coherent with Htt functions because the polyQ region is known to form a polar zipper structure that may bind transcription factors containing a polyQ region (Zuccato et al., 2010).

Highly relevant for Htt physiological and pathological function are also the interactors involved in post-translational modifications and protein degradation
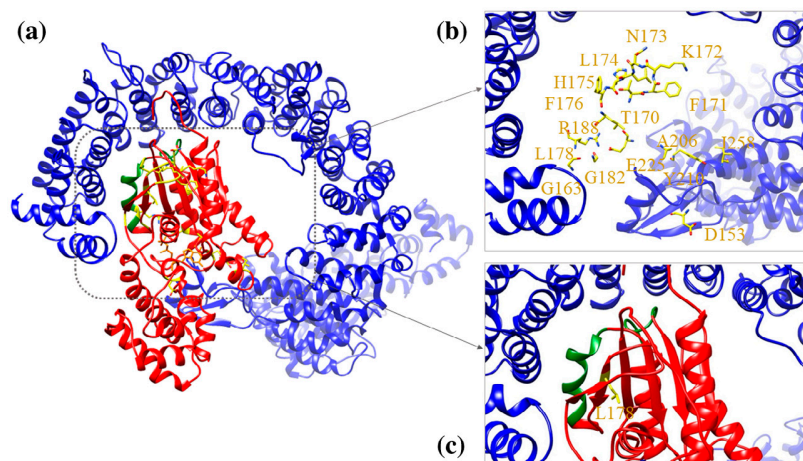
Figure 6. Structural model of the best-scoring complex between Hunt3 (in blue) and Gαo subunit (in red, GDP in orange) obtained by docking simulations (a). Residues mutated in the GNAO1 variants causing severe neurodevelopmental disorder (in yellow) are mapped onto the putative Hunt3-Gαo complex (b). The regions in green (187–196, 236–241) correspond to the areas affected by the Leu199Pro mutation (see text for details). Detail of the Hunt3-Gαo complex showing in yellow the position undergoing to the Leu199Pro mutation (corresponding to residue 178 in the structural model) and in green the regions (187–196, 236–241) affected by the mutation (c).

(a)

(b)

(c)
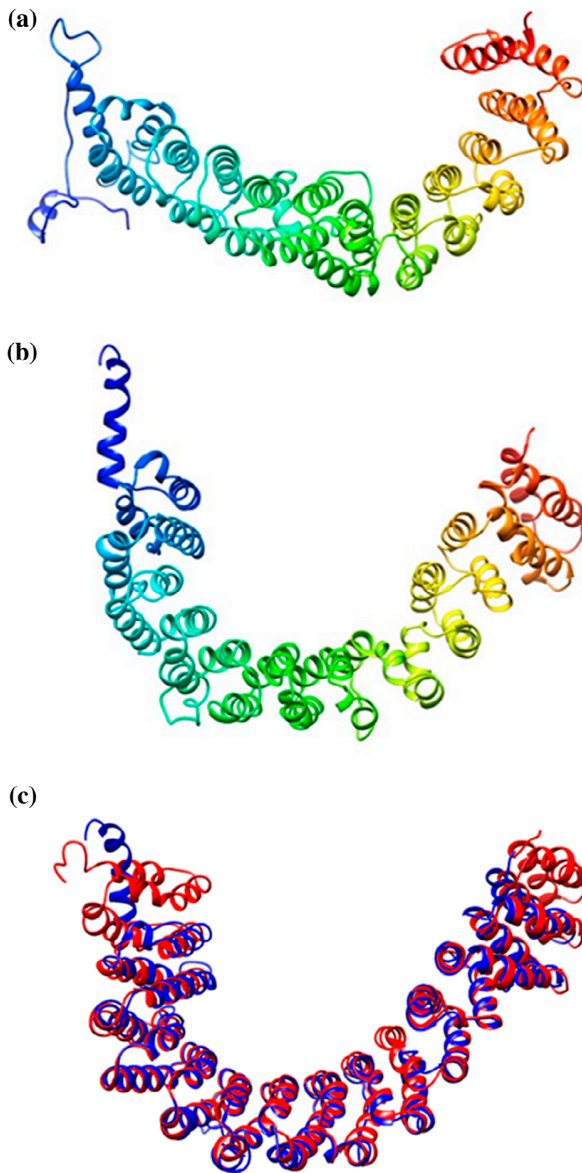
Figure 7.   Structural models of Hunt5 obtained with I-TASSER (a) and Robetta (b). Superimposition between the structural model of Hunt5 obtained with I-TASSER (blue) and the three-dimensional structure of PP2A A-subunit (red; PDB code: 1B3U) (c).

Figure 8.   Venn diagram displaying the number and percentage of Htt interactors identified by Hippie, Biogrid, IntAct and PrePPi.

(11 proteins). Indeed, Htt undergoes multiple post-translational modifications, including phosphorylation, acetylation, palmitoylation, ubiquitination, and sumoylation (Ratovitski et al., 2017; Saudou & Humbert, 2016). Ubiquitination controls the stability, function, and intracellular localization of Htt, contributing to maintain Htt homeostasis in the cells (Saudou & Humbert, 2016). Instead, phosphorylation seems to confer neuroprotective properties to wild-type Htt and is responsible for Htt-mediated transport of vesicles in neurons (Colin et al., 2008; Watkin et al., 2014). In addition, as already stated above, Htt may also modify the phosphorylation state of other proteins as mutated Htt promotes Tau hyper-phosphorylation (p-Tau) (Blum et al., 2014). Htt sumoylation modulates its subcellular localization, activity, and stability while the palmitoylation regulates vesicular trafficking (Sanders & Hayden, 2015; Sarge & Park-Sarge, 2011; Yanai et al., 2006). Finally, acetylation is required to target the protein to the macroautophagy pathway. From this viewpoint, the most interesting interactor is ZDHHC17, also known as HIP14 (Htt-interacting protein-14), which catalyzes the post-translational addition of palmitate to proteins, including Htt, and shows reduced palmitoylation of mutant Htt, leading to increased formation and toxicity of mutated Htt inclusions (Sanders, Mui, Sutton, & Hayden, 2014).

Given the Htt structure/function predictions obtained in this work, an interesting group of interactors is also the one including proteins involved in the regulation and association to cytoskeleton, in signal transduction, in endocytosis, and in vesicle trafficking. In fact, Htt regulates axonal transport by participating in the assembly of the motor complex on microtubules. In addition, it interacts with the molecular motor machinery, either directly with dynein or indirectly, through HAP1 (Huntingtin-associated protein 1), with the p150 subunit of dynactin, thus controlling the transport of organelles in both the anterograde and retrograde directions within neurons (Saudou & Humbert, 2016). Htt interacts with several proteins involved in endocytosis, including HIP1 (huntingtin-interacting protein 1), that participates in clathrin-mediated endocytosis (Saudou & Humbert, 2016). In agreement with the function hypothesized for Htt Hunt3

Table 4. List of the 80 Htt interactors identified by all methods. The gene and protein names are reported with the corresponding functions.

Gene name	Protein name	Function
AP2A2	AP-2 complex subunit alpha-2	Component of the adaptor protein complex 2 (AP-2). It is involved in protein transport via transport vesicles in different membrane traffic pathways
CCDC126	Coiled-coil domain-containing protein 126	Protein N-linked glycosylation and alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase activity
CHD3	Chromodomain-helicase-DNA-binding protein 3	Component of the histone deacetylase NuRD complex which participates in the remodeling of chromatin by deacetylating histones
CREBBP	CREB-binding protein	Acetylates histones, giving a specific tag for transcriptional activation
CRMP1	Dihydropyrimidinase-related protein 1	Necessary for signaling by class 3 semaphorins and subsequent remodeling of the cytoskeleton
CTNNB1	Catenin beta-1	Key downstream component of the canonical Wnt signaling pathway
DNALI1	Axonemal dynein light intermediate polypeptide 1	May play a dynamic role in flagellar motility
DPPA4	Developmental pluripotency-associated protein 4	May be involved in the maintenance of active epigenetic status of target genes
DUSP10	Dual specificity protein phosphatase 10	Protein phosphatase involved in the inactivation of MAP kinases
ECH1	Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase, mitochondrial	Isomerization of 3-trans,5-cis-dienoyl-CoA to 2-trans,4-trans-dienoyl-CoA
ERCC6L	DNA excision repair protein ERCC-6-like	DNA helicase that acts as an essential component of the spindle assembly checkpoint
ETV4	ETS translocation variant 4	Transcriptional activator that binds to the enhancer of the adenovirus E1A gene
EVL	Ena/VASP-like protein	EVL enhances actin nucleation and polymerization
FAM71F1	Protein FAM71F1	Unknown function
FEZ1	Fasciculation and elongation protein zeta-1	Component of the network of molecules that regulate cellular morphology and axon guidance machinery
FTL	Ferritin light chain	Important for iron homeostasis
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	It plays a role in glycolysis and nuclear functions
GGA2	ADP-ribosylation factor-binding protein GGA2	Plays a role in protein sorting and trafficking between the trans-Golgi network (TGN) and endosomes
GIT1	ARF GTPase-activating protein GIT1	GTPase-activating protein for the ADP ribosylation factor family
GOLPH3L	Golgi phosphoprotein 3-like	It is required for the process of vesicle budding at the Golgi and anterograde transport to the plasma membrane
GPRASP2	G-protein coupled receptor-associated sorting protein 2	May play a role in regulation of a variety of G-protein coupled receptors
HAP1	Huntingtin-associated protein 1	It is involved in intracellular trafficking and it is proposed to link HTT to motor proteins and/or transport cargos
HEY2	Hairy/enhancer-of-split related with YRPW motif protein 2	Downstream effector of Notch signaling which may be required for cardiovascular development
HEYL	Hairy/enhancer-of-split related with YRPW motif-like protein	Downstream effector of Notch signaling which may be required for cardiovascular development
HIP1	Huntingtin-interacting protein 1	It plays a role in clathrin-mediated endocytosis and trafficking
HMG20A	High mobility group protein 20A	Plays a role in neuronal differentiation as chromatin-associated protein
HMGA1	High mobility group protein HMG-I/HMG-Y	It binds preferentially to the minor groove of A+T rich regions in double-stranded DNA and it is involved in the transcription regulation of genes
HOXC11	Homeobox protein Hox-C11	Transcription factor which is part of a developmental regulatory system that provides cells with specific positional identities on the anterior-posterior axis
HOXC4	Homeobox protein Hox-C4	Transcription factor which is part of a developmental regulatory system that provides cells with specific positional identities on the anterior-posterior axis
IFT20	Intraflagellar transport protein 20 homolog	It is involved in ciliary process assembly
IKBKAP	Elongator complex protein 1	It acts as subunit of the RNA polymerase II elongator complex
ING5	Inhibitor of growth protein 5	It may regulate DNA replication and may function as a transcriptional coactivator
MAGEB18	Melanoma-associated antigen B18	It may enhance ubiquitin ligase activity of RING-type zinc finger containing E3 ubiquitin-protein ligases
MAGEB6	Melanoma-associated antigen B6	It code for antigens that are recognized on many human tumors by autologous cytolytic T lymphocytes

(Continued)

Table 4.   (*Continued*).

Gene name	Protein name	Function
MBD1	Methyl-CpG-binding domain protein 1	Transcriptional repressor
MBD4	Methyl-CpG-binding domain protein 4	It is involved in DNA repair
MED21	Mediator of RNA polymerase II transcription subunit 21	Involved in the regulated transcription of nearly all RNA polymerase II-dependent genes
MED31	Mediator of RNA polymerase II transcription subunit 31	Involved in the regulated transcription of nearly all RNA polymerase II-dependent genes
MIPEP	Mitochondrial intermediate peptidase	It cleaves proteins, imported into the mitochondrion, to their mature size
MRFAP1L1	MORF4 family-associated protein 1-like 1	Unknown function
NBR1	Next to BRCA1 gene 1 protein	Acts probably as a receptor for selective autophagosomal degradation of ubiquitinated targets
NCOR1	Nuclear receptor corepressor 1	Mediates transcriptional repression by certain nuclear receptors
NME4	Nucleoside diphosphate kinase, mitochondrial	Major role in the synthesis of nucleoside triphosphates other than ATP
NUPL1	Nucleoporin p58/p45	Component of the nuclear pore complex
OPTN	Optineurin	Plays an important role in the maintenance of the Golgi complex, in membrane trafficking, in exocytosis, through its interaction with myosin VI and Rab8
OSTF1	Osteoclast-stimulating factor 1	Induces bone resorption
PACSIN1	Protein kinase C and casein kinase substrate in neurons protein 1	Plays a role in the reorganization of the microtubule cytoskeleton via its interaction with MAPT
PEX11B	Peroxisomal membrane protein 11B	Involved in peroxisomal proliferation
PFN2	Profilin-2	Binds to actin and affects the structure of the cytoskeleton
PIAS1	E3 SUMO-protein ligase PIAS1	Functions as an E3-type small ubiquitin-like modifier (SUMO) ligase, stabilizing the interaction between UBE2I and the substrate
PIAS4	E3 SUMO-protein ligase PIAS4	Functions as an E3-type small ubiquitin-like modifier (SUMO) ligase, stabilizing the interaction between UBE2I and the substrate
PIK3R1	Phosphatidylinositol 3-kinase regulatory subunit alpha	Acts as an adapter, mediating the association of the p110 catalytic unit to the plasma membrane
PRMT2	Protein arginine N-methyltransferase 2	It methylates the guanidino nitrogens of arginyl residues in proteins such as STAT3, FBL, histone H4
PRPF40A	Pre-mRNA-processing factor 40 homolog A	It binds to WASL/N-WASP and suppresses its translocation from the nucleus to the cytoplasm
PSMB7	Proteasome subunit beta type-7	It is a component of the proteinase complex
PTGES2	Prostaglandin E synthase 2	Isomerase that catalyzes the conversion of PGH2 into the more stable prostaglandin E2
RNF20	E3 ubiquitin-protein ligase BRE1A	It is involved in the pathway protein ubiquitination
RPL4	60S ribosomal protein L4	It is involved in translation and rRNA processing
SETD2	Histone-lysine N-methyltransferase SETD2	It has a histone-lysine N-methyltransferase activity
SH3GL3	Endophilin-A3	Implicated in endocytosis
SP1	Transcription factor Sp1	Transcription factor that can activate or repress transcription in response to physiological and pathological stimuli
TACC1	Transforming acidic coiled-coil-containing protein 1	Likely involved in the processes that promote cell division prior to the formation of differentiated tissues
TAF4	Transcription initiation factor TFIID subunit 4	Part of the TFIID complex, a multimeric protein complex that plays a central role in mediating promoter responses to various activators and repressors
TCERG1	Transcription elongation regulator 1	Transcription factor that binds RNA polymerase II and inhibits the elongation of transcripts from target promoters
TP53	TP53	Acts as a tumor suppressor in many tumor types and it is involved in cell cycle regulation
TRAFD1	TNF receptor-associated factor 2	Regulates activation of NF-kappa-B and JNK and plays a central role in the regulation of cell survival and apoptosis
TTC23	Tetratricopeptide repeat protein 23	Unknown function
UBAC1	Ubiquitin-associated domain-containing protein 1	It is involved in the pathway protein ubiquitination
UBE2E3	Ubiquitin-conjugating enzyme E2 E3	It is involved in the pathway protein ubiquitination
UBE2 K	Ubiquitin-conjugating enzyme E2 K	It is involved in the pathway protein ubiquitination
UPF3A	Regulator of nonsense transcripts 3A	Involved in nonsense-mediated decay (NMD) of mRNAs containing premature stop codons

(*Continued*)

Table 4. (*Continued*).

Gene name	Protein name	Function
UTP14A	U3 small nucleolar RNA-associated protein 14 homolog A	May be required for ribosome biogenesis
XAGE3	X antigen family member 3	Unknown function
XRCC6	X-ray repair cross-complementing protein 6	Single-stranded DNA-dependent ATP-dependent helicase
ZDHHC17	Palmitoyltransferase ZDHHC17	Palmitoyltransferase specific for a subset of neuronal proteins. Involved in protein palmitoylation
ZFC3H1	Zinc finger C3H1 domain-containing protein	RNA processing
ZFYVE19	Abscission/NoCut checkpoint regulator	Key regulator of abscission step in cytokinesis
ZMAT2	Zinc finger matrin-type protein 2	Involved in mRNA splicing
ZNF451	E3 SUMO-protein ligase ZNF451	Protein ligase which has a preference for SUMO2 and SUMO3 and facilitates UBE2I/UBC9-mediated sumoylation of target proteins
ZNF655	Zinc finger protein 655	May be involved in transcriptional regulation

structural domain, Htt also interacts with and activates the GTPase Rab11 that takes part in vesicle recycling during endocytosis (Li et al., 2008). Further, Htt may mediate the transition of cargos between microtubules and actin cytoskeletons through recognition of the Rab8/optineurin/myosinVI complex (Zala et al., 2013). Optineurin is involved in the maintenance of the Golgi complex, membrane trafficking, and exocytosis, through its interaction with myosin VI and Rab8 (Chibalina, Roberts, Arden, Kendrick-Jones, & Buss, 2008). Optineurin plays an important role linking Htt to Rab8 and enhancing the recruitment of Htt to Rab8-positive vesicles (Hattula & Peränen, 2000). Of note, mutant Htt in HD displaces the optineurin/Rab8 complex from the Golgi apparatus, thereby affecting post-Golgi trafficking and lysosomal function (del Toro et al., 2009). GIT1 is a GTPase activating protein that modulates actin polymerization, spine morphology, and synapse formation in neuronal cells (Zhang, Webb, Asmussen, & Horwitz, 2003). It is thought to function in signaling pathways, controlling vesicle trafficking, adhesion, and cytoskeletal organization. GIT1 stimulates mutant Htt aggregation, but the mechanism by which this happens is still unclear. Experimental studies demonstrated that the expression of a *N*-terminally truncated GIT1 fragment induces the formation of membrane vesicles that contain mutant Htt. Furthermore, GIT1 is fragmented in HD but not in healthy brain (Goehler et al., 2004), suggesting that loss of the GIT1 function and/or the accumulation GIT1 fragments may contribute to HD pathogenesis (Goehler et al., 2004). PACSIN1 is a neurospecific phosphoprotein that plays a central role in synaptic vesicle recycling and clathrin-mediated endocytosis. Along with Htt, it plays a pivotal role in the transmission of impulses from one neuron to another (Gopalakrishnan, Jethi, Kalsi, & Purohit, 2016). Mutated Htt interacts with PACSIN1 and forms aggregates, disturbing the normal distribution of PACSIN1 in the cell (Gopalakrishnan et al., 2016).

Finally, the list of common interactors includes also HMG20A, a protein that plays a role in neuronal differentiation as chromatin-associated protein (Artegiani et al., 2010). The nature of its interaction with Htt is not clear but the latter could be involved in its nuclear export.

The presence, among the 80 common interactors, of proteins involved in transport and post-translational modifications is in line with the functions proposed in this work for the Htt-ordered domains. Indeed, Htt domains Hunt2 and Hunt3 are predicted to be an importin-α and a Karyopherin α homolog, respectively, involved in the transport of several cargo proteins. Further, Hunt1 and Hunt5 may be involved in the phosphorylation state of other proteins.

Since the interaction between CHD3, CBP and p53 with different importin-α has been observed and because Hunt2 has been predicted to have an importin-α-like fold, CHD3, CBP and p53 could interact with Htt through this domain.

Interestingly, HMG20A nuclear export is mediated by importin-13 (Fatima et al., 1864), which has a fold similar to that of Kap121p. Thus, the interaction between HMG20 and Htt could occur through the Htt Hunt3 domain.

An interesting link between Htt interactors and the Htt functions predicted, based on modeling results, is that PP2A mediates DNA damage-induced dephosphorylation of p53 and, in turn, the latter protein is known to interact with Htt (Li, Cai, Shouse, Piluso, & Liu, 2007). A suggestive hypothesis is that the Hunt2 domain could mediate p53 nuclear localization, whereas the Hunt5 domain could be involved in its dephosphorylation in response to DNA-damage.

Another protein, not included in the list of the common interactors, as it is identified only by prePPI, is caspase-2 (CASP2) that contains a classical NLS at the *C*-terminus of the prodomain and is recognized by the importin-α/β heterodimer (Baliga et al., 2003). Furthermore, Htt is cleaved by caspase-2 at residue 552 (Zuccato et al., 2010), which is located in the disordered region between Hunt1 and Hunt2 (Figure 1). In this regard, it can be speculated that CASP2 recruitment could occur through interaction with the Hunt2 domain of Htt.

4. Conclusions

Here, structural models of Htt-ordered domains have been proposed for the first time, providing a putative function for four out of the five predicted ordered domains. The structure prediction analysis presented in this work indicates that two of the putative ordered domains of Htt (Hunt1 and Hunt5) display a significant structural similarity with the subunit A of the protein phosphatase 2A, the most important serine/threonine phosphatase. Therefore, Hunt5, whose structural model displays structural homology with the entire PP2A subunit A, could interact with regulatory and catalytic subunits of PP2A to form a holo-enzyme with phosphatase activity. This is supported by the presence of PP2AC among the identified Htt interactors.

Hunt2 and Hunt3 models display structural similarity with two yeast importins, importin-α, and Karyopherin-α, that mediate nuclear import of protein cargoes by interacting with GTP-binding proteins. The observation that most of the residues involved in the interactions of NLS-mimicking peptides with importin-α are conserved or conservatively substituted in the Hunt2 model, strongly supports the hypothesis that Hunt2 could be one of the Htt domains involved in protein trafficking.

It is known from previous studies that Htt is involved in vesicular trafficking through the interaction with monomeric G proteins of the Rab family (Hattula & Peränen, 2000; Kobayashi & Matsuura, 2013; Pal et al., 2006). This observation, together with the results of molecular docking analysis between Hunt3 and Rab11a, presented in this work, strengthens the hypothesis that Hunt3 could be the Htt domain responsible for the interaction with Rab proteins. Further, also the guanine nucleotide-binding protein Gα$_o$ has been detected among the Htt interactors. Molecular docking simulations between Hunt3 and the Gα$_o$ protein have shown that Gα$_o$ locates in the same binding region of Rab11a. These results suggest that mutations of the Gα$_o$ subunit could impair the interaction with Htt, thus providing an explanation for the similar symptoms observed in HD and epileptic encephalopathy.

The functional diversity of the Htt interactome here reported is in line with the functions proposed for the Htt-ordered domains, especially as far as vescicular transport, nuclear targeting, and post-translational modifications are concerned, and confirms the multiplicity and complexity of Htt functions. One or more of these key cellular functions may be disrupted by abnormal interactions of expanded Htt, highlighting proteins that may be involved in HD insurgence and that could serve as therapeutic agents and/or targets.

Disclosure statement

Supplementary material

References

Arndt, J. R., Chaibva, M., & Legleiter, J. (2015). The emerging role of the first 17 amino acids of huntingtin in Huntington's disease. *Biomolecular Concepts, 6*, 33–46. doi:10.1515/bmc-2015-0001

Artegiani, B., Labbaye, C., Sferra, A., Quaranta, M. T., Torreri, P., Macchia, G., … Macioce, P. (2010). The interaction with HMG20a/b proteins suggests a potential role for beta-dystrobrevin in neuronal differentiation. *Journal of Biological Chemistry, 285*, 24740–24750. doi:10.1074/jbc.M109.090654

Atwal, R. S., Xia, J., Pinchev, D., Taylor, J., Epand, R. M., & Truant, R. (2007). Huntingtin has a membrane association signal that can modulate huntingtin aggregation, nuclear entry and toxicity. *Human Molecular Genetics, 16*, 2600–2615. doi:10.1093/hmg/ddm217

Baliga, B. C., Colussi, P. A., Read, S. H., Dias, M. M., Jans, D. A., & Kumar, S. (2003). Role of prodomain in importin-mediated nuclear localization and activation of caspase-2. *Journal of Biological Chemistry, 278*, 4899–4905. doi:10.1074/jbc.M211512200

Bannister, A. J., Miska, E. A., Görlich, D., & Kouzarides, T. (2000). Acetylation of importin-alpha nuclear import factors by CBP/p300. *Current Biology, 10*, 467–470.

Blum, D., Herrera, F., Francelle, L., Mendes, T., Basquin, M., Obriot, H., … Outeiro, T. F. (2014). Mutant huntingtin alters Tau phosphorylation and subcellular distribution. *Human Molecular Genetics, 24*, 76–85. doi:10.1093/hmg/ddu421

Cattaneo, E., Zuccato, C., & Tartari, M. (2005). Normal huntingtin function: An alternative approach to huntington's disease *Nature Reviews Neuroscience, 6*, 919–930. doi:10.1038/nrn1806

Chatr-Aryamontri, A., Breitkreutz, B., Heinicke, S., Boucher, L., Winter, A., Stark, C., … Tyers, M. (2013). The Bio-GRID interaction database: 2013 update. *Nucleic Acids Research, 41*, D816–D823. doi:10.1093/nar/gks1158

Chibalina, M. V., Roberts, R. C., Arden, S. D., Kendrick-Jones, J., & Buss, F. (2008). Rab8-optineurin-myosin VI: Analysis of interactions and functions in the secretory pathway. *Methods in Enzymology, 438*, 11–24. doi:10.1016/S0076-6879(07)38002-6

Chivian, D., Kim, D. E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., … Baker, D. (2003). Automated prediction of CASP-5 structures using the Robetta server. *Proteins: Structure, Function, and Genetics, 53*, 524–533. doi:10.1002/prot.10529

Choi, Y. J., Kim, S. I., Lee, J. W., Kwon, Y. S., Lee, H. J., Kim, S. S., & Chun, W. (2012). Suppression of aggregate formation of mutant huntingtin potentiates CREB-binding protein sequestration and apoptotic cell death. *Molecular and Cellular Neuroscience, 49*, 127–137. doi:10.1016/j.mcn.2011.11.003

Colin, E., Zala, D., Liot, G., Rangone, H., Borrell-Pagès, M., Li, X. J., … Humbert, S. (2008). Huntingtin phosphorylation acts as a molecular switch for anterograde/retrograde transport in neurons. *The EMBO Journal, 27*, 2124–2134. doi:10.1038/emboj.2008.133

Conti, E., & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin α. *Structure, 8*, 329–338.

del Toro, D., Alberch, J., Lazaro-Dieguez, F., Martin-Ibanez, R., Xifro, X., Egea, G., & Canals, J. M. (2009). Mutant Huntingtin impairs post-golgi trafficking to lysosomes by delocalizing optineurin/Rab8 complex from the Golgi apparatus. *Molecular Biology of the Cell, 20*, 1478–1492. doi:10.1091/mbc.E08-07-0726

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005a). IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics, 21*, 3433–3434. doi:10.1093/bioinformatics/bti541

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005b). The pairwise energy content estimated from amino acid composition discriminates folded and intrinsically unstructured proteins. *Journal of Molecular Biology, 347*, 827–839. doi:10.1016/j.jmb.2005.01.071

Dosztányi, Z., Mészáros, B., & Simon, I. (2009). Anchor: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics, 25*, 2745–2746. doi:10.1093/bioinformatics/btp518

Dunbrack, R. L., Jr, & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science, 6*, 1661–1681. doi:10.1002/pro.5560060807

El-Daher, M. T., Hangen, E., Bruyère, J., Poizat, G., Al-Ramahi, I., Pardo, R., … Saudou, F. (2015). Huntingtin proteolysis releases non-polyQ fragments that cause toxicity through dynamin 1dysregulation. *The EMBO Journal, 34*, 2255–2271. doi:10.15252/embj.201490808

Elias, S., McGuire, J. R., Yu, H., & Humbert, S. (2015). Huntingtin is required for epithelial polarity through RAB11A-mediated apical trafficking of PAR3-aPKC. *PLOS Biology, 13*, e1002142. doi:10.1371/journal.pbio.1002142

Fatima, S., Wagstaff, K. M., Lieu, K. G., Davies, R. G., Tanaka, S. S., Yamaguchi, Y. L., … Jans, D. A. (2017). Interactome of the inhibitory isoform of the nuclear transporter Importin 13. *Biochimica et Biophysica Acta, 1864*, 546–561. doi:10.1016/j.bbamcr.2016.12.017

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research, 42*, D222–D230. doi:10.1093/nar/gkt1223

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., … Eddy, S. R. (2015). HMMER web server: 2015 update. *Nucleic Acids Research, 43*, W30–W38. doi:10.1093/nar/gkv397

Fiser, A., Do, R. K. G., & Sali, A. (2000). Modeling of loops in protein structures. *Protein Science, 9*, 1753–1773. doi:10.1110/ps.9.9.1753

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., … Searle, S. M. J. (2014). Ensembl 2014. *Nucleic Acids Research, 42*, D749–D755. doi:10.1093/nar/gkt1196

Fontes, M. R., Teh, T., & Kobe, B. (2000). Structural basis of recognition of monopartite and bipartite nuclear localization sequences by mammalian importin-alpha. *Journal of Molecular Biology, 297*, 1183–1194. doi:10.1006/jmbi.2000.3642

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., … Jensen, L. J. (2013). STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Research, 41*, D808–D815. doi:10.1093/nar/gks1094

Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., … Wanker, E. E. (2004). A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Molecular Cell, 15*, 853–865. doi:10.1016/j.molcel.2004.09.016

Gopalakrishnan, C., Jethi, S., Kalsi, N., & Purohit, R. (2016). Biophysical aspect of huntingtin protein during polyQ: An *in silico* insight. *Cell Biochemistry and Biophysics, 74*, 129–139. doi:10.1007/s12013-016-0728-7

Hattula, K., & Peränen, J. (2000). FIP-2, a coiled-coil protein, links Huntingtin to Rab8 and modulates cellular morphogenesis. *Current Biology, 10*, 1603–1606.

Hirschman, L., Burns, G. A., Krallinger, M., Arighi, C., Cohenn, K. B., Valencia, A., … Winter, A. G. (2012). Text mining for the biocuration workflow. *Database, 2012*, bas020. doi:10.1093/database/bas020

Hügel, S., Depping, R., Dittmar, G., Rother, F., Cabot, R., Sury, M. D., … Bader, M. (2014). Identification of importin α 7 specific transport cargoes using a proteomic screening approach. *Molecular & Cellular Proteomics, 13*, 1286–1298. doi:10.1074/mcp.M112.026856

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., … Yong, S. Y. (2012). InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Research, 40*, D306–D312. doi:10.1093/nar/gkr948

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology, 338*, 1027–1036. doi:10.1016/j.jmb.2004.03.016

Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry, 49*, 2987–2998. doi:10.1021/bi902153 g

Kehrl, J. M., Sahaya, K., Dalton, H. M., Charbeneau, R. A., Kohut, K. T., Gilbert, K., … Neubig, R. R. (2014). Gain-of-function mutation in Gnao1: A murine model of epileptiform encephalopathy (EIEE17)? *Mammalian Genome, 25*, 202–210. doi:10.1007/s00335-014-9509-z

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., … Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research, 40*, D841–D846. doi:10.1093/nar/gkr1088

Kim, M. W., Chelliah, Y., Kim, S. W., Otwinowski, Z., & Bezprozvanny, I. (2009). Secondary structure of huntingtin amino-terminal region. *Structure, 17*, 1205–1212. doi:10.1016/j.str.2009.08.002

Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research, 32*, W526–W531. doi:10.1093/nar/gkh468

Kobayashi, J., & Matsuura, Y. (2013). Structural basis for cell-cycle-dependent nuclear import mediated by the karyopherin kap121p. *Journal of Molecular Biology, 425*, 1852–1868. doi:10.1016/j.jmb.2013.02.035

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden markov models in computational biology. applications to protein modeling. *Journal of Molecular Biology, 235*, 1501–1531. doi:10.1006/jmbi.1994.1104

Letunic, I., Doerks, T., & Bork, P. (2014). SMART: Recent updates, new developments and status in 2015. *Nucleic Acids Research, 43*, D257–D260. doi:10.1093/nar/gku949

Li, H. H., Cai, X., Shouse, G. P., Piluso, L. G., & Liu, X. (2007). A specific PP2A regulatory subunit, B56gamma, mediates DNA damage-induced dephosphorylation of p53 at Thr55. *The EMBO Journal, 26*, 402–411. doi:10.1038/sj.emboj.7601519

Li, X., Sapp, E., Valencia, A., Kegel, K. B., Qin, Z. H., Alexander, J., … Difiglia, M. (2008). A function of huntingtin in guanine nucleotide exchange on Rab11. *NeuroReport, 19*, 1643–1647. doi:10.1097/WNR.0b013e328315cd4c

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein disorder prediction. *Structure, 11*, 1453–1459. doi:10.1016/j.str.2003.10.002

Linding, R., Russell, R. B., Neduva, V., & Gibson, T. J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Research, 31*, 3701–3708. doi:10.1093/nar/gkm390

Lupas, A., Van Dyke, M., & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science, 252*, 1162–1164. doi:10.1126/science.252.5009.1162

Maiuri, T., Woloshansky, T., Xia, J., & Truant, R. (2013). The huntingtin N17 domain is a multifunctional CRM1 and Ran-dependent nuclear and cilial export signal. *Human Molecular Genetics, 22*, 1383–1394. doi:10.1093/hmg/dds554

Marcé-Grau, A., Dalton, J., López-Pisón, J., García-Jiménez, M. C., Monge-Galindo, L., Cuenca-León, E., … Macaya, A. (2016). GNAO1 encephalopathy: Further delineation of a severe neurodevelopmental syndrome affecting females. *Orphanet Journal of Rare Diseases, 11*, 23. doi:10.1186/s13023-016-0416-0

Mészáros, B., Simon, I., & Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS Computational Biology, 5*, e1000376. doi:10.1371/journal.pcbi.1000376

Mitchell, A., Chang, Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., … Finn, R. D. (2014). The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Research, 43*(Database issue), D213–D221. doi:10.1093/nar/gku1243

Nakamura, K., Kodera, H., Akita, T., Shiina, M., Kato, M., Hoshino, H., … Saitsu, H. (2013). De novo mutations in GNAO1, encoding a Gαo subunit of heterotrimeric G proteins, cause epileptic encephalopathy. *The American Journal of Human Genetics, 93*, 496–505. doi:10.1016/j.ajhg.2013.07.014

Nuzzo, M. T., & Marino, M. (2016). Estrogen/Huntingtin: A novel pathway involved in neuroprotection. *Neural Regeneration Research, 11*, 402–403. doi:10.4103/1673-5374.179045

Nuzzo, M. T., Fiocchetti, M., Totta, P., Melone, M. A., Cardinale, A., Fusco, F. R., … Marino, M. (2016). Huntingtin polyQ mutation impairs the 17β-estradiol/neuroglobin pathway devoted to neuron survival. *Molecular Neurobiology, 54*, 6634–6646. doi:10.1007/s12035-016-0337-x

Ochaba, J., Lukacsovich, T., Csikos, G., Zheng, S., Margulis, J., Salazar, L., … Steffan, J. S. (2014). Potential function for the Huntingtin protein as a scaffold for selective autophagy. *Proceedings of the National Academy of Sciences, 111*, 16889–16894. doi:10.1073/pnas.1420103111

Pal, A., Severin, F., Lommer, B., Shevchenko, A., & Zerial, M. J. (2006). Huntingtin-HAP40 complex is a novel Rab5 effector that regulates early endosome motility and is up-regulated in Huntington's disease. *The Journal of Cell Biology, 172*, 605–618. doi:10.1083/jcb.200509091

Palidwor, G. A., Shcherbinin, S., Huska, M. R., Rasko, T., Stelzl, U., Arumughan, A., … Foulle, Raphaele (2009). Detection of alpha-rod protein repeats using a neural network and application to Huntingtin. *PLoS Computational Biology, 5*, e1000304. doi:10.1371/journal.pcbi.1000304

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., & Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics, 7*, 208. doi:10.1186/1471-2105-7-208

Pierce, B., & Weng, Z. (2007). ZRANK: Reranking protein docking predictions with an optimized energy function. *Proteins: Structure, Function, and Bioinformatics, 67*, 1078–1086. doi:10.1002/prot.21373

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T., & Weng, Z. (2014). ZDOCK server: Interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics, 30*, 1771–1773. doi:10.1093/bioinformatics/btu097

Potenza, E., Di Domenico, T., Walsh, I., & Tosatto, S. C. (2015). MobiDB 2.0: An improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Research, 43*(D1), D315–D320. doi:10.1093/nar/gku982

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., … Sussman, J. L. (2005). FoldIndex©: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics, 21*, 3435–3438. doi:10.1093/bioinformatics/bti537

Ratovitski, T., Chighladze, E., Arbez, N., Boronina, T., Herbrich, S., Cole, R. N., & Ross, C. A. (2012). Huntingtin protein interactions altered by polyglutamine expansion as determined by quantitative proteomic analysis. *Cell Cycle, 11*, 2006–2021. doi:10.4161/cc.20423

Ratovitski, T., O'Meally, R. N., Jiang, M., Chaerkady, R., Chighladze, E., Stewart, J. C., … Ross, C. A. (2017). Post-translational modifications (PTMs), identified on endogenous Huntingtin, cluster within proteolytic domains between HEAT repeats. *Journal of Proteome Research, 16*, 2692–2708. doi:10.1021/acs.jproteome.6b00991

Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in Enzymology, 383*, 66–93. doi:10.1016/S0076-6879(04)83004-0

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins: Structure, Function and Genetics, 42*, 38–48. doi:10.1002/1097-0134(20010101)42:1

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols, 5*, 725–738. doi:10.1038/nprot.2010.5

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology, 234*, 779–815. doi:10.1006/jmbi.1993.1626

Sanders, S. S., & Hayden, M. R. (2015). Aberrant palmitoylation in Huntington disease. *Biochemical Society Transactions, 43*, 205–210. doi:10.1042/BST20140242

Sanders, S. S., Mui, K. K., Sutton, L. M., & Hayden, M. R. (2014). Identification of binding sites in huntingtin for the huntingtin interacting proteins HIP14 and HIP14L. *PLoS ONE, 9*, e90669. doi:10.1371/journal.pone.0090669

Sangodkar, J., Farrington, C. C., McClinch, K., Galsky, M. D., Kastrinsky, D. B., & Narla, G. (2016). All roads lead to PP2A: Exploiting the therapeutic potential of this phosphatase. *FEBS Journal, 283*, 1004–1024. doi:10.1111/febs.13573

Sarge, K. D., & Park-Sarge, O. K. (2011). SUMO and its role in human diseases. *International Review of Cell and Molecular Biology, 288*, 167–183.

Saudou, F., & Humbert, S. (2016). The biology of Huntingtin. *Neuron, 89*, 910–926. doi:10.1016/j.neuron.2016.02.003

Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P., Wanker, E., & Andrade-Navarro, M. A. (2012). HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE, 7*, e31826. doi:10.1371/journal.pone.0031826

Seong, I. S., Woda, J. M., Song, J. J., Lloret, A., Abeyrathne, P. D., Woo, C. J., … MacDonald, M. E. (2010). Huntingtin facilitates polycomb repressive complex 2. *Human Molecular Genetics, 19*, 573–583. doi:10.1093/hmg/ddp524

Steffan, J. S., Agrawal, N., Pallos, J., Rockabrand, E., Trotman, L. C., Slepko, N., … Marsh, J. L. (2004). SUMO modification of Huntingtin and Huntington's disease pathology. *Science, 304*, 100–104. doi:10.1126/science.1092194

Takano, H., & Gusella, J. F. (2002). The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor. *BMC Neuroscience, 3*, 15. doi:10.1186/1471-2202-3-15

Talvik, L., Møller, R. S., Vaher, M., Vaher, U., Larsen, L. H., Dahl, H. A., … Talvik, T. (2015). Clinical phenotype of de novo GNAO1 mutation: Case report and review of literature. *Child Neurology Open, 2*. doi:10.1177/2329048X15583717

Tartari, M., Gissi, C., Lo Sardo, V., Zuccato, C., Picardi, E., Pesole, G., & Cattaneo, E. (2008). Phylogenetic comparison of Huntingtin homologues reveals the appearance of a primitive polyQ in sea urchin. *Molecular Biology and Evolution, 25*, 330–338. doi:10.1093/molbev/msm258

The UniProt Consortium (2014). Activities at the universal protein resource (UniProt). *Nucleic Acids Research, 42*, D191–D198. doi:10.1093/nar/gkt1140

The UniProt Consortium (2015). UniProt: A hub for protein information. *Nucleic Acids Research, 43*, D204–D212. doi:10.1093/nar/gku989

Thompson, L. M., Aiken, C. T., Kaltenbach, L. S., Agrawal, N., Illes, K., Khoshnan, A., … Steffan, J. S. (2009). IKK phosphorylates Huntingtin and targets it for degradation by the proteasome and lysosome. *The Journal of Cell Biology, 187*, 1083–1099. doi:10.1083/jcb.200909067

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics, 41*, 415–427. doi:10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Vijayvargia, R., Epand, R., Leitner, A., Jung, T. Y., Shin, B., Jung, R., … Seong, I. S. (2016). Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. *Elife, 5*, e11184. doi:10.7554/eLife.11184

Walsh, I., Martin, A. J., Di Domenico, T., & Tosatto, S. C. (2012). ESpritz: Accurate and fast prediction of protein disorder. *Bioinformatics, 28*, 503–509. doi:10.1093/bioinformatics/btr682

Warby, S. C., Doty, C. N., Graham, R. K., Carroll, J. B., Yang, Y. Z., Singaraja, R. R., & Hayden, M. R. (2008). Activated caspase-6 and caspase-6-cleaved fragments of huntingtin specifically colocalize in the nucleus. *Human Molecular Genetics, 17*, 2390–2404. doi:10.1093/hmg/ddn139

Watkin, E. E., Arbez, N., Waldron-Roby, E., O'Meally, R., Ratovitski, T., Cole, R. N., & Ross, C. A. (2014). Phosphorylation of mutant huntingtin at serine 116 modulates neuronal toxicity. *PLoS ONE, 9*, e88284. doi:10.1371/journal.pone.0088284

Webb, B., & Sali, A. (2014). Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics,* 5.6.1–5.6.32. doi:10.1002/0471250953.bi0506s47

Xu, Y., Xing, Y., Chen, Y., Chao, Y., Lin, Z., Fan, E., … Shi, Y. (2006). Structure of the protein phosphatase 2A holoenzyme. *Cell, 127*, 1239–1251. doi:10.1016/j.cell.2006.11.033

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., … Rost, B. (2014). PredictProtein – an open resource for online prediction of protein structural and functional features. *Nucleic Acids Research, 42*(Web server issue), W337–W343. doi:10.1093/nar/gku366

Yanai, A., Huang, K., Kang, R., Singaraja, R. R., Arstikaitis, P., Gan, L., … Hayden, M. R. (2006). Palmitoylation of Huntingtin by HIP14is essential for its trafficking and function. *Nature Neuroscience, 9*, 824–831. doi:10.1038/nn1702

Yang, Z. R., Thomson, R., McNeil, P., & Esnouf, R. M. (2005). RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics, 21*, 3369–3376. doi:10.1093/bioinformatics/bti534

Zala, D., Hinckelmann, M. V., Yu, H., Lyra da Cunha, M. M., Liot, G., Cordelieres, F. P., & Saudou, F. (2013). Vesicular glycolysis provides on-board energy for fast axonal transport. *Cell, 152*, 479–491. doi:10.1016/j.cell.2012.12.029

Zhang, H., Webb, D. J., Asmussen, H., & Horwitz, A. F. (2003). Synapse formation is regulated by the signaling adaptor GIT1. *The Journal of Cell Biology, 161*, 131–142. doi:10.1083/jcb.200211002

Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L., & Honig, B. (2013). PrePPI: A structure-informed database of protein–protein interactions. *Nucleic Acids Research, 41*(Database issue), D828–D833. doi:10.1093/nar/gks1231

Zuccato, C., & Cattaneo, E. (2007). Role of brain-derived neurotrophic factor in Huntington's disease. *Progress in Neurobiology, 81*, 294–330. doi:10.1016/j.pneurobio.2007.01.003

Zuccato, C., Valenza, M., & Cattaneo, E. (2010). Molecular mechanisms and potential therapeutical targets in Huntington's disease. *Physiological Reviews, 90*, 905–981. doi:10.1152/physrev.00041.2009

SUPPLEMENTARY MATERIALS

A comprehensive *in silico* analysis of huntingtin and its interactome

Valentina Brandi[a], Valentina Di Lella[a], Maria Marino[a], Paolo Ascenzi[a,b], Fabio Polticelli[a,c]

[a]Department of Sciences, Roma Tre University, I–00146 Roma, Italy

[b]Interdepartmental Laboratory for Electron Microscopy, Roma Tre University, I–00146 Roma, Italy

[c]National Institute of Nuclear Physics, Roma Tre Section, I–00146 Roma, Italy

* **Corresponding author:** Prof. Fabio Polticelli, Department of Sciences, Roma Tre University,

Viale Guglielmo Marconi 446, I–00146 Roma, Italy

Phone: +39 06 57336362; Fax: +39 06 57336321; email: fabio.polticelli@uniroma3.it

Supplementary Tables

Table S1. C-Score of the models of Htt ordered domains analysed in this work. The models have been selected by Confidence-score, a parameter to evaluate the quality of predicted models. Generally C-score can range from -5 to 2, where a higher value corresponds to a greater level of confidence of the model.

Huntingtin domain	C-score
Hunt1	-2,37
Hunt2	-2,31
Hunt3	-1,80
Hunt5	-2,33

Table S2. Ordered and disordered regions of Htt predicted by different web services.

SERVER	SEQUENCE REGIONS IDENTIFIED
FOLDINDEX	UNFOLDED REGIONS: 1-100, 461-480, 501-629, 632-668, 1182-1212, 2328-2361, 2467-2478, 2616-2666;
GLOBPLOT	UNFOLDED REGIONS: 5-16, 62-110, 451-462, 507-521, 544-554, 564-610, 682-704, 1072-1084, 1192-1212, 1217-1256, 1353-1371, 2094-2119, 2442-2459, 2671-2688, 2958-2974; DOMINI GLOBULARI: 111-506, 705-1191, 1257-1352, 1372-2093, 2120-2441, 2460-2670, 2689-2957, 2975-3169;
PREDICTPROTEIN	UNFOLDED REGIONS: 1-96, 407-421, 427-677, 1109-1125, 1168-1225, 1721-1733, 2332-2351, 2484-2494, 2632-2660, 2933-2948;
ANCHOR	DISORDERED BINDING REGIONS: 1-23, 81-113, 126-137, 422-435, 465-486, 506-563, 573-600, 607-637, 1230-1241, 2614-2627;
HMMER	DISORDERED REGIONS: 1-100, 444-624, 643-674, 1178-1221, 2075-2090, 2332-2350, 2485-2499, 2636-2664, 2935-2949;
INTERPRO	DOMAINS: 81-401, 696-1099, 1129-1152, 1261-1305, 2624-2650, 2740-3083;
SMART	LOW COMPLEXITY REGIONS: 18-78, 413-421, 454-466, 533-547, 910-940, 1171-1182, 1444-1462, 1834-1844, 2072-2084, 2485-2496, 2635-2658, 2881-2896; HEAT REPEATS: 130-160, 253-283, 361-387, 753-778, 809-838, 854-878, 910-940;
MOBIDB	DISORDERED REGIONS: 55-98, 406-419, 428-440, 442-614, 633-676, 1102-1123, 1166-1229, 1330-1348, 1719-1731, 1857-1876, 2068-2088, 2328-2353, 2483-2496, 2632-2663, 2932-2948;
PONDR	DISORDERED REGIONS: 1-9, 23-97, 171-182, 212-230, 338-356, 396-503, 515-620, 646-671, 997-1007, 1100-1130, 1164-1223, 1337-1348, 1558-1569, 1600-1611, 1717-1736, 1861-1887, 2024-2045, 2070-2088, 2120-2134, 2321-2373,

	2477-2529, 2533-2546, 2553-2564, 2596-2610, 2628-2662, 2725-2738, 2759-2773, 2858-2877, 2889-2900, 2932-2954, 3110-3115

Table S3. Structural homologs of Hunt1 identified by I-TASSER.

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	1B3UB	0.724	2.72	0.110	0.794
2	3W3TA	0.680	4.30	0.104	0.871
3	4FGVA	0.667	4.64	0.080	0.890
4	2X1GG	0.650	4.53	0.104	0.868
5	4KNHA	0.645	5.15	0.105	0.912
6	3MLIC	0.639	4.84	0.087	0.874
7	3S4ZA	0.636	4.30	0.079	0.835
8	1QBKB	0.636	4.36	0.111	0.830
9	3A6PA	0.635	4.44	0.103	0.824

Table S4. Structural homologs of Hunt2 identified by I-TASSER.

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	1EE5A	0.770	1.39	0.101	0.787
2	3TJ3A	0.754	1.91	0.099	0.789
3	4RXHB	0.751	1.82	0.095	0.784
4	2YNSA	0.750	1.87	0.112	0.784
5	1IALA	0.731	2.11	0.081	0.774
6	4UAEA	0.723	2.27	0.083	0.774
7	3C2HB	0.698	4.38	0.077	0.860
8	4RV1A	0.696	2.70	0.148	0.760
9	4R0ZA	0.682	3.60	0.089	0.791

	10	3L6XA	0.658	3.54	0.100	0.756

Table S5. Structural homologs of Hunt3 identified by I-TASSER. The best structural homolog is the importin Kap121p (PDB code: 3W3T)

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	3W3TA	0.869	1.59	0.102	0.885
2	1QGKA	0.607	4.98	0.084	0.713
3	2BPTA	0.593	5.04	0.084	0.702
4	1WA5C	0.575	6.17	0.078	0.719
5	4C0OA	0.559	5.83	0.089	0.685
6	3ICQU	0.551	6.19	0.077	0.697
7	2X19B	0.544	5.82	0.081	0.665
8	2Z5OA	0.536	6.19	0.072	0.668
9	3GJXA	0.524	6.51	0.061	0.667
10	4FGVA	0.500	6.91	0.060	0.655

Table S6. Mutations in the GNAO1 gene causing symptoms similar to that observed in HD. Since the Gα$_o$ three-dimensional structure lacks the signal peptide at N-terminus, residue numbering is different and the corresponding residue numbers are reported for clarity.

MUTATION	CORRESPONDING RESIDUES IN Gα$_o$ STRUCTURE	MUTATION EFFECTS
L199P	L178	Neonatal-onset, refractory seizures, progressive microcephaly, oral lingual dyskinesia, nearly absent psychomotor development
R209C	R188	Involuntary movements and chorea
E246K	E225	Involuntary movements
A227V	A206	Severe chorea
D174G	D153	Epileptic encephalopathies and involuntary movements
I279N	I258	Ohtahara syndrome, no involuntary movement
T191_F197 del	T170_F176	Ohtahara syndrome, dystonia
G203R	G182	Severe chorea, athetosis
G184S	G163	partial perinatal lethality, spontaneous adult lethality, enhanced frequency of epileptiform discharges, and marked sensitization to PTZ kindling
Y231C	Y210	Stereotypic spasm-like movements

Table S7. Structural homologs of Hunt5 identified by I-TASSER.

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	1B3UA	0.908	1.97	0.080	0.948
2	3ICQU	0.693	4.84	0.054	0.886
3	4A0CA	0.677	4.99	0.074	0.884
4	4FGVA	0.674	5.24	0.059	0.903
5	3EA5B	0.673	4.94	0.063	0.867
6	3L5Q6	0.671	5.91	0.069	0.952
7	2X1GG	0.662	5.27	0.076	0.878
8	3GJXA	0.648	5.38	0.073	0.867
9	3W3TA	0.647	5.18	0.068	0.865
10	1WA5C	0.642	5.74	0.087	0.890

Table S8. List of the genes names of all the identified Htt interactors. Additional tables report the list of interactors divided by method.

AAAS	ANKRD27	ATG8	BDNF	CASP8	CHERP	CREBBP	DLD	DVL2	ETFB
AANAT	AOC3	ATL2	BID	CBLB	CHMP4B	CRK	DLG1	DYNC1H1	ETV4
ABI2	AP1B1	ATM	BIRC6	CBS	CHRD	CRKL	DLG4	DYNC1I1	EVL
ABL1	AP1G1	ATP1A2	BMP4	CBX5	CHRFAM7A	CRMP1	DLST	ECH1	EZR
ABR	AP1G2	ATP1B1	BMP7	CCDC126	CHRNB2	CRYL1	DLX2	ECSIT	F2
ACADL	AP2A1	ATP5D	BMPR1A	CCDC88A	CHUCK	CS	DNAJA1	ECT2	F2R
ACMSD	AP2A2	ATP5O	BMPR2	CCL22	CHUK	CSK	DNAJA3	EEF1D	F7
ACR	APBA1	ATP7A	BMX	CCNB1	CKM	CTBP1	DNAJB1	EFS	F8A1
ACTC1	APBA2	ATR	BRAF	CCND1	CLASP1	CTNNB1	DNAJC11	EGFR	FADS1
ACVR1	APBA3	ATXN1	BRCA1	CCNF	CLN8	CTNND2	DNAJC21	EHMT1	FAM179B
ACVR2A	APBB1	ATXN2	BRCA2	CCT6A	CLTC	CTSL2	DNAJC4	EIF2AK3	FAM71F1
ACVR2B	APBB2	AVP	BTK	CCT8	CLTCL1	CTTN	DNALI1	EIF2C2	FARP2
ADAR	API5	AXIN1	BTRC	CD2AP	CNP	CUE5	DNM1	ELANE	FBN2
ADCY5	APP	AZU1	BZW2	CD36	CNTN1	CUL2	DNM1L	ENO1	FBXL5
ADD3	APPL1	BAG3	C15orf63	CD59	CNTN4	CUL5	DNM2	ENO2	FCHSD2
ADIPOQ	AR	BAG5	C2CD3	CD74	CNTNAP2	CUL7	DNMBP	ENO3	FEM1B
ADNP	ARFGAP3	BAHCC1	C5AR2	CDH13	COL4A3BP	CUL9	DNMT3A	EP300	FER
ADORA2A	ARHGAP24	BAIAP2	CA4	CDK1	COMT	CXorf27	DNMT3B	EPB41	FEZ1
ADRBK1	ARHGAP25	BAIAP2L1	CACNA1A	CDK5	COPB1	CYB5R4	DOC2B	EPB41L1	FGF10
AGO2	ARHGAP35	BAIAP2L2	CACNA1S	CELF2	COQ7	CYC1	DOCK1	EPB41L3	FGFR1
AIFM3	ARHGAP4	BAK1	CACNA2D1	CELSR1	CORIN	CYLD	DOCK11	EPB41L5	FICD
AKAP1	ARHGEF11	BARD1	CACNB4	CENPV	COX2	DAB1	DOCK2	EPHA2	FLNA
AKT1	ARHGEF12	BARHL1	CAMK2G	CEP126	COX3	DBH	DOCK4	EPHA7	FNBP1
AKT2	ARHGEF7	BASP1	CAPN10	CEP63	COX5A	DCLK1	DOCK9	EPHB2	FNBP1L
AKTIP	ARHGEF9	BAX	CARD11	CER1	COX5B	DCTN1	DOPEY2	EPS15	FNBP4
ALDOA	ARX	BAZ1A	CASKIN2	CFLAR	COX6A1	DCTN2	DPPA4	EPS8	FOXA2
ALDOC	ASAP1	BBS2	CASP1	CFTR	COX6B1	DDB1	DPYSL2	EPS8L1	FOXP2
ALPL	ASH1L	BBS4	CASP10	CHCHD2	COX6C	DFFA	DRD1	ERBB3	FRMPD4
ALS2	ASPH	BCAN	CASP2	CHCHD3	CP	DGKD	DRD2	ERBB4	FRS2
AMFR	ATG12	BCAP31	CASP3	CHD3	CPEB1	DICER1	DTNBP1	ERCC5	FTL
AMPD1	ATG16L1	BCL2	CASP6	CHD7	CPT1B	DIRAS2	DUSP10	ERCC6L	FURIN
AMPH	ATG5	BCL2L1	CASP7	CHEK1	CREB1	DLAT	DVL1	ESR1	FYB

FYN	GRB2	HIST1H3J	ING5	KPNA3	MAGEE1	MID1	NCKIPSD	NR1H2	PARD3
GABBR1	GRIK2	HLA-B	INPP5D	KPNA4	MALT1	MIPEP	NCOR1	NR2E3	PARD6B
GABRA5	GRIN1	HMG20A	INPPL1	KPNA6	MAP1LC3A	MKRN2	NDE1	NR4A2	PARK2
GALR2	GRIN2A	HMGA1	INS	KRAS	MAP1LC3B	MOG	NDUFA13	NR4A3	PBRM1
GAPDH	GRM1	HMGB1	IRS1	LAMTOR2	MAP1S	MORC3	NDUFA8	NRAS	PCSK6
GAS7	GRM4	HNF1A	IRS2	LANCL1	MAP2K1	MPP1	NDUFA9	NRBP1	PCSK7
GCK	GRPEL2	HNRNPC	ITCH	LASP1	MAP3K10	MPP6	NDUFB10	NRBP2	PDCD4
GEMIN7	GRSF1	HOMER1	ITGA1	LDHA	MAP3K11	MPPE1	NDUFB9	NSD1	PDCD6
GFAP	GSK3B	HOMER3	ITPR1	LDHAL6A	MAP3K7	MRE11A	NDUFS1	NSG1	PDE1A
GGA1	GSX2	HOOK1	ITSN1	LDHB	MAPK1	MRFAP1	NDUFS3	NTM	PDK2
GGA2	GTF3C3	HOOK3	ITSN2	LDOC1	MAPK3	MRFAP1L1	NDUFS5	NTRK1	PDK3
GGA3	HAAO	HOXC11	JAK2	LEF1	MAPK8IP1	MSH6	NDUFV1	NUB1	PDLIM7
GHR	HADHA	HOXC4	JAKMIP1	LETM1	MAPK8IP2	MTA2	NDUFV2	NUMB	PDPK1
GHRHR	HAP1	HP	JPH2	LHX1	MAST2	MT-CO2	NEDD4	NUMBL	PDZK1
GIMAP5	HAX1	HRAS	KALRN	LHX6	MATK	MT-CO3	NEDD4L	NUP133	PEX11B
GIPC1	HBB	HSD17B10	KAT2A	LIG4	MBD1	MTOR	NEDD9	NUPL1	PEX13
GIT1	HBS1L	HSP90AB1	KAT2B	LIMK1	MBD3	MTSS1	NEGR1	ODF2L	PEX14
GNA11	HCK	HSPA4	KCNQ1	LIN7A	MBD4	MTUS1	NETO1	OPA1	PEX5
GNAO1	HCLS1	HSPA8	KIAA1377	LIN7B	MBTPS1	MTX1	NEUROD1	OPCML	PFKFB2
GNAQ	HDAC6	HSPD1	KIF1A	LIN7C	MCF2L	MTX2	NF1	OPRM1	PFN1
GNAS	HEL1	HTT	KIF1B	LIPF	MDH1	MTX3	NF2	OPTN	PFN2
GNAZ	HERC2	HUWE1	KIF23	LLGL1	MDM2	MUM1	NFKB1	OSTF1	PGAM1
GNB1	HEY2	HYPK	KIF3B	LMAN2L	MED15	MYBPC3	NLRP1	OTC	PGHA1
GNB2	HEYL	HYPM	KIF4A	LRPPRC	MED21	MYH9	NME4	OXT	PHB2
GOLPH3L	HFE	IFT20	KIF5B	LRRK2	MED31	MYL6	NOSTRIN	OXTR	PHGDH
GOPC	HGF	IKBKAP	KIT	LSAMP	MEF2D	MYL6B	NOTCH1	P4H1	PI4K2A
GORASP1	HGS	IKBKB	KLC2	LSM11	MEOX1	MYLK2	NOTCH2	P4HA1	PI4KA
GOSR2	HIP1	IKBKG	KLF11	LSM14B	MFI2	MYO1E	NOTCH3	PACSIN1	PI4KB
GPKOW	HIP1R	IKZF1	KLF4	LTBP4	MFN2	MYO5A	NOTCH4	PACSIN2	PIAS1
GPM6A	HIPK1	IL17RA	KLK14	LUC7L2	MGEA5	MYO9B	NPHP1	PADI2	PIAS4
GPRASP2	HIPK2	IL1B	KLK6	MAGEA3	MGRN1	NAGS	NPLOC4	PAFAH1B1	PIBF1
GRAP	HIST1H3A	ILK	KLK8	MAGEB18	MIA3	NAPB	NPTN	PAK2	PICK1
GRAPL	HIST1H3H	IMMT	KPNA1	MAGEB6	MIB1	NBR1	NPY2R	PALM	PIK3R1

PIK3R2	PROC	RASGRF1	SCFD1	SIPA1L1	SP1	SYN2	TNF	UBE2K	WASL
PIK3R3	PRPF40A	RBL1	SCMH1	SIRPA	SP3	SYP	TNK2	UBQLN1	WBP4
PKM	PRPF40B	RBP4	SCRIB	SKI	SPAST	SYT1	TNNC1	UBQLN2	WDFY3
PKM2	PRX	RCN2	SCYL1	SLC1A2	SPG20	SYT12	TNRC18	UBR3	WDR91
PLAT	PSEN1	RERE	SDCBP	SLC24A2	SPTAN1	SYT2	TOM1	UHMK1	WDTC1
PLCB1	PSEN2	REST	SEC22A	SLC25A11	SQSTM1	SYT7	TOPBP1	UHRF1	WFS1
PLCG1	PSMB7	RET	SEC22C	SLC25A12	SREBF2	SYT9	TP53	ULK1	WRN
PLCG2	PSMC5	REV1	SEC24A	SLC25A13	SRGAP1	SYVN1	TP53BP2	UNC13B	XAGE3
PLG	PSMD4	RFX3	SEC24B	SLC25A3	SRGAP2	TACC1	TP63	UPF3A	XIAP
PLIN4	PSMD9	RFXANK	SEC24C	SLC26A7	SRGAP3	TACR1	TPH1	UQCRC1	XRCC6
PLP1	PTGES2	RGS9	SEC24D	SLC6A1	SRI	TAF4	TPI1	UQCRC2	YWHAB
PML	PTGES3	RHOA	SEC31A	SLC9A3R1	SRP9L1	TANK	TPM1	UQCRH	YWHAE
PPA2	PTK2B	RHOB	SETD2	SLCO1A2	SRRT	TAOK2	TPP1	USP14	YWHAG
PPARD	PTK6	RHOBTB3	SETDB1	SLIT1	STAM	TBP	TPR	USP7	YWHAH
PPARG	PTPLAD1	RHOT1	SFMBT1	SLIT2	STAP1	TCERG1	TRAF2	USP8	YWHAQ
PPARGC1A	PTPN11	RHOT2	SFN	SMAD2	STAR	TERF1	TRAF3	USP9X	YWHAZ
PPIF	PTPRF	RIMS1	SFXN3	SMAD3	STAT1	TERF2	TRAF4	UTP14A	ZBTB16
PPL	PTRF	RIMS2	SH3BP4	SMURF1	STAT3	TF	TRAF5	UXT	ZBTB17
PPP1R13B	RAB11A	RNASEH1	SH3D19	SNAP25	STAT5A	TFCP2L1	TRAF6	VASP	ZDHHC17
PPP2CA	RAB14	RNF2	SH3GL1	SNCA	STAT5B	TGM2	TRAFD1	VAV1	ZEB2
PPT1	RAB2A	RNF20	SH3GL2	SNCB	STIP1	THRA	TRAPPC4	VAV2	ZFC3H1
PQBP1	RAB30	RNF40	SH3GL3	SNX1	STRADA	THY1	TRH	VAV3	ZFYVE19
PREB	RAB35	RNF8	SH3GLB1	SNX16	STUB1	TIAM1	TRIM15	VCP	ZFYVE20
PRHOXNB	RAB3D	ROBO1	SH3KBP1	SNX17	STX18	TJP1	TRIO	VDAC1	ZMAT2
PRKAB2	RAB5A	ROBO2	SH3RF1	SNX21	STX1A	TJP2	TRIP10	VDAC2	ZMYND8
PRKACA	RAB8B	ROCK1	SH3RF3	SNX9	STX1B	TKT	TSHZ3	VDAC3	ZNF133
PRKCA	RABGEF1	RPGRIP1L	SH3YL1	SOCS2	STXBP1	TLK1	TTC23	VEGFB	ZNF354A
PRKCD	RAC1	RPL4	SHB	SOCS3	STXBP4	TM4SF19	TUBA1A	VHL	ZNF451
PRKCG	RANBP2	RPS6KB1	SHC3	SOD1	SUMO1	TMEM57	TUBB	VPS4A	ZNF655
PRKCZ	RAPGEF2	S100B	SHROOM3	SORBS1	SUPT5H	TMGSF19	TUBG1	VTI1A	ZNF675
PRKDC	RARB	SAP30	SIAH1	SORBS2	SYK	TMPRSS3	TXNDC11	WAC	ZNF91
PRMT2	RARG	SASH1	SIN3A	SORBS3	SYMPK	TMPRSS6	UBAC1	WAS	ZW10
PRNP	RASA1	SCAP	SIPA1	SORL1	SYN1	TMX1	UBE2E3	WASF2	

Interactors predicted by IntAct									
ADD3	COPB1	DYNC1H1	HAP1	JAKMIP1	MTUS1	PIAS1	RPL4	TCERG1	XAGE3
AP2A2	CREBBP	ECH1	HAX1	KLF11	NAPB	PIAS4	SASH1	TMEM57	XRCC6
APBB2	CRMP1	EHMT1	HEY2	KPNA3	NBR1	PIBF1	SETD2	TMGSF19	YWHAB
API5	CTNNB1	ERCC6L	HEYL	LDOC1	NCOR1	PIK3R1	SH3GL3	TP53	ZBTB16
ARFGAP3	CUE5	ETV4	HIP1	MAGEB18	NFKB1	PIK3R2	SIN3A	TPP1	ZDHHC17
ARHGAP24	CUL2	EVL	HIST1H3A	MAGEB6	NME4	PIK3R3	SNCA	TRAFD1	ZFC3H1
ARHGAP25	CUL5	EZR	HIST1H3J	MAGEE1	NUB1	PKM	SORBS1	TSHZ3	ZFYVE19
ASPH	DCTN2	FAM179B	HMG20A	MBD1	NUPL1	PPA2	SP1	TTC23	ZMAT2
ATG8	DNAJA3	FAM71F1	HMGA1	MBD4	ODF2L	PPARG	SP3	TXNDC11	ZMYND8
BAIAP2	DNAJC11	FEM1B	HMGB1	MED15	OPTN	PPL	SPG20	UBAC1	ZNF133
BAZ1A	DNAJC21	FEZ1	HNRNPC	MED21	OSTF1	PRMT2	SREBF2	UBE2E3	ZNF451
BZW2	DNAJC4	FNBP4	HOXC11	MED31	P4H1	PRPF40A	SRGAP1	UBE2K	ZNF655
CCDC126	DNALI1	FTL	HOXC4	MEF2D	PACSIN1	PSMB7	SRGAP2	UPF3A	ZNF675
CCDC88A	DNM1	GAPDH	HSPD1	MEOX1	PAK2	PSMD4	SRGAP3	USP9X	ZNF91
CEP126	DNM2	GGA2	HYPK	MIPEP	PALM	PTGES2	SRRT	UTP14A	
CEP63	DOCK11	GIT1	HYPM	MKRN2	PDK2	PTK6	SUPT5H	WAC	
CHD3	DOCK9	GOLPH3L	IFT20	MRE11A	PEX11B	RABGEF1	TACC1	WBP4	
CHMP4B	DPPA4	GPRASP2	IKBKAP	MRFAP1L1	PFN2	RNF20	TAF4	WDFY3	
CLASP1	DUSP10	GTF3C3	ING5	MTSS1	PGHA1	RNF40	TANK	WDR91	

Interactors predicted by PrePPI									
AAAS	APBA2	BAG3	CACNA1S	CHCHD2	CTNND2	DNMBP	EP300	FEZ1	GNAO1
AANAT	APBA3	BAG5	CACNB4	CHD3	CTSL2	DNMT3A	EPB41	FGF10	GNAQ
ABI2	APBB1	BAHCC1	CAMK2G	CHD7	CTTN	DNMT3B	EPB41L1	FGFR1	GNAS
ABL1	APBB2	BAIAP2	CAPN10	CHEK1	CUL2	DOC2B	EPB41L3	FICD	GNAZ
ABR	API5	BAIAP2L1	CARD11	CHERP	CUL7	DOCK1	EPB41L5	FLNA	GNB1
ACMSD	APP	BAIAP2L2	CASKIN2	CHMP4B	CUL9	DOCK11	EPHA2	FNBP1	GNB2
ACR	APPL1	BAK1	CASP1	CHRD	CXorf27	DOCK2	EPHA7	FNBP1L	GOLPH3L
ACTC1	AR	BARD1	CASP10	CHRFAM7A	CYB5R4	DOCK4	EPHB2	FOXA2	GOPC
ACVR1	ARHGAP35	BARHL1	CASP2	CHRNB2	CYLD	DOCK9	EPS15	FOXP2	GORASP1
ACVR2A	ARHGAP4	BAX	CASP3	CHUK	DAB1	DOPEY2	EPS8	FRMPD4	GOSR2
ACVR2B	ARHGEF11	BAZ1A	CASP6	CLASP1	DBH	DPPA4	EPS8L1	FRS2	GPKOW
ADAR	ARHGEF12	BBS2	CASP7	CLN8	DCLK1	DPYSL2	ERBB3	FTL	GPRASP2
ADCY5	ARHGEF7	BBS4	CASP8	CLTC	DCTN1	DRD1	ERBB4	FURIN	GRAP
ADIPOQ	ARHGEF9	BCAN	CBLB	CLTCL1	DDB1	DRD2	ERCC5	FYB	GRAPL
ADNP	ARX	BCL2	CBS	CNP	DFFA	DTNBP1	ERCC6L	FYN	GRB2
ADORA2A	ASAP1	BCL2L1	CBX5	CNTN4	DGKD	DUSP10	ESR1	GABBR1	GRIK2
ADRBK1	ASH1L	BDNF	CCDC126	CNTNAP2	DICER1	DVL1	ETV4	GABRA5	GRIN1
AIFM3	ASPH	BID	CCL22	COL4A3BP	DIRAS2	DVL2	EVL	GALR2	GRIN2A
AKAP1	ATG12	BIRC6	CCNB1	COMT	DLAT	DYNC1H1	EZR	GAPDH	GRM1
AKT1	ATG16L1	BMP4	CCND1	COQ7	DLD	DYNC1I1	F2	GAS7	GRM4
AKT2	ATG5	BMP7	CCNF	CORIN	DLG1	ECH1	F2R	GCK	GRPEL2
AKTIP	ATL2	BMPR1A	CD2AP	COX5A	DLG4	ECSIT	F7	GEMIN7	GRSF1
ALS2	ATM	BMPR2	CD59	CP	DLX2	ECT2	F8A1	GFAP	GSK3B
AMFR	ATP1A2	BMX	CD74	CPEB1	DNAJA1	EEF1D	FADS1	GGA1	GSX2
AMPH	ATP7A	BRAF	CDH13	CREB1	DNAJA3	EFS	FAM179B	GGA2	GTF3C3
ANKRD27	ATR	BRCA1	CDK1	CREBBP	DNAJB1	EGFR	FAM71F1	GGA3	HAAO
AP1B1	ATXN1	BRCA2	CDK5	CRK	DNAJC11	EHMT1	FARP2	GHR	HAP1
AP1G1	ATXN2	BTK	CELF2	CRKL	DNAJC21	EIF2AK3	FBN2	GHRHR	HBS1L
AP1G2	AVP	BTRC	CELSR1	CRMP1	DNALI1	EIF2C2	FBXL5	GIMAP5	HCK
AP2A1	AXIN1	BZW2	CER1	CSK	DNM1	ELANE	FCHSD2	GIPC1	HCLS1
AP2A2	AZU1	C2CD3	CFLAR	CTBP1	DNM1L	ENO1	FEM1B	GIT1	HDAC6
APBA1	APBA3	CACNA1A	CFTR	CTNNB1	DNM2	ENO2	FER	GNA11	HERC2

HEY2	IKZF1	KPNA6	MAPK8IP1	MTX2	NPHP1	PARK2	PLCG2	PTK2B	RIMS2
HEYL	IL1B	KRAS	MAPK8IP2	MTX3	NPLOC4	PBRM1	PLG	PTK6	RNASEH1
HFE	ILK	LAMTOR2	MAST2	MUM1	NPTN	PCSK6	PLP1	PTPLAD1	RNF2
HGF	ING5	LASP1	MATK	MYBPC3	NR1H2	PCSK7	PML	PTPN11	RNF20
HGS	INPP5D	LDHAL6A	MBD1	MYH9	NR2E3	PDCD4	PPARD	PTPRF	RNF40
HIP1	INPPL1	LEF1	MBD3	MYL6	NR4A2	PDCD6	PPARG	RAB11A	RNF8
HIP1R	INS	LETM1	MBD4	MYL6B	NR4A3	PDK2	PPARGC1A	RAB14	ROBO1
HIPK1	IRS1	LHX1	MBTPS1	MYLK2	NRAS	PDK3	PPIF	RAB2A	ROBO2
HIPK2	IRS2	LHX6	MCF2L	MYO1E	NRBP1	PDLIM7	PPL	RAB35	ROCK1
HIST1H3A	ITCH	LIG4	MDM2	MYO5A	NRBP2	PDPK1	PPP1R13B	RAB5A	RPGRIP1L
HLA-B	ITGA1	LIMK1	MED15	MYO9B	NSD1	PDZK1	PPP2CA	RAB8B	RPL4
HLA-B	ITSN1	LIN7A	MED21	NAGS	NSG1	PEX11B	PPT1	RAC1	RPS6KB1
HMG20A	ITSN2	LIN7B	MED31	NAPB	NTM	PEX13	PREB	RANBP2	S100B
HMGA1	JAK2	LIN7C	MFI2	NBR1	NTRK1	PEX14	PRHOXNB	RAPGEF2	SAP30
HNF1A	KALRN	LIPF	MFN2	NCKIPSD	NUB1	PEX5	PRKAB2	RARB	SASH1
HNRNPC	KAT2A	LLGL1	MGEA5	NCOR1	NUMB	PFKFB2	PRKACA	RARG	SCAP
HOMER1	KAT2B	LMAN2L	MGRN1	NDE1	NUMBL	PFN1	PRKCA	RASA1	SCFD1
HOMER3	KCNQ1	LRPPRC	MIA3	NDUFA9	NUP133	PFN2	PRKCD	RASGRF1	SCMH1
HOOK1	KIAA1377	LRRK2	MIB1	NDUFS1	NUPL1	PGAM1	PRKCG	RBL1	SCRIB
HOOK3	KIF1A	LSM11	MID1	NDUFS3	OPA1	PHB2	PRKCZ	RBP4	SCYL1
HOXC11	KIF1B	LSM14B	MIPEP	NEDD4	OPCML	PHGDH	PRKDC	RCN2	SDCBP
HOXC4	KIF23	LTBP4	MORC3	NEDD4L	OPTN	PI4K2A	PRMT2	RERE	SEC22A
HP	KIF3B	MAGEA3	MPP1	NEDD9	OSTF1	PI4KA	PRNP	REST	SEC22C
HRAS	KIF4A	MAGEB18	MPP6	NETO1	OTC	PI4KB	PROC	RET	SEC24A
HSP90AB1	KIF5B	MAGEB6	MPPE1	NEUROD1	OXT	PIAS1	PRPF40A	REV1	SEC24B
HSPA4	KIT	MALT1	MRFAP1	NF1	OXTR	PIAS4	PRPF40B	RFX3	SEC24C
HSPA8	KLC2	MAP1LC3A	MRFAP1L1	NF2	P4HA1	PICK1	PRX	RFXANK	SEC24D
HTT	KLF4	MAP1S	MSH6	NLRP1	PACSIN1	PIK3R1	PSEN1	RGS9	SEC31A
HUWE1	KLK14	MAP2K1	MTA2	NME4	PACSIN2	PIK3R2	PSEN2	RHOA	SETD2
HYPK	KLK6	MAP3K10	MT-CO2	NOSTRIN	PADI2	PIK3R3	PSMB7	RHOB	SETDB1
IFT20	KLK8	MAP3K11	MT-CO3	NOTCH1	PAFAH1B1	PKM2	PSMC5	RHOBTB3	SFMBT1
IKBKAP	KPNA1	MAP3K7	MTOR	NOTCH2	PAK2	PLAT	PSMD9	RHOT1	SFN
IKBKB	KPNA3	MAPK1	MTSS1	NOTCH3	PARD3	PLCB1	PTGES2	RHOT2	SH3BP4

IKBKG	KPNA4	MAPK3	MTX1	NOTCH4	PARD6B	PLCG1	PTGES3	RIMS1	SH3D19
SH3GL1	SLC25A13	SOD1	STAT5B	SYVN1	TMPRSS6	TRAPPC4	UPF3A	VHL	YWHAQ
SH3GL2	SLC26A7	SORBS1	STIP1	TACC1	TMX1	TRH	UQCRC1	VPS4A	YWHAZ
SH3GL3	SLC6A1	SORBS2	STRADA	TACR1	TNF	TRIM15	UQCRC2	VTI1A	ZBTB16
SH3GLB1	SLC9A3R1	SORBS3	STUB1	TAF4	TNK2	TRIO	UQCRH	WAS	ZBTB17
SH3KBP1	SLCO1A2	SORL1	STX18	TAOK2	TNNC1	TRIP10	USP14	WASF2	ZDHHC17
SH3RF1	SLIT1	SP1	STX1A	TBP	TNRC18	TSHZ3	USP7	WASL	ZEB2
SH3RF3	SLIT2	SPAST	STX1B	TCERG1	TOM1	TTC23	USP8	WBP4	ZFC3H1
SH3YL1	SMAD2	SPG20	STXBP1	TERF1	TOPBP1	TUBA1A	USP9X	WDFY3	ZFYVE19
SHB	SMAD3	SQSTM1	STXBP4	TERF2	TP53	TUBB	UTP14A	WDR91	ZFYVE20
SHC3	SMURF1	SREBF2	SUMO1	TF	TP53BP2	UBAC1	UXT	WDTC1	ZMAT2
SHROOM3	SNAP25	SRGAP2	SUPT5H	TFCP2L1	TP63	UBE2E3	VASP	WFS1	ZMYND8
SIAH1	SNCA	SRGAP3	SYK	TGM2	TPM1	UBE2K	VAV1	WRN	ZNF354A
SIN3A	SNCB	SRI	SYMPK	THRA	TPR	UBQLN1	VAV2	XAGE3	ZNF451
SIPA1	SNX1	STAM	SYN1	THY1	TRAF2	UBQLN2	VAV3	XIAP	ZNF655
SIPA1L1	SNX16	STAP1	SYP	TIAM1	TRAF3	UBR3	VCP	XRCC6	ZW10
SKI	SNX17	STAR	SYT1	TJP1	TRAF4	UHMK1	VDAC1	YWHAB	
SLC1A2	SNX9	STAT1	SYT2	TJP2	TRAF5	UHRF1	VDAC2	YWHAE	
SLC24A2	SOCS2	STAT3	SYT7	TLK1	TRAF6	ULK1	VDAC3	YWHAG	
SLC25A12	SOCS3	STAT5A	SYT9	TMPRSS3	TRAFD1	UNC13B	VEGFB	YWHAH	

Interactors predicted by Hippie									
ACADL	CASP3	CPT1B	ECH1	GOLPH3L	IKBKB	MEOX1	NUPL1	PRPF40B	SNCA
ADD3	CASP6	CREB1	EEF1D	GPM6A	IKBKG	MGRN1	ODF2L	PSMB7	SORBS1
AKT1	CASP7	CREBBP	EGFR	GPRASP2	IMMT	MID1	OPCML	PSMC5	SP1
ALDOA	CASP8	CRMP1	EHMT1	GRAP	ING5	MIPEP	OPTN	PSMD4	SP3
ALDOC	CBS	CS	EIF2C2	GRB2	ITCH	MKRN2	OSTF1	PTGES2	SPG20
ALPL	CCDC126	CTBP1	ENO1	GRM1	ITPR1	MOG	P4HA1	PTK6	SPTAN1
AMFR	CCDC88A	CTNNB1	ENO3	GTF3C3	JAKMIP1	MRE11A	PACSIN1	PTRF	SQSTM1
AMPD1	CCT6A	CUL2	EPB41	HADHA	JPH2	MRFAP1	PAK2	RAB3D	SREBF2
AOC3	CCT8	CUL5	EPB41L1	HAP1	KAT2B	MRFAP1L1	PALM	RABGEF1	SRGAP1
AP2A2	CD36	CYC1	EPB41L3	HAX1	KIAA1377	MTSS1	PDE1A	RAC1	SRGAP2
APBB2	CD59	DCTN2	ERCC6L	HBB	KLF11	MTUS1	PDK2	RASA1	SRGAP3
API5	CDH13	DDB1	ETFB	HBS1L	KPNA3	MTX2	PEX11B	RCN2	SRP9L1
ARFGAP3	CENPV	DIRAS2	ETV4	HEY2	LANCL1	NAPB	PFN2	REST	SRRT
ARHGAP24	CEP63	DLAT	EVL	HEYL	LDHA	NBR1	PHB2	RNF20	STIP1
ARHGAP25	CHCHD2	DLD	EZR	HIP1	LDHAL6A	NCOR1	PIAS1	RNF40	STX1A
ASPH	CHCHD3	DLG4	F8A1	HIP1R	LDHB	NDUFA13	PIAS4	RPL4	STX1B
ATG12	CHD3	DLST	FAM179B	HIST1H3A	LDOC1	NDUFA8	PIBF1	RPS6KB1	STXBP1
ATG16L1	CHMP4B	DNAJA1	FAM71F1	HLA-B	LSAMP	NDUFA9	PIK3R1	SAP30	SUMO1
ATG5	CHUK	DNAJA3	FBN2	HMG20A	LUC7L2	NDUFB10	PIK3R2	SASH1	SUPT5H
ATP1B1	CKM	DNAJB1	FEM1B	HMGA1	MAGEA3	NDUFB9	PIK3R3	SETD2	SYMPK
ATP5D	CLASP1	DNAJC11	FEZ1	HMGB1	MAGEB18	NDUFS1	PKM2	SFXN3	SYN2
ATP5O	CLTC	DNAJC21	FICD	HNRNPC	MAGEB6	NDUFS3	PLIN4	SH3GL3	SYP
AXIN1	CNP	DNAJC4	FNBP4	HOXC11	MAP1LC3A	NDUFS5	PLP1	SH3GLB1	SYT1
BAIAP2	CNTN1	DNALI1	FTL	HOXC4	MAP1LC3B	NDUFV1	PML	SIAH1	SYT2
BASP1	COPB1	DNM1	GAPDH	HSD17B10	MAP3K10	NDUFV2	PPA2	SIN3A	SYVN1
BAZ1A	COX2	DNM2	GFAP	HSP90AB1	MBD1	NEGR1	PPARG	SIRPA	TACC1
BCAP31	COX3	DOCK11	GGA2	HSPA8	MBD4	NEUROD1	PPL	SLC1A2	TAF4
BZW2	COX5A	DOCK9	GIT1	HSPD1	MDH1	NFKB1	PPP2CA	SLC25A11	TANK
C15orf63	COX5B	DPPA4	GNAO1	HTT	MED15	NME4	PQBP1	SLC25A12	TBP
CA4	COX6A1	DPYSL2	GNAZ	HYPM	MED21	NPTN	PRMT2	SLC25A3	TCERG1
CACNA2D1	COX6B1	DUSP10	GNB1	IFT20	MED31	NTM	PRNP	SLC6A1	TGM2
CASP1	COX6C	DYNC1H1	GNB2	IKBKAP	MEF2D	NUB1	PRPF40A	SNAP25	THY1

TM4SF19	TPR	TUBA1A	UBE2K	USP14	VDAC2	XAGE3	ZFC3H1	ZNF655	
TMEM57	TRAF6	TUBB	UBQLN1	USP8	VDAC3	XRCC6	ZFYVE19	ZNF675	
TP53	TRAFD1	TUBG1	UPF3A	USP9X	WAC	YWHAB	ZMAT2	ZNF91	
TPH1	TRIP10	TXNDC11	UQCRC1	UTP14A	WBP4	YWHAG	ZMYND8		
TPI1	TSHZ3	UBAC1	UQCRC2	VCP	WDFY3	ZBTB16	ZNF133		
TPP1	TTC23	UBE2E3	UQCRH	VDAC1	WDR91	ZDHHC17	ZNF451		

Interactors predicted by Biogrid									
AGO2	CTNNB1	FTL	HLA-B	ITCH	MID1	PACSIN1	RAPGEF2	TACC1	UBQLN2
AMFR	DNAJA1	GAPDH	HMG20A	KAT2B	MIPEP	PEX11B	RASA1	TAF4	UPF3A
AP2A2	DNAJB1	GGA2	HMGA1	KIF23	MORC3	PFN2	RCN2	TCERG1	USP8
ATG5	DNALI1	GIT1	HOXC4	LLGL1	MPP1	PIAS1	RFXANK	TKT	USP14
AXIN1	DPPA4	GOLPH3L	HOXC11	MAGEA3	MRFAP1L1	PIAS4	RNF20	TP53	UTP14A
BAG3	DUSP10	GPRASP2	HSP90AB1	MAGEB6	MRFAP1	PIK3R1	RPL4	TRAF6	VCP
C5AR2	ECH1	GRB2	HSPA8	MAGEB18	NBR1	PML	RPS6KB1	TRAFD1	XAGE3
CCDC126	EGFR	GRM1	HUWE1	MAP1LC3A	NCOR1	PPP2CA	SETD2	TRIP10	XRCC6
CHD3	ERCC6L	HAP1	HYPK	MAP3K10	NEUROD1	PRMT2	SH3GL3	TTC23	ZDHHC17
CHUCK	ETV4	HBS1L	HYPM	MAPK3	NME4	PRPF40A	SIAH1	TUBA1A	ZFC3H1
CREB1	EVL	HEL1	IFT20	MBD1	NPY2R	PRPF40B	SNX21	TUBB	ZFYVE19
CREBBP	F8A1	HEY2	IKBKAP	MBD4	NUPL1	PSMB7	SP1	UBAC1	ZMAT2
CRMP1	FAM71F1	HEYL	IKBKG	MED21	OPRM1	PSMC5	SQSTM1	UBE2E3	ZNF451
CRYL1	FEZ1	HIP1	IL17RA	MED31	OPTN	PTGES2	SYT12	UBE2K	ZNF655
CTBP1	FICD	HIST1H3H	ING5	MGRN1	OSTF1	RAB30	SYVN1	UBQLN1	
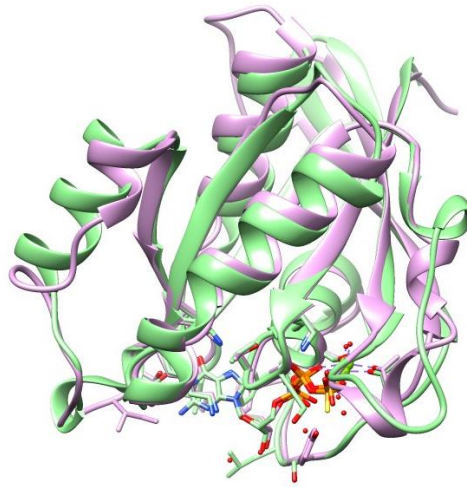
Figure S1. Superimposition of the three-dimensional structures of RanGTP (PDB code 3W3Z, in pink) and Rab11a (PDB code 1OIW, in green).

Figure S2. Structure-based sequence alignment between RanGTP and Rab11a. The positions highlighted with a star are those involved in RanGTP binding to Kap121p (red star, conserved or conservatively substituted in Rab11a; black star non conserved). Residues are coloured by percentage identity.

Score	Expect	Method	Identities	Positives	Gaps
728 bits(1878)	0.0	Compositional matrix adjust.	347/354(98%)	349/354(98%)	0/354(0%)

```
GNAO_M   1   MGCTLSAEERAALERSKAIEKNLKEDGISAAKDVKLLLLGAGESGKSTIVKQMKIIHEDG   60
             MGCTLSAEERAALERSKAIEKNLKEDGISAAKDVKLLLLGAGESGKSTIVKQMKIIHEDG
GNAO_H   1   MGCTLSAEERAALERSKAIEKNLKEDGISAAKDVKLLLLGAGESGKSTIVKQMKIIHEDG   60

GNAO_M  61   FSGEDVKQYKPVVYSNTIQSLAAIVRAMDTLGVEYGDKERKTDSKMVCDVVSRMEDTEPF  120
             FSGEDVKQYKPVVYSNTIQSLAAIVRAMDTLG+EYGDKERK D+KMVCDVVSRMEDTEPF
GNAO_H  61   FSGEDVKQYKPVVYSNTIQSLAAIVRAMDTLGIEYGDKERKADAKMVCDVVSRMEDTEPF  120

GNAO_M 121   SAELLSAMMRLWGDSGIQECFNRSREYQLNDSAKYYLDSLDRIGAGDYQPTEQDILRTRV  180
             SAELLSAMMRLWGDSGIQECFNRSREYQLNDSAKYYLDSLDRIGA DYQPTEQDILRTRV
GNAO_H 121   SAELLSAMMRLWGDSGIQECFNRSREYQLNDSAKYYLDSLDRIGAADYQPTEQDILRTRV  180

GNAO_M 181   KTTGIVETHFTFKNLHFRLFDVGGQRSERKKWIHCFEDVTAIIFCVALSGYDQVLHEDET  240
             KTTGIVETHFTFKNLHFRLFDVGGQRSERKKWIHCFEDVTAIIFCVALSGYDQVLHEDET
GNAO_H 181   KTTGIVETHFTFKNLHFRLFDVGGQRSERKKWIHCFEDVTAIIFCVALSGYDQVLHEDET  240

GNAO_M 241   TNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFGEKIKKSPLTICFPEYPGSNTYEDA  300
             TNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFGEKIKKSPLTICFPEY G NTYEDA
GNAO_H 241   TNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFGEKIKKSPLTICFPEYTGPNTYEDA  300

GNAO_M 301   AAYIQTQFESKNRSPNKEIYCHMTCATDTNNIQVVFDAVTDIIIANNLRGCGLY   354
             AAYIQ QFESKNRSPNKEIYCHMTCATDTNNIQVVFDAVTDIIIANNLRGCGLY
GNAO_H 301   AAYIQAQFESKNRSPNKEIYCHMTCATDTNNIQVVFDAVTDIIIANNLRGCGLY   354
```

Figure S3. Amino acid sequence alignment between human and mouse G(o) α subunit (GNAO_H and GNAO_M, respectively)
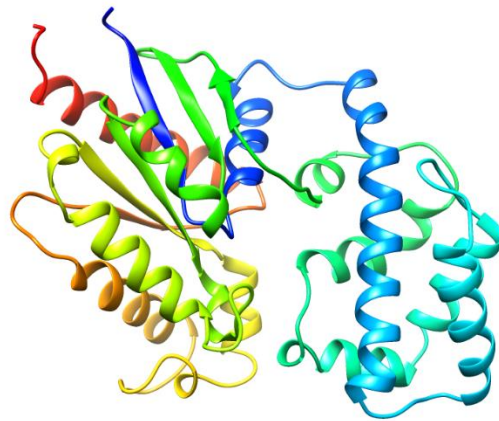
Figure S4. Structural model of the human GNAO1 Gαo subunit obtained with Modeller using the structure of the mouse counterpart (PDB code: 3C7K) as template.

Chapter 3: "Neuroglobin interactome"

Neuroglobin (Ngb), a globin found in vertebrate brain and retina, has a neuroprotective role, as was mentioned in the Introduction. Literature data indicate that htt and Ngb are both members of a neuroprotective axis triggered by the hormone 17β-estradiol (E2) to protect neurons against apoptosis (Nuzzo *et al.* 2017). In the paper "Neuroglobin and friends", the Ngb recognition properties are discussed critically to highlight the roles of this globin in health and disease. Further, an *in silico* analysis of the neuroglobin interactors has been performed to provide a comprehensive view of the whole interactome of Ngb. This work is preliminary to the investigation of the hypothesis that the interaction of Ngb with htt is mediated by common interactors (see next chapter).

WILEY Journal of Molecular Recognition

REVIEW

Neuroglobin and friends

Marco Fiocchetti[1] | **Manuela Cipolletti**[1] | **Valentina Brandi**[1] | **Fabio Polticelli**[1,2] |
Paolo Ascenzi[3] (ID)

[1] Dipartimento di Scienze, Università Roma Tre, Rome, Italy

[2] Istituto Nazionale di Fisica Nucleare, Sezione dell'Università Roma Tre, Rome, Italy

[3] Laboratorio Interdipartimentale di Microscopia Elettronica, Università Roma Tre, Rome, Italy

Correspondence
Paolo Ascenzi, Laboratorio Interdipartimentale di Microscopia Elettronica, Università Roma Tre, Via della Vasca Navale 79, I-00146 Rome, Italy.
Email: ascenzi@uniroma3.it

Abstract

In the year 2000, the third member of the globin family was discovered in human and mouse brain and named neuroglobin (Ngb). Neuroglobin overexpression significantly protects both heart and brain from hypoxic/ischemic and oxidative stress-related insults, whereas decreased Ngb levels lead to an exacerbation of tissue injuries. Moreover, Ngb overexpression protects neurons from mitochondrial dysfunctions and neurodegenerative disorders such as Alzheimer disease; however, it facilitates the survival of cancer cells. Neuroglobin, representing a switch point for cell death and survival, has been reported to recognize a number of proteins involved in several metabolic pathways including ionic homeostasis maintenance, energy metabolism, mitochondrial function, and cell signaling. Here, the recognition properties of Ngb are reviewed to highlight its roles in health and disease.

KEYWORDS

function, interactors, neuroglobin, protein-protein recognition, structure

1 | INTRODUCTION

Nerve globins are widespread within nonvertebrates and vertebrates,[1] the first nerve globin having been recognized in the nerve cord of the polychaete annelid *Aphrodite aculeata* in 1872.[2] In the year 2000, the third member of the globin family expressed in human and mouse brain, and thereby named neuroglobin (Ngb), was discovered.[3] Then, Ngb has been found in other invertebrates and vertebrates, including the nematode *Caenorhabditis elegans*,[4] the zebrafish *Danio rerio*,[5] the pufferfish *Tetraodon nigroviridis*,[5] and the Antarctic icefish *Chaenocephalus aceratus*,[6] with the exception of lampreys and ray-finned fishes.[7] More recently, androglobin, cytoglobin (Cygb), hemoglobin (Hb), and myoglobin have been reported to be expressed in the nervous system of most living organisms. The coexpression of multiple nerve globins could be at the root of specific physiological functions.[8,9]

Abbreviations: Akt, serine/threonine kinase (also named PKB); Cygb, cytoglobin; cyt, cytochrome; ETF, electron transfer flavoprotein; G protein, heterotrimeric guanine-nucleotide-binding protein; GDI, guanine nucleotide dissociation inhibitor; G_α, subunit α of G protein; Hb, hemoglobin; HTT, huntingtin; Mb, myoglobin; MOMP, mitochondrial outer membrane permeabilization; mPTP, mitochondrial permeability transition pore; Ngb, neuroglobin; OGD, oxygen and glucose deprivation; PI3K, phosphotidylinositol 3-kinase; PTEN, phosphatase and tensin homolog; ROS, reactive oxygen species; TIM23, mitochondrial inner membrane; TOM, translocase of the outer membrane; VDAC, voltage-dependent anion channel

Neuroglobin is expressed in neurons of the central and peripheral nervous system; in glioblastoma cell lines; in quiescent astrocytes of the healthy seal brain; in reactive astrocytes in neuropathological models of traumatic injury, infectious, autoimmune, and excitotoxic diseases; in the gastrointestinal tract; and in endocrine organs.[9] Neuroglobin occurs at relatively low concentration (~1 μM) in most tissues and organs including resting neurons. This suggested that the role of Ngb in O_2 supply and/or facilitated diffusion is negligible.[10] However, Ngb may act as an O_2 buffer in the retinal ganglion cell layer and the optic nerve where its concentration is up to ~100 to 200 μM.[8,9,11]

Monomeric Ngb displays the classical 3/3 globin fold adapted to host the hexa-coordinated heme-Fe atom by the proximal and distal His residues (HisF8 and HisE7, respectively) in both the ferrous and the ferric states.[12-14] Therefore, ligand (eg, O_2, CO, and NO) binding and (pseudo-)enzymatic properties (eg, NO scavenging) are possible only upon the cleavage of the distal coordination bond of the heme-Fe atom.[9,15-19]

Interestingly, upon CO binding to the ferrous murine Ngb, the heme slides toward the interior of the protein matrix and partially occupies the large internal tunnel, which has been postulated to facilitate ligand diffusion from the bulk to the metal center. This ligand-dependent structural rearrangement is associated to structural changes, especially of the EF loop, which may be coupled to the modulation of Ngb functions and recognition of partners.[20]

wileyonlinelibrary.com/journal/jmr

In human Ngb, the reversible penta- to hexa-coordination transition of the heme-Fe atom and, in turn, its reactivity, is allosterically modulated. Indeed, the redox state of the cell modulates the formation and the cleavage of the CysCD7-CysD5 disulfide bond, which, in turn, affects the coordination state of the heme-Fe atom of ferrous human Ngb, altering considerably the O_2 affinity (ie, P_{50} changes from 0.9 to 9.2 mm Hg).[12,21,22] However, this mechanism is not operative in murine Ngb because the Cys residue at the CD7 position is substituted by Gly.[14,20] Moreover, the phosphorylation of putative sites SerA7, SerA12, SerAB2, SerCD10, SerCD11, and SerDE3 by intracellular kinases (eg, ERK and PKA) increases the nitrite reductase activity of human Ngb by stabilizing the penta-coordinated ferrous heme-Fe atom. Noteworthy, binding of the scaffold protein 14-3-3 at putative sites ArgA9-ProAB3 and ArgCD7-ProD1 stabilizes phosphorylated human Ngb, inhibiting dephosphorylation.[22]

Among the most remarkable functional properties of Ngb, it has been shown in vitro and in vivo that Ngb is a stress-inducible protein that appears to respond to hypoxic/ischemic/reperfusion oxidative injuries. In the experimental stroke, the recovery after an ischemic insult correlates with the amount of Ngb expressed in the central nervous system. When Ngb expression was increased (by at least a factor of 3), the recovery of the tissue around the occlusion was ameliorated and the necrosis reduced. The opposite was observed if the Ngb expression was reduced.[23] This and other findings have been related to the efficiency of ferrous oxygenated Ngb to scavenge the NO excess produced in the penumbral area around the stroke.[15] Moreover, Ngb overexpression has been postulated to protect neurons from mitochondrial dysfunctions and neurodegenerative disorders such as Alzheimer disease and to promote cancer cell survival.[24-29]

The recent identification of Ngb interactors and inducers, including those involved in ionic homeostasis maintenance, energy metabolism, mitochondrial function, and signaling pathways for cell survival and proliferation, has enlarged the range of functions played by this stress-inducible globin.[24,28,30-34] Here, the Ngb recognition properties are reviewed critically to highlight the roles of this globin in health and disease.

2 | NEUROGLOBIN INTERACTORS

Although the analysis of protein-protein interactions is considered a good approach to investigate protein functions,[35] only few studies reported interactions between Ngb and protein interactors. Biochemical (eg, co-immunoprecipitation and mass spectrometry) and biophysical (eg, surface plasmon resonance and nuclear magnetic resonance spectroscopy) approaches identified several mitochondrial, cytoplasmic, and membrane proteins as Ngb interactors.[22,33,34,36-41] Furthermore, the Ngb interactome was investigated by using high throughput technologies such as the yeast two-hybrid screening system followed by validation in primary cultured mouse cortical neurons[42,43] and by affinity purification-mass spectrometry of recombinant mouse Ngb-immunoreactive complexes in lysates of mouse HN33 neurons.[44] The yeast two-hybrid screening system evidenced 36 proteins that interact with Ngb; however, only 6 proteins were precipitated by anti-Ngb antibody in murine neurons.[42,43] On

the other hand, 8 proteins were found in immunoreactive complexes, although this number increased up to 24 when recombinant Ngb was immunoprecipitated in hypoxia-treated HN33 neurons.[44]

Neuroglobin-protein recognition should be interpreted with circumspection principally due to the different experimental conditions (eg, hypoxia, oxidative stress, Ngb-inducer stimulation, and cell context) that could reflect different Ngb levels and/or compartmentalization. Moreover, indirect interactions may be detected, whereas genuine interactions may be missed due to low levels of Ngb and low affinity.[44]

2.1 | In silico analysis of the neuroglobin interactome

The interactome of human and murine Ngb has been analyzed through the use of different databases that collect predicted and experimentally-validated protein-protein interactions such as PrePPI,[45] IntAct,[46] Hippie,[47] Happi,[48] Biogrid,[49] and PSICQUIC[50] (see Table S1 for a schematic description of the features of these databases). To provide a comprehensive view of the whole interactome of Ngb, the above-mentioned databases have been interrogated and the results are shown in Table 1. A total of 55 and 10 Ngb interactors have been identified for human and murine Ngb, respectively. Because human and murine Ngb share a high degree of sequence identity (95%) and structure similarity[13,14] (backbone root-mean-square deviation = 0.93 Å), the identified interactors are probably relevant for both globins. However, other Ngb interactors, for which experimental data demonstrate the interaction (eg, cytochrome [cyt] c), were not found (see section 3). In this regard, it is worthwhile to note that most of these databases are manually curated, contain only direct protein-protein interactions, and are based on a scoring scheme that reflects the reliability of the techniques used for the detection of a direct interaction. Thus, if a given interaction does not reach a specific score threshold (for instance, because it has been detected by only 1 technique or the technique used does not prove a direct interaction), it is excluded from the list of interactors.

The list of all Ngb interactors shown in Table 1 includes 36, 12, 10, 24, 14, and 10 interactors predicted by PrePPI, Hippie, IntAct, Happi, Biogrid, and PSICQUIC, respectively. In particular, 8 interactors have been identified by 2 methods (ie, Hb subunits ϑ1, β, ζ, δ, and ε, Cygb, Mb, and transcriptional repressor p66-β), 3 interactors by 3 methods (ie, guanine nucleotide-binding protein G_i subunit α2, guanine nucleotide-binding protein G_k subunit α, and sodium/potassium-transporting ATPase subunit β-2), and 9 interactors by 4 methods (ie, flotillin-1, protein VAC14 homolog, lethal3 malignant brain tumor-like protein, guanine nucleotide-binding protein G_O subunit α, guanine nucleotide-binding protein G_i subunit α1, and 12 tryptophan 2,3,-dioxygenase, proto-oncogene c-Rel, uncharacterized protein C1orf94, and coiled-coil domain-containing protein 36).

For 22 of the 36 interactors predicted by PrePPI, a literature reference is not available. These are Hb subunits ϑ-1, μ, γ-1, γ-2, β, ζ, δ, and ε, Cygb, myoglobin, ammonium transporter Rh type A, estrogen receptor 1, D_2 dopamine receptor, acetylcholine receptor subunit γ, aquaporin-1 and -5, neuronal acetylcholine receptor subunit β-2, voltage-dependent T-type calcium channel subunit α-1, interactor protein for cytohesin exchange factor 1, adenosine receptor A1, amyloid β A4 protein, and neuronal acetylcholine receptor subunit α-4.

TABLE 1 Confidence score of neuroglobin (Ngb)-interacting proteins identified by using PrePPI,[45] Hippie,[47] IntAct,[46] Happi,[48] and Biogrid.[49] For mouse Ngb interactors, the PSICQUIC web service has been used[50]

Interactors Gene Name[a]	Methods Protein Name	PrePPI[b]	Hippie[c]	IntAct[b]	Happi[d]	Biogrid	PISQUIC
HBQ1	Hemoglobin subunit ϑ-1	1			MC		
CYGB	Cytoglobin	1			MC		
HBM	Hemoglobin subunit μ	1					
MB	Myoglobin	1			LC		
HBG2	Hemoglobin subunit γ-2	1					
FLOT1	Flotillin-1	1	0.84		HC	+	
HBG1	Hemoglobin subunit γ-1	1					
HBB	Hemoglobin subunit β	1			HC		
HBZ	Hemoglobin subunit ζ	1			MC		
VAC14	Protein VAC14 homolog	0.99	0.63	0.70		+	
L3MBTL3	Lethal(3)malignant brain tumor-like protein	0.99	0.63	0.56		+	
HBD	Hemoglobin subunit δ	0.98			MC		
HBE1	Hemoglobin subunit ε	0.97			MC		
RHAG	Ammonium transporter Rh type A	0.95					
GNAO1	Guanine nucleotide-binding protein G(o) subunit α	0.89	0.85		HC	+	
TDO2	Tryptophan 2,3-dioxygenase	0.89	0.63	0.70		+	
REL	Proto-oncogene c-Rel	0.89	0.63	0.56		+	
C1orf94	Uncharacterized protein C1orf94	0.89	0.63	0.56		+	
CCDC36	Coiled-coil domain-containing protein 36	0.89	0.63	0.70		+	
GNAI1	Guanine nucleotide-binding protein G(i) subunit α-1	0.89	0.85		HC	+	
ESR1	Estrogen receptor 1	0.87					
GNAI2	Guanine nucleotide-binding protein G(i) subunit α-2	0.75	0.62		MC		
GNAI3	Guanine nucleotide-binding protein G(k) subunit α	0.75	0.52		MC		
ATP1B2	Sodium/potassium-transporting ATPase subunit β-2	0.75	0.67		HC		
DRD2	D(2) dopamine receptor	0.65					
CHRNG	Acetylcholine receptor subunit γ	0.65					
AQP1	Aquaporin-1	0.63					
CHRNB2	Neuronal acetylcholine receptor subunit β-2	0.61					
CACNA1I	Voltage-dependent T-type calcium channel subunit α-1I	0.61					
IPCEF1	Interactor protein for cytohesin exchange factors 1	0.61					
ADORA1	Adenosine receptor A1	0.6					
YWHAB	14-3-3 protein β/α	0.58					
AQP5	Aquaporin-5	0.58					
APP	Amyloid β A4 protein	0.53					
CHRNA4	Neuronal acetylcholine receptor subunit α-4	0.53					
GATAD2B	Transcriptional repressor p66-β	0.49		0.49			
CCNB1	G2/mitotic-specific cyclin-B1					+	
MPP2	MAGUK p55 subfamily member 2					+	
TRIM32	E3 ubiquitin-protein ligase TRIM32					+	
AHSP	α-hemoglobin-stabilizing protein				HC		
HBA1	Hemoglobin subunit α				HC		
RAC3	Ras-related C3 botulinum toxin substrate 3				MC		
RAC1	Ras-related C3 botulinum toxin substrate 1				MC		
ARHGDIA	Rho GDP-dissociation inhibitor 1				MC		
GNAO2	Guanine nucleotide binding protein (G protein) α activating activity polypeptide O				MC		
CGB	Choriogonadotropin subunit β				LC		
RIT2	GTP-binding protein Rit2				LC		
CYB5R3	NADH-cytochrome b5 reductase 3				LC		
CHGB	Secretogranin-1				LC		

(Continues)

TABLE 1 (Continued)

Interactors Gene Name[a]	Methods Protein Name	PrePPI[b]	Hippie[c]	IntAct[b]	Happi[d]	Biogrid	PISQUIC
LYAR	Cell growth-regulating nucleolar protein				LC		
PTEN	Phosphatase and tensin homolog					+	
AKT1	RAC-α serine/threonine-protein kinase					+	
DDIT4L	DNA damage-inducible transcript 4-like protein			0.49			
BTBD2	BTB/POZ domain-containing protein 2			0.49			
NIF3L1	NIF3-like protein 1			0.49			
GABARAPL1	Gamma-aminobutyric acid receptor-associated protein-like 1						+
ETFA	Electron transfer flavoprotein subunit α, mitochondrial						+
CYC1	Cytochrome c1, heme protein, mitochondrial						+
UBC	Polyubiquitin-C						+
ATP1B1	Sodium/potassium-transporting ATPase subunit β-1						+
SYT1	Synaptotagmin-1						+
DVL1	Segment polarity protein disheveled homolog DVL-1						+
ATP1B3	Sodium/potassium-transporting ATPase subunit β-3						+
VDAC1	Voltage-dependent anion-selective channel protein 1						+
MAP1A	Microtubule-associated protein 1 A						+

[a]Interactors in bold are those for which the interaction has been experimentally demonstrated. Interactors shaded in gray have been predicted by PrePPI.

[b]The PrePPI and IntAct scores are in the form of a 0 to 1 interaction probability. The type of the interaction between NGB and interactors detected by IntAct is physical. It means an interaction between molecules within the same physical complex, often identified under conditions that suggest that the molecules are in close proximity but not necessarily in direct contact with each other.

[c]For Hippie, a score higher that 0.73 characterizes high-confidence interactors (the interactions being demonstrated by multiple experimental methods).

[d]The HAPPI ranking method works by clustering the distribution of h-scores for all interactions. The distribution of h-score ranges from 0 to 1. Based on the above binned h-score distribution, the ranking star rating for each interaction in the HAPPI database is defined with the following threshold values of h-scores:

1 star = h-score < 0.25 noisy and unsupported interactions.
2 stars = $0.25 \leq h$-score < 0.45 very-low-confidence interactions.
3 stars = $0.45 \leq h$-score < 0.75 low-confidence interactions.
4 stars = $0.75 \leq h$-score < 0.90 medium-confidence interactions.
5 stars = $0.90 \leq h$-score ≤ 1 high-confidence interactions.

Biogrid and PISQUIC do not provide any confidence value.

2.2 | In vivo and in vitro analysis of the neuroglobin interactome

Table 1 reports Ngb interactors for which sufficient experimental evidence is available to warrant a high reliability of the direct interaction and those predicted by the PrePPI pipeline. However, the abundance of Ngb and its affinity for the binding partners as well as indirect interactions, cell type, cell developmental stage, and endogenous and exogenous stimuli make challenging the identification of Ngb interactors.[51-53]

In this regard, an independent analysis of the Ngb-related literature allowed to derive a number of additional experimentally-supported Ngb interactors playing a relevant role in Ngb actions. In fact, Ngb interactors, such as cystatin c,[38] cyt c,[36,40,54] huntingtin (HTT),[90] and prion protein[39,55] have not been identified by the bioinformatic methods used here, even though there is at least 1 experimental evidence proving Ngb-protein recognition.

On the other hand, the estrogen receptors α (ESR1) and β (ESR2), for which no experimental evidence is available of a physical interaction with Ngb, are known to be involved together with Ngb in the signal transduction pathways that lead to 17β-estradiol-induced extra nervous cancer cell and neuronal survival.[24,32,56]

3 | NEUROGLOBIN INTERACTORS IN HEALTH AND DISEASE

A vast amount of intracellular functions mainly linked to Ngb neuroprotective properties has been reported.[17,57-59] Several evidence support the involvement of Ngb in metabolic, oxidative/hypoxia, and cell survival/apoptotic pathways at different steps, even if the Ngb action mechanisms are poorly defined. In this context, it is increasingly clear that most of Ngb biological functions arise from the complex interaction networks that Ngb establishes with intracellular proteins.

3.1 | Neuroglobin and mitochondrial-associated proteins

The cytosolic Ngb localization has been considered a paradigm for a long time. However, several evidence support a physical and functional association of Ngb with mitochondria.[28,29,60,61] The identification of diverse mitochondrial associated proteins as Ngb interactors further supports a role of Ngb in mitochondrial functions.

3.1.1 | Electron-transferring flavoprotein α subunit and cytochrome c_1

Electron transfer flavoprotein (ETF) acts as an obligatory electron carrier present in the mitochondrial matrix or associated with the inner mitochondrial membrane. Electron transfer flavoprotein transfers electrons from dehydrogenases to the ubiquinone pool in the inner mitochondrial membrane. In β-oxidation, ETF provides electrons for the cyt bc_1 complex (complex III) to reduce cyt c.[62,63]

Yeast two-hybrid screening and co-immunoprecipitation assays in primary mouse cortical neurons evidenced the interaction between Ngb and the subunit α of ETF as well as cyt c_1, a subunit of the mitochondrial complex III, which is a component of the respiratory chain and a major source of reactive nitrogen and oxygen species (reactive nitrogen species and reactive oxygen species [ROS]).[27] Of note, oxygen and glucose deprivation (OGD) conditions significantly increase the interaction of Ngb and cyt c_1 leading to a transient suppression of OGD-mediated mitochondrial complex III activity and ROS production; this results in cell protection against ischemic/hypoxic neurotoxicity.[34] Otherwise, Ngb knockdown in rat retinal ganglion cells impairs the activities of the respiratory chain complexes I and III and induces cell degeneration.[64] Furthermore, Ngb overexpression has been reported to rescue the ATP loss during hypoxia/re-oxygenation injury in both primary mouse cortical neurons[65] and neuroblastoma cell line SH-SY5Y.[66]

Taken together, these data indicate a specific role of Ngb in the regulation and maintenance of cellular bioenergetics.

3.1.2 | Cytochrome c

Mitochondria act as sentinels of diverse stress signals that emanate from other intracellular organelles, as well as a variety of environmental stress signals such as low nutrient levels, increased calcium levels, and oxidative stress. In response to these signals, mitochondria break leading to the release of pro-apoptotic molecules and the induction of the apoptotic cascade. This process occurs in two phases: mobilization and translocation through the outer mitochondrial membrane.[67] In the early steps of apoptosis, cardiolipin peroxidation by ROS or cyt c is considered one of the main factors responsible for the dissociation of the cardiolipin-cyt c complex and the consequent mobilization of cyt c.[67-69] Apoptotic signals result in the opening of the mitochondrial permeability transition pore (mPTP), a multiprotein complex at the contact site of the inner and outer membranes.

Mitochondrial permeability transition pore leads to the release of pro-apoptotic molecules, including cyt c that constitutes the "point of no return" in the sequence that leads to cell death. Once in the cytosol, cyt c recruits the apoptotic protease-activating factor-1, building the apoptosome macromolecular complex. Each apoptosome complex, recruiting and activating caspase-9, gives rise to the enzymatic apoptotic reaction cascade.[67,70,71]

The direct impact of Ngb on the mitochondrial intrinsic apoptotic pathway and, in particular, on the cyt c reactivity highlights the role of this globin on cell protection. A very fast electron transfer reaction from ferrous Ngb to ferric cyt c has been reported.[36,72] Hence, cytosolic Ngb might impair the cell death pathway via the reduction of ferric cyt c to ferrous cyt c, which is not active in the apoptosome assembly.[70,73,74] It is worthwhile to note that this mechanism requires a 1:3 cyt c:Ngb stoichiometry to prevent apoptosis. Therefore, the increase of cytosolic cyt c, which mimics the situation occurring in highly stressed cells, requires a higher Ngb concentration to avoid caspase activation and the consequent apoptotic cascade.[70,71,75]

Ferric cyt c reduction by ferrous Ngb occurs via direct protein-protein interaction.[36] Surface plasmon resonance and quartz nanopipette quantitative studies yielded an equilibrium dissociation constant of about 45 μM, which rises up to 200 μM at high ionic strength. The dependence of the equilibrium constant on ionic strength and temperature suggests that the interaction is largely electrostatic in nature.[36,76] Nuclear magnetic resonance analysis identified the formation of a weak Ngb-cyt c complex, and theoretical studies suggested that cyt c binds to Ngb at a site covering the area of the exposed heme (Figure 1).[36] In the absence of a crystal structure for the Ngb-cyt c complex, computational methods have been applied to further analyze Ngb-cyt c recognition. The application of soft docking algorithms allowed the identification of Ngb residues Glu60 and Glu87 and cyt c side chains Lys25 and Lys72 as pivotal in the formation of the binding interface. Of note, the Lys25 and Lys72 residues are also essential for the interaction between cyt c and the apoptotic protease-activating factor-1 in the cytosol, suggesting that the interaction of Ngb with cyt c could compete with the apoptosome formation.[36] Therefore, even in the absence of any redox reaction between Ngb and cyt c, Ngb could abolish the apoptotic cascade.[70,71,75] In addition, as demonstrated by co-immunoprecipitation assays and confocal microscopy, Ngb, recruited to mitochondria, impacts on the apoptotic cascade by interacting with cyt c and impairs its release to cytosol.[24,40,77]
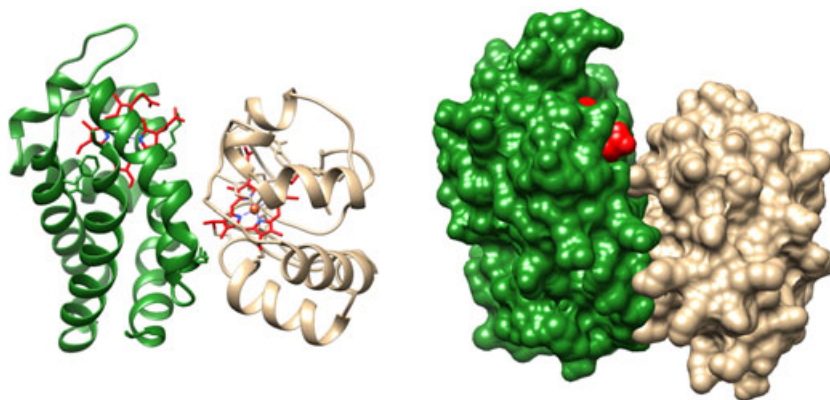

FIGURE 1 Schematic representation of the structural model of the neuroglobin (Ngb)-cyt c complex obtained by docking simulations. (left panel) Ribbon representation. (right panel) Molecular surface representation. Ngb is colored in green. The Ngb and cyt c heme groups are colored in red. See supporting information for details

Multiple evidence support the role of cytosolic Ngb in reacting with the initially-released cyt *c* during an apoptotic stimulus, avoiding the apoptotic avalanche.[70,71,75] Nevertheless, it appears that basal levels of Ngb found in neurons have the ability to protect cells against a mild challenge, which can occur during normal cell function.[75] On the other hand, during a high stress condition, such as strong oxidative stress, high levels of Ngb and its translocation into mitochondria are needed to prevent the apoptotic cascade.[24,40,77] Even if Ngb is assumed to be a stress-inducible protein, only 17β-estradiol, a well-known Ngb inducer, is able to increase both the whole cell concentration of Ngb and its level in the mitochondrial compartment to impair cell injury.[24,56,77]

As a whole, the Ngb-cyt *c* interaction appears to be a crucial event in both mitochondrial and post-mitochondrial apoptotic steps, opening the possibility that different Ngb pools synergize with each other to protect cells by affecting cyt *c* action and localization.

3.1.3 | Voltage-dependent anion channel

In the complex machinery that drives the cell through programmed cell death via the intrinsic pathway, the critical event is the mitochondrial outer membrane permeabilization (MOMP) that governs the release of pro-apoptotic molecules from mitochondria.[78,79] Most of the apoptotic stimuli require MOMP for the activation of caspases and the apoptotic program.[78] Two different models have been proposed to explain the execution of MOMP, both of which share the basic idea that membrane permeabilization is achieved via the coordinated action of both pro- and anti-apoptotic members of the BCL-2 protein family.[80] The first model assumes that MOMP is regulated by the formation of oligomeric pores (megachannels) through the homo-oligomerization of the pro-apoptotic molecules BCL-2 antagonist killer 1 and BCL-2 associated X protein.[81] The second model proposes that the modulation of mPTP may play a key role in MOMP.[79,81] Among several proteins, which seem to be part of mPTP, the voltage-dependent anion channel (VDAC) has been considered crucial for a long time.[81,82] It is located on the outer membrane of mitochondria and

constitutes the main pathway for metabolite diffusion across mitochondria.[79] Although knockout experiments of several isoforms of VDAC suggest a dispensable role of this protein in MOMP,[83] VDAC-dependent membrane permeability transition has been defined as an upstream mechanism playing a causal role in stress-induced apoptosis.[82,84]

Yeast two-hybrid, co-immunoprecipitation and immuno-cytochemistry assays in primary mouse cortical neurons evidenced the interaction between VDAC and Ngb.[34,42] Bioinformatic approaches predicted the interaction of Ngb with VDAC at level of residues located in the channel opening, supporting the hypothesis that Ngb may directly regulate the permeability of the mitochondrial outer membrane (Figure 2).[54,85] Indeed, Ngb overexpression leads to an amelioration of the loss of mitochondrial membrane potential during hypoxia/reoxygenation challenge in mouse cortical neurons[65] and in a human neuroblastoma cell line (SH-SY5Y; see Ascenzi et al[9]). Furthermore, overexpressed Ngb in mouse primary neurons inhibits the OGD-induced formation of the mPTP and the consequent mitochondrial NAD$^+$ and cyt *c* release. The increased association between Ngb and VDAC induced by OGD is crucial in the above-mentioned events, supporting the notion that this interaction may represent a molecular mechanism in Ngb neuroprotective function. Nonetheless, the overexpression of Ngb and its interaction with VDAC sustains only a partial reduction of cell death, indicating that Ngb must be involved also in other pro-survival and anti-apoptotic mechanisms.[34]

3.1.4 | Huntingtin

Huntingtin is a well-known protein because the abnormal expansion of its *N*-terminal glutamine residue stretch causes the neurodegenerative disorder Huntington disease.[86] Huntingtin participates in several pathways including cellular dynamics (endocytosis, trafficking, and adhesion), metabolism, protein turnover, and gene expression.[86] Of note, mitochondrial proteins including VDAC and cyt *c* have been reported to be among the HTT interactors.[86,87] Furthermore, mutated
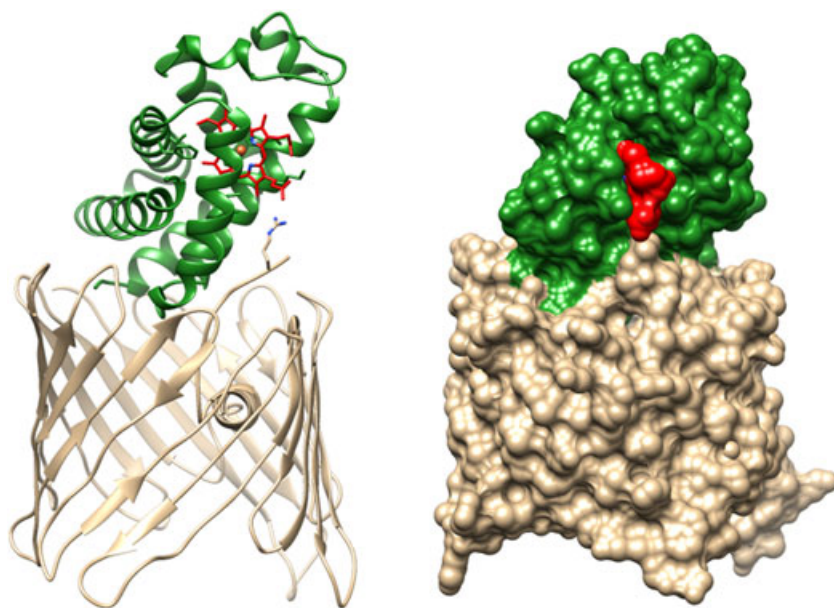
FIGURE 2 Schematic representation of the structural model of the neuroglobin (Ngb)-voltage-dependent anion channel (VDAC) complex obtained by docking simulations. (left panel) Ribbon representation. (right panel) Molecular surface representation. Ngb is colored in green. The Ngb heme group is colored in red. See supporting information for details

HTT affects mitochondria functionality, including biogenesis, protein import, complex assembly, fission and fusion, degradation (mitophagy), and transport.[88]

In human neuroblastoma and murine striatal cells, confocal microscopy analysis, co-immunoprecipitation, and proximity ligation assay demonstrated the association of Ngb with HTT.[89,90] The Ngb-HTT interaction appears to be crucial in the 17β-estradiol-dependent mitochondrial relocalization of Ngb, supporting a role of HTT as a molecular scaffold and protein carrier that promotes Ngb trafficking toward mitochondria.[89,90] However, the molecular mechanism(s) of Ngb translocation into mitochondria is not fully defined. Mitochondrial protein import generally occurs through recognition of a mitochondrial-targeting sequence by the translocase of the outer membrane (TOM)[91,92]; however, Ngb lacks any mitochondrial-targeting sequence.[27] The Ngb-VDAC interaction and the opening of the mPTP have been demonstrated as one of the possible pathways that allow Ngb translocation into mitochondria.[27] Interestingly, mutated HTT interacts with the translocase of the mitochondrial inner membrane (TIM23) and inhibits the mitochondrial protein translocation via the TOM-TIM23 complex.[88] The association among HTT, VDAC, and Ngb may be one of the plausible processes that drive Ngb to mitochondria.

3.2 | Neuroglobin and signaling proteins

The list of predicted and/or experimentally confirmed Ngb interactors (Table 1) includes several signaling proteins (eg, GNAOI, AKT, Rel, Rac1, and Rac3) that support a functional role of Ngb in the modulation of intracellular signaling pathways.[8,9,57,93] This role is further supported by Ngb trafficking into cell compartments[24,28,42,77,89] and by its structural changes upon binding of exogenous ligands or upon Ngb oxidation.[41,93]

3.2.1 | Guanine nucleotide binding proteins

The α-subunit (G_α) of a heterotrimeric guanine-nucleotide-binding protein (G protein) was the very first Ngb interactor identified by surface plasmon resonance.[33] Heterotrimeric G proteins are typically divided into 4 main classes based on the different G_α subunits: $G_{\alpha s}$, $G_{\alpha i}$, $G_{\alpha q}$, and $G_{\alpha 12}$,[94] all of which have well-defined targets and determine the specificity and temporal characteristics of the cellular responses.[95]

G proteins act as molecular switches in intracellular signaling, coupling the activation of G-protein-coupled receptors to a cellular response upon an extracellular stimulus.[96] G proteins are made up by 3 different subunits, α, β, and γ,[33,96] and the signal transduction function of the complex depends on the ability of the G_α subunit to bind guanine nucleotides.[33,94,96] The inactive heterotrimeric complex $G_{\alpha\beta\gamma}$ occurs when G_α binds GDP and the constitutive dimer $G_{\beta\gamma}$. Extracellular stimuli, including hormones, neurotransmitters, and chemokines, activate receptors and induce the release of GDP from G_α. Binding of GTP to the α subunit leads to the dissociation of the activated GTP-bound Gα subunit from the complex $G_{\beta\gamma}$, each of which interact and regulate different downstream molecular effectors.[33,94,96] The intrinsic GTPase activity of the G_α subunit determines the timing of cellular signaling activation; GTP hydrolysis to GDP causes a conformational change of the G_α subunit and leads to the reconstitution of the inactive heterotrimeric $G_{\alpha\beta\gamma}$ complex.[33,94,95]

Although both the ferric and ferrous forms of Ngb are hexacoordinated,[16] ferric human Ngb is the only form able to interact exclusively with the GDP-bound $G_{\alpha i}$ subunit.[33,41] The value of the equilibrium dissociation constant for the interaction of ferric Ngb with $G_{\alpha i}$ obtained by surface plasmon resonance was 0.6 μM[33] Even if in vivo the association between Ngb and G_α has been questioned,[97] several evidence support the crucial role of this interaction in Ngb neuroprotective activities.[41,98-101] Indeed, because of the discovery of a Ngb-$G_{\alpha i}$ complex formation, ferric Ngb has been reported to inhibit the rate of GDP/GTP exchange, exerting the role of guanine nucleotide dissociation inhibitor (GDI).[33,100] In particular, the interaction of ferric Ngb with the GDP-bound $G_{\alpha i}$ subunit inhibits G_α activity and prevents $G_{\beta\gamma}$ rebinding to G_α, thus enhancing the $G_{\beta\gamma}$-dependent pathway that favors cell survival via the activation of phosphotidylinositol 3-kinase (PI3K).[33,41,102]

Mutagenesis studies of Ngb and $G_{\alpha i}$ revealed that the negative charges of Glu53, Glu60, and Glu118 of human Ngb and the positive charges of Lys46, Lys70, Arg208, Lys209, and/or Lys210 of $G_{\alpha i}$ are crucial for complex formation.[98-100,103] The identification of these structural determinants led to a molecular docking model of the complex between ferric human Ngb and $G_{\alpha i}$. In this model, Ngb Glu60 interacts with Arg208, Lys209, and/or Lys210 of $G_{\alpha i}$, while Ngb Glu53 interacts with Lys46 of $G_{\alpha i}$. Finally, Ngb Glu118 interacts with Lys70 of $G_{\alpha i}$ (Figure 3).[98] The docking model suggests that, in the Ngb-$G_{\alpha i}$ complex, Ngb should be positioned near the $G_{\alpha i}$ GDP-binding site. Furthermore, the sequence surrounding the Glu60 residue in
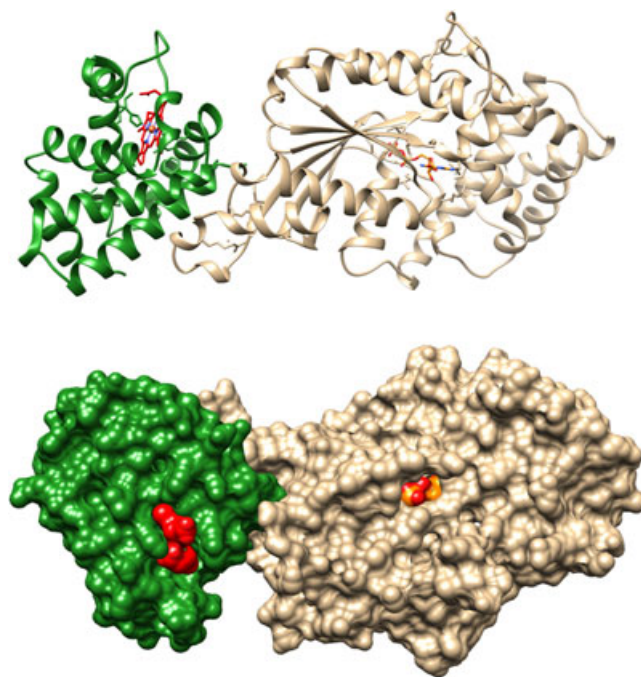

FIGURE 3 Schematic representation of the structural model of the neuroglobin (Ngb)-$G_{\alpha i}$ complex obtained by docking simulations. (left panel) Ribbon representation. (right panel) Molecular surface representation. Ngb is colored in green. The Ngb heme group is colored in red. The $G_{\alpha i}$ ligand (GTP) is coloured by atom type. See supporting information for details

human Ngb displays a similarity with the GDP-$G_{\alpha i}$ binding peptides RSA-1 and KB-752.[98-100]

Because ferrous Ngb does not exhibit any interaction with $G_{\alpha i}$ and GDI function, Ngb binding and activity are linked to its redox state.[33,41,93] Therefore, under oxidative stress, ferric Ngb would undergo a structural change and could be recruited to the plasma membrane lipid rafts by interacting with flotillin-1. Therefore, ferric Ngb binds to $G_{\alpha i}$ exerting its GDI function and preserving cells from stress-induced death.[41] These data open new possibilities about the role of Ngb in sensing the intracellular redox state and in translating a stress condition to the signal transduction pathway.[33,41] Furthermore, Ngb levels represent a marker of oxidative stress also in breast cancer cells MCF-7 in a similar way.[104]

The involvement of Ngb in intracellular signaling processes is limited not only to the interaction with heterotrimeric G proteins but also in the regulation of small GTP-binding proteins of the Rho family,[105] consisting of at least 14 distinct members including RhoA, Rac1, and Cdc42.[106] Although Rho, Rac, and Cdc42 proteins are involved in several cellular processes, including membrane trafficking as well as cell growth and development, this class of proteins primarily regulates cytoskeleton reorganization in response to extracellular signals.[106] Rho proteins mediate a large part of the cytoskeleton remodeling, mitochondrial aggregation, and lipid raft polarization in response to neuronal insults, including hypoxia, to mount a cellular response. Evidence indicate that cytoskeleton changes represent an early event in the death signal transduction that precedes cyt c release and caspase-3 activation in hypoxic neuronal death.[105]

Similarly to the G_{α} subunit of heterotrimeric G proteins, the regulation of the small G proteins function is linked to their intrinsic GTPase activity and consequent ability to cycle through inactive GDP- and active GTP-bound forms. In the cytosol, Rho GTPases are kept inactive through complexation with a GDI. Dissociation of GDI represents the first event required to convert Rho GTPase to the active GTP-bound form and to translocate it to the membrane.[105,106]

Co-immunoprecipitation assays in mouse cortical neurons demonstrated the association of Ngb with Rho-GDI, Rac1, and RhoA, which is significantly increased after hypoxia stimulation.[105] Based on this interaction, an alternative neuroprotective mechanism of Ngb has been defined relying on its inhibitory effect on the dissociation of Rac1 from its GDI. This event causes the blockage of Rac1-mediated actin assembly/membrane polarization and the consequent death signaling transduction.[105] As it occurs during an oxidative stress, hypoxia challenge increases the Ngb localization at the plasma membrane and its interaction with flotillin-1, suggesting that Ngb may interact with and regulate Rho-GDI-GTPase complex directly at the cytoplasmic face of membrane microdomains.[105]

As a whole, it is plausible that both cytosolic- and mitochondrial-associated Ngb act together in time and space to initiate a cellular response to stress conditions.

3.2.2 | Akt

The serine/threonine kinases (Akt), including Akt1, Akt2, and Akt3, are central key intermediates in downstream cell signaling of cytokines, growth factors, hormones, and other cellular stimuli. Akt activation contributes to different cellular processes including proliferation, glucose metabolism, cell size and growth, and cell survival.[107,108] Two different regulatory phosphorylation sites exist in the Akt sequence: Thr308 and Ser473. In resting conditions, Akt exists in its unphosphorylated state in the cytosolic compartment. Activation of Akt by growth factors involves the activation of PI3K, although a PI3K-independent activation has been reported.[109] At the plasma membrane, activated PI3K produces the second messenger phosphatidylinositol (3,4,5)-trisphosphate from phosphatidylinositol (4,5)-bisphosphate. This crucial phosphorylation reaction is tightly regulated by the phosphatase and tensin homolog (PTEN), known as the main upstream negative modulator of Akt-activated signaling.[109] Phosphatidylinositol (3,4,5)-trisphosphate recruits Akt to the plasma membrane and alters its conformation allowing Akt phosphorylation by the phosphoinositide-dependent kinase-1.[107,109] The activation of this pathway results in a critical survival signal that allows cells to withstand apoptotic stimuli. Evidence suggest that Akt signaling may directly regulate cell survival through the phosphorylation and inactivation of the pro-apoptotic proteins BAD and caspase 9.[108,109] Alternatively, the Akt pathway converges on diverse transcription factors, such as FOXO1, p53, and CREB, and indirectly controls cell survival by modulating the expression of both pro- (BAD) and anti- (BCL-2 and Mcl-1) apoptotic proteins.[108,109]

Given the crucial role of Akt signaling in the pro-survival activated pathway and the well-defined neuroprotective role of Ngb, a possible connection between them has been proposed. Indeed, the Ngb-dependent activation of the Akt pathway has been found to be critical in the neuroprotective effect of Ngb against oxidative stress insults,[30] β-amyloid toxicity,[110-112] and ischemic brain injury.[113] In addition, Ngb overexpression attenuates, via Akt, the hyperphosphorylation of tau protein, a pathological hallmark of Alzheimer disease.[114] Recently, a function of Ngb as a neuropeptide able to activate pro-survival pathways against oxidative stress has been hypothesized. Although further studies will be needed to dissect the mechanism by which extracellular Ngb could affect cell signaling, the Akt phosphorylation appears to be central also in Ngb neuroprotection.[115]

Bimolecular fluorescence complementation and glutathione S-transferase pull-down assays revealed the interaction of Ngb with both Akt and its upstream regulator PTEN.[116] These interactions appear to be dynamically regulated by the intracellular environment such as redox state and/or hypoxia, as shown by the differential interaction of wild type and mutant Ngb with Akt or PTEN in the modulation of neurite outgrowth.[116] Thus, several evidence support a role of Ngb as an upstream regulator of PI3K/Akt pathway. However, the exact mechanism by which direct binding of Ngb could facilitate Akt phosphorylation is not completely clear.

Interestingly, the rapid and persistent activation of PI3K/Akt pathway has been revealed as the main signal that mediates the 17β-estradiol-dependent Ngb upregulation and mitochondrial localization in human breast cancer cells.[24] This evidence suggests a possible crosstalk and feedback mechanism leading to the differential modulation of both Ngb and Akt, linked to intracellular processes of cell survival. In this regard, although the signaling function of Ngb remains largely unknown, this globin appears to be post-translationally regulated.[22] Neuroglobin phosphorylation has been suggested to serve as a sensor

signal of different stress conditions, including hypoxia, and to regulate the reaction state of the protein and macromolecular recognition properties (eg, with 14-3-3).[22] In summary, further studies will be needed to dissect if the Ngb/Akt pathway could include a direct modification of the globin phosphorylation state and impact on its function, and intracellular localization.

4 | CONCLUSION AND PERSPECTIVES

In the postgenomic era, one of the most important problems is assigning functions to novel proteins. Among the several approaches applied to solve this problem, the analysis of protein-protein interaction networks and functional data of interactors' partners represent one of the most promising ways.[117] Indeed, cellular functions, including architecture, regulation, metabolism, and signaling, are deeply affected by homotypic and heterotypic protein-protein interactions.[51] The identification and characterization of these interactions and of the entire interaction network represent a required step to understand the molecular mechanisms of the intracellular processes.[53] In this context, over the last decade, increasing efforts have been made to elucidate the molecular mechanisms at the root of Ngb intracellular protective functions. Indeed, although it has become clear that the overexpression of Ngb is needed for the protective function of this globin, the exact mechanism by which it occurs is still not properly defined. Most of the proposed mechanisms for Ngb physiological role, such O_2 carrier and reactive nitrogen species/ROS scavenger, have been prompted by the typical globin structure of the protein.[17,57-59] In addition, among the hypothesis put forward, several evidence support a role of this protein in different intracellular signaling pathways involved in the cell adaptive response to oxidative stress and/or hypoxia, ROS generation, ATP production, and apoptotic cascade.[8,29,57,70,71,76,93,118] This intriguing vision of Ngb functions is supported by the Ngb low intracellular concentration and, its expression in cytosol, inner membrane of mitochondria and cell nucleus,[24,40,77,119] and, particularly, by the broad spectrum of Ngb interacting proteins. Different conditions including hypoxia and oxidative stress challenge Ngb expression level/localization, and cell type specificity significantly affect Ngb interactions, adding obstacles to the full understanding of the globin interactome.[9,33,44] Nevertheless, the results obtained on the Ngb neuroprotective function are promising and studies on Ngb interacting proteins may be of great value for defining the mechanisms underlying Ngb activities. Finally, the definition of the Ngb interaction surface with other proteins represents a first step to devise strategies to modulate protein recognition and, consequently, Ngb intracellular functions.

REFERENCES

1. Vinogradov SN, Hoogewijs D, Bailly X, et al. A phylogenomic profile of globins. *BMC Evol Biol*. 2006;6:31.

2. Lankester ER. A contribution to the knowledge of haemoglobin. *Proc R Soc Lond*. 1872;21:70-81.

3. Burmester T, Weich B, Reinhardt S, Hankeln T. A vertebrate globin expressed in the brain. *Nature*. 2000;407(6803):520-523.

4. Tilleman L, Germani F, De Henau S, et al. Globins in *Caenorhabditis elegans*. *IUBMB Life*. 2011;63(3):166-174.

5. Awenius C, Hankeln T, Burmester T. Neuroglobins from the zebrafish *Danio rerio* and the pufferfish *Tetraodon nigroviridis*. *Biochem Biophys Res Commun*. 2001;287(2):418-421.

6. Giordano D, Boron I, Abbruzzetti S, et al. Biophysical characterisation of neuroglobin of the icefish, a natural knockout for hemoglobin and myoglobin: Comparison with human neuroglobin. *PLoS One*. 2012;7(12): e44508

7. Burmester T, Hankeln T. Function and evolution of vertebrate globins. *Acta Physiol (Oxf)*. 2014;211(3):501-514.

8. Ascenzi P, Gustincich S, Marino M. Mammalian nerve globins in search of functions. *IUBMB Life*. 2014;66(4):268-276.

9. Ascenzi P, di Masi A, Leboffe L, et al. Neuroglobin: From structure to function in health and disease. *Mol Aspects Med*. 2016;52:1-48.

10. Brunori M. Nitric oxide, cytochrome-c oxidase and myoglobin. *Trends Biochem Sci*. 2001;26(1):21-23.

11. Schmidt M, Giessl A, Laufs T, Hankeln T, Wolfrum U, Burmester T. How does the eye breathe? Evidence for neuroglobin-mediated oxygen supply in the mammalian retina. *J Biol Chem*. 2003;278(3): 1932-1935.

12. Guimarães BG, Hamdane D, Lechauve C, Marden MC, Golinelli-Pimpaneau B. The crystal structure of wild-type human brain neuroglobin reveals flexibility of the disulfide bond that regulates oxygen affinity. *Acta Crystallogr D Biol Crystallogr*. 2014;70(Pt 4): 1005-1014.

13. Pesce A, Dewilde S, Nardini M, et al. Human brain neuroglobin structure reveals a distinct mode of controlling oxygen affinity. *Structure*. 2003;11(9):1087-1095.

14. Vallone B, Nienhaus K, Brunori M, Nienhaus GU. The structure of murine neuroglobin: Novel pathways for ligand migration and binding. *Proteins*. 2004;56(1):85-92.

15. Brunori M, Giuffrè A, Nienhaus K, Nienhaus GU, Scandurra FM, Vallone B. Neuroglobin, nitric oxide, and oxygen: Functional pathways and conformational changes. *Proc Natl Acad Sci U S A*. 2005;102(24):8483-8488.

16. Dewilde S, Kiger L, Burmester T, et al. Biochemical characterization and ligand binding properties of neuroglobin, a novel member of the globin family. *J Biol Chem*. 2001;276(42):38,949-38,955.

17. Herold S, Fago A, Weber RE, Dewilde S, Moens L. Reactivity studies of the Fe(III) and Fe(II)NO forms of human neuroglobin reveal a potential role against oxidative stress. *J Biol Chem*. 2004;279(22): 22,841-22,847.

18. Trent JT III, Watts RA, Hargrove MS. Human neuroglobin, a hexacoordinate hemoglobin that reversibly binds oxygen. *J Biol Chem*. 2001;276(32):30,106-30,110.

19. Van Doorslaer S, Dewilde S, Kiger L, et al. Nitric oxide binding properties of neuroglobin. A characterization by EPR and flash photolysis. *J Biol Chem*. 2003;278(7):4919-4925.

20. Vallone B, Nienhaus K, Matthes A, Brunori M, Nienhaus GU. The structure of carbonmonoxy neuroglobin reveals a heme-sliding mechanism for control of ligand affinity. *Proc Natl Acad Sci U S A*. 2004;101(50):17,351-17,356.

21. Hamdane D, Kiger L, Dewilde S, et al. The redox state of the cell regulates the ligand binding affinity of human neuroglobin and cytoglobin. *J Biol Chem*. 2003;278(51):51,713-51,721.

22. Jayaraman T, Tejero J, Chen BB, et al. 14-3-3 binding and phosphorylation of neuroglobin during hypoxia modulate six-to-five heme pocket coordination and rate of nitrite reduction to nitric oxide. *J Biol Chem*. 2011;286(49):42,679-42,689.

23. Sun Y, Jin K, Peel A, Mao XO, Xie L, Greenberg DA. Neuroglobin protects the brain from experimental stroke in vivo. *Proc Natl Acad Sci U S A*. 2003;100(6):3497-3500.

24. Fiocchetti M, Nuzzo MT, Totta P, Acconcia F, Ascenzi P, Marino M. Neuroglobin, a pro-survival player in estrogen receptor α-positive cancer cells. *Cell Death Dis*. 2014;5:e1449

25. Khan AA, Sun Y, Jin K, et al. A neuroglobin-overexpressing transgenic mouse. *Gene*. 2007;398(1-2):172-176.

26. Szymanski M, Wang R, Fallin MD, Bassett SS, Avramopoulos D. Neuroglobin and Alzheimer's dementia: Genetic association and gene expression changes. *Neurobiol Aging*. 2010;31(11):1835-1842.

27. Yu Z, Xu J, Liu N, et al. Mitochondrial distribution of neuroglobin and its response to oxygen-glucose deprivation in primary-cultured mouse cortical neurons. *Neuroscience*. 2012;218:235-242.

28. Yu Z, Liu N, Li Y, Xu J, Wang X. Neuroglobin overexpression inhibits oxygen-glucose deprivation-induced mitochondrial permeability transition pore opening in primary cultured mouse cortical neurons. *Neurobiol Dis*. 2013;56:95-103.

29. Yu Z, Poppe JL, Wang X. Mitochondrial mechanisms of neuroglobin's neuroprotection. *Oxid Med Cell Longev*. 2013;2013:756989.

30. Antao ST, Duong TT, Aran R, Witting PK. Neuroglobin overexpression in cultured human neuronal cells protects against hydrogen peroxide insult via activating phosphoinositide-3 kinase and opening the mitochondrial K_{ATP} channel. *Antioxid Redox Signal*. 2010;13(6):769-781.

31. Cai B, Li W, Mao X, et al. Neuroglobin overexpression inhibits AMPK signaling and promotes cell anabolism. *Mol Neurobiol*. 2016;53(2):1254-1265.

32. De Marinis E, Ascenzi P, Pellegrini M, et al. 17β-estradiol—a new modulator of neuroglobin levels in neurons: Role in neuroprotection against H_2O_2-induced toxicity. *Neurosignals*. 2010;18(4):223-235.

33. Wakasugi K, Nakano T, Morishima I. Oxidized human neuroglobin acts as a heterotrimeric Gα protein guanine nucleotide dissociation inhibitor. *J Biol Chem*. 2003;278(38):36,505-36,512.

34. Yu Z, Zhang Y, Liu N, et al. Roles of neuroglobin binding to mitochondrial complex III subunit cytochrome c_1 in oxygen-glucose deprivation-induced neurotoxicity in primary neurons. *Mol Neurobiol*. 2016;53(5):3249-3257.

35. Droit A, Poirier GG, Hunter JM. Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function. *J Mol Endocrinol*. 2005;34(2):263-280.

36. Bønding SH, Henty K, Dingley AJ, Brittain T. The binding of cytochrome *c* to neuroglobin: A docking and surface plasmon resonance study. *Int J Biol Macromol*. 2008;43(3):295-299.

37. Wakasugi K, Nakano T, Kitatsuji C, Morishima I. Human neuroglobin interacts with flotillin-1, a lipid raft microdomain-associated protein. *Biochem Biophys Res Commun*. 2004;318(2):453-460.

38. Wakasugi K, Nakano T, Morishima I. Association of human neuroglobin with cystatin C, a cysteine proteinase inhibitor. *Biochemistry*. 2004;43(18):5119-5125.

39. Palladino P, Scaglione GL, Arcovito A, et al. Neuroglobin-prion protein interaction: What's the function? *J Pept Sci*. 2011;17(5):387-391.

40. De Marinis E, Fiocchetti M, Acconcia F, Ascenzi P, Marino M. Neuroglobin upregulation induced by 17β-estradiol sequesters cytocrome c in the mitochondria preventing H_2O_2-induced apoptosis of neuroblastoma cells. *Cell Death Dis*. 2013;4:e508.

41. Watanabe S, Takahashi N, Uchida H, Wakasugi K. Human neuroglobin functions as an oxidative stress-responsive sensor for neuroprotection. *J Biol Chem*. 2012;287(36):30,128-30,138.

42. Yu Z, Liu N, Wang Y, Li X, Wang X. Identification of neuroglobin-interacting proteins using yeast two-hybrid screening. *Neuroscience*. 2012;200:99-105.

43. Zhanyang Y, Ning L, Yi W, Xiaokun L, Xiaoying W. Identification of neuroglobin-interacting proteins using yeast two-hybrid screening. *Neuroscience*. 2012;200:99-105.

44. Haines B, Demaria M, Mao X, et al. Hypoxia-inducible factor-1 and neuroglobin expression. *Neurosci Lett*. 2012;514(2):137-140.

45. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B. PrePPI: A structure-informed database of protein-protein interactions. *Nucleic Acids Res*. 2013;41(Database issue):D828-D833.

46. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database. *Nucleic Acids Res*. 2012;40(Database issue):D841-D846.

47. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker E, Andrade-Navarro MA. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*. 2012;7(2):e31826.

48. Chen JY, Mamidipalli S, Huan T. HAPPI: An online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics*. 2009;10(Suppl 1):S16

49. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res*. 2015;43(Database issue):D470-D478.

50. Orchard S. Molecular interaction databases. *Proteomics*. 2012;12(10):1656-1662.

51. Lalonde S, Ehrhardt DW, Loque D, Chen J, Rhee SY, Frommer WB. Molecular and cellular approaches for the detection of protein-protein interactions: Latest techniques and current limitations. *Plant J*. 2008;53(4):610-635.

52. Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Deinhardt K, Darie CC. Protein-protein interactions: Switch from classical methods to proteomics and bioinformatics-based approaches. *Cell Mol Life Sci*. 2014;71(2):205-228.

53. Piehler J. New methodologies for measuring protein interactions in vivo and in vitro. *Curr Opin Struct Biol*. 2005;15(1):4-14.

54. Guidolin D, Agnati LF, Tortorella C, et al. Neuroglobin as a regulator of mitochondrial-dependent apoptosis: A bioinformatics analysis. *Int J Mol Med*. 2014;33(1):111-116.

55. Lechauve C, Rezaei H, Celier C, et al. Neuroglobin and prion cellular localization: Investigation of a potential interaction. *J Mol Biol*. 2009;388(5):968-977.

56. De Marinis E, Acaz-Fonseca E, Arevalo MA, et al. 17β-Oestradiol anti-inflammatory effects in primary astrocytes require oestrogen receptor β-mediated neuroglobin up-regulation. *J Neuroendocrinol*. 2013;25(3):260-270.

57. Burmester T, Hankeln T. What is the function of neuroglobin? *J Exp Biol*. 2009;212(Pt 10):1423-1428.

58. Jin K, Mao XO, Xie L, Khan AA, Greenberg DA. Neuroglobin protects against nitric oxide toxicity. *Neurosci Lett*. 2008;430(2):135-137.

59. Nicolis S, Monzani E, Ciaccio C, Ascenzi P, Moens L, Casella L. Reactivity and endogenous modification by nitrite and hydrogen peroxide: Does human neuroglobin act only as a scavenger? *Biochem J*. 2007;407(1):89-99.

60. Greenberg DA, Jin K, Khan AA. Neuroglobin: An endogenous neuroprotectant. *Curr Opin Pharmacol*. 2008;8(1):20-24.

61. Burmester T, Gerlach F, Hankeln T. Regulation and role of neuroglobin and cytoglobin under hypoxia. *Adv Exp Med Biol*. 2007;618:169-180.

62. Tsai MH, Saier MH Jr. Phylogenetic characterization of the ubiquitous electron transfer flavoprotein families ETF-α and ETF-β. *Res Microbiol*. 1995;146(5):397-404.

63. Zorov DB, Juhaszova M, Sollott SJ. Mitochondrial reactive oxygen species (ROS) and ROS-induced ROS release. *Physiol Rev*. 2014;94(3):909-950.

64. Lechauve C, Augustin S, Cwerman-Thibault H, et al. Neuroglobin involvement in respiratory chain function and retinal ganglion cell integrity. *Biochim Biophys Acta*. 2012;1823(12):2261-2273.

65. Liu J, Yu Z, Guo S, et al. Effects of neuroglobin overexpression on mitochondrial function and oxidative stress following hypoxia/reoxygenation in cultured neurons. *J Neurosci Res*. 2009;87(1):164-170.

66. Duong TTH, Witting PK, Antao ST, et al. Multiple protective activities of neuroglobin in cultured neuronal cells exposed to hypoxia re-oxygenation injury. *J Neurochem*. 2009;108(5):1143-1154.

67. Garrido C, Galluzzi L, Brunet M, Puig PE, Didelot C, Kroemer G. Mechanisms of cytochrome c release from mitochondria. *Cell Death Differ*. 2006;13(9):1423-1433.

68. Ascenzi P, Coletta M, Wilson MT, et al. Cardiolipin-cytochrome c complex: Switching cytochrome c from an electron-transfer shuttle to a myoglobin- and a peroxidase-like heme-protein. *IUBMB Life*. 2015;67(2):98-109.

69. Sinibaldi F, Milazzo L, Howes BD, et al. The key role played by charge in the interaction of cytochrome c with cardiolipin. *J Biol Inorg Chem*. 2017;22(1):19-29.

70. Brittain T, Skommer J, Henty K, Birch N, Raychaudhuri S. A role for human neuroglobin in apoptosis. *IUBMB Life*. 2010;62(12):878-885.

71. Brittain T, Skommer J, Raychaudhuri S, Birch N. An antiapoptotic neuroprotective role for neuroglobin. *Int J Mol Sci*. 2010;11(6): 2306-2321.

72. Fago A, Mathews AJ, Moens L, Dewilde S, Brittain T. The reaction of neuroglobin with potential redox protein partners cytochrome b_5 and cytochrome c. *FEBS Lett*. 2006;580(20):4884-4888.

73. Borutaite V, Brown GC. Mitochondrial regulation of caspase activation by cytochrome oxidase and tetramethylphenylenediamine via cytosolic cytochrome *c* redox state. *J Biol Chem*. 2007;282(43):31124-31130.

74. Pan Z, Voehringer DW, Meyn RE. Analysis of redox regulation of cytochrome c-induced apoptosis in a cell-free system. *Cell Death Differ*. 1999;6(7):683-688.

75. Raychaudhuri S, Skommer J, Henty K, Birch N, Brittain T. Neuroglobin protects nerve cells from apoptosis by inhibiting the intrinsic pathway of cell death. *Apoptosis*. 2010;15(4):401-411.

76. Tiwari PB, Astudillo L, Miksovska J, et al. Quantitative study of protein-protein interactions by quartz nanopipettes. *Nanoscale*. 2014;6(17):10,255-10,263.

77. Fiocchetti M, Camilli G, Acconcia F, Leone S, Ascenzi P, Marino M. ERβ-dependent neuroglobin up-regulation impairs 17β-estradiol-induced apoptosis in DLD-1 colon cancer cells upon oxidative stress injury. *J Steroid Biochem Mol Biol*. 2015;149:128-137.

78. Ow YP, Green DR, Hao Z, Mak TW. Cytochrome *c*: Functions beyond respiration. *Nat Rev Mol Cell Biol*. 2008;9(7):532-542.

79. Tait SW, Green DR. Mitochondria and cell death: Outer membrane permeabilization and beyond. *Nat Rev Mol Cell Biol*. 2010;11(9): 621-632.

80. Parsons MJ, Green DR. Mitochondria in cell death. *Essays Biochem*. 2010;47:99-114.

81. Gupta S, Kass GE, Szegezdi E, Joseph B. The mitochondrial death pathway: A promising therapeutic target in diseases. *J Cell Mol Med*. 2009;13(6):1004-1033.

82. Tomasello F, Messina A, Lartigue L, et al. Outer membrane VDAC1 controls permeability transition of the inner mitochondrial membrane in cellulo during stress-induced apoptosis. *Cell Res*. 2009;19(12): 1363-1376.

83. Baines CP, Kaiser RA, Sheiko T, Craigen WJ, Molkentin JD. Voltage-dependent anion channels are dispensable for mitochondrial-dependent cell death. *Nat Cell Biol*. 2007;9(5):550-555.

84. Abu-Hamad S, Sivan S, Shoshan-Barmatz V. The expression level of the voltage-dependent anion channel controls life and death of the cell. *Proc Natl Acad Sci U S A*. 2006;103(37):5787-5792.

85. Bayrhuber M, Meins T, Habeck M, et al. Structure of the human voltage-dependent anion channel. *Proc Natl Acad Sci U S A*. 2008;105(40):15,370-15,375.

86. Saudou F, Humbert S. The biology of huntingtin. *Neuron*. 2016;89(5):910-926.

87. Ratovitski T, Chighladze E, Arbez N, et al. Huntingtin protein interactions altered by polyglutamine expansion as determined by quantitative proteomic analysis. *Cell Cycle*. 2012;11(10):2006-2021.

88. Guedes-Dias P, Pinho BR, Soares TR, de Proenca J, Duchen MR, Oliveira JM. Mitochondrial dynamics and quality control in Huntington's disease. *Neurobiol Dis*. 2016;90:51-57.

89. Nuzzo MT, Fiocchetti M, Servadio M, Trezza V, Ascenzi P, Marino M. 17β-estradiol modulates huntingtin levels in rat tissues and in human neuroblastoma cell line. *Neurosci Res*. 2016;103:59-63.

90. Nuzzo MT, Fiocchetti M, Totta P, et al. Huntingtin polyQ mutation impairs the 17β-estradiol/neuroglobin pathway devoted to neuron survival. *Mol Neurobiol*. 2016. https://doi.org/10.1007/s12035-016-0337-x

91. Chacinska A, Koehler CM, Milenkovic D, Lithgow T, Pfanner N. Importing mitochondrial proteins: Machineries and mechanisms. *Cell*. 2009;138(4):628-644.

92. Truscott KN, Brandner K, Pfanner N. Mechanisms of protein import into mitochondria. *Curr Biol*. 2003;13(8):R326-R337.

93. Giuffré A, Moschetti T, Vallone B, Brunori M. Is neuroglobin a signal transducer? *IUBMB Life*. 2008;60(6):410-413.

94. Oldham WM, Hamm HE. Structural basis of function in heterotrimeric G proteins. *Q Rev Biophys*. 2006;39(2):117-166.

95. McCudden CR, Hains MD, Kimple RJ, Siderovski DP. Willard FS (2005) G-protein signaling: Back to the future. *Cell Mol Life Sci*. 2005;62(5):551-577.

96. Oldham WM, Hamm HE. Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat Rev Mol Cell Biol*. 2008;9(1):60-71.

97. Burmester T, Hankeln T. Neuroglobin: A respiratory protein of the nervous system. *News Physiol Sci*. 2004;19:110-113.

98. Takahashi N, Wakasugi K. Identification of residues crucial for the interaction between human neuroglobin and the α-subunit of heterotrimeric G_i protein. *Sci Rep*. 2016;6:24,948.

99. Wakasugi K, Morishima I. Preparation and characterization of a chimeric zebrafish-human neuroglobin engineered by module substitution. *Biochem Biophys Res Commun*. 2005;330(2):591-597.

100. Wakasugi K, Morishima I. Identification of residues in human neuroglobin crucial for guanine nucleotide dissociation inhibitor activity. *Biochemistry*. 2005;44(8):2943-2948.

101. Watanabe S, Wakasugi K. Neuroprotective function of human neuroglobin is correlated with its guanine nucleotide dissociation inhibitor activity. *Biochem Biophys Res Commun*. 2008;369(2): 695-700.

102. Brittain T, Skommer J. Does a redox cycle provide a mechanism for setting the capacity of neuroglobin to protect cells from apoptosis? *IUBMB Life*. 2012;64(5):419-422.

103. Takahashi N, Watanabe S, Wakasugi K. Crucial roles of Glu60 in human neuroglobin as a guanine nucleotide dissociation inhibitor and neuroprotective agent. *PLoS One*. 2013;8(12): e83698.

104. Fiocchetti M, Cipolletti M, Leone S, et al. Neuroglobin in breast cancer cells: Effect of hypoxia and oxidative stress on protein level, localization, and anti-apoptotic function. *PLoS One*. 2016;11(5): e0154959.

105. Khan AA, Mao XO, Banwait S, et al. Regulation of hypoxic neuronal death signaling by neuroglobin. *FASEB J*. 2008;22(6):1737-1747.

106. Takai Y, Sasaki T, Matozaki T. Small GTP-binding proteins. *Physiol Rev*. 2001;81(1):153-208.

107. Altomare DA, Testa JR. Perturbations of the AKT signaling pathway in human cancer. *Oncogene*. 2005;24(59):7455-7464.

108. Manning BD, Toker A. AKT/PKB signaling: Navigating the network. *Cell*. 2017;169(3):381-405.

109. Song G, Ouyang G, Bao S. The activation of Akt/PKB signaling pathway and cell survival. *J Cell Mol Med*. 2005;9(1):59-71.

110. Li RC, Pouranfar F, Lee SK, Morris MW, Wang Y, Gozal D. Neuroglobin protects PC12 cells against β-amyloid-induced cell injury. *Neurobiol Aging*. 2008;29(12):1815-1822.

111. Zara S, De Colli M, Rapino M, et al. Ibuprofen and lipoic acid conjugate neuroprotective activity is mediated by Ngb/Akt intracellular signaling pathway in Alzheimer's disease rat model. *Gerontology*. 2013;59(3): 250-260.

112. Li Y, Dai YB, Sun JY, et al. Neuroglobin attenuates beta amyloid-induced apoptosis through inhibiting caspases activity by activating PI3K/Akt signaling pathway. *J Mol Neurosci*. 2016;58(1):28-38.

113. Zhang B, Ji X, Zhang S, et al. Hemin mediated neuroglobin induction exerts neuroprotection following ischemic brain injury through PI3K/Akt signaling. *Mol Med Rep*. 2013;8(2):681-685.

114. Chen LM, Xiong YS, Kong FL, et al. Neuroglobin attenuates Alzheimer-like tau hyperphosphorylation by activating Akt signaling. *J Neurochem*. 2012;120(1):157-164.

115. Amri F, Ghouili I, Amri M, Carrier A, Masmoudi-Kouki O. Neuroglobin protects astroglial cells from hydrogen peroxide-induced oxidative stress and apoptotic cell death. *J Neurochem*. 2017;140(1):151-169.

116. Li L, Liu QR, Xiong XX, et al. Neuroglobin promotes neurite outgrowth via differential binding to PTEN and Akt. *Mol Neurobiol*. 2014;49(1):149-162.

117. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. *Proc IEEE Comput Soc Bioinform Conf*. 2002;1:197-206.

118. Brittain T. The anti-apoptotic role of neuroglobin. *Cell*. 2012;1(4): 1133-1155.

119. Hundahl CA, Allen GC, Hannibal J, et al. Anatomical characterization of cytoglobin and neuroglobin mRNA and protein expression in the mouse brain. *Brain Res*. 2010;1331:58-73.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Supplementary Material

Neuroglobin and Friends

Marco Fiocchetti [a], Manuela Cipolletti [a], Valentina Brandi [a], Fabio Polticelli [a,b],

and Paolo Ascenzi [c,*]

[a] Dipartimento di Scienze, Università Roma Tre, Viale Guglielmo Marconi 446, I-00146 Roma, Italy

[b] Istituto Nazionale di Fisica Nucleare, Sezione dell'Università Roma Tre, Via della Vasca Navale 84, I-00146 Roma, Italy

[c] Laboratorio Interdipartimentale di Microscopia Elettronica, Università Roma Tre, Via della Vasca Navale 79, I-00146 Roma, Italy

Running title: Recognition properties of neuroglobin

Keywords: neuroglobin; interactors; protein-protein recognition; structure; function.

* Address correspondence to:

Paolo Ascenzi, Laboratorio Interdipartimentale di Microscopia Elettronica, Università Roma Tre, Via della Vasca Navale 79, I-00146 Roma, Italy.

Tel: +39-06-57333621; Fax: +39-06-57336321; E-mail: ascenzi@uniroma3.it

Table 1. Bioinformatics tools used to collect human Ngb interactors.

Method name	Description	URL	Reference
PrePPI	It combines the prediction of protein-protein interactions (PPI) with those coming from experimental evidence. Predicted interactions are derived from calculated likelihood ratios (LRs) by combining structural, functional, evolutionary and expression information with the most important contribution coming from structure. LRs are also assigned to the experimentally determined interactions, collected from six publicly available databases (MIPS, DIP, IntAct, MINT, HPRD and BioGRID). A final probability is then assigned to every interaction by combining the LRs for both predicted and experimentally determined interactions.	http://bhapp.c2b2.columbia.edu/PrePPI	1
IntAct	It contains molecular interactions curated from the literature or from direct data depositions. Each entry in IntAct is peer reviewed by a senior curator, and not released until accepted by that curator. Each interaction is then scored, by addition of the cumulated value of a weighted score for the interaction detection method and the interaction type for each interaction evidence.	http://www.ebi.ac.uk/intact/	2
Hippie	It is a human protein-protein interaction (PPI) dataset with a normalized scoring scheme that integrates multiple experimental PPI datasets. The scoring takes into account the reliability of different experimental evidence in the definition of a PPI combining three types of information: experimental techniques used, number of studies finding the PPI, and reproducibility in model organisms. For each interaction a score S between 0 and 1 is calculated as a weighted sum of three different subscores.	http://cbdm.mdc-berlin.de/tools/hippie	3
Happi	It extracts and integrates protein interactions from multiple data sources including both experimental and computationally- derived PPI. Each protein interaction in HAPPI is assigned a PPI confidence grade of 1, 2, 3, 4, or 5 to facilitate the evaluation of the reliability and the confidence of reported interactions	http://bio.informatics.iupui.edu/HAPPI/	4
Biogrid	It is an open access database directed to the annotation of genetic and physical interactions between genes or gene products across all major model organism species. Curation for BioGRID is performed by a dedicated team of PHD-level curators who use text-mining tools to prioritize the relevant literature for each curation	http://www.thebiogrid.org	5
PSIQUIC	The Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) provides computational access to molecular-interaction data resources, supplying more than 150 million interactions.	http://www.ebi.ac.uk/Tools/webservices/psicquic/view/home.xhtml	6

References

1. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B. PrePPI: a structure-informed database of protein-protein interactions. Nucleic Acids Res. 2013;41(Database issue):D828–D833.
2. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database. Nucleic Acids Res. 2012;40(Database issue):D841-D846;
3. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker E, Andrade-Navarro MA. HIPPIE: integrating protein interaction networks with experiment based quality scores. PLoS One. 2012;7(2):e31826.
4. Chen JY, Mamidipalli S, Huan T. HAPPI: an online database of comprehensive human annotated and predicted protein interactions. BMC Genomics. 2009;10(Suppl 1):S16.
5. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43(Database issue):D470-D478.
6. Orchard S. (2012) Molecular interaction databases. Proteomics. 12(10):1656-1662.

Chapter 4: Huntingtin, Neuroglobin and HD-like syndromes

4.1 Huntingtin and Neuroglobin interactomes

Literature data have shown that the interaction between htt and Ngb could be indirect (Cardinale *et al.*, 2018). Therefore, Ngb interactome (see Chapter 3, Table 1 in Fiocchetti *et al.* 2017) has been compared with that of htt in order to find common interacting proteins. The analysis of Ngb and htt interactome has yielded 10 common interactors (**Figure 10**). Among these, there is the subunit α of G_o1 ($G\alpha_o1$), a heterotrimeric G protein, already examined (Table S8, Supplementary material in Brandi *et al.* 2017; see Chapter 2), where molecular docking simulations have supported the hypothesis that hunt3 is the putative htt interaction site with G proteins. In fact, $G\alpha_o1$ is highly abundant in the CNS, constituting about 1% of brain membrane proteins. Moreover, as already mentioned, heterozygous mutations in the GNAO1 gene (Nakamura *et al.*, 2013; Kehrl *et al.*, 2014; Talvik *et al.*, 2015; Marcé-Grau *et al.*, 2016) are the cause of a severe neurodevelopmental disorder, featuring early infantile seizures, profound cognitive dysfunction and, occasionally, movement disorders (early infantile epileptic encephalopathy-17), the latter symptoms being similar to those observed in HD.

Ngb	**Htt**

45 (4.4%)	10 (1%)	969 (94.6%)
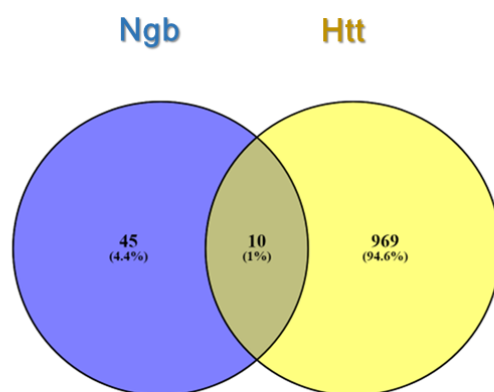
GENE NAME	PROTEIN NAME
HBB	Hemoglobin subunit beta
GNAO1	Guanine nucleotide-binding protein G(o) subunit alpha
ESR1	Estrogen receptor 1
DRD2	D(2) dopamine receptor
CHRNB2	Neuronal acetylcholine receptor subunit beta-2
YWHAB	14-3-3 protein beta/alpha
APP	Amyloid beta A4 protein
CCNB1	G2/mitotic-specific cyclin-B1
RAC1	Ras-related C3 botulinum toxin substrate 1
AKT1	RAC-alpha serine/threonine-protein kinase

Figure 10. Common interacting proteins between htt and Ngb

Interestingly a protein closely related to $G\alpha_o1$ ($G\alpha_i1$, with 73% identity) has been demonstrated experimentally to interact with Ngb (Takahashi and Wakasugi, 2016). Moreover, under oxidative stress conditions, ferric human Ngb is recruited to lipid rafts by interacting with flotillin-1, where it binds exclusively to the GDP-bound form of the α -subunits of heterotrimeric $G_{i/o}$ proteins ($G\alpha_{i/o}$), and acts as a guanine nucleotide dissociation inhibitor (GDI) by inhibiting the rate of exchange of GDP for GTP on $G\alpha_{i/o}$. The interaction of GDP-bound $G\alpha_{i/o}$ with ferric Ngb induces the release of the β and γ subunits, and leads to the intracellular signal transduction cascade that causes the inhibition of the reduction of intracellular cAMP concentration to protect against cell death (Watanabe *et al.*, 2012; Takahashi and Wakasugi, 2016; Watanabe and Wakasugi, 2008; Wakasugi *et al.*, 2003; Wakasugi and Morishima, 2005; Takahashi *et al.*, 2013).

Thus the structural model of hunt3-$G\alpha_o1$ subunit complex (see Figure 6 in Brandi *et al.,* 2017) has been used to carry out a molecular docking simulation against Ngb, to investigate the possible formation of a ternary complex between hunt3, $G\alpha_o1$ and Ngb (**Figure 11**). Indeed, docking

simulations indicate that such a ternary complex is plausible leading to the hypothesis that G proteins mediate the interaction between htt and Ngb.
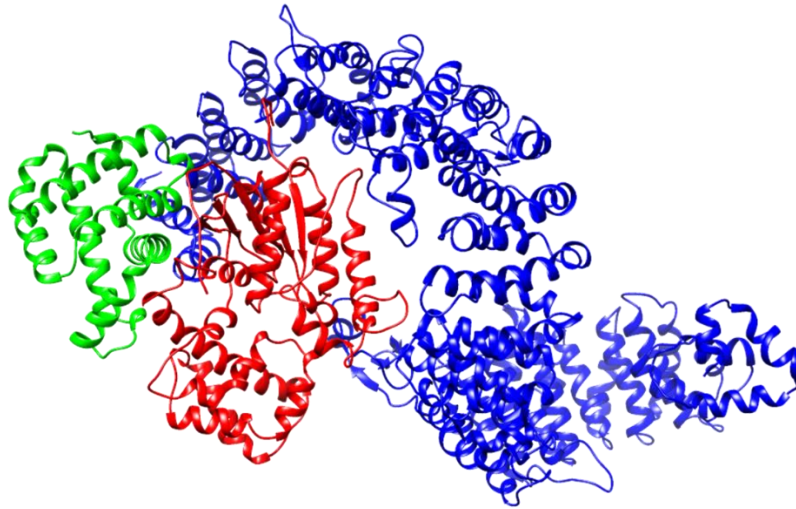
Figure 11. The best-ranked ZDOCK complex between htt (depicted in blue), $G\alpha_o1$ (in red) and Ngb (in green)**.**

In Brandi *et al*., 2017 (Figure 6) the $G\alpha_o1$ mutations causing epileptic encephalopathy-17 have been mapped onto the predicted Hunt3-$G\alpha_o1$ complex. Since some $G\alpha_o1$ mutations are located in the putative $G\alpha_o1$-hunt3 interface, such mutations could impair the interaction with htt, thus providing an explanation for the similar symptoms observed in HD and epileptic encephalopathy, *i.e.* impairment of a physiologically relevant interaction between htt and $G\alpha_o1$ when either of the two is mutated, and a link between these two pathologies. Also, $G\alpha_o1$ mutations could affect the interaction with Ngb, inhibiting its neuroprotective function.

4.2 Huntington-disease like syndromes

The epileptic encephalopathy-17 is not the only one to be considered a HD-like syndrome. In the last 17 years, several other distinct genetic disorders have been identified that can present with a clinical picture indistinguishable from that of HD (Margolis *et al.*, 2001; Mencacci and Carecchio, 2017; Mencacci *et al.*, 2015; Anderson *et al.*, 2017; Vijayvargia *et al.*, 2016; Gövert and Schneider, 2013). Interestingly some genes, whose mutations cause HD-like syndromes, code for htt-interacting proteins. Pathological expansions in the genes encoding the prion protein (PRNP), TATA box-binding protein (TBP; also responsible for the dominant spinocerebellar ataxia type 17), the ferritin light chain gene (FTL), Serine-protein kinase ATM (ATM), Synaptotagmin-1 (SYT1), Adenylate cyclase type 5 (ADCY5) result into HD mimics (Mencacci and Carecchio, 2017). This complex network of interactions complicates the clinical picture of the HD.

4.3 Methods

Docking simulations between hunt3-Gα$_o$1 complex and Ngb (PDB code: 4MPM) were performed using the protein–protein docking server ZDOCK (https://zdock.umassmed.edu; version 3.0.2; Pierce *et al.* 2014). ZDOCK implements a Fast Fourier Transform algorithm and a scoring system based on a combination of shape complementarity, electrostatics and statistical potential terms (Pierce *et al.*, 2014). The 2000 complexes predicted by ZDOCK were re-ranked using ZRANK (Pierce and Weng, 2007), which uses a more detailed potential including electrostatics, van der Waals, and desolvation terms (Pierce and Weng, 2007).

Venny 2.1 (http://bioinfogp.cnb.csic.es/tools/venny/) has been used to compare and display the two lists of interactors in a Venn diagram.

4.4 References

Anderson,D.G. *et al.* (2017) A systematic review of the Huntington Disease-Like 2 Phenotype. *J. Huntingtons. Dis.*, **6**, 37–46.

Brandi,V. *et al.* (2017) A comprehensive in silico analysis of huntingtin and its interactome. *J. Biomol. Struct. Dyn.*, **1102**, 1–17.

Cardinale,A. *et al.* (2018) Localization of neuroglobin in the brain of R6/2 mouse model of Huntington's disease. *Neurol. Sci.*, **39**, 275–285.

Fiocchetti,M. *et al.* (2017) Neuroglobin and friends. *J. Mol. Recognit.*, **30**, 1–12.

Gövert,F. and Schneider,S.A. (2013) Huntington,s disease and Huntington,s disease-like syndromes: An overview. *Curr. Opin. Neurol.*, **26**, 420–427.

Kehrl,J.M. *et al.* (2014) Gain-of-function mutation in Gnao1: A murine model of epileptiform encephalopathy (EIEE17)? *Mamm. Genome*, **25**, 202–210.

Marcé-Grau,A. *et al.* (2016) GNAO1 encephalopathy: further delineation of a severe neurodevelopmental syndrome affecting females. *Orphanet J. Rare Dis.*, **11**, 38.

Margolis,R.L. *et al.* (2001) A disorder similar to Huntington's disease is associated with a novel CAG repeat expansion. *Ann. Neurol.*, **50**, 373–380.

Mencacci,N.E. *et al.* (2015) ADCY5 mutations are another cause of benign hereditary chorea. *Neurology*, **85**, 80–88.

Mencacci,N.E. and Carecchio,M. (2017) Recent advances in genetics of chorea. *Curr. Opin. Neurol.*, **29**, 486–495.

Nakamura,K. *et al.* (2013) De novo mutations in GNAO1, encoding a gαosubunit of heterotrimeric g proteins, cause epileptic encephalopathy. *Am. J. Hum. Genet.*, **93**, 496–505.

Pierce,B. and Weng,Z. (2007) ZRANK: Reranking protein docking predictions with an optimized energy function. *Proteins Struct. Funct. Genet.,* **67**, 1078-86

Pierce,B.G. *et al.* (2014) ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*, **30**, 1771–1773.

Takahashi,N. *et al.* (2013) Crucial roles of Glu60 in human neuroglobin as a guanine nucleotide dissociation inhibitor and neuroprotective agent. *PLoS One*, **8**.

Takahashi,N. and Wakasugi,K. (2016) Identification of residues crucial for the interaction between human neuroglobin and the α-subunit of heterotrimeric Giprotein. *Sci. Rep.*, **6**, 3–8.

Talvik,I. *et al.* (2015) Clinical Phenotype of De Novo *GNAO1* Mutation. *Child Neurol. Open*, **2**, 2329048X1558371.

Vijayvargia,R. *et al.* (2016) Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. *Elife*, **5**, 1–16.

Wakasugi,K. *et al.* (2003) Oxidized human neuroglobin acts as a heterotrimeric Galpha protein guanine nucleotide dissociation inhibitor. *J. Biol. Chem.*, **278**, 36505–36512.

Wakasugi,K. and Morishima,I. (2005) Identification of residues in human neuroglobin crucial for guanine nucleotide dissociation inhibitor activity. *Biochemistry*, **44**, 2943–2948.

Watanabe,S. *et al.* (2012) Human neuroglobin functions as an oxidative stress-responsive sensor for neuroprotection. *J. Biol. Chem.*, **287**, 30128–30138.

Watanabe,S. and Wakasugi,K. (2008) Neuroprotective function of human neuroglobin is correlated with its guanine nucleotide dissociation inhibitor activity. *Biochem. Biophys. Res. Commun.*, **369**, 695–700.

Chapter 5: Molecular dynamics simulations of htt hunt3 domain

5.1 Comparison between htt experimental structure and molecular models

After htt structure has been determined by cryo-electron microscopy (Guo *et al.*, 2018), the structural models of the ordered domains have been compared to the corresponding regions of the structure to assess the reliability of the models obtained (**Figure 12**).
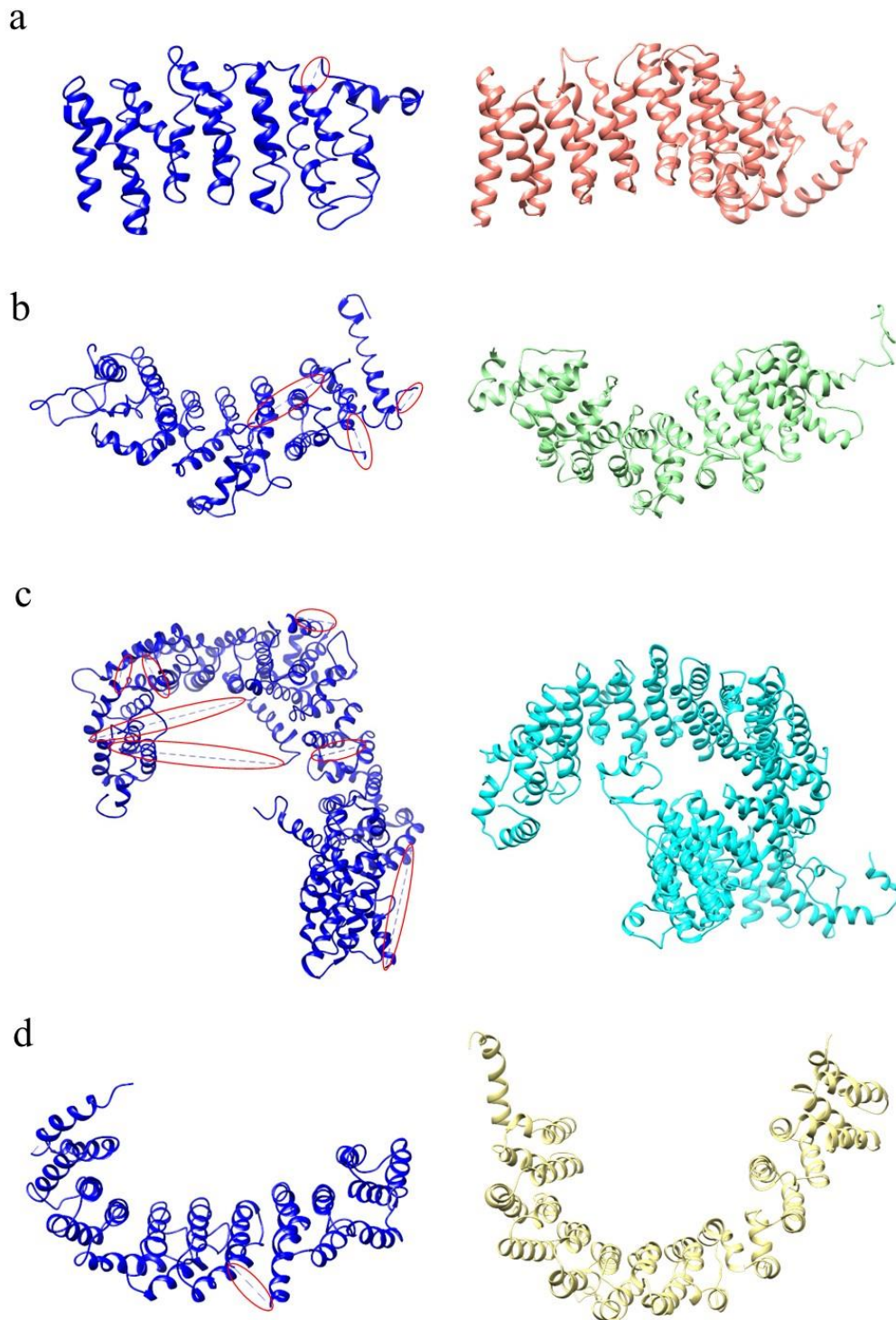
Figure 12. Comparison between the models of htt ordered domains and the corresponding regions of the structure determined by cryo-electron microscopy (PDB code: 6EZ8). The structure lacks several atomic details, depicted as dotted lines and highlighted in red. In each panel the region of the htt structure corresponding to the model is represented in blue. Hunt1 is depicted in coral (**a**), hunt2 in light green (**b**), hunt3 in cyan (**c**), hunt5 in khaki (**d**).

Unfortunately, htt structure lacks atomic details of several unstructured regions and this prevents a complete superimposition between the structure and the models.

Nevertheless, the overall fold of the models is very similar to the structure of the corresponding htt regions, especially for hunt1, hunt2 and hunt5. Even if in hunt3 there are several missing regions, the fold of the N-terminal portion of the model resembles that of the structure. The region of htt structure corresponding to hunt4 (not shown here) is characterized by a long loop and some α-helices, not properly arranged as HEAT repeats.

The cryo-electron microscopy structure of htt evidences that the putative interaction site of hunt3 with the G proteins is occupied by an α-helix (from 1348 to 1371 residues) connected to the opposite side of the domain through disordered loop regions and stabilised by electrostatic interactions (**Figure 12c, Figure 13**). Moreover, this α-helix contains two known phosphorylation sites at Ser1345 and Tyr1351 residues (**Figure 13**).
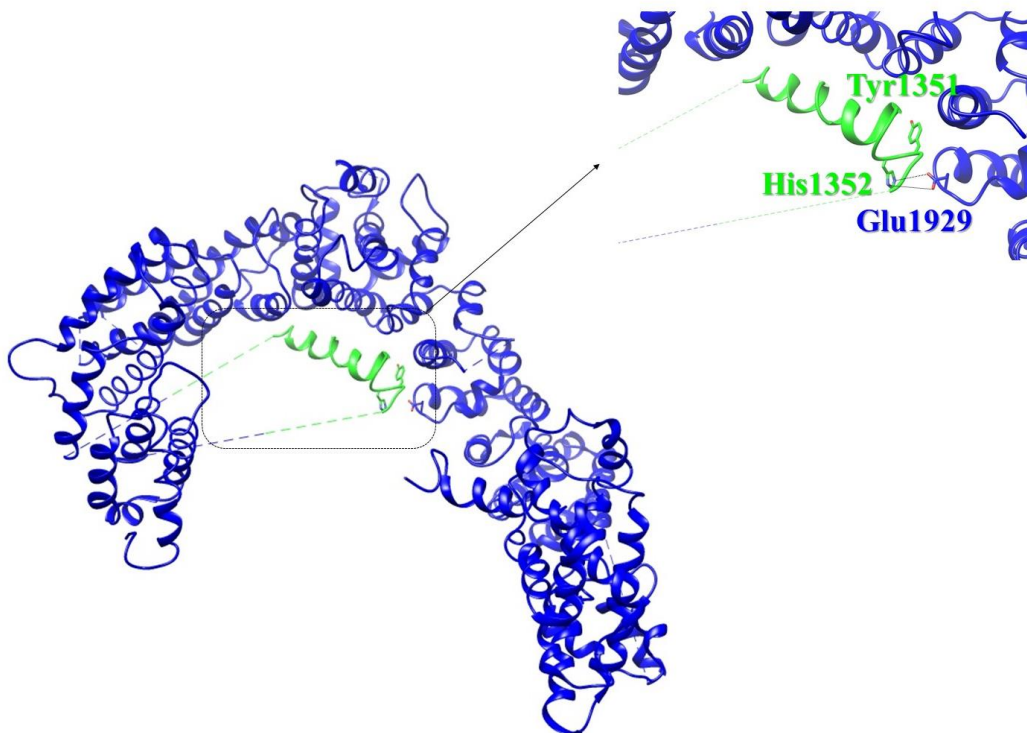
Figure 13. Structure of the htt region corresponding to hunt3. The α-helix (depicted in green) is stabilised by an electrostatic interaction between His1352 and Glu1929. Interestingly, the α-helix contains two phosphorylation sites, one at Tyr1351 and the other at Ser1345 (Huang *et al.*, 2015), the latter located in one missing region.

Unfortunately, the functional role of these phosphorylations is not known. However, molecular docking simulations carried out using the experimental structure of hunt3 have shown that when the α-helix is removed from the structure, G proteins are predicted to bind to htt replacing the α-helix (**Figure 14**).
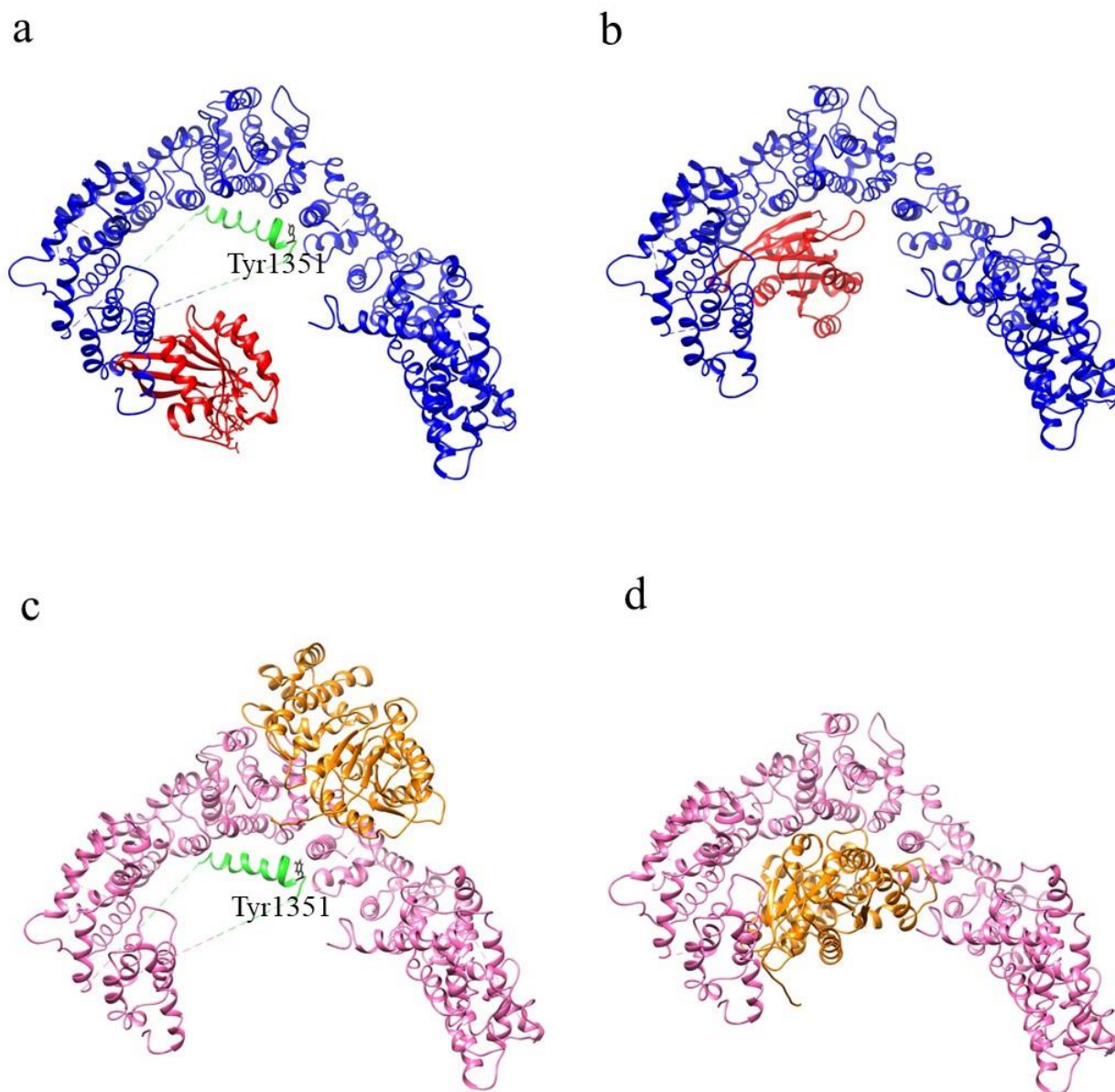
Figure 14. Best ranking-ZDOCK complexes between the cryo-electron microscopy structure of hunt3 domain (in blue, PDB code: **6ez8**) and Rab11a (in red, PDB code: **1OIW**) (**a**, **b**) and between hunt3 (in magenta) and Gα$_o$1 (in orange) (**c**, **d**). In **a** and **c** the α-helix is depicted in green, while in **b** and **d** it has been manually removed. The tyrosine residue which is phosphorylated is highlighted in black. The dotted lines indicate unresolved regions.

Interestingly, a BLAST search against the non-redundant protein sequences database using the sequence of the α-helix as a bait evidences a high local sequence similarity with some G proteins, such as Gα$_i$1, closely related to Gα$_o$1. The alignment (**Figure 15**) evidences the conservation of a motif "Y-HFT-A".
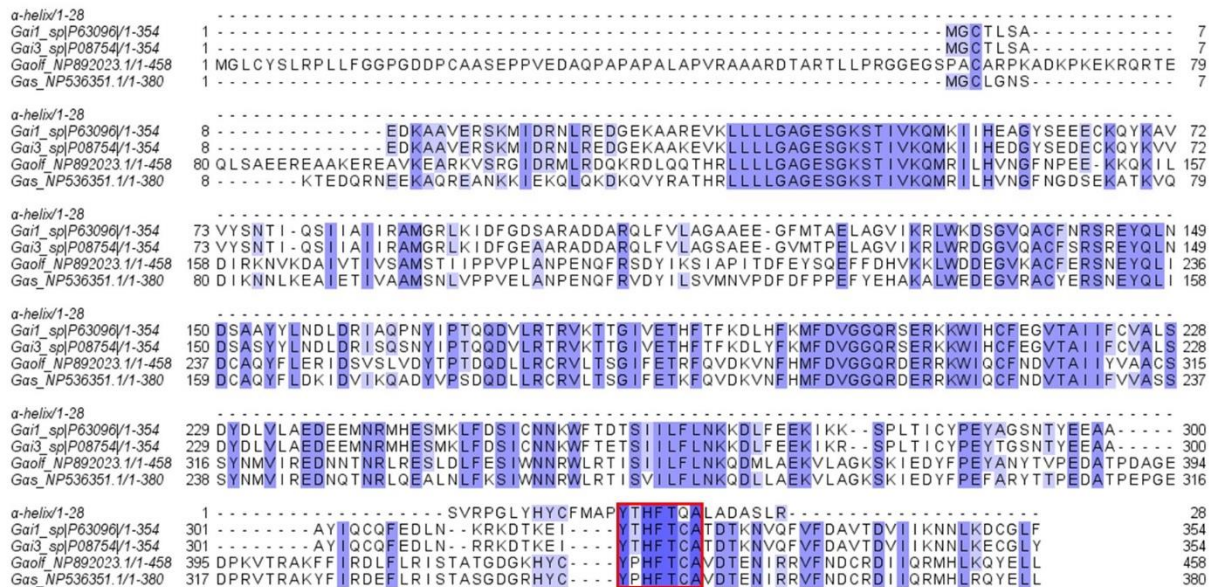
Figure 15. Amino acid sequence alignment between htt α-helix and some G proteins, closely related to Gα$_o$1. The conserved motif is highlighted by the red box. Residues are coloured by percentage identity.

5.2 Molecular dynamics simulations

Given the molecular docking results in the presence and in the absence of the α-helix, the presence on this α-helix of a sequence motif found in some G proteins, and the presence of two phosphorylation sites, 500 ns molecular dynamics simulations have been performed in order to validate the hypothesis that the α-helix could act as a molecular switch for the binding of G proteins when phosphorylation occurs. The analysis has been focused on 4 systems: hunt3 wild type (Htt3 wt), hunt3 with single phosphorylation at Ser1345 (Htt3 SP) and at Tyr1351 (Htt3 YP) and htt phosphorylated at both sites (Htt3 SYP).

RMSD analyses indicate a conformational instability of the hunt3 domain during the simulation, when it is phosphorylated on Ser1345 as compared to wild type, single phosphorylated on Tyr1351 and double phosphorylated domain (**Figure 16**). The radius of gyration plot (**Figure 17**) indicates a tendency of Htt3 SP and Htt3 YP to adopt a less compact structure, while interestingly an opposite trend is observed for the double phosphorylated domain. In all simulations, the highest RMSF values

(**Figure 18a**) correspond to loops, some of which have been modelled since they were unresolved in the experimental structure.

Further, RMSF values reveal a high flexibility of hunt3. This is in agreement with the fact that the protein is composed of HEAT repeats, which are known to be flexible because are involved in protein-protein interactions. However, as observed for RMSD, consistently higher values are observed for Htt3 SP. RMSF values of the α-helix (**Figure 18b**) are very similar except for the Pro1348, which shows a peak in Htt3 SYP.
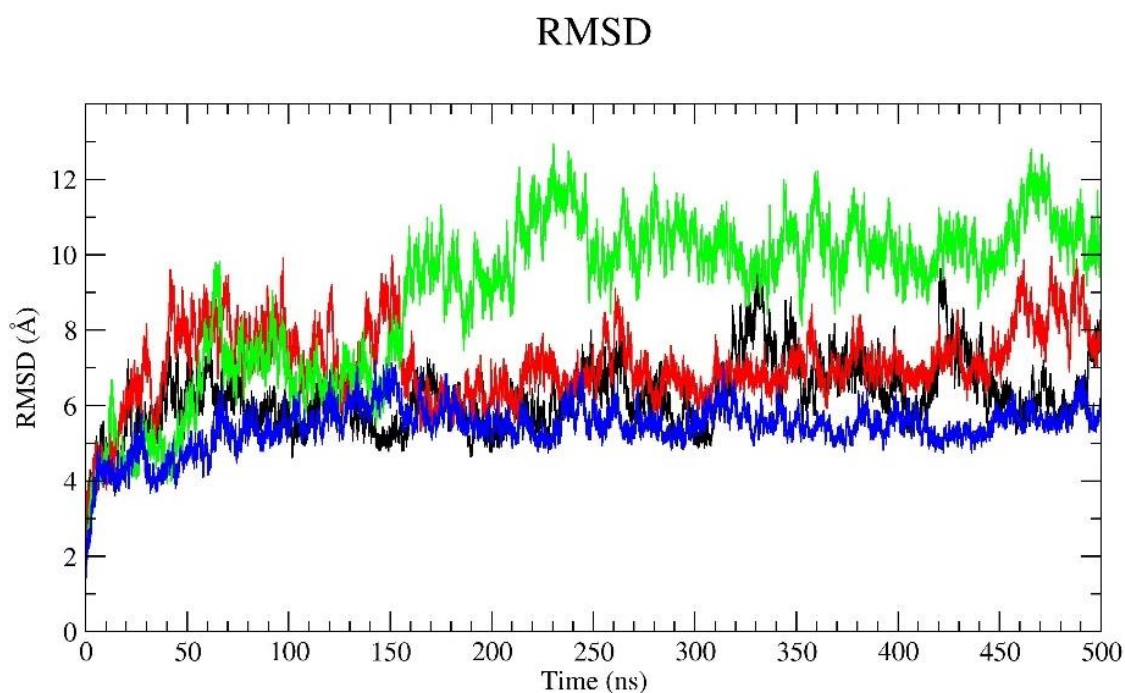
Figure 16. RMSD plot calculated on backbone heavy atoms for Htt3 wt (in black), Htt3 YP (in red), Htt3 SP (in green), Htt3 SYP (in blue).

Radius of gyration
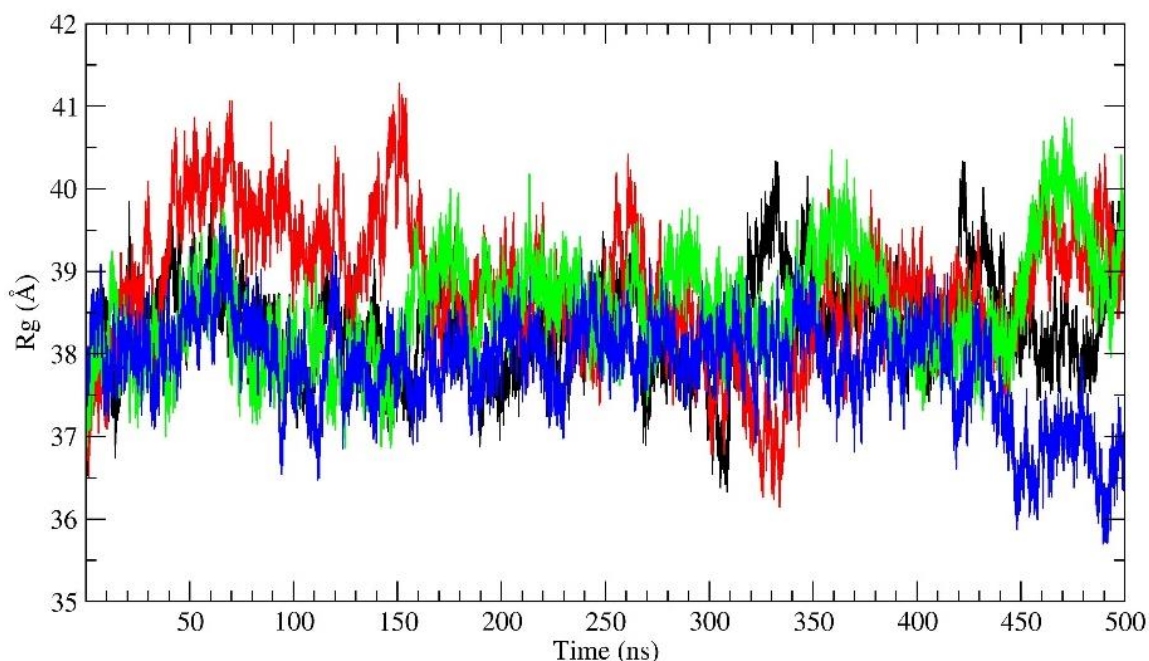
Figure 17. Radius of gyrations calculated on C$_\alpha$ atoms. Htt3 wt is depicted in black, Htt3 YP in red, Htt3 SP in green, Htt3 SYP in blue.

Intramolecular hydrogen bonds (paired within 0.35 nm) between the α-helix and the remaining portion of the protein were calculated during the 500 ns of all simulations (**Figure 19**).

The number of hydrogen bonds during the simulations remains almost unchanged for all systems except for Htt3 SP (**Table S1**, Supplementary material), in which the α-helix forms the highest number of hydrogen bonds. Instead the number of salt bridges is slightly larger in Htt3 YP (**Figure 20**, **Table S2**, Supplementary material). The salt bridge between Arg1372 and Glu1635 is the only one that occurs in all simulations. The distance between these residues increases significantly in Htt3 wt especially after 300ns (**Figure 21**). Probably the single phosphorylation tends to stabilize the α-helix and in particular the formation of the above mentioned salt bridge.
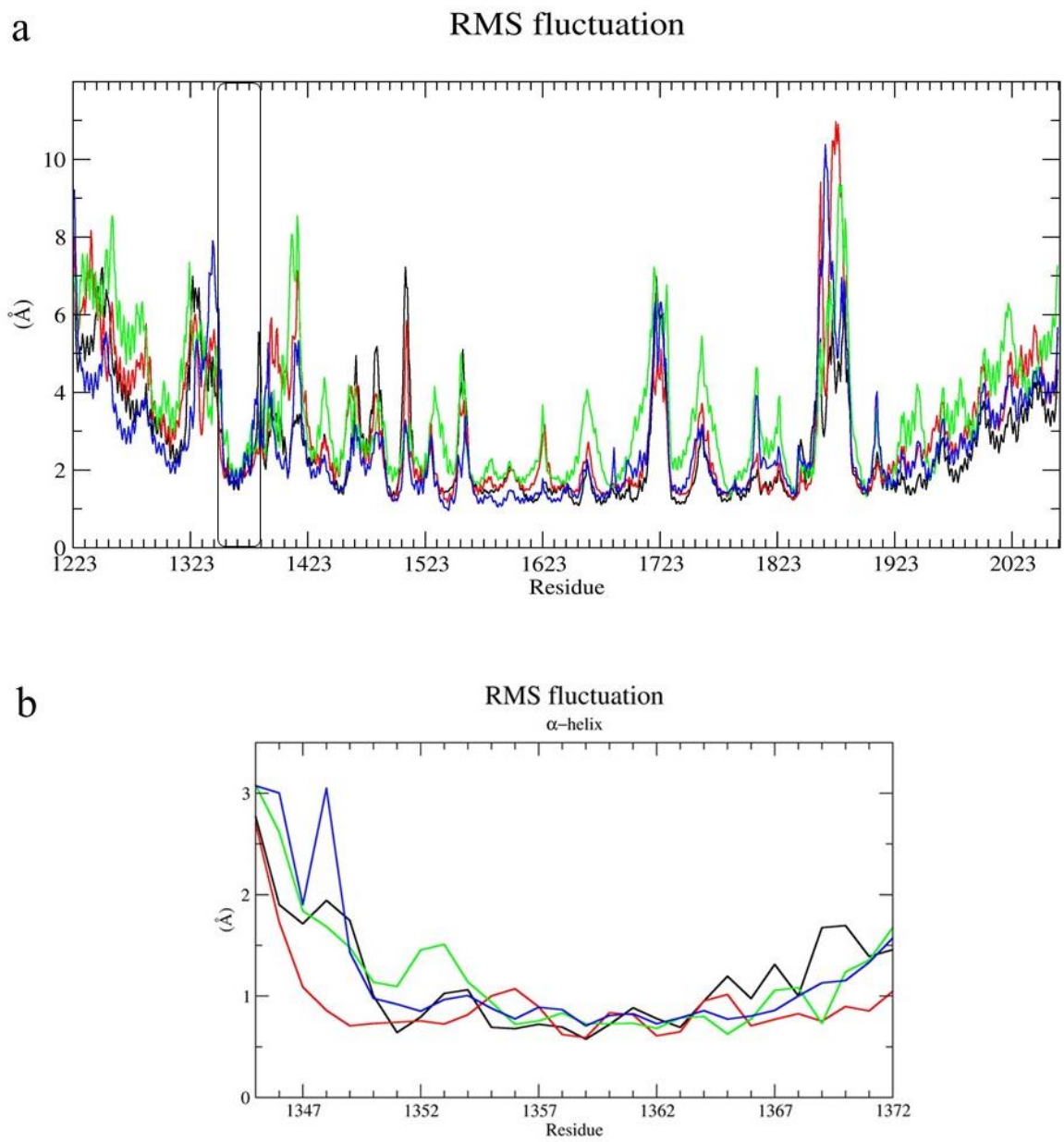
Figure 18. RMS fluctuations of Cα atoms of hunt3 in (**a**) and of α-helix in (**b**). Htt3 wt in black, Htt3 YP in red, Htt3 SP in green and Htt3 SYP in blue.
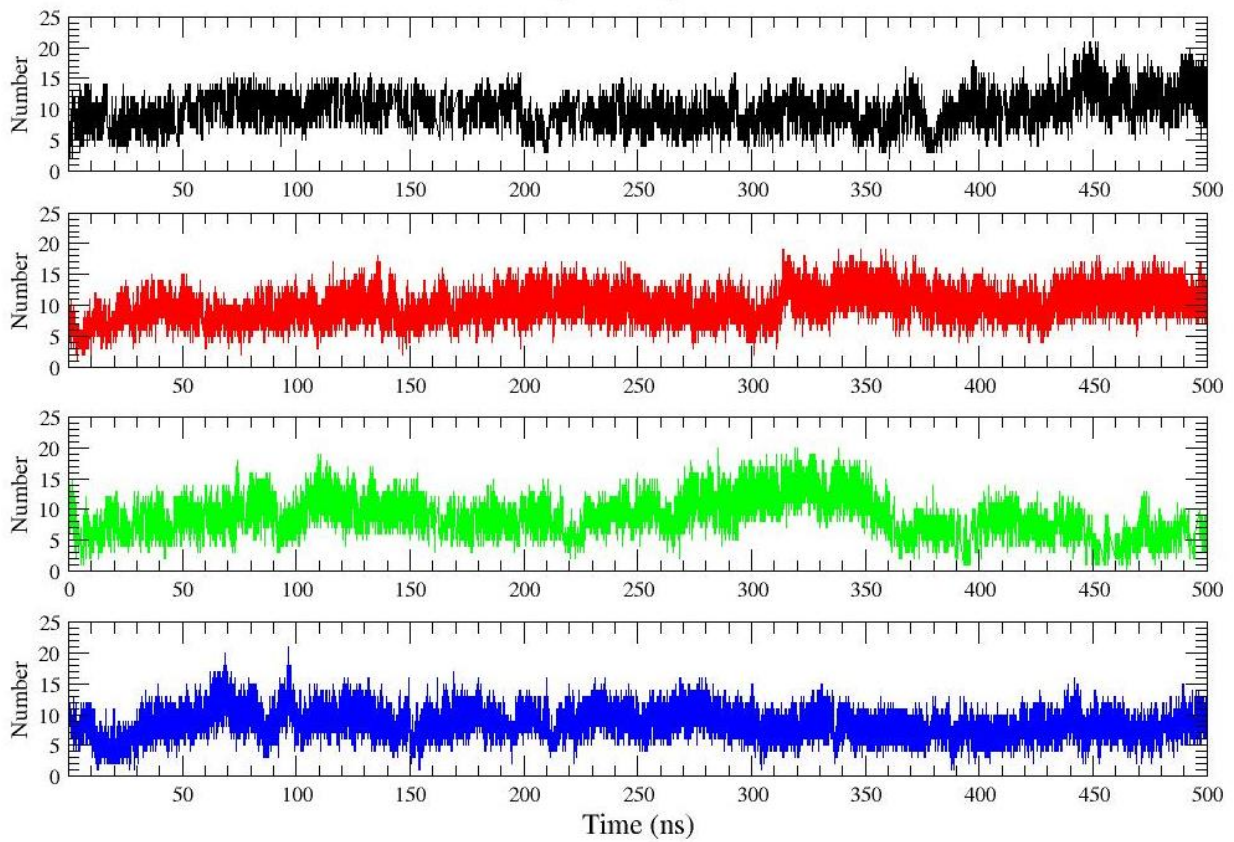
Figure 19. Number of hydrogen bonds during the four simulations. Htt3 wt in black, Htt3 YP in red, Htt3 SP in green, Htt3 SYP in blue.
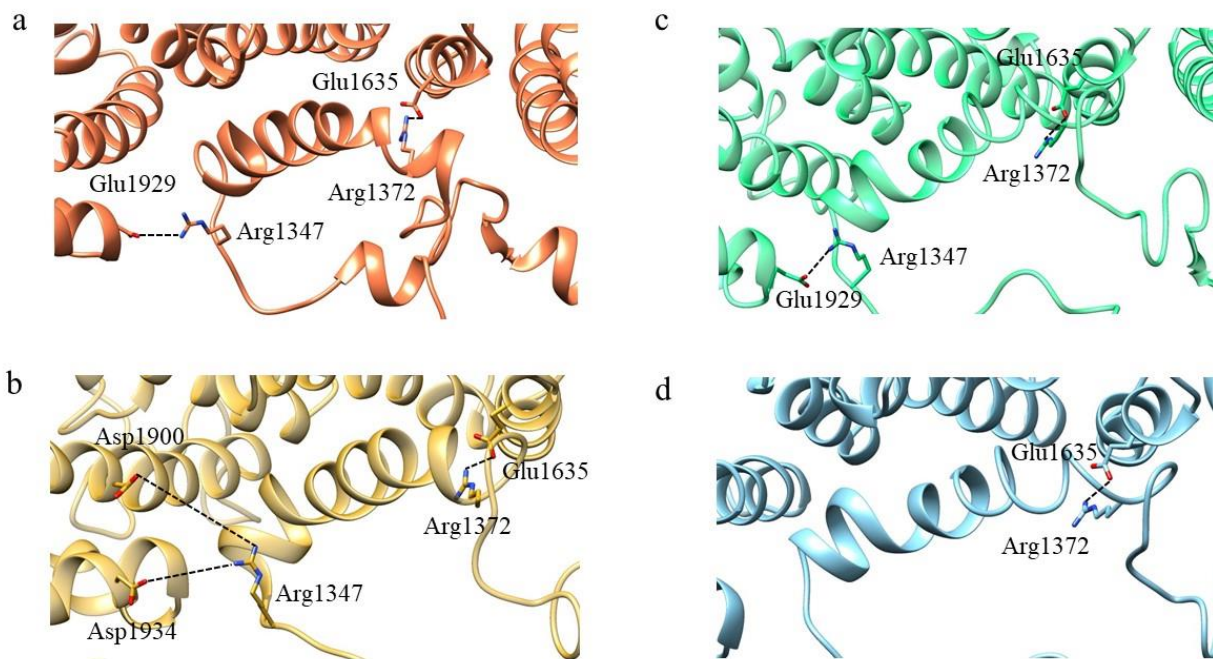
Figure 20. Salt bridges between the α-helix and the remaining portion of hunt3 have been identified throughout the trajectories of all systems. Here the first frame of each trajectory has been used only to show the localization of the salt bridges in Htt3 wt (**a**), in Htt3 YP (**b**), in Htt3 SP (**c**) and in Htt3 SYP (**d**).

Arg1372-Glu1635 Salt bridge
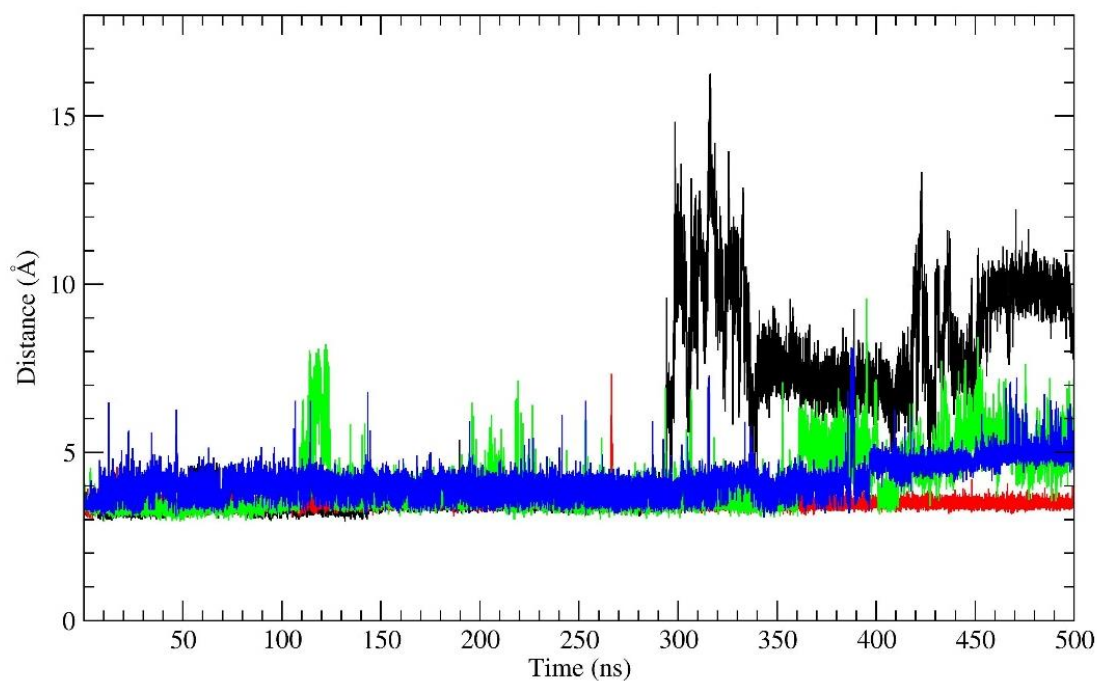
Figure 21. Time-dependent plot of the distance between the center of mass of the oxygen atoms in the acidic side chain of Glu1635 and the center of mass of the nitrogen atoms in the basic side chain of Arg1372 during the simulation time. This is the only salt bridge that occurs in all the simulations. Htt3 wt in black, Htt3 YP in red, Htt3 SP in green and Htt3 SYP in blue.

Principal component analysis (PCA) provides insight into the nature of conformational changes in protein structures by identifying conformational motions and underlying fluctuations (Jolliffe and Cadima, 2016). The major constituents of correlated motions of the protein structures were obtained from the covariance matrix diagonalization for the Cα atoms. In order to gain insight into the significance of such conformational motions, the conformational behaviour of all systems was projected along the first two principal eigenvectors (**Figure 22**).

The displacements of Htt3 wt are reduced, while that of Htt3 SP are wider than the others, suggesting greater flexibility, consistent with the analyses shown above.
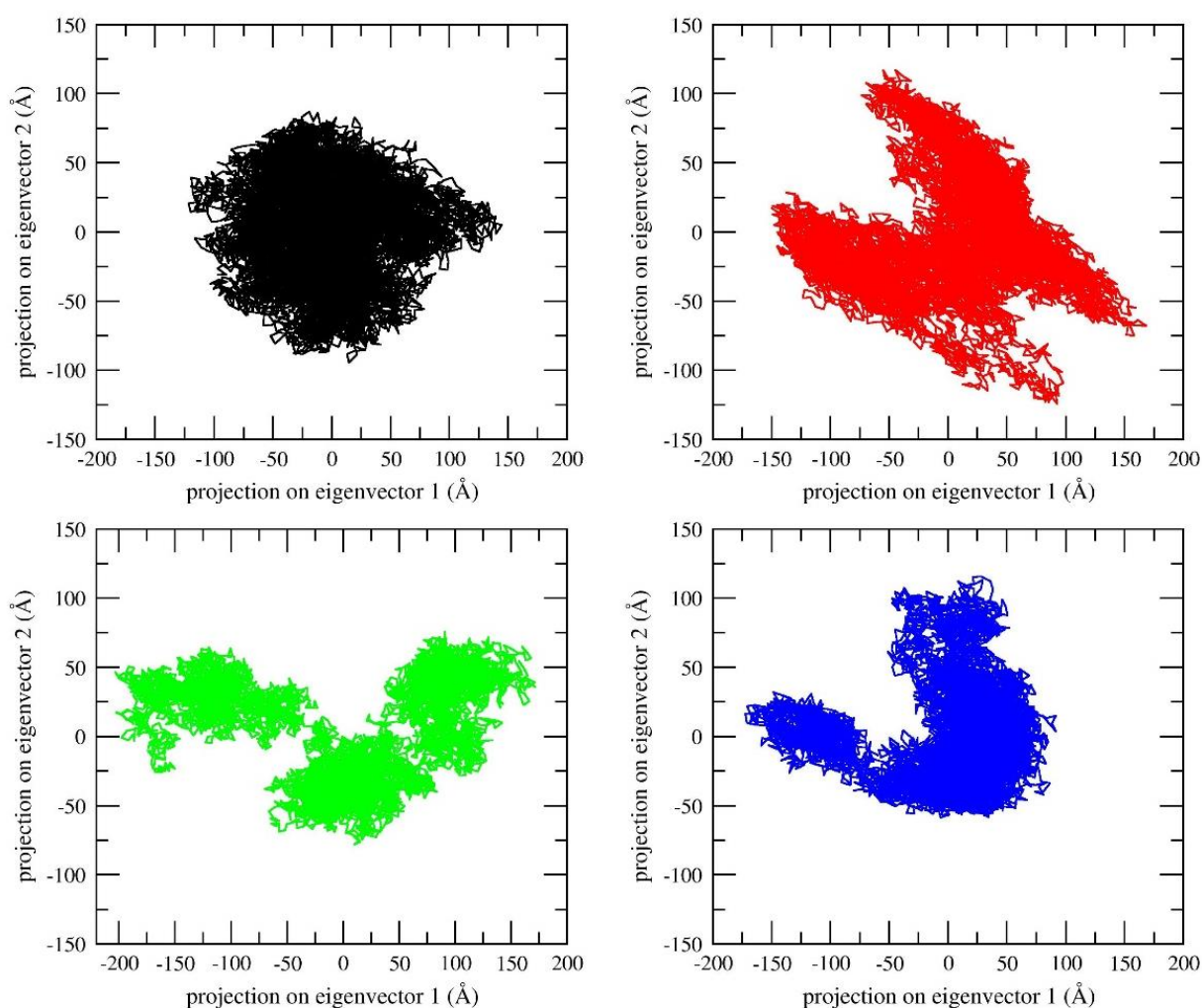
Figure 22. PCA projection of Cα atoms motion constructed by plotting the first two principal components (eigenvector 1 and 2) in the conformational subspace. Htt3 wt is depicted in black, Htt3 YP in red, Htt3 SY in green and Htt3 SYP in blue.

The visualization of the first principal component (**Figure 23**) shows that hunt3 undergoes an opposite movement mode when it is phosphorylated on tyrosine (**Figure 23b**), with respect to the other conditions. Further, it seems to adopt a more elongated shape, with the N and C-termini moving

in opposite directions. This is consistent with the cross-correlation matrix analysis (**Figure 24**), that highlights anticorrelated movements of the N and C-terminal regions when single phosphorylation occurs on tyrosine (**Figure 24 b**). Anticorrelated movements are also observed between 1673-1900 residues and the N-terminal, in Htt3 SP (**Figure 23 c**). The α-helix exhibits more anticorrelated movements with the N-terminal and correlated with the C-terminal in each system.

These data suggest that the phosphorylation of Ser1345 increases the flexibility of the entire Hunt3 domain with respect to the wild type. An opposite effect is observed on the double phosphorylated domain. Thus, at variance to what has been hypothesized, only the phosphorylation on Ser might act as a positive regulator of htt binding to G proteins, while double phosphorylation would revert the effect. This points to a sort of phosphorylation code regulating hunt3 functions, as observed in other systems (Salazar and Höfer, 2009). However, it must be taken into account that these are preliminary data obtained on an isolated domain of htt, and it cannot be excluded that this region displays a different behaviour in the context of the entire protein.
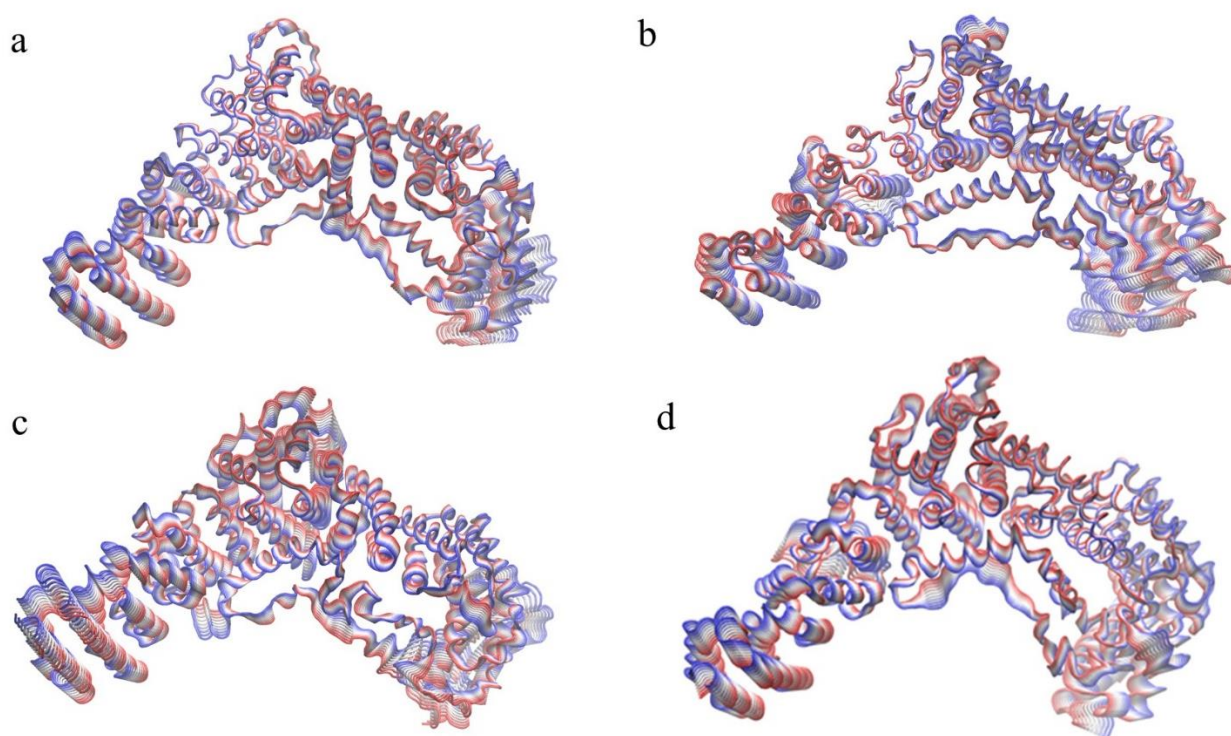
Figure 23. Visualization of the first principal component, eigenvector 1. In (**a**) is shown Htt3 wt, in (**b**) Htt3 YP, in (**c**) Htt3 SP, in (**d**) Htt3 SYP. The direction of the motion is from red to blue.
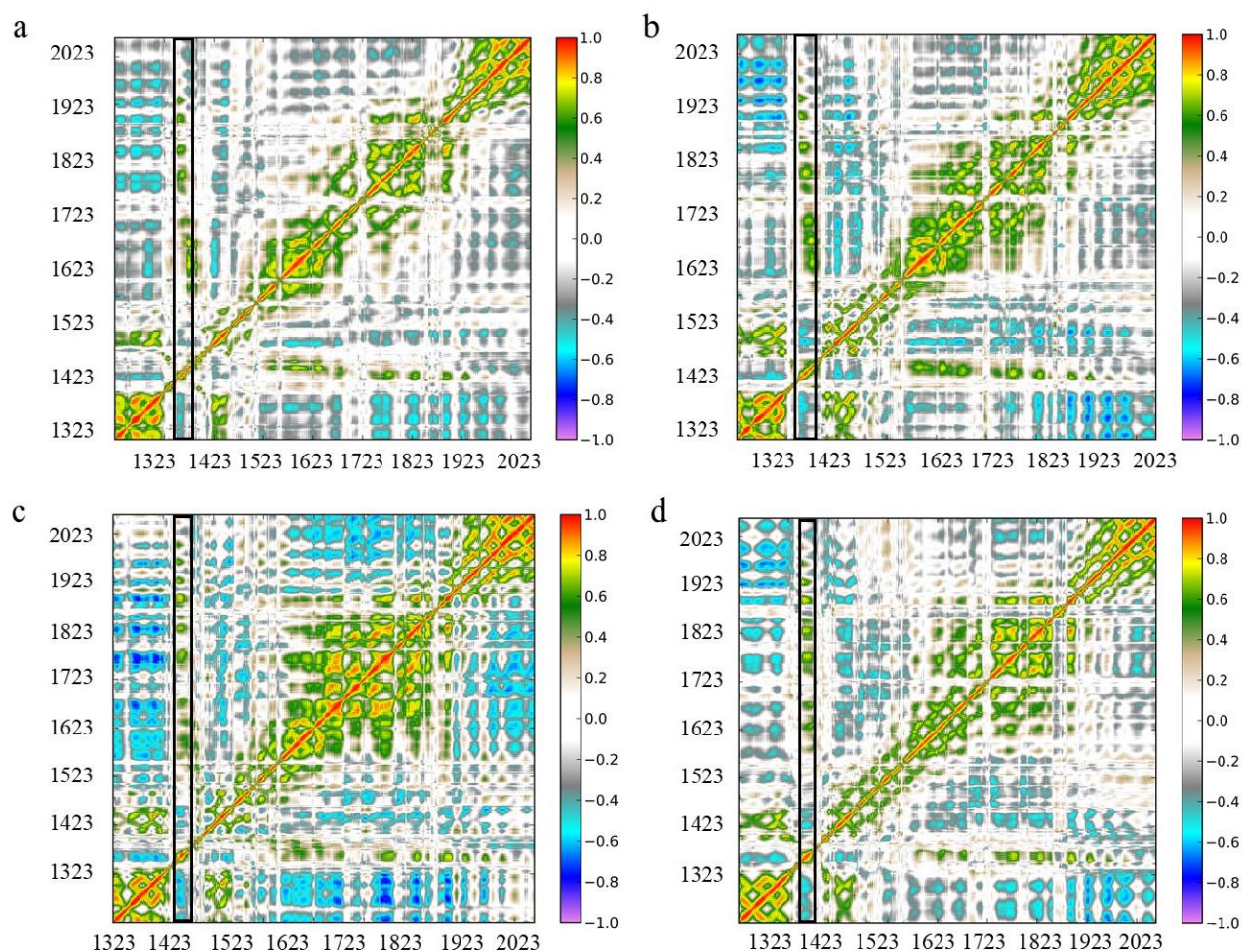
Figure 24. Cross-correlation matrix of Cα atoms during 500 ns of simulation for Htt3 wt (**a**), Htt3 YP (**b**), Htt3 SP (**c**) and Htt3 SYP (**d**). The range of motions is indicated by various colours in the sidebar. Red indicates positive correlation whereas blue indicates anticorrelation. The α-helix region is highlighted in black.

5.3 Methods

Docking simulations between Hunt3 and G proteins were performed using the protein–protein docking server ZDOCK (https://zdock.umassmed.edu; version 3.0.2; Pierce *et al.* 2014). The 2000 complexes predicted by ZDOCK were re-ranked using ZRANK, which uses a more detailed potential including electrostatics, van der Waals, and desolvation terms (Pierce and Weng, 2007).

The multiple alignment between α-helix and G-proteins sequences has been performed using Clustal Omega (Sievers *et al.*, 2011) and displayed with Jalview (Waterhouse *et al.*, 2009).

In Clustal Omega, the alignments are computed using the very accurate HHalign package (Söding, 2005), which aligns two hidden Markov models-based profiles (Eddy, 1998).

The C-terminal region of the structure of hunt3 has been removed because it is connected to the rest of the protein structure by an unresolved region, in order to reduce the number of missing regions to model and thus the sources of potential errors/uncertainties.

Four missing regions within hunt3 domain have been modelled with MODELLER (Webb and Sali, 2017; Fiser *et al.*, 2000). MODELLER is a program for comparative structure modelling (Fiser *et al.*, 2000; Šali and Blundell, 1993) which, starting from the alignment of the query protein sequence with the template structures and the atomic coordinates of the templates, builds a model containing all non-hydrogen atoms (Webb and Sali, 2017).

Molecular dynamics simulations of hunt3 domain with and without phosphorylated residues have been performed by NAMD 2.12 MD program (Phillips *et al.*, 2005).

AMBER16 tleap module (Case *et al.*, 2005; Salomon-Ferrer *et al.*, 2013) has been used to obtain system topologies and the coordinates. Hunt3 has been parameterized with AMBER ff14SB force field (Maier *et al.*, 2015). The system has been inserted in a rectangular box and filled with TIP3P (Jorgensen *et al.*, 1983) water molecules, imposing a minimal distance between the solute and the box walls of 12.0 Å. Two Na^+ or two Cl^- ions have been added to neutralize the net charge of the protein with and without the two phosphorylated sites, respectively. Instead no ions have been added to the protein with single phosphorylated sites. Before starting the simulation, an energy minimization of 2000 steps imposing a constraint of 1 kcal·mol^{-1}·A^{-1} on the protein atoms has been run, initially using the steepest descent algorithm, while switching to the conjugate gradient algorithm after 500 cycles.

Three thermalization procedures have been performed on the minimized structures increasing the temperature from 0 to 303 K. Subsequently, the force constant for restraints has been gradually reduced from 0.3 to 0.0 kcal/mol/Å^2. The canonical ensemble (NVT) with a time step of 2.0 fs has been employed. For the control of temperature the Langevin thermostat has been used (Loncharich *et al.*, 1982).

A molecular dynamics simulation of 0.5 ns has been carried out using the isobaric-isothermal ensemble (NPT) to equilibrate the pressure. Electrostatic interactions have been taken into account through the Particle Mesh Ewald's method (Darden *et al.*, 1993). The systems have been simulated in periodic boundary conditions, using a cut off radius of 1.2 nm for the non-bonded interactions.

In the equilibration phase a constant pressure of 1 bar has been set using the Berendsen barostat (Berendsen *et al.*, 1984). The SHAKE algorithm (Ryckaert *et al.*, 1977) has been used to constrain covalent bonds involving hydrogen atoms. Each system has been simulated for 500 ns with a timestep of 2 fs.

Molecular dynamics analyses have been carried out using the GROMACS 2018.2 package (Berendsen *et al.*, 1995; Abraham *et al.*, 2015). Only the existence of salt bridges has been evaluated with the VMD plugin (Humphrey *et al.*, 1996).

Plots have been generated with the Grace program (http://plasma-gate.weizmann.ac.il/Grace/) and the images have been generated using the VMD (Humphrey *et al.*, 1996) and Chimera packages (Pettersen *et al.*, 2004).

5.4 References

Abraham,M.J. *et al.* (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1**–**2**, 19–25.

Berendsen,H.J.C. *et al.* (1995) GROMACS : A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, **91**, 43–56.

Berendsen,H.J.C. *et al.* (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys*, **3684**.

Case,D.A. *et al.* (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.

Darden,T. *et al.* (1993) Particle mesh Ewald : An Nlog (N) method for Ewald sums in large systems Particle mesh Ewald : An N -log (N) method for Ewald sums in large systems. *J. Chem. Phys.*, **10089**.

Eddy,S. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fiser,A. *et al.* (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.

Guo,Q. *et al.* (2018) The cryo-electron microscopy structure of huntingtin. *Nature*, **555**, 117–120.

Huang,B. *et al.* (2015) Scalable production in human cells and biochemical characterization of full-length normal and mutant huntingtin. *PLoS One*, **10**, 1–23.

Humphrey,W. *et al.* (1996) VMD : Visual Molecular Dynamics. *J. Mol. Graph.*, **14**, 33–38.

Jolliffe,I.T. and Cadima,J. (2016) Principal component analysis : a review and recent developments Subject Areas. *Philos Trans A Math Phys Eng Sci.*, **374**.

Jorgensen,W.L. *et al.* (1983) Comparison of simple potential functions for simulating liquid water Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*, **79**.

Loncharich,R. *et al.* (1982) Langevin Dynamics of Peptides: The Frictional Dependence of lsomerization Rates of N-Acetylalanyl-WMethylamide. *Biopolymers.*, **32**, 523–535.

Maier,J.A. *et al.* (2015) ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput.*, **11**, 3696–3713.

Pettersen,E.F. *et al.* (2004) UCSF Chimera — A Visualization System for Exploratory Research and Analysis. *J Comput Chem 25*, **25**, 1605–2004.

Phillips,J.C. *et al.* (2005) Scalable Molecular Dynamics with NAMD. *J Comput Chem*, **26**, 1781–

1802.

Pierce,B. and Weng,Z. (2007) ZRANK: Reranking protein docking predictions with an optimized energy function. *Proteins Struct. Funct. Genet.,* **67**, 1078-86

Pierce,B.G. *et al.* (2014) ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*, **30**, 1771–1773.

Ryckaert,J. *et al.* (1977) Numerical integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.*, **23**, 327–341.

Salazar,C. and Höfer,T. (2009) Multisite protein phosphorylation - From molecular mechanisms to kinetic models. *FEBS J.*, **276**, 3177–3198.

Šali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Salomon-Ferrer,R. *et al.* (2013) An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 198–210.

Sievers,F. *et al.* (2011) Fast , scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol. 7*, **7**.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. *BIOINFORMATICS*, **25**, 1189–1191.

Webb,B. and Sali,A (2017) Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinforma.*, **54**.

Chapter 6: Study of Huntingtin sequences in less complex organisms

The study of new animal models opens the possibility of a better understanding of the evolution and function of htt. Therefore, in this study the domain composition of htts present in basal chordates has been investigated focussing the attention on *Ciona intestinalis* and *Branchiostoma floridae*. Moreover, the possibility that htt-like proteins are present also in lower organism has been probed focussing on the nematode *Caenorabditis elegans*.

6.1 Analysis of *Ciona intestinalis* htt

Ciona intestinalis is a tunicate (sea squirt) belonging to the phylum Chordata, which has a great relevance in evolutionary studies because it has the advantage of being a chordate-invertebrate. Indeed as a chordate it shows a body plan and an embryonic development very similar to those of vertebrates (Passamaneck and Di Gregorio, 2005) but, as an invertebrate, it shows enough genetic divergence from vertebrates to allow evolutionary and comparative analyses at the protein level. Thus, the large evolutionary distance separating tunicates and vertebrates (about 520 million years) (Chen *et al.*, 2003) could allow the identification of a htt "signature" related to the ancestral chordate function of the gene/protein (Gissi *et al.*, 2006).

The htt protein of *C. intestinalis* is 2945 amino acids long, notably shorter than its vertebrate homologs, which are 3130 amino acids long on average. This length difference can be due to deletions in the N-terminal region of the protein (**Figure S1,** Supplementary material).

Moreover, the N-terminal region of the *C. intestinalis* protein lacks the polyQ domain, or any kind of simple repeat (**Figure S1**). Even the proline-rich region typical of mammalian htt is absent. However, the amino acid sequence identity between human and ascidian htt is 34% (**Table S3**, Supplementary material).

A total of 8 HEAT repeats are present in the ascidian htt. These are located as tandem arrays or as single elements in the N-terminal (4 repeats), central (2 repeats), and C-terminal (2 repeats) regions (Gissi *et al.*, 2006) (**Table 1**).

Table 1. HEAT repeats within human (Saudou and Humbert, 2016) ascidian and amphioxus htt (Candiani *et al.*, 2007).

HEAT repeats location in *H. sapiens* htt	HEAT repeats location in *C. intestinalis* htt	HEAT repeats location in *B. floridae* htt
N-terminal 114-413	N-terminal 58–96	N-terminal 75-113
N-terminal 672-1176	N-terminal 139–177	N-terminal 156-194
Central 1289-1710	N-terminal 181–219	N-terminal 198-236
Central 2175-2325	N-terminal 682–720	N-terminal 306-344
C-terminal 2355-2475	Central 867–905	N-terminal 802-840
C-terminal 2667-2937	Central 1341–1378	Central 1371-1409
C-terminal 2975-3107	C-terminal 2771–2809	Central 1556-1595
	C-terminal 2864-2904	Central 1618-1656
		C-terminal 2746-2784
		C-terminal 2927-2965
		C-terminal 3020-3038

The analysis of htt protein alignment (**Figure S1**, Supplementary material) shows that all ascidian HEAT repeats are also conserved in the human homolog.

An identification and analysis of the ordered domains similar to that performed in human htt has also been carried out for the ascidian htt sequence.
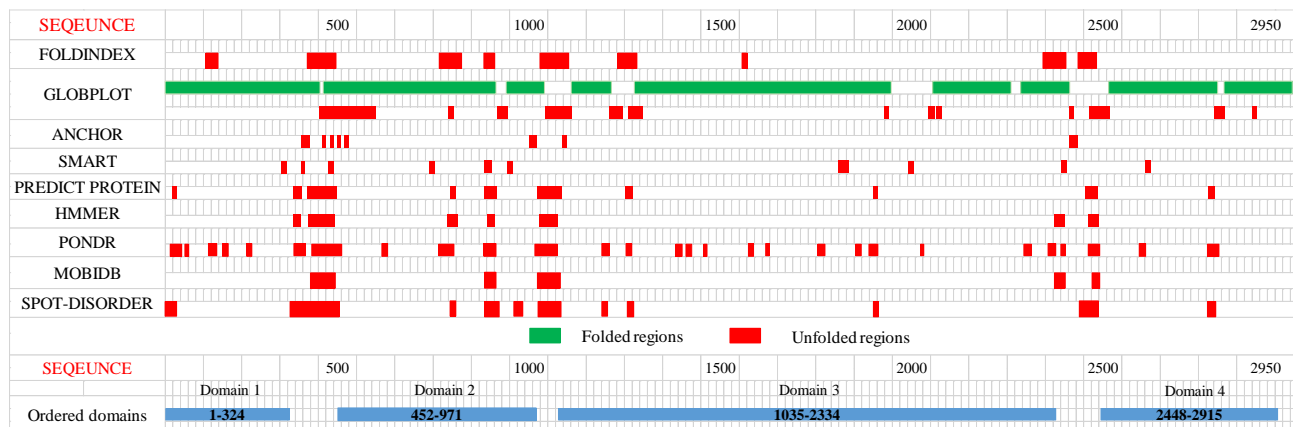
Figure 25. Predicted ordered (green) and disordered (red) regions of ascidian htt. Htt domains predicted as ordered by consensus among the different methods are displayed in the bottom part of the figure as light blue bars.

Using several order/disorder prediction methods (**Figure 25**, **Table S4**, Supplementary material), four ordered domains have been identified (hereafter named domain 1-4). Then, the structure prediction of each ordered domain has been performed with I-TASSER.

All the models show an α-helical structure (**Figure 26**), due to the presence of HEAT repeats. Domain 1, 2 and 4 are characterized by a concave shape. Instead the third ordered domain exhibits structural

similarity with hunt3, including the presence of an α-helix stemming from the C-terminal region of the domain and interacting with a concave region on the opposite side (**Figure 26 c**).

Since I-TASSER identifies structural templates from the PDB by a multiple threading approach to build structural models, all the models of ascidian htt ordered domains have been built using the structure of human htt as a template and selected by C-score (**Table S5**, Supplementary material). As expected, human htt is the best structural homolog of each model (**Tables S6**, **S7**, **S8**, **S9**, Supplementary material). **Figure 27** displays the superimposition between the models of ordered domains of the ascidian protein and the corresponding domains of human htt.

All the models (**Figure 27 b-d**) overlap with different regions of the human protein, covering almost the entire structure, even though the ascidian protein is slightly shorter.
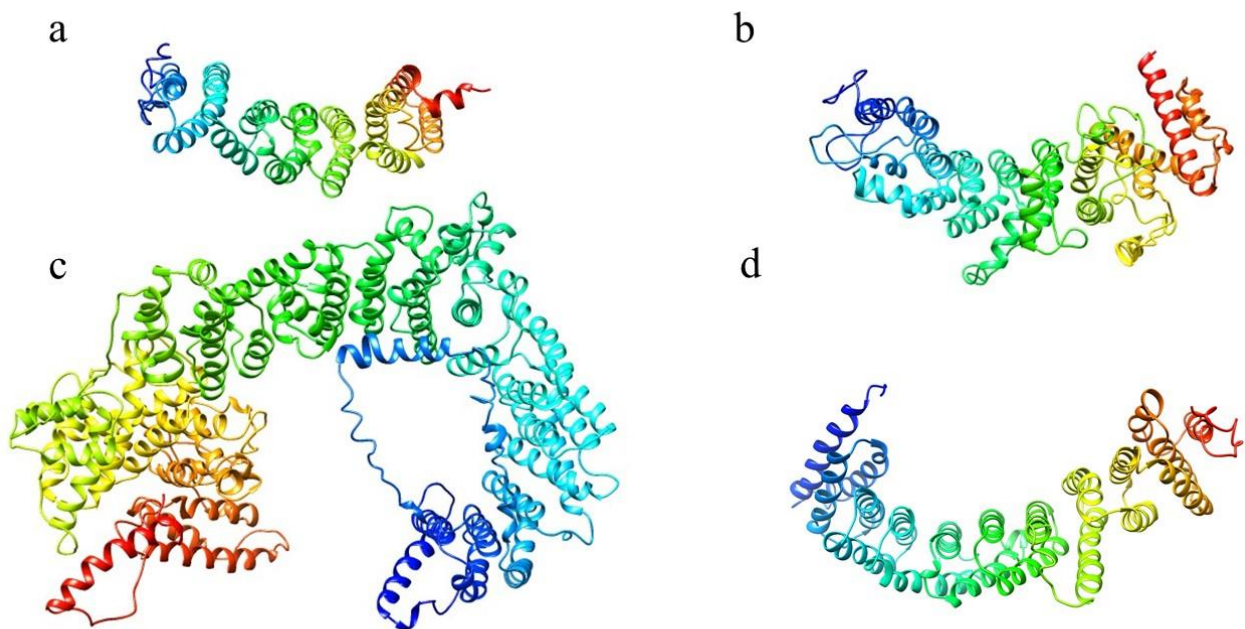
Figure 26. Structural models of the ordered domains of ascidian htt. Domain 1 in (**a**), domain 2 in (**b**), domain 3 in (**c**) and domain 4 in (**d**).
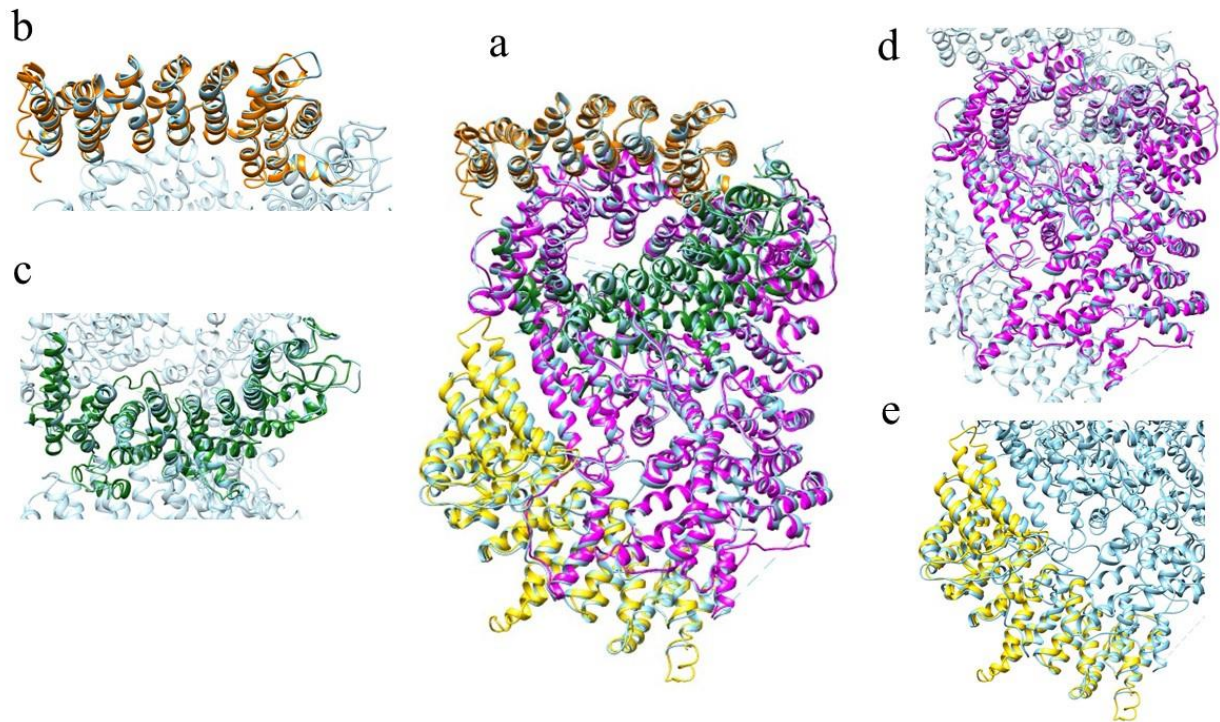
Figure 27. (**a**) Superimposition between the models of the four ordered domains of ascidian htt (domain 1 in orange, domain 2 in dark green, domain 3 in magenta, domain 4 in yellow) and the structure of human protein (in sky blue, PDB code: **6EZ8**). In the side panels particulars of the superimpositions (domain 1 in **b**, domain 2 in **c**, domain 3 in **d**, and domain 4 in **e**) are shown.

Given the sequence identity (34%) and the structural homology between the models of the ordered domains of ascidian htt and the human structure, the subdivision of the ascidian sequence in four ordered domains and the models obtained appear to be reliable.

6.2 Analysis of *Branchiostoma floridae* htt

Branchiostoma floridae is a lancelet (also known as amphioxus) that belongs to the subphylum Cephalochordata of the phylum Chordata.

The nervous system development of the amphioxus is particularly close to that of vertebrates as it includes vertebrate-like anatomical characteristics such as a dorsal nerve cord, a notochord and segmentally arranged muscles, but it lacks the typical subcellular and tissue specialization of the vertebrates nervous system (Candiani *et al.*, 2007). Thus, this organism is particularly useful to deduce features already present in the last common ancestor of chordates.

The amphioxus htt (AmphiHtt) comes from an invertebrate chordate whose phylogenetic node of divergence is thought to go back 540 million years, while *Ciona intestinalis* seems to have diverged more recently (Blair and Hedges, 2005; Bourlat *et al.*, 2006; Delsuc *et al.*, 2006).

AmphiHtt is mainly abundant in the neural compartment, indicating that htt, in amphioxus, could be involved in neuronal functions (Candiani *et al.*, 2007).

AmphiHtt protein has two glutamines (Q17 and Q18) in the same polyQ tract position of human htt (**Figure S1**, Supplementary material), thus suggesting that a polyQ tract was already present 540 million years ago (Candiani *et al.*, 2007) (see the Introduction for the htt and polyQ evolution). Therefore, Amphioxus is the first known non-vertebrate species to contain glutamine residues in htt, thus dating the presence of glutamines in non-vertebrates and demonstrating that the common ancestor of cephalochordates and vertebrates already had this signature (**Figure 2**).

The differences in the length of the polyQ tract between amphioxus and vertebrates suggest that the function of htt may have evolved different biochemical properties in both lineages (Candiani *et al.*, 2007).

One further characteristic of AmphiHtt is the complete absence of the polyP-rich region, such as in ascidian htt, indeed the polyP tract is only present in the mammalian protein (**Figure S1**, Supplementary material). On the contrary, the first 17 amino acids of AmphiHtt, with its three lysines involved in the intracellular distribution of the protein between the cytoplasm and nucleus in vertebrates (Rockabrand *et al.*, 2007), are also strongly conserved in human htt (**Figure S1**). With respect to the latter, conservation of the primary amino acid sequence in amphioxus protein (46%) is greater than that in *C. intestinalis* (34%) (**Figure S3**, Supplementary material). The comparison of the gene structure of AmphiHtt with the human and ascidian homologues highlights that amphioxus htt is closer to vertebrate htt than the ascidian one and leads to the hypothesis that also its functions are possibly closer to those of the vertebrate protein (Candiani *et al.*, 2007).

Almost all HEAT repeats in amphioxus (**Table 1**, **Figure S1**, Supplementary material) seem to be conserved in the human homolog, an exception being the last HEAT consensus at 3020-3038 in amphioxus sequence that has no correspondence in the human protein.

AmphiHtt sequence has been analyzed with the same tools used for human and ascidian htt (**Table S10**, Supplementary material). Using a consensus between different methods, which predict disorder regions starting from the amino acid sequence, four ordered domains have been identified, depicted in **Figure 28.**
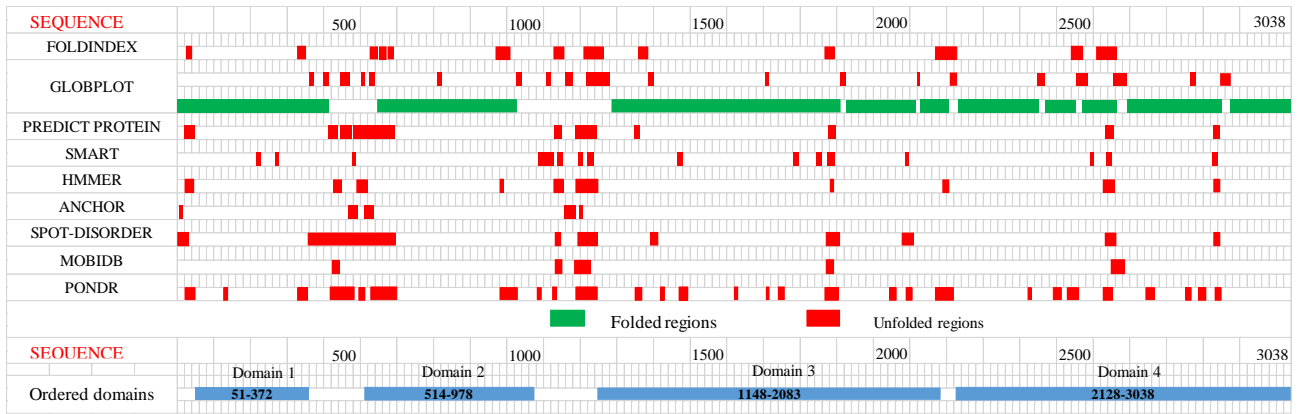
Figure 28. Predicted ordered (green) and disordered (red) regions of AmphiHtt. Htt domains predicted as ordered by consensus among the different methods are displayed in the bottom part of the figure as light blue bars.

The structure prediction has been performed for all ordered domains using I-TASSER (**Table S11**, Supplementary material). The four domains show the typical topology of HEAT repeats (**Figure 29**) and structural similarity with human htt (**Tables S12-S15**, Supplementary material). Interestingly domain 3 has structural similarity with the human hunt3 domain (**Figure 29 c**), that includes the α-helix located in the putative site of interaction with G proteins.
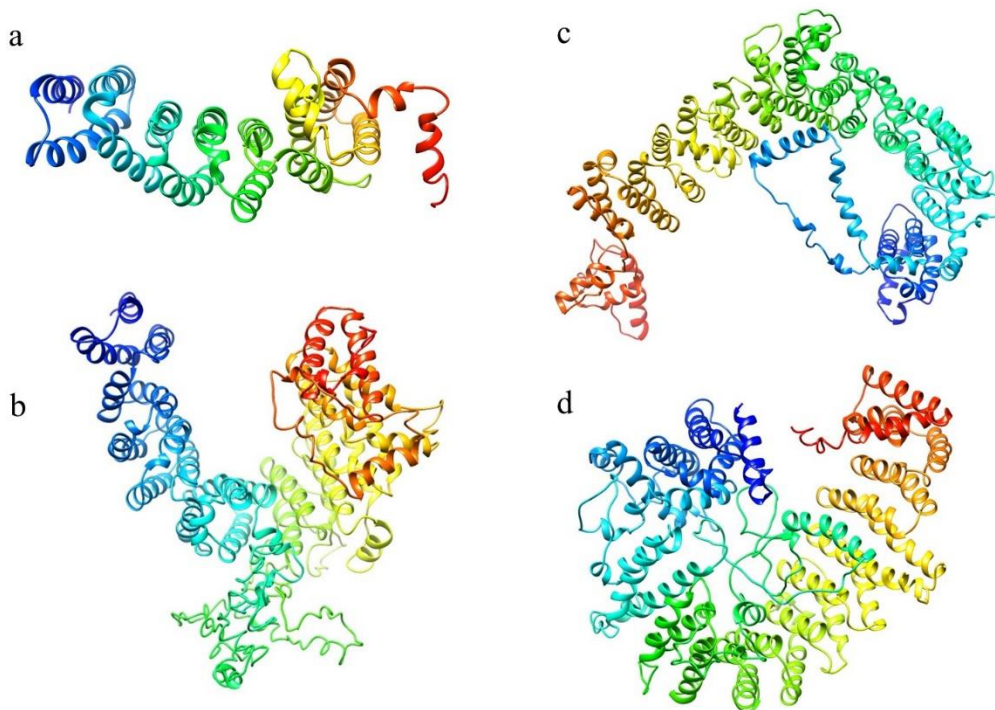
Figure 29. Structural models of the ordered domains of AmphiHtt. Domain 1 (**a**), domain 2 (**b**), domain 3 (**c**), domain 4 (**d**).
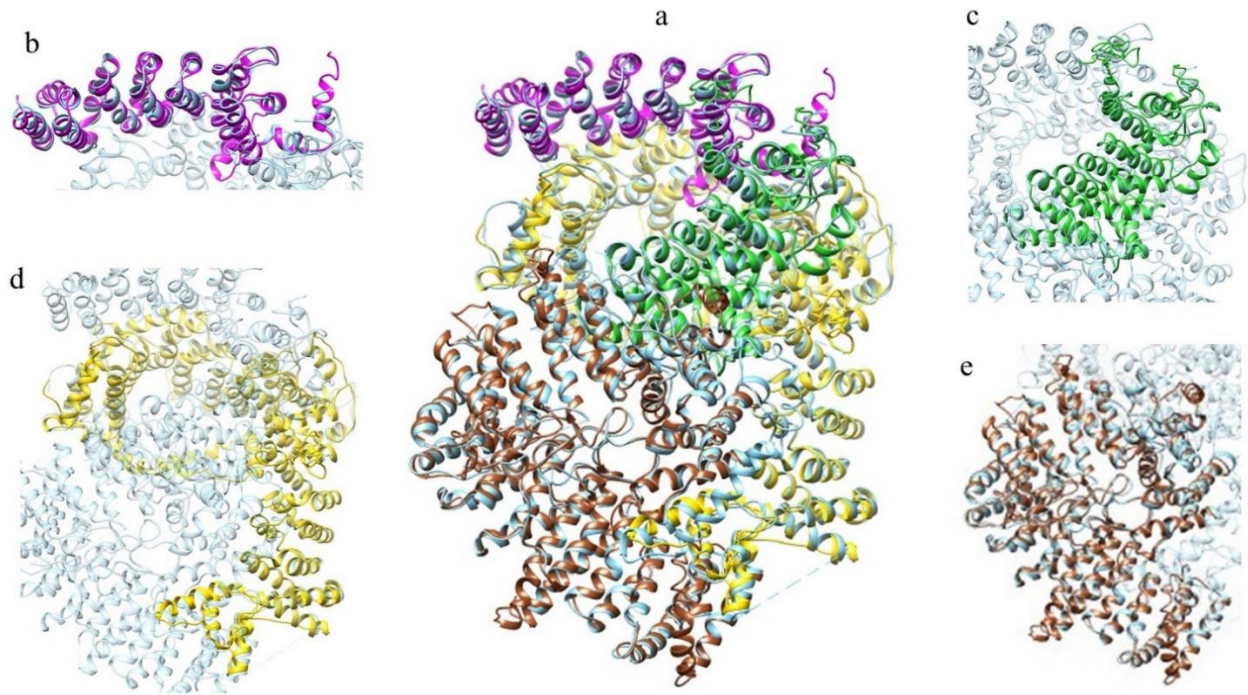
Figure 30. (**a**) Superimposition between the models of the four ordered domains of AmphiHtt (domain 1 in magenta, domain 2 in dark green, domain 3 in yellow and domain 4 in sienna) and the structure of human protein (in cyan, PDB code: **6EZ8**). In the side panels particulars of the superimpositions (domain 1 in **b**, domain 2 in **c**, domain 3 in **d**, and domain 4 in **e**) are shown.

Figure 30 displays the superimposition between the models of the ordered domains of AmphiHtt and the human ones. As for the ascidian htt the models of AmphiHtt domains cover almost the entire structure of human htt, suggesting that the subdivision of the domains is rather reliable.

6.3 Identification of an htt-like protein in *Caenorhabditis elegans*

Caenorhabditis elegans is a nematode that lives in temperate soil environments and is one of the "supermodels" of modern biology as it possesses many genes with a significant similarity to those implicated in human diseases. Further, a large number of human parasites, which have a relevant impact on world health are included in the phylum Nematoda (Aboobaker and Blaxte, 2000).

No literature data are available on the existence of a htt-like protein in *C. elegans*. Interestingly a delta-BLAST search using human htt as a bait against the *C. elegans* genome retrieved an uncharacterized protein (CELE_F21G4.6) of 2022 residues, which displays 21% sequence identity over a 13% of query coverage (**Table S3**, **Figure S1**, Supplementary material). This sequence is shorter than the human one, with a deletion observed in the N-terminal region with respect to the human protein.

The uncharacterized protein sequence has been analysed adopting the same approach used to study human, ascidian and AmphiHtt. Through the use of several web services (**Table S16**, Supplementary material) two putative ordered domains have been identified (**Figure 31**).
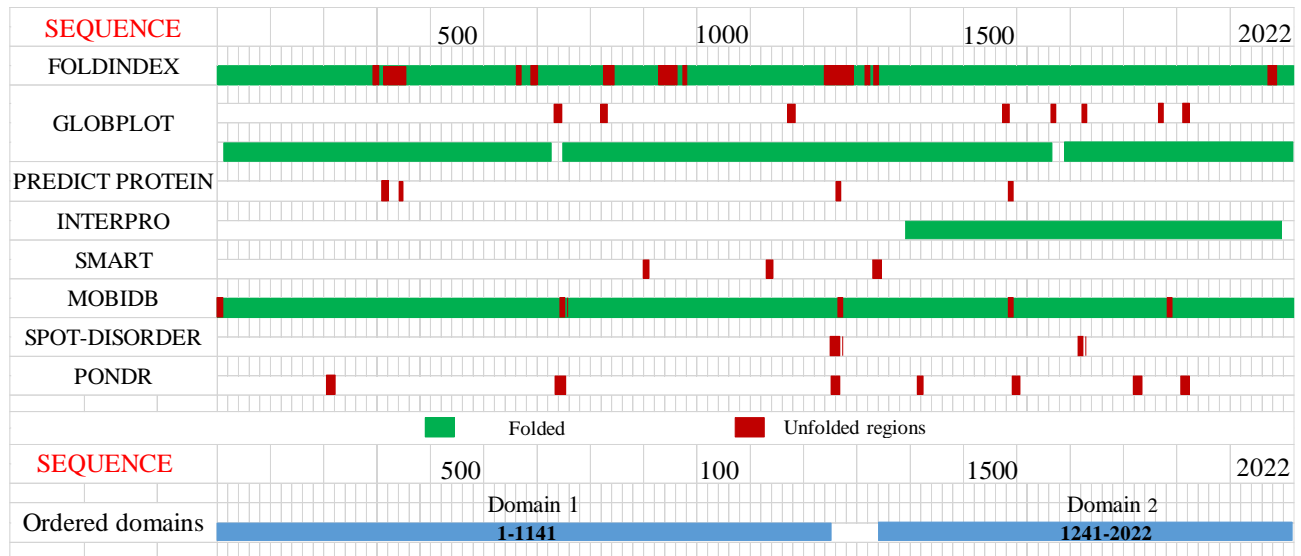
Figure 31. Predicted ordered (green) and disordered (red) regions of the putative *C. elegans* htt-like protein. Htt domains, predicted as ordered by consensus among the different methods, are displayed in the bottom part of the figure as light blue bars.

A structure prediction has been conducted for both ordered regions using I-TASSER (**Figure 32**, **Table S17**, Supplementary material). Both domains are made up of α-helices and have structural similarity with the structure of the human htt (**Figure 33**, **Table S18**, **Table S19**, Supplementary material). The model of the first domain overlaps with the N-terminal region of human htt, while the second domain with the C-terminal region, leaving the bridge region of human htt structure uncovered (**Figure 33**). Moreover domain 1 overlaps with the portion of hunt3 containing the α-helix, analyzed in the previous chapter.

The presence of htt in the amoeba *Dictyostelium discoideum* has suggested that HTT is an evolutionarily old gene that has been lost in some later animals, such as *C. elegans* (Cattaneo, *et al.*, 2005). No htt-like protein has been previously reported in *C. elegans*, but it has just been predicted. Thus, the uncharacterized protein of *C. elegans* analyzed in this work could be an htt-like protein which lacks two portions (the N-terminal and the corresponding "bridge region" of human htt). This is in agreement with the length of the sequence, which is shorter than human one.
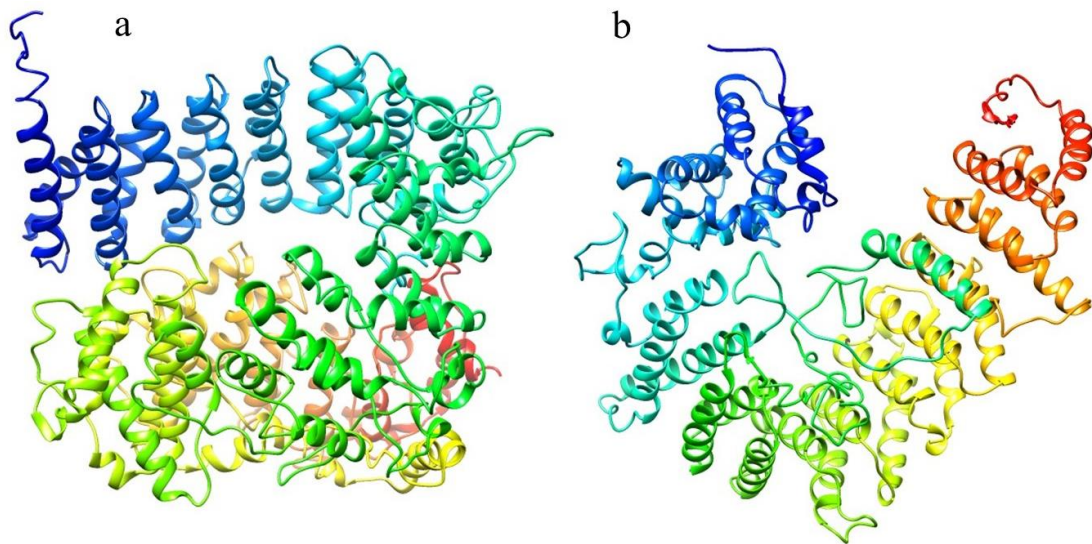
Figure 32. Structural models of domain 1 (**a**) and domain 2 (**b**) of a putative htt-like protein identified in *C. elegans*.
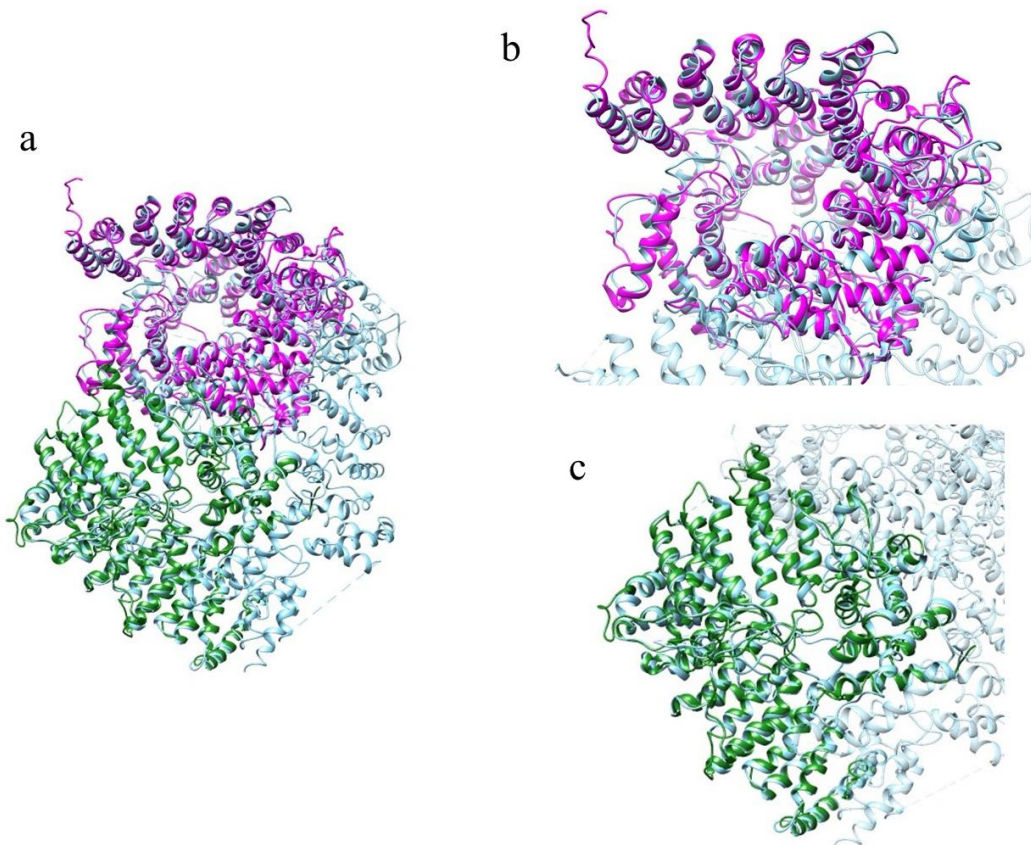
Figure 33. (**a**) Superimposition between the models of the two ordered domains of an htt-like protein (in magenta domain 1 and in dark green domain 2) identified in *C. elegans* and the structure of human htt (in sky blue, PDB code: **6EZ8**). In the two side panels particulars of the superimpositions (domain 1 in **b**, domain 2 in **c**) are depicted.

In conclusion, the analyses of htt sequences of *C.intestinalis*, *B. floridae* and *C. elegans* have shown the conservation of hunt3 domain in these organisms, in which it could have a similar function in vesicular transport, but not necessarily in neuronal tissues. On the contrary, the other domains of human htt identified in this work seem to be conserved only in the proteins from *C. intestinalis* and *B. floridae*, which display a length comparable to that of human protein. The lack of two portions in the nematode htt-like protein suggests that the ancestral htt sequence could be as long as that of the nematode one and that the proteins found in higher complexity organisms have acquired additional domains and, consequently, additional functions.

6.4 Methods

The sequences of the ascidian and the amphioxus htts, and of the htt-like protein identified in *C. elegans* have been retrieved through a delta-BLAST (Domain Enhanced Look-up Time Accelerated BLAST) (Boratyn *et al.*, 2012) search using human htt as a bait against the non-redundant protein sequences database. Delta-BLAST aligns a query sequence to conserved domains in CDD (Conserved Domain Database) (Marchler-Bauer *et al.*, 2011) through RPS-BLAST (which stands for "Reverse Position-Specific BLAST and is a variant of the popular PSI-BLAST program, "Position-Specific Iterated BLAST") and then performs a sequence database search using a PSSM (position specific score matrix) (Taylor, 1986; Gribskov *et al.*, 1987; Patthy, 1987; Altschul *et al.*, 1997) derived from the aligned domains. A PSSM is obtained from a multiple sequence alignment (MSA) of related proteins. Delta-BLAST uses aligned domains to compute a PSSM and to find more homologs (Boratyn *et al.*, 2012).

The multiple sequence alignment between *H. sapiens*, *C. intestinalis*, *B. floridae* Htt and *C. elegans* htt-like proteins has been performed using Clustal Omega (Sievers *et al.*, 2011) and visualized with Jalview (Waterhouse *et al.*, 2009). In Clustal Omega, the alignments are computed using the very accurate HHalign package (Söding, 2005), which aligns two hidden Markov model profiles (Eddy, 1998).

Disordered regions and ordered domains have been identified by several tools, described below.

Foldindex (https://fold.weizmann.ac.il/fldbin/findex) predicts if a given protein sequence is intrinsically unfolded using the algorithm of Uversky and co-workers (Uversky *et al.*, 2000). This algorithm is based on average hydrophobicity and net charge and provides a single 'foldability' score (Prilusky *et al.*, 2005).

Globplot (http://globplot.embl.de/cgiDict.py) detects disordered regions within protein sequences. It measures and displays the propensity of a sequence region to be ordered or disordered (Linding, Russell, *et al.*, 2003).

PredictProtein (https://www.predictprotein.org) is a meta-service for sequence analysis that predicts structural and functional features of proteins. For most prediction methods, the prediction quality is estimated by a reliability score (Yachdav *et al.*, 2014).

ANCHOR (https://anchor.enzim.hu/) is a web service that predicts protein binding regions that are disordered in isolation but can become ordered upon binding of a molecular partner (Dosztányi *et al.*, 2009). ANCHOR deduces the basic biophysical properties of disordered binding regions using estimated energy calculations (Mészáros *et al.*, 2009). ANCHOR's output is a score, which represents the probability of each residue of the input sequence to be in a disordered binding region. Regions with a score higher than 0.5 are predicted as disordered binding regions (Dosztányi *et al.*, 2009). IUPred and ANCHOR have been currently replaced by IUPred2A (https://iupred2a.elte.hu/) (Mészáros *et al.*, 2018).

InterPro database (https://www.ebi.ac.uk/interpro/) classifies sequences into protein families and predicts the presence and function of important domains and sites, making use of diagnostic models known as signatures (Hunter *et al.*, 2012; Mitchell *et al.*, 2015; Finn *et al.*, 2017).

SMART (http://smart.embl-heidelberg.de/) is a web resource for the identification of protein domains and the analysis of their architectures; it is synchronized with Uniprot (The UniProt Consortium, 2014), ENSEMBL (Flicek *et al.*, 2014) and STRING (Franceschini *et al.*, 2013). The SMART database integrates manually-curated hidden Markov models for many domains (Krogh *et al.*, 1994) and offers a variety of analysis and visualization tools (Letunic and Bork, 2018; Letunic *et al.*, 2015).

HMMER (http://www.ebi.ac.uk/Tools/hmmer/) is a suite for sequence similarity searches using profile hidden Markov models (HMMs) (Potter *et al.*, 2018) focused primarily on UniProtKB (The UniProt Consortium, 2017). The query sequence is scanned against the Pfam profile HMM library using hmmscan, to detect any Pfam family (Finn *et al.*, 2014) and analyzed for the presence of disordered regions, using IUPred (Dosztányi, Csizmok, *et al.*, 2005; Dosztányi, Csizmók, *et al.*, 2005), signal peptides and transmembrane regions, using Phobius (Käll *et al.*, 2004), and coiled-coil regions (Lupas *et al.*, 1991). Currently, the maximum length of the input sequence has been reduced to 1000 amino acids. Therefore it has not been possible to use this method for the sequence of *C. elegans*, the last and recently analyzed sequence.

MobiDB 3.0 (http://mobidb.bio.unipd.it/), (Piovesan *et al.*, 2018) is an update of the previous database of intrinsically disordered and mobile proteins, MobiDB 2.0, (Potenza *et al.*, 2015), that provides a complete picture of different types of protein disorder covering all Uniprot sequences. MobiDB 3.0 returns a consensus of different predictors: ESpritz (Walsh *et al.*, 2012), IUPred (Dosztányi, Csizmok, *et al.*, 2005), DisEMBL (Linding, Jensen, *et al.*, 2003), GlobPlot (Linding, Russell, *et al.*, 2003), VSL2b (Peng *et al.*, 2006), DynaMine (Cilia *et al.*, 2013), Anchor (Mészáros *et al.*, 2009), FeSS (Piovesan *et al.*, 2017). Consensus generation is handled by MobiDB-lite (Necci *et al.*, 2017).

PONDR (http://www.pondr.com) is a neural network predictor of natural protein disordered regions. It includes different predictors for short, medium, and long disordered regions. In this work PONDR VL-XT has been used. It integrates the VL1 predictor (for internal regions), the N-terminus predictor (XN), and the C-terminus predictor (XC). The XN and XC predictors are called XT because trained using X-ray crystallographic data (Romero *et al.*, 2001).

SPOT-DISORDER (http://sparks-lab.org/server/SPOT-disorder/) is available both as a web server and a standalone program and predicts both short and long disordered regions. It implements deep bidirectional LSTM (Long Short-Term Memory) recurrent neural networks to capture nonlocal interactions (long-range interactions between amino acid residues that are structural but not sequence neighbours) that are essential for determining whether a protein will fold (structured) or will not fold (intrinsically disordered) into a unique three-dimensional structure (Hanson *et al.*, 2017).

The I-TASSER server (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) has been used to carry out the structure prediction of protein domains. I-TASSER is an integrated platform to perform automated protein structure and function prediction. It generates three-dimensional models of the query protein using a threading approach to identifying suitable modelling templates for the query sequence (Roy *et al.*, 2011). The quality of the template alignments is judged accounting for the statistical significance of the best threading alignment, the Z-score (Roy *et al.*, 2011). Models are built using a combination of threading and *ab initio* techniques for the template-aligned and template-unaligned regions, respectively. The function of the query protein is predicted by structurally matching the models obtained against the protein structures present in the PDB (Roy *et al.*, 2011) and ranking the functional analogs through a combination of TM-score, RMSD, sequence identity, and coverage of the structure alignment (Roy *et al.*, 2011). The reliability of the structural models can be assessed by the C-score parameter provided by the server. Supplementary Material Table S5, Table S11 and Table S17 report the C-score values obtained for the models of the predicted ordered domains of htt sequences in *C. intestinalis* and *B. floridae* and in *C. elegans*, respectively.

6.5 References

Aboobaker,A. and Blaxte,M. (2000) Medical Significance of Phylum Arthropoda. *Ann. Med.*, **32**, 23–30.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Blair,J.E. and Hedges,S.B. (2005) Molecular phylogeny and divergence times of deuterostome animals. *Mol. Biol. Evol.*, **22**, 2275–2284.

Boratyn,G.M. *et al.* (2012) Domain enhanced lookup time accelerated BLAST. *Biol. Direct*, **7**.

Bourlat,S.J. *et al.* (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature*, **444**, 85–88.

Candiani,S. *et al.* (2007) Characterization, developmental expression and evolutionary features of the huntingtin gene in the amphioxus Branchiostoma floridae. *BMC Dev. Biol.*, **7**, 1–16.

Cattaneo,E. *et al.* (2005) Normal huntingtin function: An alternative approach to Huntington's disease. *Nat. Rev. Neurosci.*, **6**, 919–930.

Chen,J.Y. *et al.* (2003) The first tunicate from the Early Cambrian of South China. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 8314–8318.

Cilia,E. *et al.* (2013) From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.*, **4**, 1–10.

Delsuc,F. *et al.* (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.

Dosztányi,Z. *et al.* (2009) ANCHOR: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, **25**, 2745–2746.

Dosztányi,Z., Csizmok,V., *et al.* (2005) IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Dosztányi,Z., Csizmók,V., *et al.* (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Eddy,S. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Finn,R.D. *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.

Finn,R.D. *et al.* (2014) Pfam: The protein families database. *Nucleic Acids Res.*, **42**, 222–230.

Flicek,P. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, 749–755.

Franceschini,A. *et al.* (2013) STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, 808–815.

Gissi,C. *et al.* (2006) Huntingtin gene evolution in Chordata and its peculiar features in the ascidian Ciona genus. *BMC Genomics*, **7**, 1–16.

Gribskov,M. *et al.* (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **84**, 4355–4358.

Hanson,J. *et al.* (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685–692.

Hunter,S. *et al.* (2012) InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, 306–312.

Käll,L. *et al.* (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.

Krogh,A. *et al.* (1994) Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Letunic,I. *et al.* (2015) SMART: Recent updates, new developments and status in 2015. *Nucleic Acids Res.*, **43**, D257–D260.

Letunic,I. and Bork,P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.*, **46**, D493–D496.

Linding,R., Russell,R.B., *et al.* (2003) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.

Linding,R., Jensen,L.J., *et al.* (2003) Protein disorder prediction: Implications for structural proteomics. *Structure*, **11**, 1453–1459.

Lupas,A. *et al.* (1991) Predicting coiled coils from proteins sequences. *Science (80-.).*, **252**, 1162–1164.

Marchler-Bauer,A. *et al.* (2011) CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, 225–229.

Mészáros,B. *et al.* (2018) IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, **46**, W329–W337.

Mészáros,B. *et al.* (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*

Mitchell,A. *et al.* (2015) The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.

Necci,M. *et al.* (2017) MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, **33**, 1402–1404.

Passamaneck,Y.J. and Di Gregorio,A. (2005) Ciona intestinalis: Chordate development made simple. *Dev. Dyn.*, **233**, 1–19.

Patthy,L. (1987) Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.*, **198**, 567–577.

Peng,K. *et al.* (2006) Length-dependent prediction of protein in intrinsic disorder. *BMC Bioinformatics*, **7**, 1–17.

Piovesan,D. *et al.* (2017) FELLS: Fast estimator of latent local structure. *Bioinformatics*, **33**, 1889–1891.

Piovesan,D. *et al.* (2018) MobiDB 3.0: More annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.*, **46**, D471–D476.

Potenza,E. *et al.* (2015) MobiDB 2.0: An improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, **43**, D315–D320.

Potter,S.C. *et al.* (2018) HMMER web server: 2018 update. *Nucleic Acids Res.*, **46**, W200–W204.

Prilusky,J. *et al.* (2005) FoldIndex©: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.

Rockabrand,E. *et al.* (2007) The first 17 amino acids of Huntingtin modulate its sub-cellular localization, aggregation and effects on calcium homeostasis. *Hum. Mol. Genet.*, **16**, 61–77.

Romero,P. *et al.* (2001) Sequence complexity of disordered protein. *Proteins Struct. Funct. Genet.*, **42**, 38–48.

Roy,A. *et al.* (2011) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.

Saudou,F. and Humbert,S. (2016) The Biology of Huntingtin. *Neuron*, **89**, 910–926.

Sievers,F. *et al.* (2011) Fast , scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Taylor,W.R. (1986) Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, **188**, 233–258.

The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, 191–198.

The UniProt Consortium (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Uversky,V.N. *et al.* (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins Struct. Funct. Genet.*, **41**, 415–427.

Walsh,I. *et al.* (2012) Espritz: Accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. *BIOINFORMATICS*, **25**, 1189–1191.

Yachdav,G. *et al.* (2014) PredictProtein - An open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.*, **42**, 1–7.

Chapter 7: Conclusion

Huntington's disease (HD) is a rare neurodegenerative and autosomal dominant disorder (Saudou and Humbert 2016). The neuropathology consists in the dysfunction and death of the neurons of the striatum and the neurons that project from the cortex to the striatum. The symptoms vary between individuals but are usually characterized by a triad of motor, cognitive, and psychiatric symptoms. From its earliest stages, HD dramatically affects patients quality of life and prevents their functional abilities. In many cases patients die because of fatal aspiration pneumonia (Saudou and Humbert 2016).

HD is caused by a mutation in the IT15 gene located on chromosome 4, which codes for a protein called huntingtin (htt) (HD Collaborative research group 1993). The result is the production of a mutant htt (mhtt) with an abnormally long polyglutamine (polyQ) repeat that leads to htt aggregation into fibrils and a variety of other aggregates (Legleiter *et al.* 2010; McColgan and Tabrizi 2018; Poirier *et al.* 2002).

Htt is a protein of 348 kDa and 3144 amino acids. The N-terminal region of the protein includes the polyQ stretch, which begins at the 18th amino acid and it is followed by a proline-rich domain (PRD) (Zuccato, Valenza, and Cattaneo 2010). Downstream of the polyQ there are several HEAT repeats, sequences of 40 amino acids, made up of two antiparallel α-helices linked by a short loop and packed together to form a flexible rod (denoted alpha-rod). The HEAT repeat domains can act as a scaffold for diverse protein complexes and mediate inter- and intra-molecular interactions (Palidwor *et al.* 2009). The first known life form in the evolutionary lineage to carry the htt gene is the *Dictyostelium discoideum*, an amoeba, and the presence of htt has been predicted in the nematode *Caenorhabdtis elegans*, but not in the yeast *Saccharomyces cerevisiae*, indicating that HTT is an evolutionarily old gene (Palidwor *et al.* 2009). Htt is also present in an old deuterostome (sea pineapple) and in the sea urchin (Cattaneo, Zuccato, and Tartari 2005). Two CAG triplets occur in the sea urchin and amphioxus sequences but tend to increase in organisms with a more developed nervous systems, until they reach their maximum extension in humans (Zuccato and Cattaneo 2016). Htt undergoes to proteolysis at numerous sites, mostly found in disordered regions. The proteolysis of wild-type htt causes the toxicity of the protein. Htt is subjected to various post-translational modifications, such as phosphorylation, acetylation, palmitoylation, ubiquitynation, and sumoylation, and some of these can have a therapeutic significance since they regulate mhtt toxicity (Saudou and Humbert 2016). Htt has been involved in several cellular events during development and adulthood, mostly in the nervous system (Saudou and Humbert 2016).

Htt is present predominantly in the cytoplasm of neurons and is enriched in compartments containing vesicle-associated proteins (Velier *et al.* 1998). Htt controls the transport of organelles, in axons and dendrites within neurons, in both the anterograde and retrograde directions (Engelender 1997; Li *et al.* 1998). Htt is located not only in neurons but also at the spindle poles, mitotic spindles, and astral microtubules in dividing cells (Elias *et al.* 2014; Godin *et al.* 2010; Gutekunst *et al.* 1995) and regulates the cortical localization of proteins essential for mitotic spindle positioning (Elias *et al.* 2014). Further, htt mediates the transport of proteins to the pericentriolar material (PCM), including PCM1, a protein that is required for ciliogenesis (Saudou and Humbert 2016).

Htt binds several proteins involved in endocytosis acting as a scaffold and regulating the first step of endocytosis though membrane coating and invagination (El-Daher *et al.* 2015; Kaltenbach *et al.* 2007; Moreira Sousa *et al.* 2013). Htt may directly regulate autophagy (Ochaba *et al.* 2014; del Toro *et al.* 2009; Wong and Holzbaur 2014) and transcriptional regulation by the interaction with transcriptional activators and repressors (Dunah *et al.* 2002; Futter *et al.* 2009; Holbert *et al.* 2001; Kegel *et al.* 2002; Yohrling IV *et al.* 2003; Zuccato *et al.* 2003) and may also act as a transcriptional cofactor of itself (Benn *et al.* 2008). Different studies have shown that htt is essential for embryonic development (Duyao *et al.* 1995; Nasir *et al.* 1995; Zeitlin *et al.* 1995) and for the formation of the nervous system (White *et al.* 1997), for embryogenesis and development (Godin *et al.* 2010).

Some studies indicate that htt is a regulator of tissue maintenance (Elias *et al.* 2014, 2015) and that has a prosurvival role (Dragatsis, Efstratiadis, and Zeitlin 1998; Dragatsis, Levine, and Zeitlin 2000; Leavitt *et al.* 2006; Nasir *et al*. 1995; O'Kusky *et al.* 1999; Rigamonti *et al*. 2000, 2001; Zhang *et al*. 2003; Zuccato *et al*. 2001).

Other experimental studies have highlighted a possible involvement of htt in the neuroprotective role mediated by 17β-estradiol (E2) (Cardinale *et al*. 2018; De Marinis *et al*. 2013; De Marinis, Marino, and Ascenzi 2011; Nuzzo *et al*. 2016, 2017). E2 protective effects depend on the over-expression of neuroglobin (Ngb), a brain globin, promoted by ERβ (estrogen receptor β) (De Marinis *et al.* 2013, 2011). E2 stimulation induces the over-expression of both htt and Ngb, promoting the formation of the htt/Ngb complex at the mitochondria (Nuzzo *et al.* 2017) where Ngb binds to cytochrome c, reducing the release of cytochrome c into the cytosol. This event elicits a decrease of caspase-3 activation and consequently of the apoptosis activation (De Marinis *et al.* 2013). However, how Ngb translocation to mitochondria occurs is completely unknown. Given a possible role of htt in vesicular trafficking, htt may be a putative Ngb carrier.

Htt huge size still doesn't allow crystallization and X-ray diffraction studies. However, recently the structure of full-length human htt in a complex with HAP40 (Htt-associated protein 40) has been

determined by cryo-electron microscopy, but this structure lacks atomic details of several regions (Guo *et al.* 2018).

Due to the lack of structural data when this work began, the aim of this thesis has been that of obtaining information on the structure-function relationships of htt through a structural bioinformatics approach. As a result of this work, five htt regions predicted to fold into ordered domains have been detected, leading to the identification of a previously undetected HEAT repeats region in the third ordered domain (hunt3). Structural models of htt ordered domains have been proposed for the first time, providing a putative function for four out of the five predicted ordered domains. The third ordered domain has shown structural similarity with a Karyopherin, a protein involved in nuclear import by interacting with GTP-binding proteins. Surprisingly the analysis of the htt interactome has revealed that it interacts with several GTP binding proteins. Molecular docking simulations have confirmed that hunt3 can be considered the putative interaction site between htt and G proteins. Among htt interactors, the G protein $G\alpha_o1$ (encoded by the GNAO1 gene) is particularly interesting. In fact, different heterozygous mutations in the GNAO1 gene are the cause of a severe neurodevelopmental disorder, the epileptic encephalopathy type 17, which is characterized by symptoms similar to those observed in HD.

Moreover, the analyses of htt and Ngb interactome revealed that $G\alpha_o1$ interacts both with htt and with Ngb. Molecular docking simulations indicated that a ternary complex between htt, $G\alpha_o1$ and Ngb is plausible, leading to the hypothesis that G proteins mediate the interaction between htt and Ngb, and possibly the transport of the latter protein to mitochondria.

In addition to the epileptic encephalopathy type 17 other distinct genetic disorders have been identified that can present with a clinical picture indistinguishable from that of HD (Anderson *et al*. 2017; Gövert and Schneider 2013; Margolis *et al*. 2001; Mencacci *et al.* 2015; Mencacci and Carecchio 2017; Vijayvargia *et al*. 2016). Interestingly, some genes whose mutations cause HD-like syndromes code for htt-interacting proteins (Mencacci and Carecchio 2017), complicating the neuropathology of HD.

The analysis of the cryo-electron microscopy structure of htt evidenced that the putative interaction site of hunt3 with the G proteins is occupied by an α-helix connected to the opposite side of the domain through disordered loop regions and stabilized by electrostatic interactions. Moreover, this α-helix contains two phosphorylation sites, whose function is not known, and a signature sequence shared with several G proteins. Molecular docking simulations have shown that when the α-helix is removed from the structure, G proteins can bind to htt replacing the α-helix. To validate the hypothesis that the α-helix could act as a molecular switch for the binding to the G proteins when phosphorylation occurs, 0.5 μs molecular dynamics simulations of the hunt3 domain in different

phosphorylation states have been performed. The results obtained in this work suggest that the phosphorylation of Ser1351 increases the flexibility of the entire hunt3 domain with respect to the wild type. An opposite effect is observed on the double phosphorylated domain. Thus, at variance to what has been hypothesized, only the phosphorylation on Ser might act as a positive regulator of htt binding to G proteins, while double phosphorylation would revert this effect, pointing to a sort of phosphorylation code that could regulate the functions of hunt3 domain (Salazar and Höfer, 2009). As a cautionary note, it must be taken into account that these are preliminary data, because replicas of the simulations of all systems are currently still in progress. Moreover, molecular dynamics results have been obtained on an isolated domain of htt, and it cannot be excluded that hunt3 domain displays a different behaviour in the context of the entire protein. The hypothesis that the α-helix could act as a molecular switch upon phosphorylation is not fully confirmed by the simulations but it must be taken into account that this process could occur on a longer timescale with respect to the simulated time.

The outcome of this work is currently being validated experimentally by recombinant expression of hunt3 domain in *E. coli* and proteomic analysis of its interactome. The vector encoding full-length htt, pARIS-htt (**A**daptable, **R**NAi **I**nsensitive & **S**ynthetic), has been selected for the cloning because it contains unique restriction sites (URSs) dispersed throughout the entire sequence without modifying the translated amino acid sequence and includes multiple cloning sites at the N- and C-terminal ends (Pardo *et al.* 2010).

Another line of research of this work has been the study of htt sequences and structures of less complex model organisms. The htt sequences of *C. intestinalis* and *B. floridae* have been analysed, identifying ordered domains and performing structure prediction. All the models have shown structural similarity with the human ones. Finally, an htt-like protein, not previously identified, has been detected for the first time in *C. elegans*.

7.1 References

Anderson,D.G. *et al.* (2017) A systematic review of the Huntington Disease-Like 2 Phenotype. *J. Huntingtons. Dis.*, **6**, 37–46.

Benn,C.L. *et al.* (2008) Huntingtin Modulates Transcription, Occupies Gene Promoters In Vivo, and Binds Directly to DNA in a Polyglutamine-Dependent Manner. *J. Neurosci.*, **28**, 10720–10733.

Cardinale,A. *et al.* (2018) Localization of neuroglobin in the brain of R6/2 mouse model of Huntington's disease. *Neurol. Sci.*, **39**, 275–285.

Cattaneo,E. *et al.* (2005) Normal huntingtin function: An alternative approach to Huntington's disease. *Nat. Rev. Neurosci.*, **6**, 919–930.

De Marinis,E. *et al.* (2011) Neuroglobin, estrogens, and neuroprotection. *IUBMB Life*, **63**, 140–145.

De Marinis,E. *et al.* (2013) Neuroglobin upregulation induced by 17β-estradiol sequesters cytocrome c in the mitochondria preventing H2O2-induced apoptosis of neuroblastoma cells. *Cell Death Dis.*, **4**, e508.

del Toro,D. *et al.* (2009) Mutant Huntingtin Impairs Post-Golgi Trafficking to Lysosomes by Delocalizing Optineurin/Rab8 Complex from the Golgi Apparatus. *Mol. Biol. Cell*, **20**, 1478–1492.

Dragatsis,I. *et al.* (2000) Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. *Nat. Genet.*, **26**, 300–306.

Dragatsis,I. *et al.* (1998) Mouse mutant embryos lacking huntingtin are rescued from lethality by wild-type extraembryonic tissues. *Development*, **125**, 1529–1539.

Dunah,A.W. *et al.* (2002) Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. *Science*, **296**, 2238–2243.

Duyao,M. *et al.* (1995) Inactivation of the Mouse Huntington ' s Disease Gene Homolog Hdh. *Science,* **269**, 407–411.

El-Daher,M.-T. *et al.* (2015) Huntingtin proteolysis releases non-polyQ fragments that cause toxicity through dynamin 1 dysregulation. *EMBO J.*, **34**, 2255–2271.

Elias,S. *et al.* (2015) Huntingtin Is Required for Epithelial Polarity through RAB11A-Mediated Apical Trafficking of PAR3-aPKC. *PLoS Biol.*, **13**, 1–27.

Elias,S. *et al.* (2014) Huntingtin regulates mammary stem cell division and differentiation. *Stem Cell Reports*, **2**, 491–506.

Engelender,S. (1997) Huntingtin-associated protein 1 (HAP1) interacts with the p150Glued subunit of dynactin. *Hum. Mol. Genet.*, **6**, 2205–2212.

Futter,M. *et al.* (2009) Wild-type but not mutant huntingtin modulates the transcriptional activity of liver X receptors. *J. Med. Genet.*, **46**, 438–446.

Godin,J.D. *et al.* (2010) Huntingtin Is Required for Mitotic Spindle Orientation and Mammalian Neurogenesis. *Neuron*, **67**, 392–406.

Gövert,F. and Schneider,S.A. (2013) Huntington,s disease and Huntington,s disease-like syndromes: An overview. *Curr. Opin. Neurol.*, **26**, 420–427.

Guo,Q. *et al.* (2018) The cryo-electron microscopy structure of huntingtin. *Nature*, **555**, 117–120.

Gutekunst,C. a *et al.* (1995) Identification and localization of huntingtin in brain and human lymphoblastoid cell lines with anti-fusion protein antibodies. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 8710–4.

HD Collaborative research group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.

Holbert,S. *et al.* (2001) The Gln-Ala repeat transcriptional activator CA150 interacts with huntingtin: Neuropathologic and genetic evidence for a role in Huntington's disease pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 1811–1816.

Kaltenbach,L.S. *et al.* (2007) Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet.*, **3**, 689–708.

Kegel,K.B. *et al.* (2002) Huntingtin is present in the nucleus, interacts with the transcriptional corepressor C-terminal binding protein, and represses transcription. *J. Biol. Chem.*, **277**, 7466–7476.

Leavitt,B.R. *et al.* (2006) Wild-type huntingtin protects neurons from excitotoxicity. *J. Neurochem.*, **96**, 1121–1129.

Legleiter,J. *et al.* (2010) Mutant huntingtin fragments form oligomers in a polyglutamine length-dependent manner in Vitro and in Vivo. *J. Biol. Chem.*, **285**, 14777–14790.

Li,S.H. *et al.* (1998) Interaction of huntingtin-associated protein with dynactin P150Glued. *J. Neurosci.*, **18**, 1261–1269.

Margolis,R.L. *et al.* (2001) A disorder similar to Huntington's disease is associated with a novel CAG repeat expansion. *Ann. Neurol.*, **50**, 373–380.

McColgan,P. and Tabrizi,S.J. (2018) Huntington's disease: a clinical review. *Eur. J. Neurol.*, **25**, 24–34.

Mencacci,N.E. *et al.* (2015) ADCY5 mutations are another cause of benign hereditary chorea. *Neurology*, **85**, 80–88.

Mencacci,N.E. and Carecchio,M. (2017) Recent advances in genetics of chorea. *Curr. Opin. Neurol.*, **29**, 486–495.

Moreira Sousa,C. *et al.* (2013) The Huntington disease protein accelerates breast tumour development and metastasis through ErbB2/HER2 signalling. *EMBO Mol. Med.*, **5**, 309–325.

Nasir,J. *et al.* (1995) Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell*, **81**, 811–823.

Nuzzo,M.T. *et al.* (2016) 17β-Estradiol modulates huntingtin levels in rat tissues and in human neuroblastoma cell line. *Neurosci. Res.*, **103**, 59–63.

Nuzzo,M.T. *et al.* (2017) Huntingtin polyQ Mutation Impairs the 17β-Estradiol/Neuroglobin Pathway Devoted to Neuron Survival. *Mol. Neurobiol.*, **54**, 6634–6646.

O'Kusky,J.R. *et al.* (1999) Neuronal degeneration in the basal ganglia and loss of pallido-subthalamic synapses in mice with targeted disruption of the Huntington's disease gene. *Brain Res.*, **818**, 468–479.

Ochaba,J. *et al.* (2014) Potential function for the Huntingtin protein as a scaffold for selective autophagy. *Proc. Natl. Acad. Sci.*, **111**, 16889–16894.

Palidwor,G.A. *et al.* (2009) Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput. Biol.*, **5**.

Pardo,R. *et al.* (2010) PARIS-htt: An optimised expression platform to study huntingtin reveals functional domains required for vesicular trafficking. *Mol. Brain*, **3**, 1–17.

Poirier,M.A. *et al.* (2002) Huntingtin spheroids and protofibrils as precursors in polyglutamine fibrilization. *J. Biol. Chem.*, **277**, 41032–41037.

Rigamonti,D. *et al.* (2001) Huntingtin's Neuroprotective Activity Occurs via Inhibition of Procaspase-9 Processing. *J. Biol. Chem.*, **276**, 14545–14548.

Rigamonti,D. *et al.* (2000) Wild-type huntingtin protects from apoptosis upstream of caspase-3. *J. Neurosci.*, **20**, 3705–13.

Salazar,C. and Höfer,T. (2009) Multisite protein phosphorylation - From molecular mechanisms to kinetic models. *FEBS J.*, **276**, 3177–3198.

Saudou,F. and Humbert,S. (2016) The Biology of Huntingtin. *Neuron*, **89**, 910–926.

Velier,J. *et al.* (1998) Wild-type and mutant huntingtins function in vesicle trafficking in the secretory and endocytic pathways. *Exp. Neurol.*, **152**, 34–40.

Vijayvargia,R. *et al.* (2016) Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. *Elife*, **5**, 1–16.

White,J.K. *et al.* (1997) Huntington is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nat. Genet.*, **17**, 404–410.

Wong,Y.C. and Holzbaur,E.L.F. (2014) The Regulation of Autophagosome Dynamics by Huntingtin and HAP1 Is Disrupted by Expression of Mutant Huntingtin, Leading to Defective Cargo Degradation. *J. Neurosci.*, **34**, 1293–1305.

Yohrling IV,G.J. *et al.* (2003) Mutant huntingtin increases nuclear corepressor function and enhances

ligand-dependent nuclear hormone receptor activation. *Mol. Cell. Neurosci.*, **23**, 28–38.

Zeitlin,S. *et al.* (1995) Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. *Nat. Genet.*, **11**, 155–163.

Zhang,Y. *et al.* (2003) Start Cast Breakouts Preventative Prediction Using Multi-Way PCA Technology. *IFAC Proc. Vol.*, **36**, 101–106.

Zuccato,C. *et al.* (2003) Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat. Genet.*, **35**, 76–83.

Zuccato,C. *et al.* (2001) Loss of Huntingtin-Mediated BDNF Gene Transcription in Huntington ' s Disease. **293**, 493–498.

Zuccato,C. *et al.* (2010) Molecular Mechanisms and Potential Therapeutical Targets in Huntington ' s Disease. *Physiol Rev*, **90**, 905–981.

Zuccato,C. and Cattaneo,E. (2016) The Huntington's Paradox. *Sci. Am.*, **315**, 56–61.

Supplemetary material

Table S1. Hydrogen bonds between α-helix and the remaining portion of hunt3. The number of bonds is indicated between parentheses.

Htt3 wt (45)		Htt3 YP (42)		Htt3 SP (63)		Htt3 SYP (42)	
Ser1345	His1825	Ser1345	Gln1933	Ser1345	Gln1763	Ser1345	Gln1763
Arg1347	His1928	Ser1345	His1928	Ser1345	Gln1770	Ser1345	His1825
Arg1347	Glu1929	Ser1345	Asp1934	Ser1345	Lys1881	Arg1347	His1825
Pro1348	His1928	Arg1347	Asp1900	Val1346	Gly1883	Arg1347	Met662
Gly1349	His1928	Arg1347	Asp1934	Arg1347	Gly1883	Arg1347	Gly1883
Gly1349	Arg1892	Gly1349	Glu1929	Arg1347	His1928	Arg1347	His1928
Leu1350	Glu1929	Leu1350	Glu1929	Arg1347	Glu1929	Arg1347	Asn1886
Tyr1351	Glu1929	Leu1350	Gln1834	Arg1347	Ala1880	Pro1348	His1928
Tyr1351	Gly1893	Tyr1351	Gln1834	Arg1347	Lys1881	Gly1349	His1928
Tyr1351	Asp1900	Tyr1351	Gln1904	Arg1347	His1928	Gly1349	Glu1929
Tyr1351	Asp1900	Tyr1351	Glu1929	Arg1347	Glu1929	Leu1350	Glu1929
His1352	Glu1929	Tyr1353	His1825	Arg1347	Gln1933	Tyr1351	Glu1929
Tyr1353	His1825	Tyr1353	Ala1879	Arg1347	Leu1882	Tyr1351	Lys1784
Tyr1353	Pro1826	Tyr1353	Ala1880	Pro1348	His1928	Tyr1351	Gln1834
Tyr1353	Ile1889	Tyr1353	Glu1874	Gly1349	Ala1879	Tyr1351	Gln1904
Cys1354	Gln1770	Tyr1353	Asp1875	Gly1349	Ala1880	Tyr1351	Glu1929
Tyr1359	Ile1780	Tyr1353	Asp1877	Gly1349	Lys1881	Tyr1353	His1825
Tyr1359	His1781	Tyr1359	Lys1784	Gly1349	Glu1929	Tyr1353	Ile1889
Tyr1359	Lys1784	Tyr1359	Ser1785	Gly1349	Ala1879	Tyr1353	Ala1827
Tyr1359	Gln1834	Tyr1359	Met1777	Gly1349	Ala1880	Tyr1353	Arg1892
His1361	Gln1770	Tyr1359	Ile1780	Gly1349	Lys1881	Cys1354	Gln1834
His1361	Thr1774	Tyr1359	His1781	Leu1350	Ala1879	Tyr1359	Lys1784
His1361	Arg1677	Tyr1359	Lys1784	Leu1350	Glu1929	Tyr1359	Ser1785
Phe1362	Arg1677	His1361	Gln1770	Tyr1351	Glu1929	Tyr1359	His1781
Thr1363	Lys1784	Phe1362	Arg1677	Tyr1351	Asp1875	Tyr1359	Lys1784
Gln1364	Lys1784	Thr1363	His1781	Tyr1351	Ala1879	Tyr1359	Ser1785
Ala1365	Arg1677	Ala1365	Arg1677	Tyr1351	Glu1888	His1361	Gln1770
Ala1365	Ser1681	Ala1365	Ser1681	Tyr1351	Ile1889	Ala1365	Arg1677
Leu1366	Arg1677	Leu1366	Arg1677	Tyr1351	Arg1892	Leu1366	Arg1677
Leu1366	Ser1681	Leu1366	Ser1681	Tyr1351	Gly1893	Leu1366	Ser1681
Leu1366	Gln1682	Ala1367	Ser1681	Tyr1351	Asp1900	Leu1366	Gln1682
Ala1367	Ser1681	Ala1369	Arg1677	Tyr1351	His1928	Ala1367	Ser1681
Ala1367	Gln1682	Ala1369	Ser1681	His1352	Glu1929	Ala1369	Ser1681
Ala1369	Arg1677	Ala1369	Gln1682	His1352	Gln1834	Ala1369	Gln1682
Ser1370	Arg1643	Ser1370	Arg1643	His1352	Gly1893	Ser1370	Arg1643
Ser1370	Arg1677	Ser1370	Ser1681	His1352	Gln1834	Ser1370	Ser1681
Ser1370	Ser1681	Ser1370	Gln1682	His1352	Gly1893	Ser1370	Gln1682
Ser1370	Gln1682	Ser1370	Glu1635	Tyr1353	Gln1770	Arg1372	Gln1589
Ser1370	Glu1635	Arg1372	Arg1677	Tyr1353	His1825	Arg1372	His1592
Ser1370	Leu1637	Arg1372	Glu1635	Phe1355	Gln1834	Arg1372	Thr1632
Ser1370	Ser1681	Arg1372	Thr1632	Tyr1359	Lys1784	Arg1372	Arg1677
Arg1372	Arg1677	Arg1372	Glu1635	Tyr1359	Gln1834	Arg1372	Glu1635
Arg1372	Glu1635			Tyr1359	Met1777		
Arg1372	Asn1631			Tyr1359	Ile1780		
Arg1372	Thr1632			Tyr1359	His1781		
				Tyr1359	Lys1784		
				Tyr1359	Ser1785		
				Tyr1359	Gln1834		
				Phe1362	Arg1677		

				Thr1363	His1781		
				Ala1365	Arg1677		
				Leu1366	Arg1677		
				Leu1366	Ser1681		
				Leu1366	His1781		
				Ala1367	Ser1681		
				Ala1369	Ser1681		
				Ala1369	Gln1682		
				Ser1370	Arg1643		
				Ser1370	Ser1681		
				Ser1370	Gln1682		
				Arg1372	Thr1632		
				Arg1372	Arg1677		

Table S2. Salt bridges between the α-helix and the remaining portion of hunt3, formed during the simulations.

Htt3 wt		Htt3 YP		Htt3 SP		Htt3 SYP	
Arg1347	Glu1929	Arg1347	Asp1900	Arg1347	Glu1929	Arg1372	Glu1635
Arg1372	Glu1635	Arg1347	Asp1934	Arg1372	Glu1635		
		Arg1372	Glu1635				
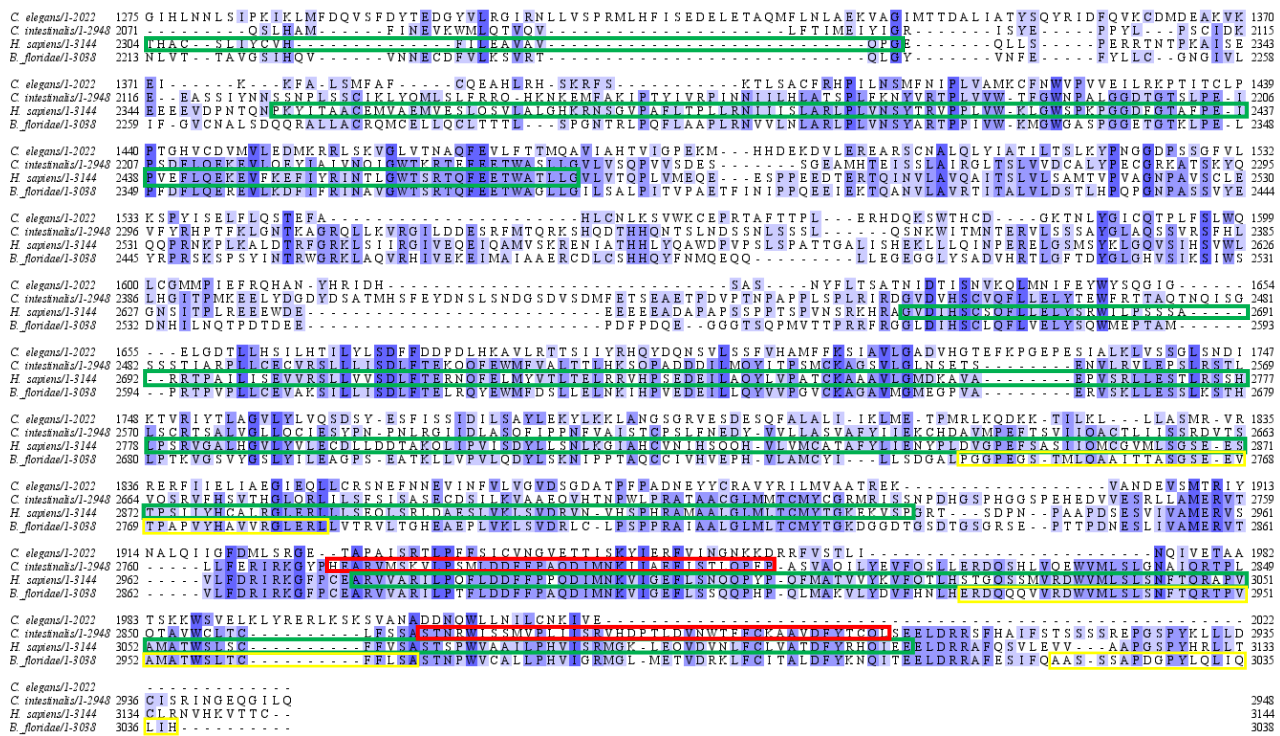
Figure S1. Multiple sequence alignment between *H. sapiens*, *C. intestinalis*, *B. floridae* htt and *C. elegans* htt-like. Residues are coloured by percentage identity. The red, green and yellow boxes indicate HEAT repeats in ascidian, in human and amphioxus htt, respectively.

Table S3. Query coverage, *E*-value and percentage identity between human, ascidian, amphioxus htt sequences and *C. elegans* htt-like sequence, obtained through distinct BLAST searches.

Blast hit	Query coverage	*E*-value	Identity	Accession number
Ciona intestinalis htt	88%	0.0	34%	CAJ87484.4
Branchiostoma floridae htt	99%	0.0	46%	CAM12495.1
Caenorhabditis elegans	13%	3×10^{-6}	21%	NP509663.3

Table S4. Disordered regions identified in *C. intestinalis* htt sequence.

Method	Disordered regions
Foldindex	102-139, 370-445, 726-776, 833-863, 981-1056, 1182-1231, 1510-1523, 1546-1550, 1867-1871, 2121-2130, 2133-2138, 2209-2213, 2298-2356, 2383-2387, 2389-2436
Predictprotein	18-29, 336-356, 373-449, 745-758, 835-865, 974-1036, 1204-1222, 1853-1863, 2408-2438, 2729-2743
Anchor	357-378, 411-420, 434-443, 452-461, 470-481, 951-969, 1040-1051, 2366-2387
SMART	303-317, 355-363, 431-440, 693-705, 834-854, 896-908, 1760-1784, 1946-1957, 2341-2355, 2562-2575
HMMER	339-354, 375-444, 835-840, 842-861, 981-1027, 2328-2352, 2415-2442
MOBIDB	380-447, 836-866, 876-1037, 2329-2350, 2420-2444
GLOBPLOT	404-456, 742-751, 863-897, 999-1061, 1165-1185, 1215-1229, 1880-1891, 1998-2011, 2215-2231, 2362-2377, 2418-2474, 2751-2769, 2944-2954; Globular domains: 2-403, 442-862, 898-998, 1062-1164, 1230-1997, 2012-2214, 2232-2361, 2475-2750, 2770-2950
SPOT-DISORDER	1-30, 330-452, 746-760, 836-867, 923-935, 976-1036, 1142-1157, 1209-1222, 1853-1864, 2391-2440, 2726-2745
INTERPRO	23-324, 471-988, 2334-2915
PONDR VLXT	25-43, 50-60, 112-132, 154-164, 305-313, 337-363, 369-463, 567-576, 719-759, 838-863, 968-1024, 1142-1157, 1206-122, 1338-1351, 1358-1377, 1405-1415, 1524-1539, 1567-1579, 1706-1725, 1802-1820, 1851-1864, 1972-1981, 2243-2266, 2310-2324, 2340-2352, 2414-2446, 2547-2562, 2727-2753

Table S5. C-Score of the models of ascidian htt ordered domains, obtained with I-TASSER. The models have been selected by Confidence-score, a parameter to evaluate the quality of predicted models. Generally, C-score can range from -5 to 2, where a higher value corresponds to a greater level of confidence of the model.

Ordered domains of ascidian htt	C-score
Domain 1	-0,92
Domain 2	-0,28
Domain 3	0,59
Domain 4	0,33

Table S6. Structural homologs of domain 1 of ascidian htt, identified by I-TASSER. TM-score and RMSD are known standards for measuring structural similarity between two structures which are usually used to measure the accuracy of structure modelling when the native structure is known.

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.857	0.86	0.397	0.870
2	5YZ0	0.734	4.18	0.071	0.948
3	4A0C	0.700	3.43	0.101	0.849
4	5VCH	0.697	4.19	0.108	0.910
5	5IFE	0.694	4.13	0.079	0.901
6	1VSY	0.691	4.39	0.091	0.914
7	2IAE	0.687	4.26	0.088	0.910
8	3W3T	0.686	4.19	0.103	0.901
9	2JAK	0.685	4.42	0.077	0.910
10	5DLQ	0.682	4.81	0.112	0.957

Table S7. Structural homologs of domain 2 of ascidian htt, identified by I-TASSER.

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.828	1.13	0.288	0.840
2	5DLQ	0.525	6.19	0.085	0.744
3	4UVK	0.518	6.04	0.103	0.717
4	1VSY	0.515	5.72	0.079	0.696
5	2X1G	0.508	5.54	0.076	0.686
6	3ICQ	0.503	6.07	0.083	0.702
7	1IBR	0.503	4.89	0.084	0.638
8	1U6G	0.499	5.75	0.102	0.683
9	2XWU	0.499	6.40	0.067	0.721
10	4KNH	0.498	6.28	0.083	0.719

Table S8. Structural homologs of domain 3 of ascidian htt, identified by I-TASSER.

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.866	0.81	0.354	0.870
2	5T8V	0.389	8.26	0.060	0.541
3	5HB4	0.372	8.11	0.044	0.516
4	6BQ1	0.365	8.96	0.037	0.531
5	2X1G	0.362	6.33	0.063	0.449
6	2XWU	0.357	7.15	0.057	0.463
7	1VSY	0.356	7.60	0.065	0.478
8	4C0O	0.354	6.86	0.061	0.455
9	3ICQ	0.352	7.15	0.052	0.460
10	1WA5	0.351	6.55	0.044	0.441

Table S9. Structural homologs of domain 4 of ascidian htt, identified by I-TASSER.

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.870	1.07	0.458	0.882
2	3ZKV	0.651	5.23	0.061	0.863
3	2XWU	0.638	5.44	0.059	0.861
4	4C0O	0.632	5.10	0.088	0.846
5	1W63	0.630	4.40	0.075	0.784
6	3DW8	0.628	4.67	0.084	0.803
7	3NBZ	0.627	5.48	0.078	0.874
8	3ICQ	0.619	5.30	0.071	0.833
9	1U6G	0.618	4.68	0.078	0.791
10	5DLQ	0.618	5.72	0.071	0.863

Table S10. Disordered regions identified in *B. floridae* htt sequence

Method	Disordered regions
Foldindex	27-39, 430-452, 517-532, 545-575, 577-592, 868-908, 1025-1059, 1108-1162, 1260-1285, 1777-1798, 2083-2127, 2441-2472, 2515-2569
Predictprotein	19-50, 416-439, 447-479, 490-532, 534-592, 1030-1051, 1088-1147, 1254-1269, 1780-1799, 2537-2560, 2831-2848
Anchor	7-16, 474-496, 514-539, 1058-1088, 1097-1107
SMART	217-230, 268-278, 476-488, 984-1024, 1039-1053, 1096-1108, 1123-1141, 1365-1381, 1686-1697, 1747-1758, 1774-1794, 1991-2002, 2494-2505, 2538-2552, 2827-2843
HMMER	27-43, 426-450, 493-520, 883-892, 1031-1056, 1095-1146, 1782-1791, 2091-2103, 2537-2560, 2831-2849
MOBIDB	422-441, 1031-1051, 1091-1150, 1777-1798, 2536-2561
GLOBPLOT	373-385, 402-416, 450-474, 505-514, 528-542, 714-722, 927-941, 1008-1021, 1060-1080, 1118-1182, 1284-1301, 1606-1618, 1806-1827, 2020-2034, 2116-2135, 2357-2375, 2459-2483, 2565-2595, 2771-2781, 2854-2878 Globular domain: 1-401, 543-926, 1183-1805, 1828-2019, 2035-2115, 2136-2356, 2376-2458, 2477-2564, 2596-2853, 2879-3038
SPOT-DISORDER	1-50, 370-594, 1033-1050, 1092-1148, 1254-1268, 1298-1308, 1778-1805, 1990-2006, 2538-2564, 2831-2848
PONDR	21-56, 128-139, 349-357, 419-482, 491-516, 527-600, 880-935, 981-995, 1024-1034, 1089-1147, 1256-1271, 1322-1331, 1479-1498, 1520-1533, 1607-1619, 1642-1662, 1773-1809, 1944-1963, 1990-2008, 2071-2120, 2320-2334, 2398-2417, 2428-2464, 2537-2557, 2658-2670, 2742-2776, 2788-2805, 2832-2855

Table S11. C-Score of the models of AmphiHtt ordered domains obtained with I-TASSER

Ordered domains of AmphiHtt	C-score
Domain 1	0.04
Domain 2	-0.67
Domain 3	0.50
Domain 4	0.49

Table S12. Structural homologs of domain 1 model of AmpiHtt, identified by I-TASSER

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.897	1.32	0.589	0.916
2	4A0C	0.679	3.46	0.091	0.826
3	2BKU	0.677	3.86	0.102	0.845
4	1VSY	0.661	4.46	0.089	0.867
5	4XRI	0.661	3.82	0.068	0.829
6	3FGA	0.656	4.26	0.110	0.863
7	5IFE	0.649	4.27	0.087	0.845
8	3S4Z	0.642	4.34	0.089	0.867
9	5DLQ	0.638	4.76	0.090	0.891

Table S13. Structural homologs of domain 2 model of AmpiHtt, identified by I-TASSER

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.890	2.74	0.366	0.956
2	4KF7	0.579	5.10	0.071	0.774
3	4KNH	0.542	5.54	0.080	0.755
4	1IBR	0.540	5.52	0.101	0.744
5	3NBY	0.538	5.16	0.071	0.737
6	4C0O	0.538	5.26	0.107	0.730
7	5VCH	0.538	5.54	0.077	0.741
8	3ND2	0.538	5.63	0.092	0.741
9	4XRI	0.537	4.82	0.059	0.693
10	1VSY	0.535	6.08	0.062	0.797

Table S14. Structural homologs of domain 3 model of AmpiHtt, identified by I-TASSER

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.824	0.89	0.562	0.829
2	3ICQ	0.495	6.38	0.070	0.642
3	2XWU	0.482	7.04	0.077	0.660
4	6EMK	0.474	7.17	0.054	0.648
5	4C0O	0.472	6.74	0.047	0.633
6	3EA5	0.470	6.76	0.067	0.630
7	5HB4	0.469	7.75	0.071	0.673
8	6BCU	0.461	6.55	0.088	0.607
9	4A0C	0.461	6.45	0.065	0.604
10	5DLQ	0.461	6.65	0.052	0.610

Table S15. Structural homologs of domain 4 model of AmpiHtt identified by I-TASSER

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.839	0.61	0.528	0.842
2	5HB4	0.503	6.50	0.061	0.663
3	4KF7	0.502	6.22	0.085	0.644
4	5IJO	0.488	6.20	0.068	0.628
5	5DLQ	0.480	7.53	0.080	0.685
6	5YZ0	0.479	6.59	0.055	0.632
7	5GM6	0.477	6.34	0.065	0.616
8	3M1I	0.473	6.60	0.076	0.625
9	3NBY	0.473	6.51	0.062	0.620
10	3W3T	0.472	5.89	0.075	0.586

Table S16. Disordered regions identified in the htt-like sequence of *C. elegans*

Method	Disordered regions
Foldindex	299-303, 317-347, 561-566, 594-598, 728-741, 835-864, 871-880, 1142-1199, 1224-1228, 1231-1240, 1971-1989
Predictprotein	1, 129, 314-325, 340-348, 1161-1164, 1488-1489, 1494
SMART	801-812, 1027-1041, 1232-1243
MOBIDB	1-5, 646-651, 653-654, 731, 1168-1169, 1495-1496, 1793-1794
ESPRITZ X-RAY	1-23, 343-345, 393-396, 1152-1171, 1186-1190, 1522-1525, 1789-1795, 2001-2007, 2017-2020
GLOBPLOT	630-644, 729-733, 1482-1487, 1564-1578, 1621-1632, 1769-1774, 1917-1926 DOMAINS: 3-629, 645-1563, 1579-2068
SPOT-DISORDER	1-2, 1159-1170, 1172-1173, 1611-1625, 1627-1628
INTERPRO	Htt family domain: 1283-1999
PONDR VLXT	204-219, 634-651, 1155-1178, 1318-1333, 1487-1502, 1815-1833, 1989-2002

Table S17. C-Score of the models of the ordered domains of an htt-like protein identified in *C.elegans*

Ordered domains of an Htt-like protein	C-score
Domain 1	-1.26
Domain 2	0.04

Table S18. Structural homologs of domain 1 of the htt-like protein in *C. elegans*, identified by I-TASSER

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.845	1.88	0.145	0.862
2	1Z3H	0.504	6.30	0.082	0.633
3	3W3T	0.497	6.86	0.075	0.637
4	3ND2	0.496	6.79	0.071	0.644
5	1QGK	0.484	6.71	0.054	0.626
6	6EMK	0.481	6.91	0.063	0.631
7	5YZ0	0.463	6.61	0.048	0.595
8	4XRI	0.459	7.32	0.068	0.621
9	4C0O	0.451	6.85	0.079	0.589
10	5VCH	0.438	7.17	0.070	0.582

Table S19. Structural homologs of domain 2 of the htt-like protein in *C. elegans*, identified by I-TASSER

Rank	PDB hit	TM-score	RMSD (Å)	Identity (%)	Coverage
1	6EZ8	0.874	2.09	0.174	0.903
2	4KF7	0.576	6.32	0.061	0.767
3	5IJO	0.550	6.34	0.076	0.735
4	5HB4	0.543	6.27	0.070	0.725
5	5DLQ	0.534	7.40	0.061	0.772
6	3A6P	0.532	6.91	0.065	0.737
7	3GJX	0.530	6.76	0.065	0.733
8	5YZ0	0.523	6.12	0.051	0.692
9	3M1I	0.519	6.80	0.067	0.719
10	4KNH	0.503	6.86	0.062	0.711

List of publications

1.   Pennisi R., Salvi D., **Brandi V**., Angelini, R, Ascenzi, P., Polticelli, F. (2016) Molecular evolution of alternative oxidase proteins: A phylogenetic and structure modeling approach. *Journal of Molecular Evolution* **82**(4-5), pp. 207-218

2.   **Brandi V**, Di Lella V, Marino M, Ascenzi P, Polticelli F. (2017) A comprehensive in silico analysis of huntingtin and its interactome. *J Biomol Struct Dyn*. [Epub ahead of print]

3.   Fiocchetti M., Cipolletti M., **Brandi V**., Polticelli F., Ascenzi P (2017) Neuroglobin and friends. *Journal of Molecular Recognition*, **30**(12)

4.   Caprari, S., Minervini, G., **Brandi V**. Polticelli, F. (2017) In silico study of the structure and function of Streptococcus mutans plasmidic proteins. *Bio-Algorithms and Med-Systems* **13**(2):51-61

5.   Cruciani M, Etna MP, Camilli R, Giacomini E, Percario ZA, Severa M, Sandini S, Rizzo F, **Brandi V**, Balsamo G, Polticelli F, Affabris E, Pantosti A, Bagnoli F, Coccia EM. (2017) Staphylococcus aureus Esx Factors Control Human Dendritic Cell Functions Conditioning Th1/Th17 Response. *Frontiers in Cellular and Infection Microbiology,* **7**:330

6.   Toti D., Viet Hung L., Tortosa V., **Brandi V**., Polticelli, F. (2018) LIBRA-WA: A web application for ligand binding site detection and protein function recognition. *Bioinformatics* **34**(5), pp. 878-880

7.   Tortosa V., Bonaccorsi Di Patti M.C., **Brandi V**., Musci G., Polticelli F. (2017) An improved structural model of the human iron exporter ferroportin. Insight into the role of pathogenic mutations in hereditary hemochromatosis type 4, *Bio-Algorithms and Med-Systems* **13**(4), pp. 215-222

CrossMark

ORIGINAL ARTICLE

Molecular Evolution of Alternative Oxidase Proteins: A Phylogenetic and Structure Modeling Approach

Rosa Pennisi[1] · Daniele Salvi[2] · Valentina Brandi[1] · Riccardo Angelini[1] ·
Paolo Ascenzi[1] · Fabio Polticelli[1,3]

Abstract Alternative oxidases (AOXs) are mitochondrial cyanide-resistant membrane-bound metallo-proteins catalyzing the oxidation of ubiquinol and the reduction of oxygen to water bypassing two sites of proton pumping, thus dissipating a major part of redox energy into heat. Here, the structure of *Arabidopsis thaliana* AOX 1A has been modeled using the crystal structure of *Trypanosoma brucei* AOX as a template. Analysis of this model and multiple sequence alignment of members of the AOX family from all kingdoms of Life indicate that AOXs display a high degree of conservation of the catalytic core, which is formed by a four-α-helix bundle, hosting the di-iron catalytic site, and is flanked by two additional α-helices anchoring the protein to the membrane. Plant AOXs display a peculiar covalent dimerization mode due to the conservation in the *N*-terminal region of a Cys residue forming the inter-monomer disulfide bond. The multiple sequence alignment has also been used to infer a phylogenetic tree of AOXs whose analysis shows a polyphyletic origin for the AOXs found in Fungi and a monophyletic origin of the AOXs of Eubacteria, Mycetozoa, Euglenozoa, Metazoa, and Land Plants. This suggests that AOXs evolved from a common ancestral protein in each of these kingdoms. Within the Plant AOX clade, the AOXs of monocotyledon plants form two distinct clades which have unresolved relationships relative to the monophyletic clade of the AOXs of dicotyledonous plants. This reflects the sequence divergence of the *N*-terminal region, probably due to a low selective pressure for sequence conservation linked to the covalent homo-dimerization mode.

Abbreviations

Δ9	Desaturase
Stearoyl	Acyl carrier desaturase
AOX	Alternative oxidase
ML	Maximum Likelihood
MMO	Methane–monooxygenase
RNR	R2 subunit from ribonucleotide reductase
ROS	Reactive oxygen species
TAO	Trypanosomal alternative oxidase
AtAOX	*Arabidopsis thaliana* AOX 1A

✉ Fabio Polticelli
fabio.polticelli@uniroma3.it

[1] Department of Sciences, Roma Tre University,
Viale Guglielmo Marconi 446, 00146 Rome, Italy

[2] CIBIO-InBIO, Centro de Investigação em Biodiversidade e
Recursos Genéticos, Universidade do Porto, Campus Agrário
de Vairão, 4485–661 Vairão, Portugal

[3] National Institute of Nuclear Physics, Roma Tre Section,
00146 Rome, Italy

Introduction

Alternative oxidases (AOXs) are mitochondrial cyanide—insensitive membrane-bound proteins involved in redox reactions. These metallo-proteins, recalling the structural organization of members of the "ferritin" family, display a

🍨 Springer

four-α-helix bundle, a common structural element for binuclear metal centers. In general, the two iron atoms are ligated by four conserved Glu or Asp residues and two His side chains, which form two universally conserved Glu–X–X–His motifs (Siedow et al. 1995; Berthold et al. 2000; McDonald 2008; Albury et al. 2009; Moore et al. 2013; Neimanis et al. 2013).

AOX branches from the cytochrome pathway and catalyzes the oxidation of ubiquinol and the four-electron reduction of oxygen to water. In contrast to cytochrome *c* oxidase, the main mitochondrial terminal oxidase, AOX does not pump protons. Bypassing two sites of proton pumping, at complexes III and IV, AOX activity dissipates a major part of the redox energy into heat and thus results in a lower synthesis of ATP (Vanlerberghe and McIntosh 1997; Berthold et al. 2000; Millenaar and Lambers 2003; McDonald and Vanlerberghe 2004; Rasmusson et al. 2008; McDonald et al. 2009; Moore et al. 2013; Vishwakarma et al. 2014).

Although AOX is an ubiquitous enzyme within the Plant kingdom, it is also found in many other Eukaryote kingdoms, including fungi, protists, and few animal species, as well as Bacteria but not Archaea (Chaudhuri and Hill 1996; Van Hellemond et al. 1998; Ajayi et al. 2002; Stenmark and Nordlund 2003; Veiga et al. 2003; Roberts et al. 2004; Stechmann et al. 2008; Albury et al. 2009; McDonald et al. 2009). AOX is encoded by a small gene family, consisting of two distinct subfamilies termed *AOX1* and *AOX2*. In the model plant *Arabidopsis thaliana*, AOX is encoded by five genes, namely *AOX1a*, *AOX1b*, *AOX1c*, *AOX1d*, and *AOX2* (Supplementary Fig. 1) (Clifton et al. 2006). The expression of AOX is primarily induced when the mitochondrial respiration is impaired. Moreover, the expression of *AOX* gene(s) is tissue-specific and evolutionarily regulated (Thirkettle-Watts et al. 2003). AOX seems to play a role in controlling the levels of reactive oxygen species (ROS) generated by the respiratory chain (Purvis and Shewfelt 1993). In fact, AOX could act to reduce ROS generation and could prevent the over-reduction of the ubiquinone pool, not only in plants but also in fungi and lower invertebrates (Maxwell et al. 1999; Parsons et al. 1999; McDonald et al. 2009; Van Aken et al. 2009; Li et al. 2011). Then, the functions of AOX include thermogenesis, stress tolerance, and the maintenance of mitochondrial and cellular homeostasis (Finnegan et al. 2004; Vanlerberghe 2013).

Two plant AOX molecular models have been built (Umbach and Siedow 1993; Moore et al. 1995; Siedow et al. 1995; Andersson and Nordlund 1999; Berthold et al. 2000) and the crystal structure of the *Trypanosoma brucei* AOX (TAO) has been solved (Shiba et al. 2013). However, the plant AOX molecular models are conflicting.

In the first model published in 1995, AOX was postulated to show a covalent homo-dimeric assembly through the occurrence of a disulfide bridge, each monomer displaying a hydroxo-bridged di-iron center residing within a four-α-helix bundle. The two di-iron binding Glu–X–X–His motifs were predicted to be located on the first and fourth α-helix of the bundle. Moreover, it was proposed that AOX is anchored to the mitochondrial membrane by two transmembrane α-helices (Umbach and Siedow 1993; Moore et al. 1995; Siedow et al. 1995). The second model published in 1999 predicted a different length of the α-helices and postulated that the Glu–X–X–His motifs involved in di-iron binding were located on the second and the fourth α-helix of the bundle. A putative membrane-binding motif formed by two short hydrophobic helices, which would insert in a single leaflet of the lipid bilayer, was also proposed (Andersson and Nordlund 1999; Berthold et al. 2000).

Very recently, the crystal structure of the *Trypanosoma brucei* alternative oxidase (TAO) has been solved (Shiba et al. 2013). The catalytic center is formed by two His residues, each coordinating one iron atom on the same side of the bi-metal cluster, two Glu residues connecting the two ferric ions, and two additional Glu residues each coordinating one iron ion. In addition to the di-iron active site, TAO contains a hydrophobic cavity for ubiquinol binding. TAO is a non-covalent homo-dimer, the *N*-terminal region of each monomer being pivotal for assembly in a domain-swapping fashion (Shiba et al. 2013).

The availability of the TAO crystal structure provides for the first time the opportunity to analyze the structural details of AOXs from different organisms through homology modeling. In addition, structural phylogenomic analyses can provide important insights into AOX function and molecular evolution. Here, the structural model of *Arabidopsis thaliana* AOX 1A (AtAOX), built by a combination of ab initio/threading strategy is reported. Furthermore, in order to put the available structural information in an evolutionary context, 111 protein sequences homologous to *A. thaliana* AOX 1A were collected and assembled in a multiple amino acid sequence alignment. This allowed to map the conservation of residues essential for AOXs structure and function along the evolutionary ladder and to infer a phylogenetic framework for the structural evolution of AOXs.

Results and Discussion

The Three-Dimensional Model of AOX 1A

Plant AOXs contain a mitochondrial targeting pre-sequence (residues 1–62), which is cleaved after translocation of the protein on the mitochondrial membrane (Tanudji et al. 1999). We thus removed the first 62 residues

of the AtAOX amino acid sequence and used the mature sequence to build the structural model shown in Fig. 1.

Analysis of the model shows that the catalytic core of AtAOX is formed by a four-α-helix bundle. As expected, likewise TAO, the AtAOX active site is located in a hydrophobic cavity and contains four strictly conserved Glu residues at positions 183, 222, 273, and 324, which coordinate the two iron atoms. The stability of the di-iron center is also due to the presence of two strictly conserved His residues at positions 225 and 327 (Fig. 2). All these conserved residues form the two Glu–X–X–His motifs (*i.e.*, Glu222–X–X–His225 and Glu324–X–X–His327), located on the second and the fourth helix of the four α-helix
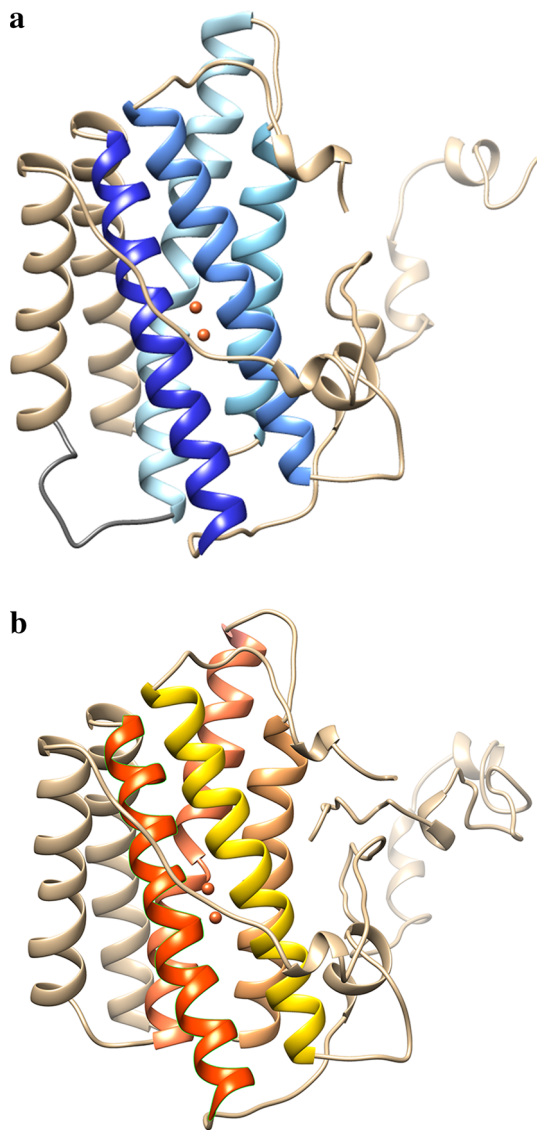
Fig. 1 Schematic representation of **a** the three-dimensional structure of the TAO dimer (PDB code 3VV9; [36]) and **b** of the molecular model of the *A. thaliana* AOX 1A dimer. Iron atoms are represented by *orange spheres* (Color figure online)

bundle, as previously predicted (Andersson and Nordlund 1999). The two His residues are too far away from the di-iron center to coordinate the two metal ions, but they form hydrogen bonds with Glu183, Glu222, Glu273, and with Asn221, another conserved residue within the catalytic center (Moore et al. 2013; Shiba et al. 2013). Like in TAO, Asn221 forms a hydrogen bond network with conserved residues Tyr304 and Asp323 (Shiba et al. 2013). In TAO, a second putative hydrophobic cavity connects the di-iron active site with a single leaflet of the lipid bilayer facing the mitochondrial matrix (Shiba et al. 2013). This groove is present also in the AtAOX model, being formed by the highly conserved residues Arg164, Asp168, Arg178, Ala186, Leu182, Leu272, Glu183, Glu275, and Ala276. Moreover, three Tyr residues, pivotal in TAO ubiquinol binding and catalysis (Shiba et al. 2013), are also conserved in AtAOX as Tyr258, Tyr280, and Tyr304. As in TAO, a fourth tyrosine residue, Tyr271, is present in AtAOX. However, this Tyr residue is not strictly conserved across all AOXs sequences (see below; Shiba et al. 2013).

Plant AOXs have been reported to be covalent homo-dimers (Umbach and Siedow 1993; Umbach et al. 1994); accordingly, the *N*-terminal arm of AtAOX appears to be involved in a disulfide bond-mediated homo-dimeric assembly. Indeed, the modeled *N*-terminal region of AtAOX contains highly conserved amino acid residues (Met191, Met195, His198, Ser201, Leu202, Arg203, Leu216, Arg223, Leu226, Arg240, and Gln247), which in
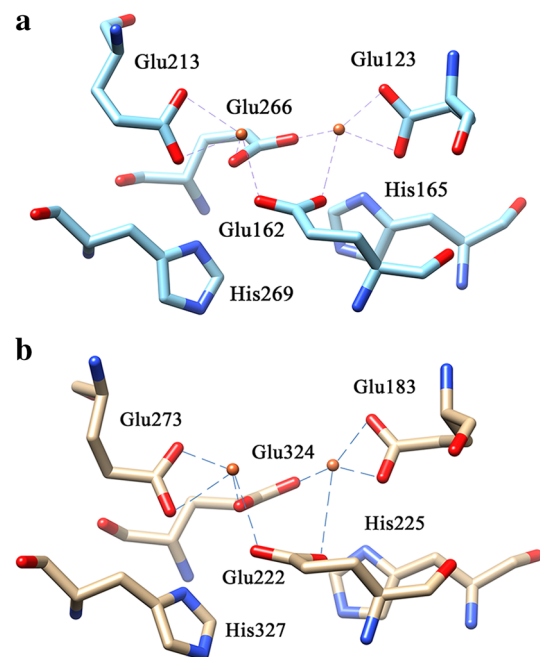
Fig. 2 Structural details of the AOXs catalytic center. **a** TAO di-iron center. **b** *A. thaliana* AOX 1A di-iron center. Iron atoms are represented by *orange spheres* (Color figure online)

TAO mediate the monomer–monomer contacts (Shiba et al. 2013). Interestingly, in the AtAOX dimer model Cys127 of each monomer is perfectly positioned to form a disulfide bond, despite the non-covalent assembly of the template TAO structure (see below). This residue is conserved in all plant AOXs and together with a second conserved cysteine, Cys177 in AtAOX, it is well known to play a role in the post-translational regulation of AOX activity (Polidoros et al. 2009; Moore et al. 2013; Neimanis et al. 2013).

The N-terminal 31 residues of AtAOX (residues 63–93) are missing from the structural model as the corresponding region of TAO is unstructured and not resolved in the crystal structure of the latter protein. Therefore, to get insight into the structure and function of this region, structure prediction was carried out using two alternative ab initio protein structure prediction programs (Rosetta ab initio and QUARK; Kim et al. 2004; Xu and Zhang 2012). In both cases, the N-terminal 31 residues of AtAOX are predicted to form a single α-helix displaying an amphiionic character with four lysine residues lining one side of the helix (Supplementary Fig. 2). Therefore, this helix could contribute to stabilize the interaction of AtAOX with the mitochondrial membrane through interaction with the negatively charged phosphate groups of phospholipids.

In order to put the structural information derived from the TAO three-dimensional structure and the AtAOX molecular model in an evolutionary context, protein sequences homologous to *A. thaliana* AOX 1A were collected by exhaustive BLASTP searches and assembled in a multiple amino acid sequence alignment (see below). The retrieved AOXs protein sequences dataset contained 111 homologous protein sequences from 37 Land Plants, three Rhodophyta, 27 Fungi, three Metazoa, four Euglenozoa, five Stramenopiles, one Alveolate, two Mycetozoa, one Heterolobosea, and 28 Eubacteria (see Supplementary Materials Table 1 for details). The AtAOX molecular model was then used to map the conservation of residues essential for AOXs structure and function (Moore et al. 2013) along the evolutionary ladder (Fig. 3). Panel A of Fig. 3 shows the location of the three strictly conserved Tyr residues (Tyr258, Tyr280, Tyr304), plus that of a fourth, almost strictly conserved tyrosine residue (Tyr271). Given the proximity to the iron cluster, Tyr280 (Tyr220 in TAO) is the most likely candidate for the amino acid radical involved in catalysis, while Tyr304 (Tyr246 in TAO) supports the hydrogen bonding network which stabilizes the iron cluster ligands (Moore et al. 2013). The strict conservation of Tyr258 (Tyr198 in TAO) is less straightforward to explain. However, both in TAO three-dimensional structure and in AtAOX structural model, this Tyr residue forms a hydrogen bond with a His residue (His206 in TAO, His266 in AtAOX) which links helices α5 and α6
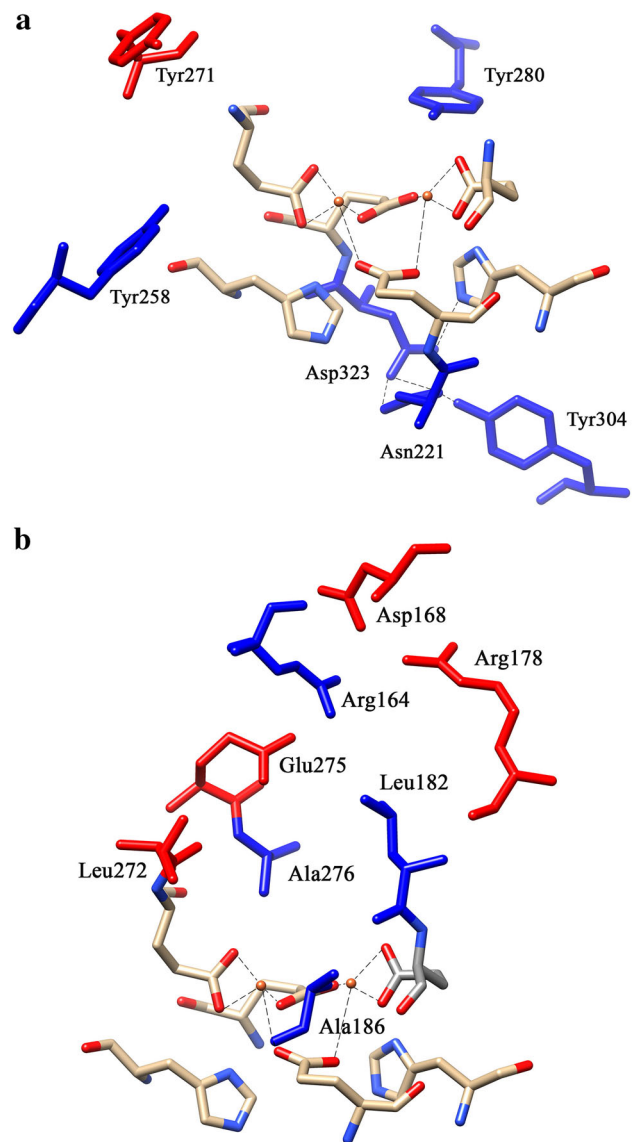
Fig. 3 Conservation of residues important for the structure/function of AOXs mapped onto the AtAOX structural model. **a** Active site cavity. **b** Second hydrophobic cavity. Strictly conserved residues are colored in *blue*, residues not strictly conserved in *red*. Residues forming the catalytic site, shown as a reference, are colored by atom type (Color figure online)

and contributes to shape the cavity that connects the active site to the inner leaflet of the membrane. The almost strict conservation of Tyr271 (Tyr211 in TAO) is apparently linked to his role in dictating the conformation of the N-terminal region which folds back onto the protein helical core. In fact, Tyr271 links helix α4 to the unstructured N-terminal region via an hydrogen bond with the protein backbone. Interestingly, in our dataset, the only substitutions of Tyr 271 observed are with a His (in *Ceriporiopsis subvermispora*, *Sphaerulina musiva*, *Zymoseptoria tritici*, and *Nannochloropsis gaditana*) and a Cys (in *Dacryopinax*

sp.), residues which retain the hydrogen bonding capability.

Panel B of Fig. 3 shows the residues essential for AOXs structure and function which build up the second hydrophobic cavity observed in the vicinity of the AOX active site (Moore et al. 2013). All the residues are conserved or conservatively substituted. In particular, the three hydrophobic residues closest to the iron cluster (Leu182, Ala186 and Ala276; Leu122, Ala126 and Ala216 in TAO, respectively) are strictly conserved in all the dataset members. Leu272 (Leu212 in TAO) is conservatively substituted by a Phe residue in 20 bacterial AOX sequences, while Glu275 (Glu215 in TAO) conservative substitution into Gln is observed only in *Acidovorax sp.* (Bacteria) and *Angomonas deanei* (Trypanosomatidae). Interestingly, Asp168 and Arg178 (Asp100 and Arg118 in TAO), which participate to an extensive hydrogen bonding network centered around the strictly conserved Arg164 (Arg96 in TAO) and involving also Glu275, are substituted into residues able to preserve the above-mentioned network of electrostatic interactions. In fact, Asp168 is substituted by a His in *Naegleria gruberi* and in *Chondrus crispus*, and by an Asn in *Dictyostelium purpureum* and *Polysphondylium pallidum* (Eukarya), while Arg178 is substituted by a His in *Naegleria gruberi* and *Chondrus crispus*, and by a Tyr in *Dictyostelium purpureum* and *Polysphondylium pallidum*. Thus, the Asp168–Arg178 interaction becomes a His–His interaction in *Naegleria gruberi* and in *Chondrus crispus*, and an Asn–Tyr interaction in *Dictyostelium purpureum* and *Polysphondylium pallidum*.

Phylogenesis and Molecular Evolution of AOXs

The full-length alignment of the AOXs protein sequences forming our dataset includes 556 or 585 positions (depending on the alignment algorithm), of which 222 positions are included in the catalytic site alignment. The alignment trimmed by poorly aligned positions with the Gblock analysis was 225 in length and was identical to the catalytic site alignment plus three additional residues at its C-terminus. Therefore, while the catalytic residues show a high degree of conservation across the taxa analyzed, the remaining region of the AOX protein proves difficult to align as a consequence of the high evolutionary rate both in terms of amino acid substitutions and insertions or deletions. These poorly aligned regions had no impact on the phylogenetic reconstruction as their removal produced a phylogenetic tree which has identical supported nodes to that based on the full alignment (Figs. 4, 5). Moreover, phylogenetic reconstructions were consistent irrespective of the algorithm used to build multiple sequence alignments and of the number of seeds used during tree searches (Supplementary Fig. 3).

The inferred Maximum Likelihood (ML) trees show that AOXs of Eubacteria, Land Plants, Metazoa, Mycetozoa, and Euglenozoa form five derived and monophyletic clades with high bootstrap support (BS > 79; Fig. 4). Thus, the main isoform of AOX in each of these organism groups likely evolved from the ancestral AOX carried by their common ancestor. This result is in good agreement with the phylogenetic analysis of Suzuki et al. (2005). Compared to this latter study, our phylogeny includes an extended sampling of AOXs across both the bacterial and eukaryote domains, thus uncovering additional phylogenetic relationships and clades, especially within Eubacteria, Land plants, and Fungi. In the Fungi kingdom, at least two independent AOX lineages are present. One lineage includes the reciprocally monophyletic AOX clades of Ascomycota and Basidiomycota, the other one including the monophyletic AOX clades of Zygomycota, Chytridiomycota, and Microsporidia (Fig. 4). However, the support for these two main lineages is not very strong in the ML trees (Figs. 4, 5). Previous studies supported the monophyly of Ascomycota, Basidiomycota, and Microsporidia AOX clades (Williams et al. 2010; Sun et al. 2007), but with no support for their sister relationships (Williams et al. 2010). Thus, the possibility that the AOX variants of each of these Fungi phyla may have had an independent origin cannot be ruled out. Within Stramenopiles, AOXs of Oomycota and Phaeophytes form two monophyletic groups (BS > 84; Figs. 4, 5) with poorly resolved phylogenetic relationships.

Although the AOX lineages corresponding to Eubacteria and most Eukaryote kingdoms and phyla are well supported, their relationships are poorly resolved. By applying a midpoint rooting, we observed two main clusters in the ML trees, one containing AOXs of Ascomycota and Basidiomycota Fungi, Metazoa, Euglenozoa, Stramenopiles, and Alveolata, and the other including AOXs of Eubacteria, Land Plants, Rhodophyta, Mycetozoa, Heterolobosea, and Zygomycota, Microsporidia, and Chytridiomycota Fungi (Figs. 4, 5). An interesting relationship is found between Florideophyceae red algae and Land Plants AOXs suggesting a common ancestral AOX for Archaeoplastida. However, the lack of support for this node and other basal nodes in the ML trees prevents clear-cut conclusions about the phylogenetic relationships among the main AOX lineages.

AOX, initially considered to be limited to plant species, is expressed in all kingdoms, except Archaebacteria (McDonald et al. 2009). Besides, since 1971, cyanide-tolerant O_2 consumption has been reported in animal mitochondria (Hall et al. 1971). AOX was found for the first time in animals which belong to the phyla Mollusca, Nematoda, and Chordata (McDonald and Vanlerberghe 2004), but the taxonomic distribution in this kingdom is
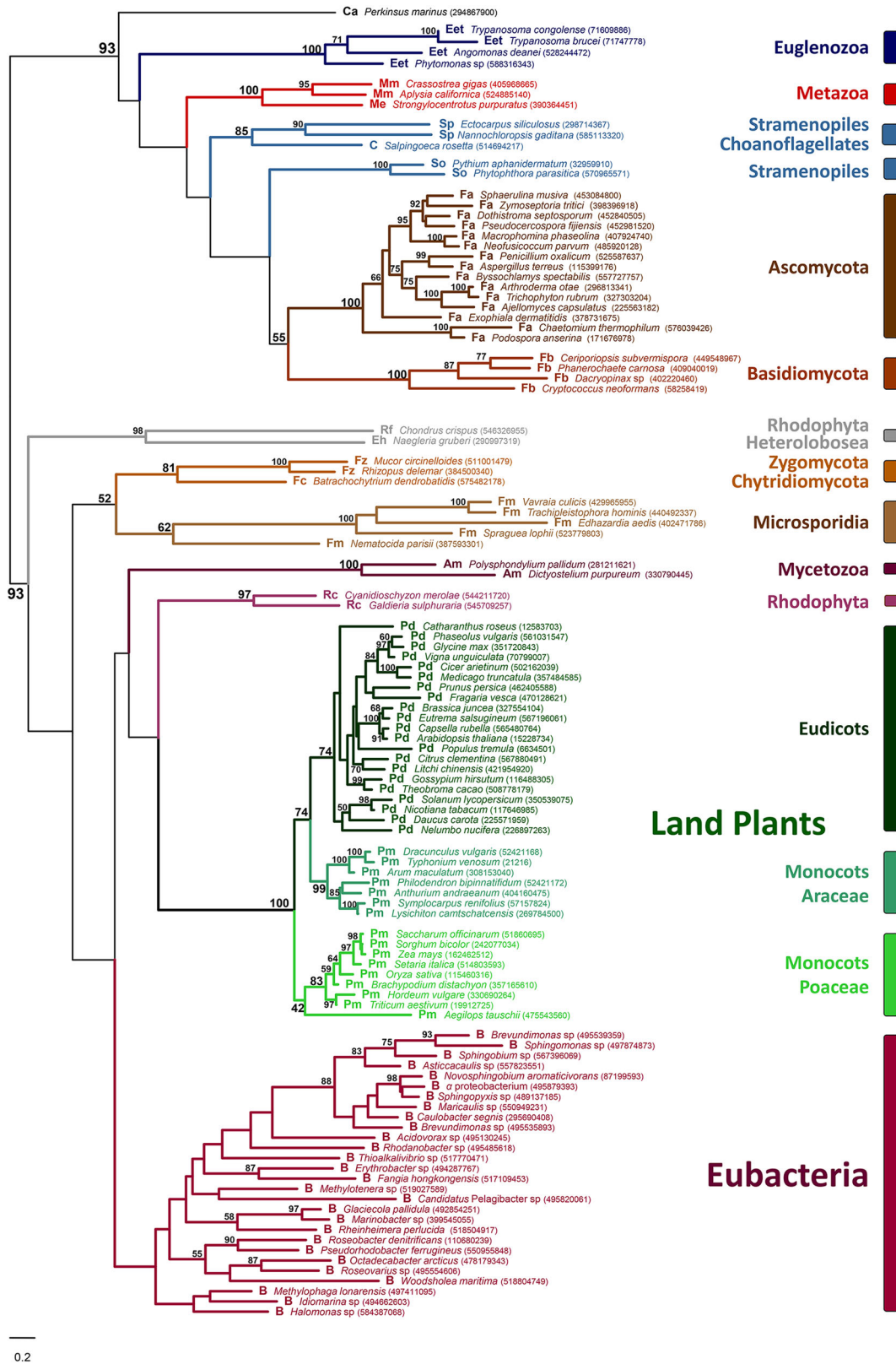
Fig. 4 Maximum likelihood (ML) phylogenetic tree based on the full-length sequence of AOX proteins from member of Eubacteria and Eukaryote kingdoms. Multiple sequence alignment was calculated using Clustal W. The ML tree search implemented the LG + G + I substitution model selected by ProtTest under the AIC criterion. Bootstrap values over 100 replicates are reported in correspondence to the nodes
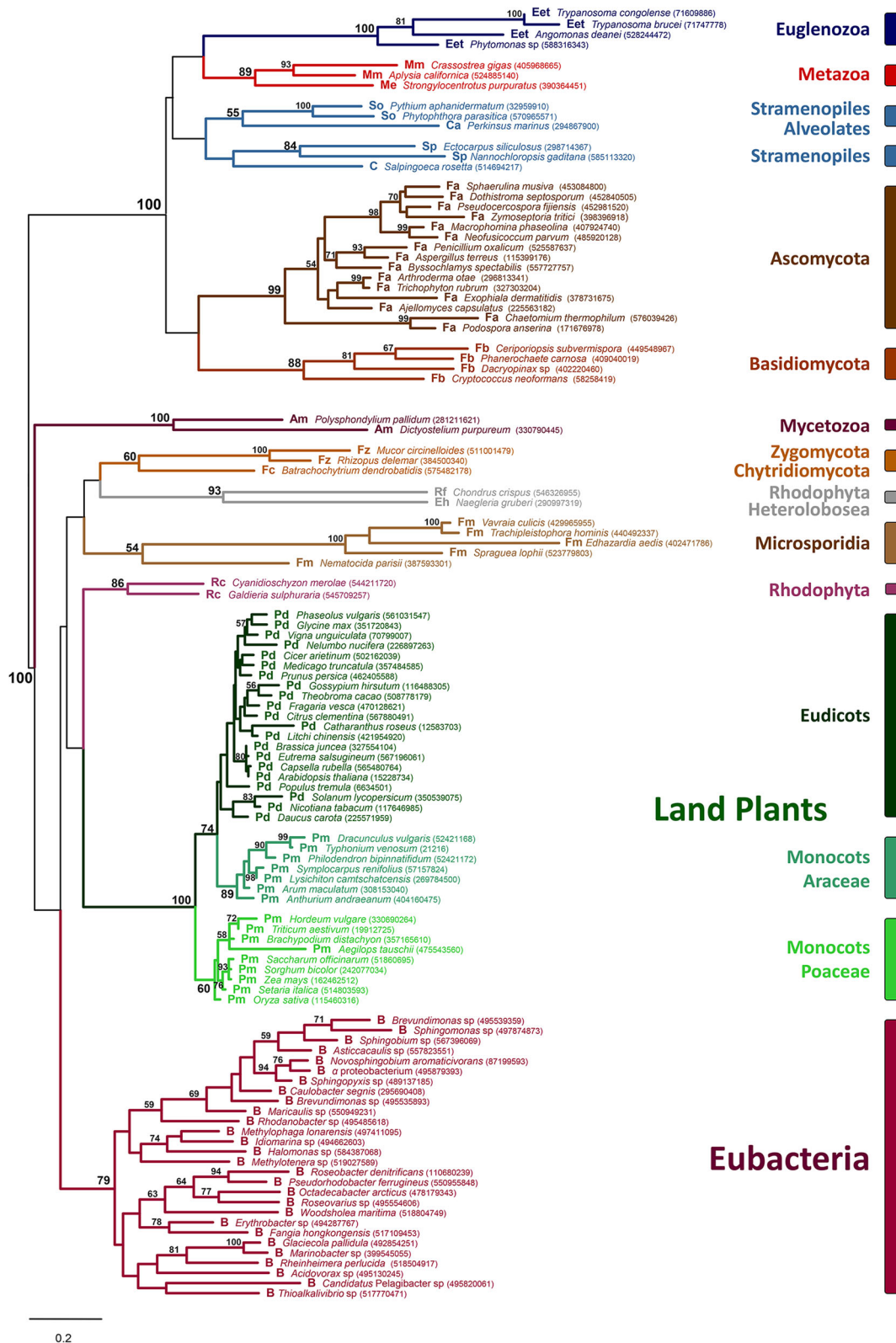
Fig. 5 Maximum likelihood (ML) phylogenetic tree based on the catalytic core sequence of AOX proteins from member of Eubacteria and Eukaryotes kingdoms. Multiple sequence alignment was calculated using Clustal W. The ML tree search implemented the LG + G substitution model selected by ProtTest under the AIC criterion. Bootstrap values over 100 replicates are reported in correspondence to the nodes

Fig. 6 Sequence and structural features of land plant AOXs N-terminal region. **a** Multiple amino acid sequence alignment of the 94–128 region of Land Plants AOXs (numbering refers to *A. thaliana* AOX). The position of the conserved Cys residue involved in the covalent homo-dimer formation (Cys127 in *A. thaliana* AOX) is indicated by the *black star* on *top* of the alignment. Eudicots, Araceae, and Poaceae sequence blocks are indicated by the *letters E, A,* and *P* on the *right side* of the alignment. For clarity, the Araceae and Poaceae sequence blocks are enclosed in a semi-transparent *green* and *yellow rectangle*, respectively. **b** Schematic representation of the structural model of the *A. thaliana* AOX homo-dimer. The 94–127 regions of each monomer, mediating monomer–monomer interactions, are highlighted by a *green border*. Cys127 residue of each monomer, mediating covalent dimer formation through a disulfide bridge, is represented in spacefill and colored in *yellow*. Plant AOXs sequence identity is mapped onto the backbone by coloring from *red* (high sequence conservation) to *blue* (low sequence conservation). *Yellow circles* highlight the low sequence conservation regions corresponding to the *N*-terminal segment and contacting regions (Color figure online)
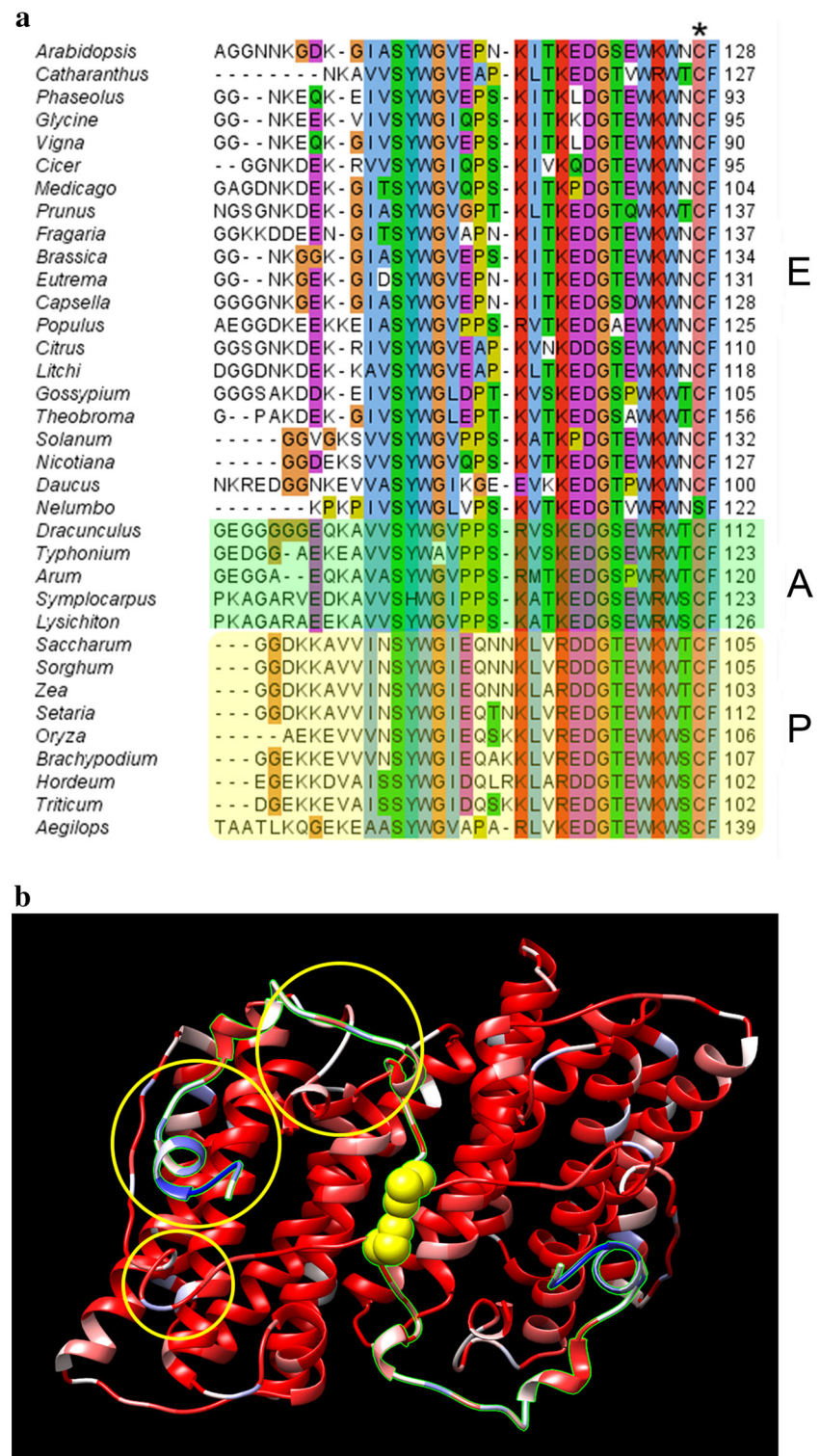
broad. The phylogenetic analysis in this work evidenced the presence of a Metazoa monophyletic clade grouping marine benthic organisms, *i.e.*, *Crassostrea gigas*, *Aplysia californica*, and *Strongylocentrotus purpuratus*. The presence of AOX in these species is probably related to the fact that marine organisms may be confronted with rapidly varying oxygen concentrations, hypoxic conditions, and increase of hydrogen sulfide concentration (Turrens 2003; Abele et al. 2007; Sussarellu et al. 2013). Interestingly, Eubacteria AOXs identified in this study all come from

sea-water bacterial species in which AOX likely exerts a protective role toward oxidative stress and hydrogen sulfide-mediated inhibition of cytochrome *c* oxidase.

In the Fungi kingdom, the two AOX clades of Ascomycota and Basidiomycota include species which are mammalian parasites and plant pathogens (*e.g.*, *Arthroderma otae* and *Sphaerulina musiva*); in Fungi, one of the putative roles of AOX is to decrease ROS generation (Yukioka et al. 1998; Li et al. 2011). Moreover, in these pathogens, AOXs may have evolved as a defense against the host toxic metabolites, such as carbon monoxide, nitric oxide, cyanide, and hydrogen sulfide that inhibit cytochrome *c* oxidase (Cooper and Brown 2008).

The AOX lineage of Land Plants is structured in three monophyletic clusters with AOXs of the Poaceae Monocots either lying outside a clade comprising AOXs of Araceae Monocots and Eudicots (Figs. 4, 5, and Supplementary Fig. 3A) or in a sister relationship with the AOXs clade of Araceae Monocots (Supplementary Fig. 3B). The support for two distinct clades of Monocots AOXs is high in all the ML trees either based on the full sequence or on the catalytic residues (Figs. 4, 5, Supplementary Figs. 3A, B). In order to analyze in a deeper detail the differences between these two clades, a multiple sequence alignment of full-length AOX proteins from Land Plants was constructed (Fig. 6). This analysis shows that AOXs of Poaceae possess a characteristic portion of the *N*-terminal sequence which differs from the corresponding region of the other Land Plants.

The inter-monomers disulfide bond mediating covalent dimerization (see above) is shared by all three Land Plants clades as demonstrated by analysis of the structural models built for two representative species of Araceae and Poaceae, *Arum maculatum* and *Zea mays*, respectively (not shown). This result provides some clues for the divergence of the *N*-terminal sequence of Poaceae with respect to Araceae and Eudicots. In fact, the covalent nature of the plant AOXs homo-dimer likely releases the selective pressure for sequence conservation of the *N*-terminal region of plant AOXs, leading to the accumulation of mutations in the relatively recent Poaceae group. This is not limited to the *N*-terminal region involved in the domain-swapping interactions but involves also positions in the catalytic core region that are involved in monomer–monomer contacts (Fig. 6). This finding is in agreement with the separation of Araceae and Poaceae also in the phylogenetic trees based on the multiple sequence alignment of the catalytic site region (Fig. 5, and Supplementary Fig. 3A). An alternative explanation is that, despite the conservation of the disulfide bond, Poaceae may have developed a different pattern of interactions mediated by the *N*-terminal region with respect to the other Land Plants for a different requirement of physiological regulation. In fact, AOX activity depends on the redox state of the disulfide bridge, with the reduced form being significantly more active than the oxidized form (Umbach et al. 1994; Rhoads et al. 1998). Indeed, in the Poaceae present in our dataset, a specific regulation of AOX activity may be related to the peculiar physiology acquired during their evolution from primitive grasses, living near forest margins, to plants growing in open dry habitats and resisting to drought and other environmental stress (Kellogg 2001).

Conclusion

The recent availability of the crystal structure of TAO (Shiba et al. 2013) has allowed for the first time to analyze the structural details of Plant AOXs through molecular modeling of *A. thaliana* AOX 1A. Furthermore, the integration of structure prediction with the analyses of amino acid sequences of AOXs from all kingdoms of Life (except Archaea, which do not encode AOXs) allowed to identify conserved elements of AOXs and to infer a phylogenetic tree for the molecular evolution of this protein family.

A very high degree of conservation of the catalytic core of AOXs has been observed, with main differences restricted to the highly variable *N*-terminal region and contacting residues involved in the homo-dimer formation. In this regard, it is worthwhile to mention the peculiar covalent homo-dimerization mode of Plant AOXs due to the conservation of a cysteine residue which participates to the formation of the intermolecular disulfide bond.

Phylogenetic relationships of AOXs suggest that AOXs proteins have a monophyletic origin in most Eukaryotes kingdoms and in Eubacteria, while AOXs of Fungi seems to have had an independent origin in most phyla. Within plants, AOXs of monocotyledons form two distinct clades with Poaceae AOXs showing loose sequence conservation in the *N*-terminal region and in positions of the catalytic core region involved in monomer–monomer contacts (Fig. 5). This may be due to the fact that the covalent nature of the plant AOXs homo-dimer releases the selective pressure for sequence conservation or rather to structural requirements for physiological regulation related to environmental stress, evolved during evolution of Poaceae in dry open habitats.

Methods

Structural Model and Amino Acid Sequence Searches

The *A. thaliana* AOX 1A, the *Arum maculatum* AOX 1B, and the *Zea mays* AOX 1 amino acid sequences were obtained from the National Center for Biotechnology

Information (NCBI) website (http://www.ncbi.nlm.nih.gov/). These sequences were used to build the corresponding structural models using I-TASSER (Roy et al. 2010) (http://zhanglab.ccmb.med.umich.edu/). The UCSF Chimera program (Pettersen et al. 2004) was used to analyze the structural models and to compare them with the three-dimensional structure of TAO (PDB code 3VV9) (Shiba et al. 2013).

The structural model of the *N*-terminal 31 residues of AtAOX (residues 63–93) were modeled using the ab initio protein structure prediction programs Rosetta ab initio and QUARK (Kim et al. 2004; Xu and Zhang 2012). Electrostatic potential calculations have been carried out using Swiss PDB viewer (Guex and Peitsch 1997).

Amino acid sequences from the different taxa were obtained using the *A. thaliana* AOX 1A amino acid sequence as a bait in BLASTP (Altschul et al. 1997) searches. Searches were carried out in the NCBI non-redundant amino acid sequences (NR) database using an *E*-value threshold of 1×10^{-10}. The searches were conducted with a maximum target of 50 aligned sequences. Besides, subsets of the NR database were used to retrieve AOX sequences from specific taxa: Bacteria (taxid:2), Trypanosomatidae (taxid:5654), Fungi (taxid:4751), Monocotyledoneae (taxid:4447), Dicotyledoneae (taxid:71240), as well as non-Plant and non-Fungi Eukarya (taxid:2759).

Phylogenetic Analyses

Phylogenetic analyses were conducted using the Maximum Likelihood (ML) method. Since multiple sequences alignment and tree search parameters have a great impact in phylogenetic reconstructions, we tested for consistency of phylogenetic reconstructions based on different datasets, alignment algorithms, and starting trees.

First, to assess whether highly variable regions in the alignment introduced noise in the phylogenetic reconstruction, we built three distinct amino acid sequence alignments using Clustal W 2.0.1 (Thompson et al. 1994). A first alignment included the entire length sequences of AOXs (full-length alignment); then the full-length alignment was trimmed by (1) removing poorly aligned positions (conserved alignment) with the software Gblocks 0.91b (Castresana 2000) using a relaxed selection of blocks (Talavera and Castresana 2007), and (2) selecting only the portion of the alignment corresponding to the catalytic site (catalytic site alignment). These trimmed alignments resulted identical for 99 % of the positions (see "Results" section), therefore downstream analyses were performed only on the full-length alignment and on the catalytic site alignment. Second, we performed multiple sequence alignment of the full-length and catalytic site datasets using Clustal Omega 1.2.1 (Sievers et al. 2011), which has a

similar performance to that of high-quality aligners on small datasets, while it outperforms most packages, in terms of execution time and quality, on large datasets.

The Pairwise Homoplasy Index (PHI) test, implemented in SPLITSTREE v. 4 (Huson and Bryant 2006), did not find statistically significant evidence for recombination in our alignments ($p > 0.05$). For each of the four alignments, the fitness among 112 models of protein evolution was estimated trough PhyML 3.0 (Le and Gascuel 2008) by ProtTest 2.0 (Abascal et al. 2005) and the optimal model was selected under the corrected Akaike Information Criterion (Akaike 1974). The LG model (Le and Gascuel 2008) with gamma distributed rates across site (+G) and a proportion of invariant sites (+I) was selected for the full-length alignments and the LG + G model for the catalytic site alignments (with the observed amino acid frequencies (+F) for the alignment based on Clustal W).

ML tree searches were performed via PhyML using the model selected by ProtTest, the best swapping strategy between the Nearest Neighbor Interchanges (NNI) and the Subtree Pruning and Regrafting (SPR), with either a single starting tree or multiple (100) random starting trees. Nodes support for the resulting phylogenetic tree was evaluated by 100 bootstrap replications in single starting-tree searches and by Shimodaria–Hasegawa approximate likelihood ratio tests of bipartition support (SH-aLRT) (Guindon et al. 2010) for multiple-seed searches. Phylogenetic analyses were carried out in the PALM integrated framework for phylogenetic reconstruction with automatic likelihood model selectors (Chen et al. 2009) and in the Lifeportal computational resource (https://lifeportal.uio.no).

Phylogenetic analyses based on different multiple sequence alignment algorithms and with single- or multiple-seed searches gave identical results in terms of supported nodes; therefore we discuss in detail trees based on the PALM pipeline using Clustal W alignment algorithm, single starting tree, and 100 bootstrap replicates (Figs. 3, 4) and we provide in Supplementary Fig. 3, the trees calculated in the Lifeportal, based on Clustal Omega, 100 starting tree, and SH-aLRT test of support.

Author Contributions RP collected the AOXs sequences, carried out the molecular modeling studies and drafted the manuscript. DS carried out the phylogenetic analysis and helped to draft the manuscript. VB carried out ab initio molecular modeling studies and electrostatic potential calculations. RA participated in the interpretation of the results and helped to draft the manuscript. PA and FP

conceived the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105

Abele E, Philip E, Gonzalez PM, Puntarulo S (2007) Marine invertebrate mitochondria and oxidative stress. Front Biosci 12:933–946

Ajayi WU, Chaudhuri M, Hill GC (2002) Site-directed mutagenesis reveals the essentiality of the conserved residues in the putative diiron active site of the trypanosome alternative oxidase. J Biol Chem 277:8187–8193

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr AC-19:716–723

Albury MS, Elliott C, Moore AL (2009) Towards a structural elucidation of the alternative oxidase in plants. Physiol Plant 137:31–327

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402

Andersson ME, Nordlund P (1999) A revised model of the active site of alternative oxidase. FEBS Lett 449:17–22

Berthold DA, Andersson ME, Nordlund P (2000) New insight into the structure and function of the alternative oxidase. Biochim Biophys Acta 1460:241–254

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Chaudhuri M, Hill GC (1996) Cloning, sequencing and functional activity of the *Trypanosoma brucei* brucei alternative oxidase. Mol Biochem Parasitol 83:125–129

Chen SH, Su SY, Lo CZ, Chen KH, Huang TJ, Kuo BH, Lin CY (2009) PALM: a paralleled and integrated framework for phylogenetic inference with automatic likelihood model selectors. PLoS One 4:e8116

Clifton R, Millar AH, Whelan J (2006) Alternative oxidase in *Arabidopsis*: a comparative analysis of differential expression in gene family provides new insights into function of non-phosphorylating bypasses. Biochim Biophys Acta 1757:730–741

Cooper CE, Brown GC (2008) The inhibition of mitochondrial cytochrome oxidase by the gases carbon monoxide, nitric oxide, hydrogen cyanide and hydrogen sulfide: chemical mechanism and physiological significance. J Bioenerg Biomembr 40:533–539

Finnegan PM, Soole KL, Umbach AL (2004) Alternative mitochondrial electron transport proteins in higher plants. In: Day DA, Millar AH, Whelan J (eds) Plant mitochondria: from gene to function, vol 17., Advances in photosynthesis and respirationKluwer, Dordrecht, pp 163–230

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. Syst Biol 59:307–321

Hall FR, Hollingworth RM, Shankland DL (1971) Cyanide tolerance in millipedes: the biochemical basis. Comp Biochem Physiol 38:723–737

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267

Keeling P, eander BS, Simpson A (2009) Eukaryotes. Eukaryota, organisms with nucleated cells. Version 28 Oct 2009. http://tolweb.org/Eukaryotes/3/2009.10.28 in The Tree of Life Web Project, http://tolweb.org/

Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol 125:1198–1205

Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucl Acids Res 32:526–531

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320

Li Q, Bai Z, O'Donnell A, Harvey LM, Hoskisson PA, McNeil B (2011) Oxidative stress in fungal fermentation processes: the roles of alternative respiration. Biotechnol Lett 33:457–467

Maxwell DP, Wang Y, McIntosh L (1999) The alternative oxidase lowers mitochondrial reactive oxygen production in plant cells. Proc Natl Acad Sci USA 96:8271–8276

McDonald AE (2008) Alternative oxidase: an inter-kingdom perspective on the function and regulation of this broadly distributed 'cyanide resistant' terminal oxidase. Funct Plant Biol 35:535–552

McDonald A, Vanlerberghe G (2004) Branched mitochondrial electron transport in the animalia: presence of alternative oxidase in several animal phyla. IUBMB Life 56:333–341

McDonald AE, Vanlerberghe GC, Staples JF (2009) Alternative oxidase in animals: unique characteristics and taxonomic distribution. J Exp Biol 212:2627–2634

Millenaar FF, Lambers H (2003) The alternative oxidase: in vivo regulation and function. Plant Biol 5:2–15

Moore AL, Umbach AL, Siedow JN (1995) Structure-function relationship of the alternative oxidase of plant mitochondria: a model of the active site. J Bioenerg Biomembr 27:367–377

Moore AL, Shiba T, Young L, Harada S, Kita K, Ito K (2013) Unraveling the heater: new insights into the structure of the alternative oxidase. Annu Rev Plant Biol 64:637–663

Neimanis K, Staples JF, Hüner NP, McDonald AE (2013) Identification, expression, and taxonomic distribution of alternative oxidases in non-angiosperm plants. Gene 526:275–286

Parsons HL, Yip JY, Vanlerberghe GC (1999) Increased respiratory restriction during phosphate-limited growth in transgenic tobacco cells lacking alternative oxidase. Plant Physiol 121:1309–1320

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612

Polidoros AN, Mylona PV, Arnholdt-Schmitt B (2009) AOX gene structure, transcript variation and expression in plants. Physiol Plant 137:342–353

Purvis AC, Shewfelt RL (1993) Does the alternative pathway ameliorate chilling injury in sensitive plant tissue? Physiol Plant 88:712–718

Rasmusson AG, Geisler DA, Moller IM (2008) The multiplicity of dehydrogenases in the electron transport chain of plant mitochondria. Mitochondrion 8:47–60

Rhoads DM, Umbach AL, Sweet CR, Lennon AM, Rauch GS, Siedow JN (1998) Regulation of the cyanide-resistant alternative oxidase of plant mitochondria. Identification of the cysteine residue involved in α-keto acid stimulation and intersubunit disulfide bond formation. J Biol Chem 273:30750–30756

Roberts CW, Roberts F, Henriquez FL, Akiyoshi D, Samuel BU, Richards TA, Mihous W, Kyle D, McIntosh L, Hill GC, Chaudhuri M, Tzipori S, McLeod R (2004) Evidence for mitochondrial-derived alternative oxidase in the apicomplexan parasite *Cryptosporidium parvum*: a potential anti-microbial agent target. Int J Parasitol 34:297–308

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738

Shiba T, Kido Y, Sakamoto K, Inaoka DK, Tsuge C (2013) Structure of the trypanosome cyanide insensitive alternative oxidase. Proc Natl Acad Sci USA 110:4580–4585

Siedow JN, Umbach AL, Moore AL (1995) The active site of the cyanide-resistant oxidase from plant mitochondria contains a binuclear iron center. FEBS Lett 362:10–14

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539

Stechmann A, Hamblin K, Pérez-Brocal V, Gaston D, Richmond GS, van der Giezen M, Clark CG, Roger AJ (2008) Organelles in Blastocystis that blur the distinction between mitochondria and hydrogenosomes. Curr Biol 18:580–585

Stenmark P, Nordlund P (2003) A prokaryotic alternative oxidase present in the bacterium Novosphingobium aromaticivorans. FEBS Lett 552:189–192

Sun J, Lu X, Rinas U, Ping Zeng A (2007) Metabolic peculiarities of Aspergillus niger disclosed by comparative metabolic genomics. Genome Biol 8(9):R182

Sussarellu R, Dudognon T, Fabioux C, Soudant P, Moraga D, Kraffe E (2013) Rapid mitochondrial adjustments in response to short-term hypoxia and re-oxygenation in the Pacific oyster, Crassostrea gigas. J Exp Biol 216:1561–1569

Suzuki T, Hashimoto T, Yabu Y, Majiwa PA, Ohshima S, Suzuki M, Lu S, Hato M, Kido Y, Sakamoto K, Nakamura K, Kita K, Ohta N (2005) Alternative oxidase (AOX) genes of African trypanosomes: phylogeny and evolution of AOX and plastid terminal oxidase families. J Eukaryot Microbiol 52(4):374–381

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577

Tanudji M, Sjöling S, Glaser E, Whelan J (1999) Signals required for the import and processing of the alternative oxidase into mitochondria. J Biol Chem 15:1286–1293

Thirkettle-Watts D, McCabe TC, Clifton R, Moore C, Finnegan PM, Day DA, Whelan J (2003) Analysis of the alternative oxidase promoters from soybean. Plant Physiol 133:1158–1169

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. Nucl Acids Res 22:4673–4680

Turrens JF (2003) Mitochondrial formation of reactive oxygen species. J Physiol 552:335–344

Umbach AL, Siedow JN (1993) Covalent and noncovalent dimers of the cyanide-resistant alternative oxidase protein in higher plant mitochondria and their relationship to enzyme activity. Plant Physiol 103:845–854

Umbach AL, Wiskich JT, Siedow JN (1994) Regulation of alternative oxidase kinetics by pyruvate and intermolecular disulfide bond redox status in soybean seedling mitochondria. FEBS Lett 348:181–184

Van Aken O, Giraud E, Clifton R, Whelan J (2009) Alternative oxidase: a target and regulator of stress responses. Physiol Plant 137:354–361

Van Hellemond JJ, Simons B, Millenaar FF, Tielens AG (1998) A gene encoding the plant-like alternative oxidase is present in Phytomonas but absent in Leishmania spp. J Eukaryot Microbiol 45:426–430

Vanlerberghe GC (2013) Alternative oxidase: a mitochondrial respiratory pathway to maintain metabolic and signaling homeostasis during abiotic and biotic stress in plants. Int J Mol Sci 14:6805–6847

Vanlerberghe GC, McIntosh L (1997) Alternative oxidase: from gene to function. Annu Rev Plant Physiol Plant Mol Biol 48:703–734

Veiga A, Arrabaça JD, Loureiro-Dias MC (2003) Cyanide–resistant respiration, a very frequent metabolic pathway in yeasts. FEMS Yeast Res 3:239–245

Vishwakarma A, Bashyam L, Senthilkumaran B, Scheibe R, Padmasree K (2014) Physiological role of AOX1a in photosynthesis and maintenance of cellular redox homeostasis under high light in Arabidopsis thaliana. Plant Physiol Biochem 81:44–53

Williams BAP, Elliot C, Burri L, Kido Y, Kita K et al (2010) A broad distribution of the alternative oxidase in microsporidian parasites. PLoS Pathog 6(2):e1000761

Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80:1715–1735

Yukioka H, Inagaki S, Tanaka R, Katoh K, Miki N, Mizutani A, Masuko M (1998) Transcriptional activation of the alternative oxidase gene of the fungus Magnaporthe grisea by a respiratory-inhibiting fungicide and hydrogen peroxide. Biochim Biophys Acta 1442:161–169

Silvia Caprari, Giovanni Minervini, Valentina Brandi and Fabio Polticelli*

In silico study of the structure and function of *Streptococcus mutans* plasmidic proteins

Abstract: The Gram-positive bacterium *Streptococcus mutans* is the principal causative agent of human tooth decay, an oral disease that affects the majority of the world's population. Although the complete *S. mutans* genome is known, approximately 700 proteins are still annotated as hypothetical proteins, as no three-dimensional structure or homology with known proteins exists for them. Thus, the significant portion of genomic sequences coding for unknown-function proteins makes the knowledge of pathogenicity and survival mechanisms of *S. mutans* still incomplete. Plasmids are found in virtually every species of *Streptococcus*, and some of these mediate resistance to antibiotics and pathogenesis. However, there are strains of *S. mutans* that contain plasmids, such as LM7 and UA140, to which no function has been assigned yet. In this work, we describe an *in silico* study of the structure and function of all the *S. mutans* proteins encoded by pLM7 and pUA140 plasmids to gain insight into their biological function. A combination of different structural bioinformatics methodologies led to the identification of plasmidic proteins potentially required for the bacterial survival and pathogenicity. The structural information obtained on these proteins can be used to select novel targets for the design of innovative therapeutic agents towards *S. mutans*.

Keywords: *ab initio* molecular modeling; bacterial pathogenicity; function recognition; plasmidic proteins; *Streptococcus mutans*.

───────
*Corresponding author: Fabio Polticelli, Department of Sciences, University of Roma Tre, Viale G. Marconi 446, 00146 Rome, Italy, Phone: +39-06-57336362, Fax: +39-06-57336321, E-mail: fabio.polticelli@uniroma3.it; and National Institute of Nuclear Physics, Roma Tre Section, 00146 Rome, Italy. http://orcid.org/0000-0002-7657-2019
Silvia Caprari and Valentina Brandi: Department of Sciences, University of Roma Tre, 00146 Rome, Italy
Giovanni Minervini: Department of Sciences, University of Roma Tre, 00146 Rome, Italy; and Department of Biology, University of Padua, 35131 Padua, Italy

Introduction

Dental caries, an oral disease directly connected to the development of dental biofilms, is still a major burden for majority of the world's population [1]. Although, during the last few years, the World Health Organization (WHO) Global Oral Health Programme has tried to increase awareness on oral health worldwide, dental caries still represents a serious health problem, especially in the poorest developing areas [2]. Furthermore, even in developed countries, dental caries is still a significant health problem afflicting 60%–90% of school children, as well as majority of adults [3].

Dental biofilm formation results from the action of a wide spectrum of acidogenic micro-organisms. Among these, *Streptococcus mutans* is considered the principal causative agent of human dental caries [4]. Its high pathogenic potential is due to its ability to strongly adhere on the tooth surface, to the synthesis of elevate amounts of extracellular polysaccharides, representing the main components of the dental biofilm, and to its highly acidogenic potential, which leads to the dissolution of enamel and the initiation of the dental caries process [5]. Furthermore, the pathogenic potential of *S. mutans* is not limited to dental caries, as it can also enter the bloodstream, adhere to endothelial cells, and lead to infective endocarditis [6].

Although a biological function has been assigned to many proteins coded by the *S. mutans* genome, approximately 700 proteins (≈35% of the whole *S. mutans* proteome) are annotated as putative proteins, as no three-dimensional structure or homology with known proteins exists for them [7]. Therefore, the significant percentage of genomic sequences coding for unknown-function proteins makes the knowledge about pathogenicity and survival mechanisms of *S. mutans* still incomplete, with a negative impact on the development of novel antibacterial compounds. For these reasons, carrying on structural and functional genomics studies on this bacterium is very useful in order to get a complete knowledge about the structure and function of all the *S. mutans* proteins.

Most of the *Streptococcus* species contain plasmids involved in a high variety of biological activities, such as resistance to antibiotics, production of toxins against closely related bacterial strains, immunity and

pathogenesis mechanisms [8]. Some of these are designated as cryptic because their function is unknown, and also some *S. mutans* plasmids fall within this category. Indeed, there are some strains of *S. mutans* that contain plasmids, such as LM7 and UA140, to which no specific function has been assigned yet. Currently, there are no available literature data concerning the pathogenic relevance of these plasmids, which are named, respectively, pLM7 and pUA140. However, it is known that *S. mutans* LM7 strain is among the most virulent strains, being able to invade human coronary artery endothelial cells [9].

The plasmid-containing strain UA140 has a clinical relevance as well because it was isolated from the caries-active of an African-American child from Birmingham, Alabama [10]. Thus, it can be hypothesized that the pathogenic features of *S. mutans* LM7 and UA140 strains might be due also to the proteins encoded by the plasmids. The studies aimed to characterize the pUA140 sequence using restriction enzymes led to define a physical map model of pUA140 [10]. The authors of this study deduced that pUA140 displays all the typical characteristics of those plasmids, which replicate by a rolling circle mechanism. One of these is the presence of a putative replication protein (rep protein) corresponding to the pUA140_p1 protein.

For the plasmidic proteins pLM7p01, pLM7p03, pUA140_p2 and pUA140_p5, a generic assignment at the level of protein functional family is available. In fact, when their sequences are submitted to a BLAST (Basic Local Alignment Search Tool [11]) search, the retrieved sequences indicate that pLM7p01 and pUA140_p5 both belong to the plasmid stabilization system family and that pLM7p03 and pUA140_p2 both belong to the mob proteins family (mobilization proteins). No information regarding the biological function of all the other *S. mutans* plasmidic proteins can be inferred on the basis of the sequence similarity with proteins with known function.

In this work, we describe an *in silico* study of the structure of *S. mutans* pLM7 and pUA140 plasmidic proteins in order to gain insight into their biological function. Knowledge of the structure and function of plasmidic proteins can lead to the discovery of potential pharmacological targets on the basis of which new drugs can be designed. The results here described indicate that a combination of different bioinformatics methodologies can lead to relevant results, such as the identification of proteins required for bacterial survival and pathogenicity. Therefore, the structural information obtained on these proteins might be used to design novel and efficient therapeutic agents towards *S. mutans*.

Methods

Molecular modeling of the pLM7 and pUA140 plasmidic proteins

All the sequences of the plasmidic proteins belonging to both plasmids were first submitted to a BLAST search [11] against the non-redundant NCBI protein sequence database (from now on, referred simply as the NR database) in order to check if these displayed any sequence homology with proteins of known structure and/or function.

With the exception of two 555 amino acid residues long plasmidic proteins, the structural models of all the plasmidic proteins were predicted using the Rosetta *ab initio* software [12]. In detail, Rosetta *ab initio* generates a certain number of structural models or "decoys" for each protein. The models are then subjected to a clustering and ranking procedure, using the dedicated Rosetta *ab initio* modules, which lead to the selection of the lowest energy structural model presented in the Results section.

As the computing time required by Rosetta *ab initio* execution dramatically increases with the length of the amino acid sequence, the prediction of the three-dimensional structure of the two 555 amino acid residues long proteins, i.e. pLM7p03 and pUA140_p2, was carried out using the I-TASSER software [13].

The three-dimensional structures of pLM7p04, pUA140_p1, pUA140_p3 and pUA140_p4 proteins were also predicted by homology modeling using the program NEST, a fast model building program that applies an "artificial evolution" algorithm to construct a model from a given template and alignment [14]. The NEST option – *tune 2* – was used to refine the alignment avoiding the unlikely occurrence of insertions and deletions within template secondary structure elements and the presence of "zig-zag" gaps in the alignment.

The quality of the final models was evaluated using PROCHECK v.3.5.4 [15]. For all models, with the only exception of that of pUA140_p1 (see below), the average G-value calculated by PROCHECK was equal or higher than −0.5, indicating good quality models (Table S1).

Protein function prediction

In order to get information on the biological function of the plasmidic proteins, the Dali web server was used [16], which allows the comparison of the predicted structures with all the protein structures present in the Protein Data Bank (http://www.rcsb.org) and the retrieval of protein

structures, which display significant structural similarity, and hence similar function, with the input structure.

Results

pUA140 plasmid proteins

The pUA140 plasmid is 5640-bp long and contains five genes coding for five different proteins, namely, pUA140_p1, pUA140_p2, pUA140_p3, pUA140_p4 and pUA140_p5, which are 258, 555, 188, 105 and 120 amino acid residues long, respectively. For each of these proteins, a BLAST search in the NR database was performed and a structural model was built starting only from their amino acid sequence. Then, it was attempted to assign them a putative biological function on the basis of sequence and structure homology with known proteins already present in the Protein Data Bank (http://www.rcsb.org).

pUA140_p1

As reported in the introduction, pUA140_p1 is known as a rep protein. When the pUA140_p1 sequence was submitted to a BLAST query, the only significant alignment was obtained with the pLM7p04 protein (E value = 0.00), belonging to the other plasmid analyzed in this study, pLM7 (Table 1). No compact and structured model was obtained for pUA140_p1 using Rosetta. The predicted three-dimensional structure of pUA140_p1, obtained using I-TASSER, is shown in Figure 1.

This structural model of pUA140_p1 does not appear to be reliable, as a substantial part of the protein appears unstructured. The results obtained by submitting the structural model of pUA140_p1 to the Dali server are reported in Table 2, and these show that the most significant structural similarity (Z-score = 13.5, rmsd = 3.0 Å, lali = 208, nres = 280 and %id = 13) is detected with the dimeric enzyme phosphopantothenoylcysteine synthetase (PDB code: 1P9O). In order to find out what the likely function of pUA140_p1 could be, the structure of the enzyme phosphopantothenoylcysteine synthetase was used as a template to build the structural model of pUA140_p1 using NEST [14]. Although the resulting model appears more structured than that predicted by Rosetta, it is not possible to identify a protein region that shows features of a site responsible for binding of metal ions, an essential characteristic for the catalytic activity of rep proteins. Furthermore, rep proteins display the conserved

Table 1: BLAST [11] search results for the *S. mutans* plasmidic protein sequences against the NR database.

Query protein	Blast hit	Query coverage	E value	% Identity
pUA140_p1	Hypothetical protein pLM7p04 [*S. mutans*]	100%	0.00	99%
	Uncharacterized protein [uncultured *Ruminococcus* sp.]	84%	6×10^{-12}	26%
pUA140_p2	Hypothetical protein pLM7p03 [*S. mutans*]	100%	0.00	99%
	Type IV secretory pathway 2C VirD4 components [uncultured *Ruminococcus* sp.]	67%	4×10^{-58}	33%
	pST0 hypothetical protein [*S. thermophilus*]	69%	1×10^{-48}	30%
	MOB-like transmembrane protein [*S. poulsonii*]	63%	5×10^{-19}	26%
pUA140_p3	Hypothetical protein pLM7p02 [*S. mutans*]	74%	3×10^{-96}	98%
pUA140_p4	Hypothetical protein [*Streptococcus pneumoniae*]	99%	1×10^{-32}	57%
	Multispecies: pilus assembly protein HicB [*Mycobacterium avium* complex (MAC)]	37%	0.42	41%
	DNA-binding protein [*Klebsiella pneumoniae*]	25%	1.80	56%
pUA140_p5	Hypothetical protein pLM7p01 [*S. mutans*]	100%	6×10^{-80}	96%
	Plasmid stabilization protein [*Streptococcus* sp. M334]	93%	7×10^{-37}	52%
	Plasmid stabilization protein [*S. equi*]	91%	1×10^{-29}	47%
	Plasmid stabilization protein [*Streptococcus suis*]	92%	7×10^{-26}	45%
pLM7p01	Hypothetical protein pUA140_p5 [*S. mutans*]	100%	6×10^{-80}	96%
	Plasmid stabilization protein [*Streptococcus* sp. M334]	93%	6×10^{-36}	52%
	Plasmid stabilization protein [*S. equi*]	91%	1×10^{-29}	48%
	Plasmid stabilization protein [*S. suis*]	92%	5×10^{-25}	45%
pLM7p02	Hypothetical protein pUA140_p3 [*S. mutans*]	100%	2×10^{-96}	98%
pLM7p03	Mob protein pUA140_p2 [*S. mutans*]	100%	0.00	99%
	Type IV secretory pathway 2C VirD4 components [uncultured *Ruminococcus* sp.]	67%	2×10^{-58}	33%
	MOB-like transmembrane protein [*S. poulsonii*]	63%	6×10^{-19}	26%
pLM7p04	Hypothetical protein pUA140_p1 [*S. mutans*]	100%	0.00	99%
	Uncharacterized protein [uncultured *Ruminococcus* sp.]	84%	6×10^{-12}	26%

Figure 1: Structural model of pUA140_p1 obtained using I-TASSER.

sequence motif His-Hydr-His-Hydr-Hydr-Hydr (where Hydr stands for a bulky hydrophobic amino acid) [17]. However, the presence of such a motif was not evident in pUA140_p1.

pUA140_p2

pUA140_p2 is 555 amino acid residues long protein. The results of the BLAST search using this protein as a bait are reported in Table 1 and show a significant sequence

similarity with the pLM7p03 protein encoded by pLM7 plasmid (E value = 0.00). In addition to this, pUA140_p2 also displays a significant similarity with mob proteins such as the MOB-like transmembrane protein from *Spiroplasma poulsonii* (E value = 5×10^{-19}) (Table 1).

The predicted structural model of pUA140_p2 obtained using I-TASSER is shown in Figure 2. The Dali results indicate that the majority of significant structural alignments found are relative to proteins involved in bacterial conjugation (Table 2; Z-score = 39.7, rmsd = 1.1 Å, lali = 393, nres = 416 and %id = 16) and, in particular, to the *Escherichia coli* protein TrwB (PDB code: 1E9R) (Figure 2).

TrwB is a multimeric membrane protein responsible for the recruitment of the protein complex that unwinds DNA, as well as for the coupling of this complex with the structures involved in the transfer of a single DNA strand during the bacterial conjugation [18]. This process requires energy; consequently, each TrwB subunit has a binding site for ATP whose hydrolysis is required for the DNA strand transfer. It is important to note that the TrwB three-dimensional structure reported in Figure 2 is relative to the soluble domain of the conjugation protein and lacks the transmembrane anchoring domain [18].

As can be observed in Figure 2, the pUA140_p2 structural model, with the exception of the amino terminal part, shares a remarkable structural similarity with any of TrwB subunits. Furthermore, a structure-based sequence alignment between pUA140_p2 and TrwB shows that the amino acid sequence involved in ATP binding in the latter protein is conserved in pUA140_p2 (Figure 2). From these results, it can be concluded that also pUA140_p2 has an ATP binding site.

Table 2: Dali [16] search results for the structural models of *S. mutans* plasmidic proteins.

Query protein	Known structure proteins found	Z-score[a]	rmsd, Å	lali	nres	% id
pUA140_p1	1P9O (all the chains): phosphopantothenoylcysteine synthetase	13.5	3.0	208	280	13
pUA140_p2	1E9R (all the chains): conjugal transfer protein TrwB	39.7	1.1	393	416	16
pUA140_p3	1Q46 (A): translation initiation factor 2 α subunit	8.2	3.1	103	174	17
pUA140_p4	1PAR (all the chains): Arc repressor	4.5	1.7	49	53	27
	2K29 (A,B) antitoxin RelB	2.7	3.5	47	50	17
pUA140_p5	2A6R (all the chains): toxin YoeB	10.2	1.9	83	84	12
	2KC9 (A): toxin RelE	9.8	2.2	86	95	14
pLM7p01	2A6R (all the chains): toxin YoeB	10.2	2.0	81	84	12
	2KC9 (A): toxin RelE	9.6	2.2	87	95	14
pLM7p02	1Q46 (A): translation initiation factor 2 α subunit	8.0	3.4	103	174	17
pLM7p03	1E9R (all the chains): conjugal transfer protein TrwB	48.0	1.1	390	415	16
pLM7p04	1P9O (all the chains): phosphopantothenoylcysteine synthetase	13.5	3.0	208	280	13

[a]Hits with a Z-score higher than 2 are considered significant, those with Z-score higher than 4 are considered highly reliable [16]. The rmsd value is the root mean square deviation calculated for the matched Cα pairs between the query protein and the known three-dimensional structure, lali is the length of the aligned substructures, nres is the number of residues of the aligned structure, and % id is the percentage amino acid sequence identity between the query protein and the known three-dimensional structure over the aligned region [16].

N.	N-terminal	Transmembrane region	C-terminal	Type	Length
1	63	IFLLISFLLAPFAGFISILCYGL	85	Primary	23
2	95	ATLLLFFTGVIMIISLTFSIRFP	117	Primary	23
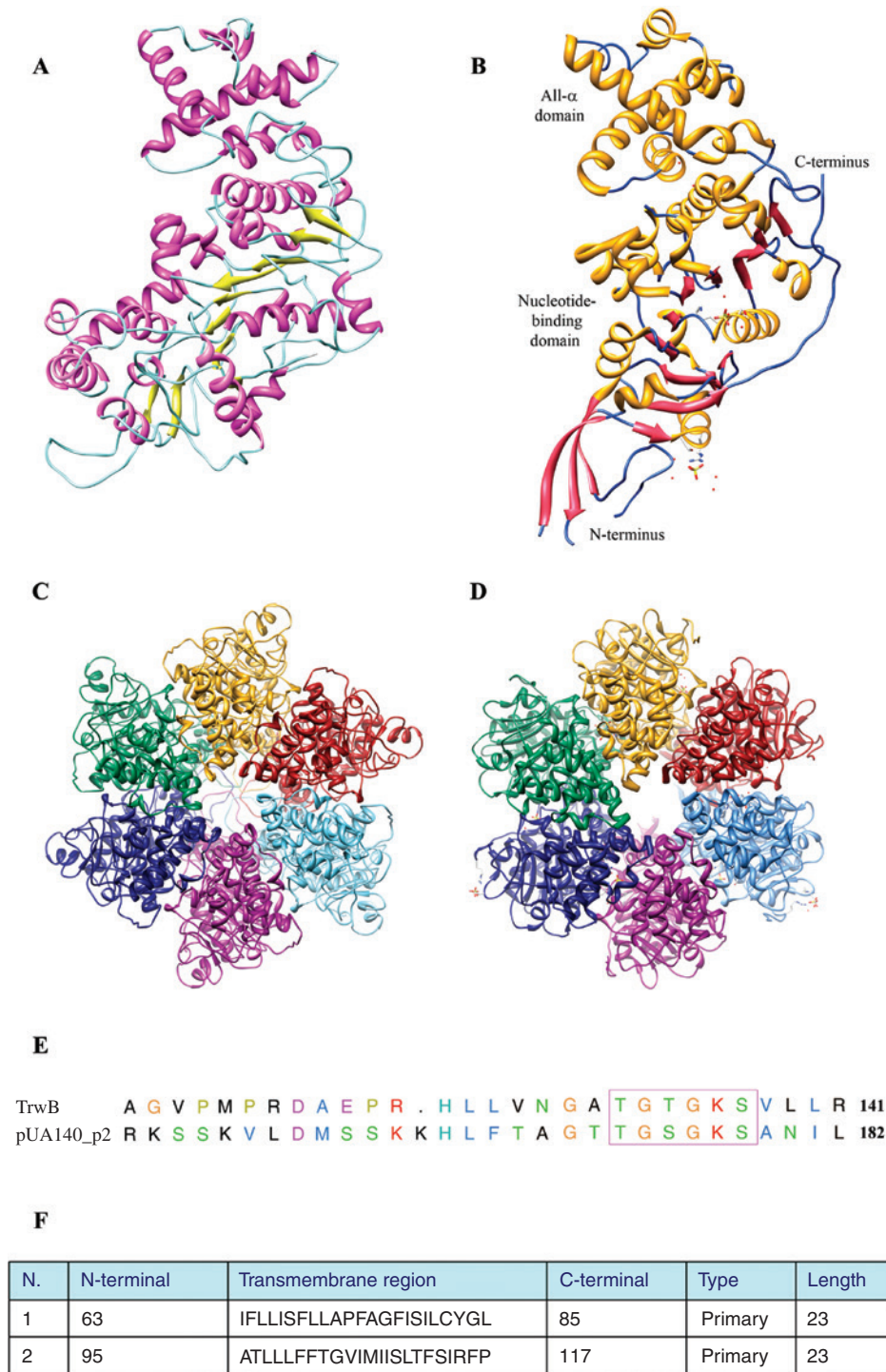
Figure 2: Predicted structural features of pUA140_p2.
(A) Structural model of pUA140_p2 obtained using I-TASSER. (B) Crystal structure of one TrwB subunit (PDB code 1E9R). (C) Structure of TrwB hexamer. (D) Hypothetical structural model resulting from the interaction between six pUA140_p2 monomers. (E) Sequence alignment between the ATP binding region of TrwB and the corresponding region of pUA140_p2. Residues involved in ATP binding (Thr132, Lys136, and Ser137 in TrwB, and Thr173, Lys177, and Ser178 of pUA140_p2) are enclosed in the magenta square. (F) Results obtained by SOSUI server analysis for pUA140_p2. Two N-terminal transmembrane helix segments are predicted; the first spans residues 63–85, and the second, residues 95–117.

The interaction between six TrwB subunits, requested for the formation of the multimeric structure, is mainly hydrophilic and leads to the formation of a channel that runs from the cytoplasm side to the membrane side and is responsible for transferring the DNA through the bacterial membrane (Figure 2) [18]. By analogy, it can be

hypothesized that six pUA140_p2 monomers interact in order to form a multimeric structure, whose model is reported in Figure 2.

Thus, the putativepUA140_p2 hexamer would display six ATP binding sites and a channel responsible for transferring DNA. Additionally, the amino terminal part of pUA140_p2 structural model that is not superimposable to the TrwB subunit displays a significant proportion of hydrophobic amino acids and could form a transmembrane region that, by analogy with TrwB, would attach the plasmidic protein to the bacterial membrane.

To verify this hypothesis, pUA140_p2 amino acid sequence was submitted to the SOSUI server (http://bp.nuap.nagoya-u.ac.jp/sosui/) [19], which allows the prediction of the secondary structure of membrane proteins, and was found to be able to form two transmembrane helices (Figure 2).

All these findings reinforce the hypothesis that pUA140_p2 is a protein that binds to the membrane through two hydrophobic helices. By analogy with TrwB, the putative pUA140_p2 hexamer would display 12 transmembrane helices, two for each monomer.

All these data strongly support the hypothesis that pUA140_p2 is a membrane-anchored hexameric ATPase involved in the bacterial conjugation process. We also explored the hypothesis that the presence of pUA140_p2 in *S. mutans* could be the result of a horizontal gene transfer event by performing a search for homologous proteins in all the bacterial genomes using BLAST. Indeed, this search evidenced a high sequence similarity between pUA140_p2 and a hypothetical protein coded by pST0 plasmid of *Streptococcus thermophilus* (E value = 1×10^{-48}) (Table 1). It appears thus likely that the pUA140 plasmid might have been transferred from *S. mutans* to other *Streptococcus* species or *vice versa*.

pUA140_p3

The third ORF of pUA140 encodes pUA140_p3, a 188 amino acid residues long protein. The results obtained by a BLAST search are not noteworthy, as the only significant sequence similarity is observed with the pLM7p02 protein (E value = 3×10^{-96}) (Table 1). The structural model of pUA140_p3 obtained using Rosetta is made up of a series of α-helices joined together by connecting loops plus a region with segments in extended conformation that evidently do not acquire a proper fold in the modeling procedure (Figure 3).
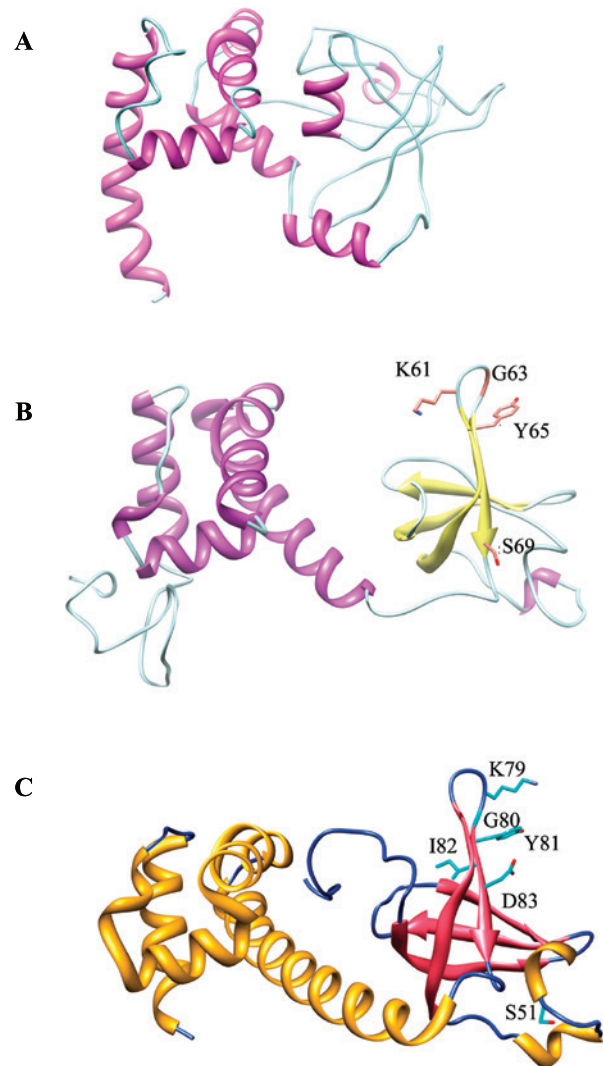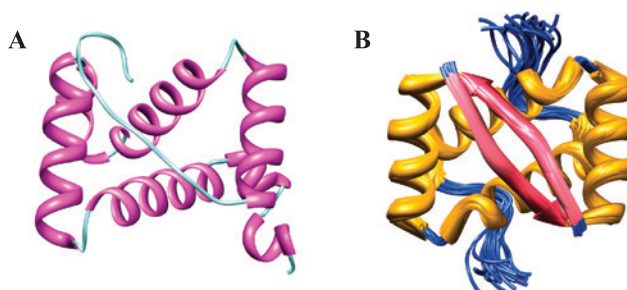

Figure 3: Predicted structural features of pUA140_p3.
(A) Structural model of pUA140_p3 obtained using Rosetta *ab initio*. (B) pUA140_p3 structural model obtained with homology modeling using the crystal structure of the α subunit of the *Saccharomyces cerevisiae* translation initiation factor 2 (eIF2B) (PDB code 1Q46) as template. Residues forming the putative pUA140_p3 phosphorylation site on the surface of the β-domain (Lys61, Gly63, Tyr65 and Ser69) are colored in pink. (C) Crystal structure of eIF2B. Residues forming the eIF2B phosphorylation site (Lys79, Gly80, Tyr81, Ile82, Asp83 and Ser51) are colored in cyan.

Dali results show a significant homology with the α subunit of translation initiation factor 2 (eIF2B) of *Saccharomyces cerevisiae* (PDB code:1Q46) (Table 2).

Structure-based sequence alignment between eIF2B and pUA140_p3 highlights a significant similarity between the two proteins (Z-score = 8.2, rmsd = 3.1 Å, lali = 103, nres = 174 and %id = 17). Thus a homology model of pUA140_p3 was built using eIF2B as template. The resulting model of pUA140_p3 is significantly more ordered

than the one obtained with Rosetta *ab initio* and obviously similar to the structure eIF2B (Figure 3).

An attempt to predict the function of pUA140_p3 was then made on the basis of the known biological function of eIF2B. eIF2B is a multifunctional protein made up of three subunits, one of which, the α subunit, regulates the start of the translation following phosphorylation of a serine residue catalyzed by a protein kinase [20]. In particular, the phosphorylation site is Ser51 residue located on the loop that connects strands β3 and β4 of the amino terminal region. A set of amino acid residues are essential for recognition by the kinase: Lys79, Gly80, Tyr81, Ile82 and Asp83 (Figure 3).

A similar amino acid pattern is present in the β domain of the pUA140_p3 structural model. These residues are Lys61, Gly63 and Tyr65. Furthermore, a serine residue, Ser69, is located in close proximity of the above-mentioned residues (Figure 3). Although these amino acid residues are not conserved in the corresponding positions in pUA140_p3, it has to be considered that the putative *S. mutans* protein kinase involved in the phosphorylation of pUA140_p3 could recognize a slightly different epitope with respect to the eukaryotic one.

Thus, by analogy with the α subunit of eIF2B, it can be hypothesized that these residues constitute a Ser-specific protein kinase recognition site. In order to investigate this issue, a search for the presence of potential phosphorylation sites in pUA140_p3 was carried out using PROSITE [21]. The PROSITE output showed that pUA140_p3 displays two putative phosphorylation sites, one of which includes Ser69 as a potential target phosphorylated by casein kinase 2.

To verify whether, in *S. mutans* genome, there are genes coding for casein kinase 2-like proteins, human casein kinase 2 was used as a bait in a BLAST search on the *S. mutans* strain UA159 genome. This search revealed the presence of a serine-treonine protein kinase, which displays a high sequence similarity with casein kinase 2 (E value = 7×10^{-17}), widespread through all the amino acid sequence. The results obtained reinforce the hypothesis that pUA140_p3 (and hence the pLM7p02 protein given its high sequence similarity with pUA140_p3) is a transcription factor whose function can be regulated by a serine-treonine kinase able to phosphorylate it at the level of Ser69.

pUA140_p4

pUA140_p4 is a 105 amino acid residue long protein that does not display any significant sequence similarity with proteins present in the NR database (Table 1).

The pUA140_p4 structural model obtained using Rosetta is reported in Figure 4.

The results of Dali search indicate that there is a high structural homology (Z-score = 4.5, rmsd = 1.7 Å, lali = 49, nres = 53 and %id = 27) with chains of Arc transcriptional repressor (PDB code: 1PAR) (Table 2), responsible for the transcription of the *ant* gene in P22 bacteriophage [22]. In addition, a significant structural alignment (Z-score = 2.7, rmsd = 3.5 Å, lali = 47, nres = 50 and %id = 17) is also obtained with the RelB antitoxin protein (PDB code: 2K29) (Table 2) (Figure 4).

RelB is an antitoxin participating in the RelB-RelE antitoxin-toxin system [23]. As a dimer, RelB can bind a DNA regulatory region with its N-terminal domain and inhibit the transcription of the genes regulated by it.

pUA140_p4 monomer displays a significant structural similarity both with the Arc repressor and RelB dimeric structures (Table 2 and Figure 4). Thus it is possible that pUA140_p4 evolved from duplication of a gene encoding a single protein chain followed by fusion of two originally identical genes.

If pUA140_p4 has a function similar to that of RelB, it should bind to a DNA regulatory region and repress the transcription of the genes controlled by it. No DNA-bound RelB structure is present in the PDB. However, a DNA-bound structure was determined for the Arc repressor (PDB code 1PAR), a structural homolog of RelB. Thus, the DNA-bound Arc repressor structure was used to probe the stereochemical feasibility of the interaction of pUA140_p4 with DNA. Arc repressor binds to DNA as a tetramer inserting in the DNA major groove two two-stranded β sheets formed by two monomers of each dimer (Figure 5). The amino acid residues involved in contact with the DNA bases in the major groove are Gln9, Asn11 and Arg13 belonging to both β strands of each β sheet. A model of a DNA bound dimer of pUA140_p4 was built through structural superimposition using the DNA bound
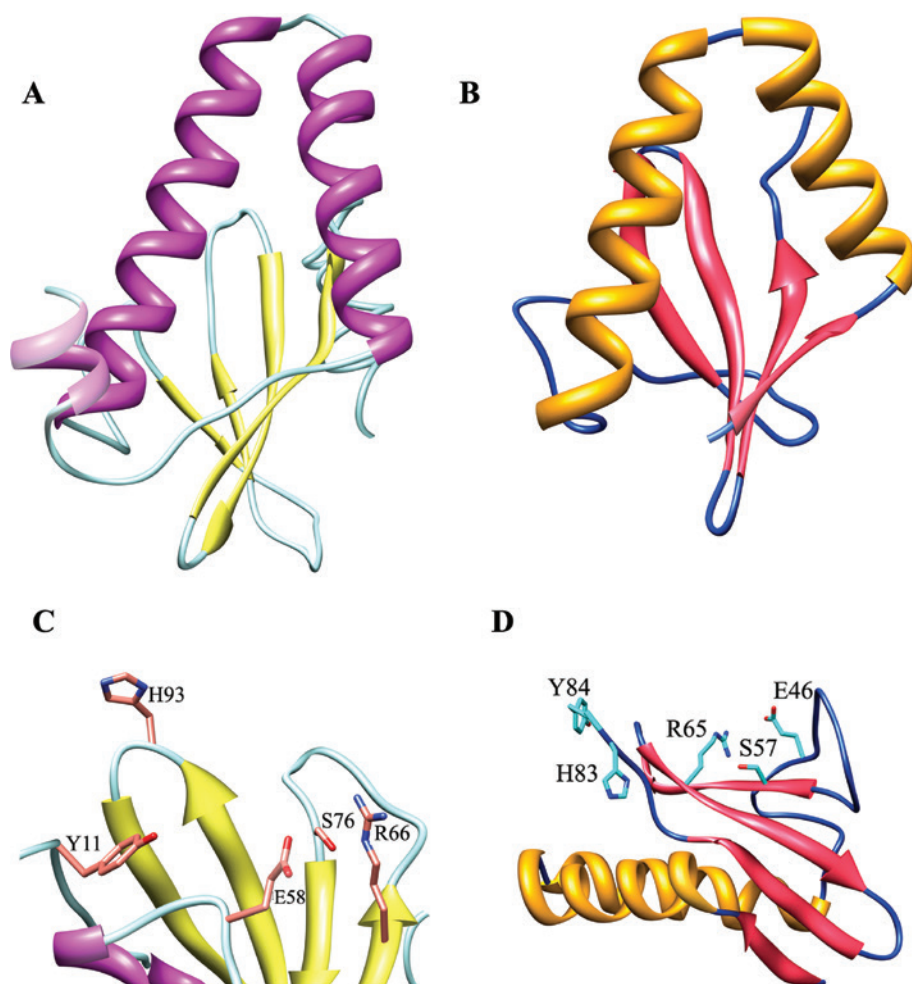

Figure 4: Predicted structural features of pUA140_p4.
(A) Structural model of pUA140_p4 obtained using Rosetta *ab initio*.
(B) Crystal structure of RelB antitoxin dimer (PDB code 2K29).
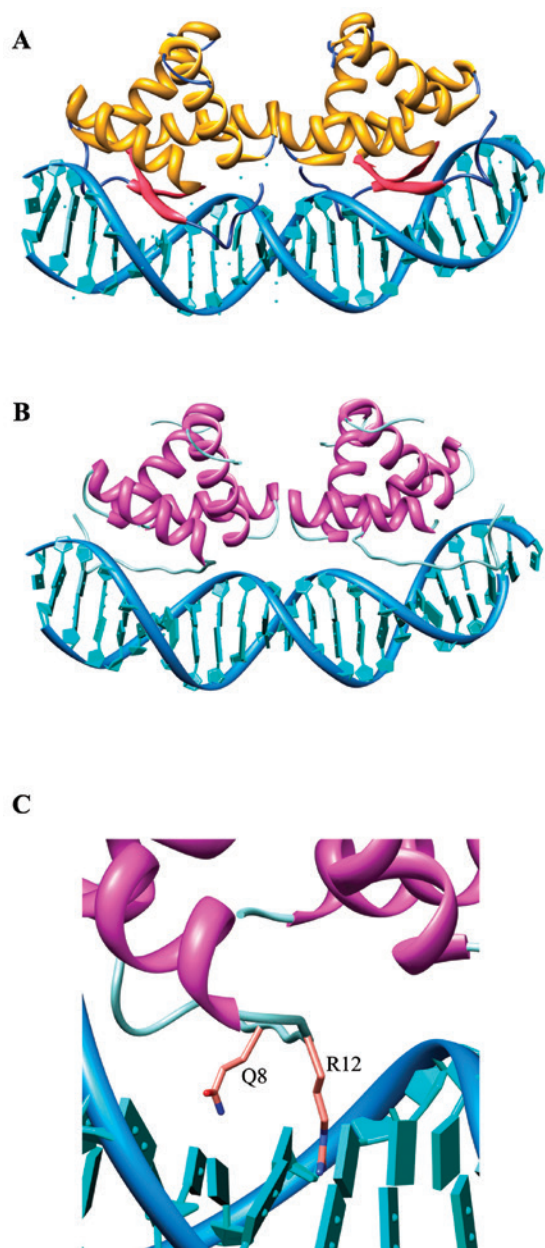
Figure 5: Comparison between the ARC repressor-DNA complex and the putative pUA140_p4-DNA complex.
(A) Tetrameric structure of the Arc repressor bound to DNA (PDB code 1PAR). (B) Hypothetical dimeric structure of pUA140_p4 bound to DNA obtained by homology modeling using the DNA-bound tetrameric structure of the Arc repressor as a template. (C) Enlarged view of the pUA140_p4-DNA interface showing the position of Gln8 and Arg12 within the major groove of DNA.

Arc repressor tetramer as a template (Figure 5). Analysis of the model obtained indicates that pUA140_p4 can form a dimeric structure on DNA without any significant steric hindrance. Unlike the Arc repressor, DNA bound pUA140_p4 model displays the insertion of a single β strand in the DNA major groove, which could be involved

in the hypothetical interaction with the bases. Furthermore, the side chains of Gln8 and Arg12 are located in the DNA major grove in positions similar to those of the amino acids responsible for binding of the Arc repressor to DNA (Figure 5).

pUA140_p5

pUA140_p5 is a 120-amino-acid residue long protein. BLAST results clearly show that pUA140_p5 is highly homologous to pLM7p01 (E value = 6×10^{-80}) (Table 1). In addition, as can be seen in Table 1, pUA140_p5 shows a significant sequence similarity with proteins belonging to the plasmids stabilizing systems, such as the plasmidic protein from *Streptococcus* sp. M334 (E value = 7×10^{-37}) and *Streptococcus equi* (E value = 1×10^{-29}). It is thus likely that also pUA140_p5 is part of one of these systems.

The pUA140_p5 structural model obtained with Rosetta, shown in Figure 6, is made up by a 5-stranded β sheet covered on one side by two α helices.

The Dali results reported in Table 2 reveal a high structural homology (Z-score = 10.2, rmsd = 1.9 Å, lali = 83, nres = 84 and %id = 12) with YoeB (PDB code: 2A6R), a toxin belonging to an *E. coli* toxin-antitoxin system [24] (Figure 6). These systems have been discovered both in plasmids where they constitute a system necessary for maintaining the plasmid into the host bacterial cell and in bacterial chromosomes where they act as response elements to nutritional and environmental stresses to which a bacterium can be subjected [23]. A toxin-antitoxin system is always made up by two close genomic ORFs of the same operon.

In order to get a deeper insight on the putative function of pUA140_p5, given the significant structural homology between pUA140_p5 and the YoeB toxin, YoeB function and the amino acid residues responsible for its catalitic activity were first analyzed. YoeB is an RNAse, and its active site is made up by Glu46, Ser57, Arg65, His83 and Tyr84, positioned on the rim of the central β sheet [24] (Figure 6).

Although YoeB catalytic residues are not conserved in orthologous positions in pUA140_p5, an identical set of residues is present in pUA140_p5 in the same protein region and with inter-residues distances very similar to those observed in YoeB. pUA140_p5 putative catalytic residues are Tyr11, Glu58, Arg66, Ser76 and His93, positioned on the rim of the central β sheet, as observed in YoeB (Figure 6). These observations suggest that pUA140_p5 could be a toxin with RNAse catalytic activity.

Figure 6: Predicted structural features of pUA140_p5.
(A) Structural model of pUA140_p5 obtained using Rosetta *ab initio*. (B) Crystal structure of the YoeB toxin from *E. coli* (PDB code 2A6R). (C) Detailed view of the pUA140_p5 putative active showing residues Arg66, Ser76, Glu58, His93 and Tyr11 in pink. (D) Detailed view of YoeB active site showing residues Glu46, Ser57, Arg65, His83 and Tyr84 in cyan.

pLM7 plasmid proteins

The pLM7 plasmid displays a length of 5658 bp and contains four genes encoding four different proteins, pLM7p01, pLM7p02, pLM7p03, pLM7p04, which are 120, 140, 555 and 258 amino acid residues long, respectively. The analyses carried out are identical to those performed in the pUA140 plasmid proteins study. As mentioned above, all the four pLM7 plasmid proteins display a high sequence similarity with the corresponding proteins of the pUA140 plasmid.

pLM7p01 displays 96% sequence identity with pUA140_p5 (Table 1), and its structural model obtained with Rosetta is identical to that of pUA140_p5. Thus it can be concluded that also pLM7p01 is the toxin component of a toxin-antitoxin system.

pLM7p02 is 140 amino acid residues long. No significant sequence similarity with proteins present in the NR database has been evidenced using BLAST, with the exception of a very high sequence similarity with pUA140_p3 protein (98% sequence identity over 140 amino acids, Table 1). However, pLM7p02 lacks 48 C-terminal amino acid residues with respect to pUA140_p3 and 16 C-terminal amino acid residues with respect to eIF2B. Nonetheless, the core of the structural model of pLM7p02 is essentially (and obviously) identical to that of pUA140_p3, and residues identified in the latter protein as part of the putative kinase recognition and phosphorylation site are all conserved in pLM7p02.

pLM7p03 is a 555 amino acid residues long protein practically identical to that of pUA140_p2, their sequence identity being 99% (Table 1). Thus, the conclusion drawn for pUA140_p2 that it is a membrane-anchored hexameric ATPase involved in the bacterial conjugation process is valid also for pLM7p03.

Similarly also, pLM7p04 displays an amino acid sequence practically identical to that of pUA140_p1. The two proteins have exactly the same length and a sequence identity of 99% (Table 1). Unfortunately, as for pUA140_p1, our analyses do not allow to draw reliable conclusions concerning the three-dimensional structure and function of pLM7p04.

Discussion

The results reported in the present work indicate that a computational approach based on the combination of various bioinformatics strategies can lead to biologically relevant results in the study of bacterial proteins with unknown structure and function and with no sequence similarity with structurally characterized proteins. In detail, a combination of *ab initio* protein structure prediction strategies, using Rosetta and I-TASSER, coupled with homology modeling techniques, allowed to assign a putative function to most of the proteins encoded by the pUA140 and pLM7 plasmids.

All the proteins encoded by the pUA140 plasmid are almost identical to the proteins encoded by the pLM7 plasmid, with the exception of pUA140_p4. This result suggests that pUA140 and pLM7 derive from an ancestral plasmid, which replicated by a rolling circle process, typical of conjugative plasmids, and was transferred to another *S. mutans* cell, eventually leading to the evolution of a different *S. mutans* strain. In line with this observation, pUA140_p2 and pLM7p03 are predicted with a high reliability to be conjugation proteins responsible for ATP binding and DNA transfer through the bacterial membrane. Furthermore, BLAST results reveal a high sequence similarity between these proteins and a *S. thermophilus* plasmidic protein (Table 1), indicating that plasmid transfer happened also among different species of Streptococci. Indeed, gene transfer among different species of oral Streptococci is known to happen with high frequency when these are in close contact with each other [25]. An interesting result of the present study is the identification of *S. mutans* plasmidic proteins, which display significant structural similarity with toxin-antitoxin system proteins. As reported in the Results section, a cluster of catalytic residues similar to that responsible for the RNAse activity of the *E. coli* YoeB toxin was found in pUA140_p5. This finding strongly supports the hypothesis that this plasmidic protein is a member of a toxin-antitoxin system. In addition, pUA140_p4 displays a high structural similarity with the RelB antitoxin and displays amino acid residues that could interact with DNA. Thus a complete

toxin-antitoxin system appears to be encoded by the pUA140 plasmid. A similar system is also encoded by the pLM7 plasmid, although limited to the toxin component, which is pLM7p01. It is likely that the antitoxin is not present on the plasmid genome but on the bacterial chromosome. Indeed, the *S. mutans* chromosome-encoded DNA-damage protein was found to be a close structural homolog of pUA140_p4. Thus, the former protein could represent a chromosome-encoded antitoxin that inhibits the plasmidic pLM7p01 toxin.

The pLM7p02 and pUA140_p3 are proteins whose activity is predicted to be regulated by a casein kinase-like bacterial kinase, as they display a phosphorylation site, that is, Ser69, and a cluster of residues probably involved in the recognition of the kinase protein. In this regard, over the last 20 years, many bacterial and archeobacterial proteins, containing phosphoserine, phosphotyrosine, and phosphothreonine residues, have been discovered [26].

Finally, no significant result regarding pUA140_p1 and pLM7p04 proteins was obtained. One might conclude that either the hypothesis in literature that pUA140_p1 is a rep protein is wrong or, in this case, Rosetta and I-TASSER failed to produce a reliable structural model.

All these results are relevant in order to identify plasmidic proteins that may represent targets for the structure-based development of new antibacterial drugs. From this viewpoint, the most promising targets appear to be the toxin-antitoxin systems and the conjugation proteins, which are likely responsible for conferring important pathogenic features to *S. mutans*. In particular, the toxin-antitoxin systems are essential for the plasmid stabilization, and drugs targeting the antitoxin can be developed to inhibit its association with the toxin, which, in the free state, would exert its toxic activity on the bacterium. Furthermore, conjugation proteins, which pLM7p03 and pUA140_p2 are predicted to be with high reliability, are proteins that are essential for the transfer of plasmidic genes from one *S. mutans* strain to the other and to other *Streptococcus* species, thus contributing to the spread of pathogenicity features among bacterial microorganisms. Thus, these two protein systems can be considered potential pharmacological targets and the structural information obtained in this study can be exploited for molecular docking studies aimed to identify lead compounds able to inhibit their activity.

In conclusion, this study highlights the usefulness of combined molecular modeling approaches in obtaining reliable structural and functional information on bacterial proteins with unknown function and no sequence similarity with structurally characterized proteins. This is particularly important in view of the massive amount of sequence data available on bacterial genomes and the

much lower amount of structural data. The approach presented in this study could, in principle, be extended to larger sets of protein sequences and even to entire viral and/or bacterial proteomes. Furthermore, reliable structural models can be successfully exploited by protein crystallographers in molecular replacement procedures to speed up the resolution of protein structure from X-ray diffraction data. To this aim, we have deposited the models of pUA140_p2 and pUA140_p5in the Protein Models Database (http://mi.caspur.it/PMDB/) with accession numbers PM0077907 and PM0077906, respectively.

References

1. Bowen WH. Dental caries – not just holes in teeth! A perspective. Mol Oral Microbiol 2016;31:228–33.
2. Petersen PE. World Health Organization global policy for improvement of oral health-World Health Assembly. Int Dent J 2008;58:115–21.
3. Jin LJ, Lamster IB, Greenspan JS, Pitts NB, Scully C, Warnakulasuriya S. Global burden of oral diseases: emerging concepts, management and interplay with systemic health. Oral Dis 2016;22:609–19.
4. Koo H, Jeon JG. Naturally occurring molecules as alternative therapeutic agents against cariogenic biofilms. Adv Dent Res 2009;21:63–8.
5. Jeon JG, Klein MI, Xiao J, Gregoire S, Rosalen PL, Koo H. Influences of naturally occurring agents in combination with fluoride on gene expression and structural organization of *Streptococcus mutans*in biofilms. BMC Microbiol 2009;9:228.
6. Moreillon P, Que YA. Infective endocarditis. Lancet 2004;363:139–49.
7. Reference genome: Streptococcus mutans UA159 (https://www.ncbi.nlm.nih.gov/genome/856).
8. Caufield PW, Childers NK, Allen DN, Hansen JB. Distinct bacteriocin groups correlate with different groups of *Streptococcus mutans* plasmids. Infect Immun 1984;48:51–6.
9. Abranches J, Miller JH, Martinez AR, Simpson-Haidaris PJ, Burne RA, Lemos JA. The collagen-binding protein Cnm is required for *Streptococcus mutans* adherence to and intracellular inva-

sion of human coronary artery endothelial cells. Infect Immun 2011;9:2277–84.
10. Zhou X, Caufield PW, Li Y, Qi F. Complete nucleotide sequence and characterization of pUA140, a cryptic plasmid from *Streptococcus mutans*. Plasmid 2001;46:77–85.
11. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
12. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Method Enzymol 2004;383:66–93.
13. Yang Z. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:40.
14. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, et al. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modelling. Proteins 2003;53:430–5.
15. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283–91.
16. Holm L, Kääriäinen S, Rosenström P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics 2008;24:2780–1.
17. Ilyina TV, Koonin EV. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archeobacteria. Nucleic Acids Res 1992;20:3279–85.
18. Gomith-Rüth FX, Moncalián G, Luque-Pérez R, Gonzáles A, Cabezón E, de La Cruz F, et al. The bacterial conjugation protein TrwB resembles ring helicases and F$_1$-ATPase. Nature 2001;409:637–41.
19. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 1998;14:378–9.
20. Dhaliwal S, Hoffman DW. The crystal structure of the N-terminal region of the alpha subunit of translation initiation factor 2 (eIF2alpha) from *Saccharomyces cerevisiae* provides a view of the loop containing serine 51, the target of the eIF2alpha-specific kinases. J Mol Biol 2003;334:187–95.
21. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, et al. The 20 years of PROSITE. Nucleic Acids Res 2008;36:D245–9.
22. Raumann BE, Rould MA, Pabo CO, Sauer RT. DNA recognition by β-sheets in the Arc repressor-operator crystal structure. Nature 1994;367:754–7.
23. Li G, Zhang Y, Inouye M, Ikura M. Structural mechanism of transcriptional autorepression of the *Escherichia coli* RelB/RelE antitoxin/toxin module. J Mol Biol 2008;380:107–19.
24. Kamada K, Hanaoka F. Conformational change in the catalytic site of the ribonucleaseYoeB toxin by YefM antitoxin. Mol Cell 2005;19:497–509.
25. Le Blanc DJ, Hawley RJ, Lee LN, St Martin EJ. "Conjugal" transfer of plasmid DNA among oral streptococci. Proc Natl Acad Sci USA 1978;75:3484–7.
26. Leonard CJ, Aravind L, Koonin EV. Novel families of putative protein kinases in bacteria and Archaea: evolution of the "eukaryotic" protein kinase superfamily. Genome Res 1998;8:1038–47.

Staphylococcus aureus Esx Factors Control Human Dendritic Cell Functions Conditioning Th1/Th17 Response

Melania Cruciani[1,2†], Marilena P. Etna[2†], Romina Camilli[2], Elena Giacomini[2], Zulema A. Percario[1], Martina Severa[2], Silvia Sandini[2], Fabiana Rizzo[2], Valentina Brandi[1], Giuliana Balsamo[3], Fabio Polticelli[1,4], Elisabetta Affabris[1], Annalisa Pantosti[2], Fabio Bagnoli[3] and Eliana M. Coccia[2]*

[1] Department of Science, University Roma Tre, Rome, Italy, [2] Department of Infectious Diseases, Istituto Superiore di Sanità, Rome, Italy, [3] Research Center, GSK Vaccines, Siena, Italy, [4] National Institute of Nuclear Physics, Roma Tre University, Rome, Italy

The opportunistic pathogen *Staphylococcus aureus* (*S. aureus*) is a major cause of nosocomial- and community-acquired infections. In addition, many antibiotic-resistant strains are emerging worldwide, thus, there is an urgent unmet need to pinpoint novel therapeutic and prophylactic strategies. In the present study, we characterized the impact of infection with the pandemic methicillin-resistant USA300 *S. aureus* strain on human primary dendritic cells (DC), key initiators and regulators of immune responses. In particular, among staphylococcal virulence factors, the function of EsxA and EsxB, two small acidic dimeric proteins secreted by the type VII-like secretion system Ess (ESAT-6-like secretion system), was investigated in human DC setting. A comparative analysis of bacterial entry, replication rate as well as DC maturation, apoptosis, signaling pathway activation and cytokine production was performed by using wild type (wt) USA300 and three isogenic mutants carrying the deletion of *esxA* (Δ*esxA*), *esxB* (Δ*esxB*), or both genes (Δ*esxAB*). The *S. aureus* mutant lacking only the EsxA protein (Δ*esxA*) stimulated a stronger pro-apoptotic phenotype in infected DC as compared to wt USA300, Δ*esxAB,* and Δ*esxB* strains. When the mutant carrying the *esxB* deletion (Δ*esxB*) was analyzed, a higher production of both regulatory and pro-inflammatory mediators was found in the infected DC with respect to those challenged with the wt counterpart and the other *esx* mutants. In accordance with these data, supernatant derived from Δ*esxB*-infected DC promoted a stronger release of both IFN-γ and IL-17 from CD4+ T cells as compared with those conditioned with supernatants derived from wild type USA300-, Δ*esxAB*-, and Δ*esxA*-infected cultures. Although, the interaction of *S. aureus* with human DC is not yet fully understood, our data suggest that both cytokine production and apoptotic process are modulated by Esx factors, thus indicating a possible role of these proteins in the modulation of DC-mediated immunity to *S. aureus*.

Keywords: *S. aureus*, Ess, dendritic cells, vaccine, cytokine, apoptosis, Th response

INTRODUCTION

Staphylococcus aureus (*S. aureus*) is a highly adaptive Gram-positive coccus that can be either a human commensal or a potentially lethal opportunistic pathogen. Approximately one third of the entire human population worldwide is asymptomatically colonized by *S. aureus*, which mainly inhabits epithelial surfaces (Missiakas and Schneewind, 2016). However, under certain circumstances, this bacterium can eventually become a life-threatening pathogen causing infections in skin and soft tissues, sepsis, endocarditis, osteomyelitis, pneumonia, and toxic shock syndrome (David and Daum, 2010; Olaniyi et al., 2016). *S. aureus* is responsible for infections both in the healthcare setting and in the community and, also due to its impressive capacity to develop antibiotic resistance, constitutes a major burden on human society (Monaco et al., 2016). Indeed, the increasing frequency of both methicillin-resistant *S. aureus* strains (MRSA) and strains with decreased susceptibility to vancomycin has complicated disease treatment (Gould et al., 2012), thus the need to develop new therapeutic strategies and vaccines against this dangerous pathogen becomes urgent.

The remarkable ability of *S. aureus* to cause a wide range of infections is related to its extensive armamentarium of virulence factors, which includes secreted proteins, cell wall anchored proteins and cell surface components that allow the bacteria to adhere to cell surface, invade or escape host immune system and cause harmful toxic effects (Gordon and Lowy, 2008; Bien et al., 2011; Pozzi et al., 2015). In addition, several evidences showed that *S. aureus* possesses a specialized secretion system, namely ESAT-6-like secretion system (Ess), similar to the ESX-1 secretion system described in *Mycobacterium tuberculosis* (Sundaramoorthy et al., 2008; Anderson et al., 2013), which contributes to *S. aureus* virulence and pathogenesis (Burts et al., 2005, 2008). Ess is encoded within the *ess* gene cluster, a region highly conserved in the genomes of both community- and hospital-acquired *S. aureus* strains (Korea et al., 2014). Ess consists of structural components, EssA, EssB, EssC, EssD, EssE, and accessory molecules, including the membrane protein EsaA and the cytosolic proteins EsaB and EssI (Anderson et al., 2016; Warne et al., 2016; Ohr et al., 2017). Secreted substrates of the *S. aureus* Ess machinery include EsxA and EsxB, two small acidic dimeric proteins carrying a distinctive WXG motif, and two additional proteins, EsxC and EsxD, which lack the WXG motif (Jager et al., 2016). Moreover, recent evidences showed that Ess may also secrete the structural factor EssD, which acts as a nuclease to cleave DNA and whose activity is inhibited by EssI (Ohr et al., 2017).

Among the Ess secreted substrates, EsxA and EsxB show similarity to ESAT-6 and CFP-10, two well characterized virulence factors encoded by *M. tuberculosis* (Burts et al., 2005), and are reported to be bacterial effectors that are involved in host-pathogen interaction (Groschel et al., 2016). Mutants unable to secrete EsxA and EsxB displayed defects in *S. aureus*-mediated abscess formation and persistence in host tissue (Burts et al., 2008). Moreover, Ess has also been shown to manipulate host immune responses by modifying the production of specific cytokines (Anderson et al., 2016) and to impact nasal colonization and pneumonia, although in a strain-dependent manner (Kneuper et al., 2014). The mechanisms by which EsxA and EsxB exert their pathogenic function remains to be elucidated, although a recent study performed in epithelial cells indicates a role in the modulation of apoptosis and release of intracellular *S. aureus* from the host cell (Korea et al., 2014). In the present study, we further investigated the mechanism of action of EsxA and EsxB using an *in vitro* human dendritic cells (DC)-based infection model. These studies suggested that Esx factors might modulate DC apoptosis and cytokine production and, in turn, T cell expansion, highlighting the key role played by these factors in regulating the host immune response.

MATERIALS AND METHODS

DC Preparation and Infection

Istituto Superiore di Sanità Review Board approved the present research project (CE/13/387). DC were prepared as previously described (Etna et al., 2015). DC were generated by culturing $CD14^+$ monocytes (purified from human peripheral blood mononuclear cells of anonymous healthy blood donors) with 50 ng/ml GM-CSF (R&D Systems, Minneapolis, MN, USA) and 200 U/ml IL-4 (Miltenyi, Bergisch Gladbach, Germany) for 5 days at 0.5×10^6 cells/ml in RPMI 1640 (BioWhittaker Europe, Verviers, Belgium) supplemented with 2 mM L-glutamine (BioWhittaker) and 15% Fetal Bovine Serum (FBS) (Lonza, Basel, Switzerland). At day 5, cells were tested for their differentiation status by evaluating CD1a expression (>90% $CD1a^+$) and lack of CD14 (>95% $CD14^-$). Before infection, the medium was replaced with RPMI without antibiotics and supplemented with 2 mM L-glutamine and 15% FBS. Cytokine deprivation did not affect DC survival rate, which was >90%.

DC were then infected with *S. aureus* cultures grown as described above using a multiplicity of infection of 0.1 bacteria/cell. After 2 h of infection, extracellular bacteria were killed by adding gentamicin (50 μg/ml, Sigma Aldrich) for 2 h at 37°C. Then, DC cultures were washed with growth medium before plating in antibiotic-free medium.

Bacterial Strains and Growth Conditions

S. aureus wt USA300 strain and three isogenic mutants, namely Δ*esxAB*, Δ*esxA*, and Δ*esxB* carrying the deletions in both *esxA* and *esxB*, or only *esxA* or *esxB* genes, respectively, were used to infect DC. The cloning strategy has been previously described (Korea et al., 2014). Briefly, Δ*esxA and* Δ*esxB* single deleted mutants were generated by cloning the respective *esx* gene into the *Escherichia coli-S. aureus* shuttle/suicide vector pKOR1; while

esxA gene was deleted from Δ*esxB* strain to obtain the Δ*esxAB* double mutant. All isogenic mutants were tested for the correct expression of *esxA* and *esxB* genes by PCR, using external primers targeting flanking regions, and by sequencing the genomic region corresponding to the deleted genes (*esxA* and *esxB*), including upstream and downstream regions (500–700 bp; Korea et al., 2014). Green Fluorescent Protein (GFP)-expressing bacterial strains were instead obtained by cloning pOS1CK-GFP vector into wt and *esx* mutants (Korea et al., 2014).

Fresh bacterial preparations were cultured for each experiment. Briefly, wt USA300 and *esx* mutants were grown in tryptic soy broth (Becton Dickinson, Sparks, MD, USA) overnight at 37°C. The next day, bacterial broths were diluted 1:100 in fresh tryptic soy broth, cultured until the exponential phase of growth (OD_{600} of 0.6), and then washed in RPMI 1640 and re-suspended in RPMI 1640 supplemented with L-glutamine (2 mM) and 15% FBS for DC infection.

Colony Forming Units (CFU) Count

DC were infected with wt USA300, Δ*esxAB*, Δ*esxA*, and Δ*esxB* at multiplicity of infection of 0.1. At the indicated time points, cells were harvested and washed in RPMI 1640 by centrifugation at $150 \times g$ for 10 min to selectively spin down cells, while extracellular bacteria remained in the supernatants. Collected cells were lysed with distilled water containing 0.1% saponin for 5 min at room temperature and then plated at serial dilutions on tryptic soy agar (Becton Dickinson) plates to obtain the number of CFU/ml.

Confocal Microscopy

DC were infected for 4 h with wt USA300 and *esx* mutants expressing GFP. After infection, DC were harvested, washed and then fixed with 4% paraformaldehyde. Cells were subsequently washed with PBS and incubated with phalloidin-Tritc (Sigma Aldrich) for 1 h to highlight the actin filaments (red). Nucleic acids (blue) were labeled with far-red fluorescent dye RedDot™2 (Biotium, Fremont, CA, USA) following an incubation of 30 min in PBS solution. Vectashield Antifade Mounting Medium (Vector Laboratories, Burlingame, CA, USA) diluted at 80% in PBS was used to prepare specimens for confocal microscopy observation. Images were acquired with Leica TCS SP5 confocal microscope and processed with LAS AF software (version 1.6.3, Leica Microsystems). Objective 63.0X. Lasers activated: He/Ne laser at 543 nm to phalloidin-Tritc's excitation, He/Ne laser at 633 nm to dye RedDot™2's excitation and Argon laser at 488 nm to visualize GFP-*S. aureus* (green). Images were acquired activating single laser in sequential mode to prevent fluorescence overlay. Image magnification 1575X, pixels 1,024 × 1,024.

Antibodies and Other Reagents

Monoclonal antibodies (Abs) specific for cluster of differentiation (CD)1a, CD14, CD38, CD86, CD83, HLA-DR, IgG1, IgG2a (BD Bioscience, San Diego, CA, USA), and Annexin V (Abcam, Cambridge, UK) were used as direct conjugates to Fluorescein Isothiocyanate (FITC), Phycoerythrin (PE) as needed. To exclude dead cells from analysis Fixable Viability Dye eFluor®780 (FvDye, eBioscience, San Diego, CA, USA) was used.

For immunoblotting analysis rabbit anti-phospho-NF-kB p65 (Ser536), rabbit anti-phospho-p44/42 MAPK (Thr202/Tyr204), rabbit anti-phospho-p38 MAPK (Thr180/Tyr182) (Cell Signaling Technology, Danvers, MA), mouse anti-β-actin (Sigma-Aldrich, St. Louis, MO, USA) and horseradish peroxidase-conjugate anti-mouse (Santa Cruz Biotechnology, Dallas, TX, USA) and anti-rabbit (Santa Cruz Biotechnology) secondary Abs were used. Staurosporine (1 μM, Sigma Aldrich) was utilized as pharmacological inducer of apoptosis.

Flow Cytometry Analysis

Cells (10^5) were washed once in PBS containing 2% FBS and incubated with indicated Abs at 4°C for 30 min. DC were then washed and fixed with 2% paraformaldehyde (Panreac Quimica, Castellar del Valles, Spain) before analysis on Gallios cytometer (Beckman Coulter, Brea, CA, USA). A total of 30,000 events were analyzed per sample in FvDye negative live cells. The expression of cell surface molecules was evaluated in viable DC using the median fluorescence intensity (MFI) after subtraction of the respective isotype control Ab values.

Apoptosis Detection

Phosphatidylserine exposure and membrane integrity were analyzed by using Annexin V-FITC and FvDye according to manufacturing protocols. Twenty-four hours post infection, DC were stained with FvDye and then labeled for 15 min with 5X Annexin V-FITC, as previously described (Etna et al., 2014). Finally, cells were fixed overnight with 4% paraformaldehyde before analysis on Gallios cytometer (Beckman Coulter). Data were analyzed by Kaluza software (Beckman Coulter).

Cytokine Determination

Supernatants of DC cultures were harvested 24 h after infection, filtered with 0.2 μm devices and then stored at −80°C. The production of IL-12, TNF-α, IL-10, IL-1β, IL-6, and IL-8 was measured by human Inflammatory Cytokine kit (Cytometric Bead Array, BD Bioscience), while IL-23 release was quantified by ELISA (R&D Systems). Supernatants from stimulated CD4$^+$ T cells were collected after 5 days, filtered and stored at −80°C. IFN-γ and IL-17 production was determined by specific ELISA kits (R&D Systems).

Immunoblot Analysis

Western blot were performed as previously described (Etna et al., 2015). Briefly, 25 μg of total protein extracts were separated on 10% SDS-PAGE gel and blotted onto nitrocellulose membranes (Millipore). Blots were incubated with rabbit polyclonal anti-phospho-p38 MAPK (Thr180/Tyr182) (p-p38), rabbit polyclonal anti-phospho-p44/42 MAPK (Thr202/Tyr204) (p-p44/42) and rabbit polyclonal anti-phospho-NF-kB p65 (Ser536) (p-p65) Abs (Cell Signaling Technology, Danvers, MA). Detection was achieved using anti-rabbit horseradish peroxidase-conjugate secondary Ab (Santa Cruz Biotechnology), and visualized with Enhanced Chemiluminescence plus kit (GE Healthcare Bio-Sciences, Pittsburgh, PA, USA). A ChemiDoc XRS (Bio-Rad, Hercules, CA, USA) instrument and ImageLab software (Bio-Rad) were used to reveal and analyze the chemiluminescence

signal. For loading control, β-actin levels were quantified by using a mouse anti-β-actin Ab (Sigma Aldrich). Protein amount was normalized to the actin level by ImageLab software. Fold changes of each analyzed protein were calculated by dividing values obtained in infected conditions by those of the uninfected counterpart.

T Cell Activation and Expansion

Total CD4$^+$ T cells were isolated from freshly collected buffy coat as previously described (Etna et al., 2014). Briefly, CD4$^+$ T cells were purified by indirect magnetic sorting with a CD4$^+$ T-cell isolation kit (Miltenyi, Teterow, Germany) from allogeneic PBMC. Purified cells were plated in 96-well U-bottomed tissue culture plates at the density of 0.4×10^6 cells/ml with supernatants obtained from control or 24 h-infected DC. Conditioned T cells were stimulated with anti-CD3/CD28 beads (Invitrogen life technology, Carlsbad, CA, USA) (ratio 1:1) for 5 days and then supernatants were harvested for IFN-γ and IL-17 detection.

Bioinformatics Analysis

Prediction of the three-dimensional structure of EsxB has been performed by using *Rosetta* (http://robetta.bakerlab.org/queue.jsp) and I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) software packages (Supplementary Figure 1). *Rosetta* parses the sequence into putative domains, looking for regions into the query sequence that are homologous to experimentally determined structures. Then proceeds with multiple sequence alignment (MSA) to predict putative domains. If a confident match to a protein of known structure is found, that is used as a template for comparative modeling. If no match is found, structure predictions are made using the *de novo* Rosetta fragment insertion method (Kaufmann et al., 2010). Once the protein chain is completely assembled, the side-chains of the final model are repacked using a Monte Carlo algorithm with a backbone-dependent sidechain rotamers library (Kaufmann et al., 2010). I-TASSER identifies template proteins, which are predicted to display a fold similar to that of the protein of interest using several threading techniques implemented in LOMETS. Structure fragments excised from the templates are then assembled through replica-exchange Monte Carlo simulations. Finally, the models obtained are iteratively refined to optimize their free energy and global topology (Roy et al., 2010).

Dali server has been used to detect structural homologs of EsxA and EsxB in the Protein Data Bank (see Supplementary Materials). The program has retrieved a list of 500 structural neighbors, ranked by highest Z-score, and the alignment data (Holm and Rosenstrom, 2010). Among all structural homologs, only human proteins have been further analyzed.

Human interactors of the proteins selected by the Dali server have been identified using the PrePPI server (http://bhapp.c2b2.columbia.edu/PrePPI). PrePPI combines the prediction of protein-protein interactions (PPI) with those coming from experimental evidence (Zhang et al., 2013). Predicted interactions are derived from calculated likelihood ratios (LRs) by combining structural, functional, evolutionary and expression information,

with the most important contribution coming from structure. The experimentally determined interactions are collected from six publicly available databases (MIPS, DIP, IntAct, MINT, HPRD, and BioGRID) that manually collect PPIs from the literature and are also assigned LRs. A final probability is then assigned to every interaction by combining the LRs for both predicted and experimentally determined interactions (Zhang et al., 2013).

cNLS Mapper (http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi) has been used for the prediction of the classical importin-alpha/beta nuclear localization sequence (NLS) in EsxA and EsxB amino acid sequences. cNLS mapper calculates NLS activity scores by using activity-based, but not sequence-based, profiles for different classes of importin-α-dependent NLSs. Higher scores indicate stronger NLS activities. Score values between 8 and 10 indicate that the protein is exclusively localized in the nucleus, score values of 7 or 8 that it is partially localized in the nucleus, score values between 3 and 5 that it is localized both in the nucleus and in the cytoplasm, and score values between 1 and 2 that it is localized exclusively to the cytoplasm (Kosugi et al., 2009).

Protein-Protein docking simulations to evaluate the feasibility of formation of EsxA-EsxB heterodimers and heterotetramers were performed using the ZDOCK online server (http://zdock.umassmed.edu, version 3.0.2), which is based on the rigid-body protein-protein docking program ZDOCK (Pierce et al., 2014). ZDOCK implements a Fast Fourier Transform algorithm and a scoring system based on a combination of shape complementarity, electrostatics and statistical potential terms (Pierce et al., 2014). The 2,000 complexes predicted by ZDOCK were re-ranked using ZRank (Pierce and Weng, 2007), a program that rescores the ZDOCK predictions using a more detailed potential which includes electrostatics, van der Waals, and desolvation terms (Pierce and Weng, 2007). Results of the above described bioinformatics analyses are reported in the Supplementary Material.

Statistical Analysis

Statistical analysis was carried-out using a two-tailed Student's *t*-test for paired data. A $P < 0.05$ was considered statistically significant.

RESULTS

Infection of Human DC with wt USA300 and *esx* Deletion Mutants

In light of the importance of DC in the surveillance of peripheral sites and in the induction of an effective immune response, we sought to explore the role of *S. aureus* EsxA and EsxB proteins by infecting DC with the isogenic Δ*esxA* and Δ*esxB* single mutants as well as the Δ*esxAB* double mutant generated in *S. aureus* USA300 background (Korea et al., 2014).

In agreement with a previous report demonstrating the presence of *S. aureus* inside DC or on their cell surface (Jin et al., 2014), we observed that human DC actively phagocytized *S. aureus* (**Figure 1**). In particular, the comparative analysis of the intracellular CFU number showed that DC

efficiently internalized wt USA300 as well as the isogenic *esx* mutants (**Figure 1A**). A time course analysis showed that CFU counts within DC increased up to 2 h post infection and then gradually decreased during the following 22 h. Importantly, CFU counts showed that the number of the internalized wt USA300 and mutant strains was similar at all analyzed time points, suggesting that EsxA and EsxB do not directly affect the ability of *S. aureus* to infect DC cultures. Confocal microscopy experiments performed with GFP-expressing wt USA300 and *esx* recombinants, confirmed CFU data showing no differences among the internalized bacteria (**Figure 1B**).

Induction of DC Maturation and Apoptosis in Response to the Infection with wt USA300 and *esx* Deletion Mutants

Having found that wt USA300 and *esx* deletion mutants displayed a similar ability to infect human DC cultures, next we evaluated how the infection impacts on DC immune-phenotype. In particular, DC were infected for 24 h with wt USA300, ΔesxAB, ΔesxA, and ΔesxB mutants and, then, the surface expression of the co-stimulatory molecule CD86, the maturation marker CD83, the MHC class II molecule HLA-DR and the activation marker CD38 was examined by flow cytometry. All strains induced a robust expression of the analyzed molecules as compared to the uninfected counterpart irrespective of the presence of Esx-encoded factors (**Figure 2**), suggesting that Esx proteins do not interfere with this process.

Given the importance of the apoptosis in antigen presentation and, consequently, in the commitment of T cell response, we analyzed the percentage of Annexin-V and FvDye labeled DC 24 h after infection with wt USA300 or ΔesxAB, ΔesxA, and ΔesxB mutants.

Infection with all analyzed strains increased the percentage of both early (annexin$^+$/FvDye$^-$) and late (annexin$^+$/FvDye$^+$) apoptotic cells as compared to the uninfected counterpart (**Figure 3A**). However, ΔesxA mutant infected DC displayed a significantly stronger apoptotic phenotype respect to wt USA300 infected cultures (**Figures 3B,C**).

Cytokine Production by DC Infected with wt USA300 and *esx* Deletion Mutants

In an attempt to evaluate whether the presence or absence of Esx proteins in USA300 bacteria would affect the ability of DC to secrete soluble immune mediators, DC culture supernatants were collected 24 h after infection with wt USA300 or the *esx* isogenic mutants, and the cytokine production was analyzed. Both wt USA300 and all isogenic mutants drove production of regulatory cytokines (IL-23, IL-12, and IL-10) and pro-inflammatory mediators (TNF-α, IL-6, and IL-1β), as well as the chemokine IL-8, although at different extent (**Figure 4**). In particular, the single *esxB* deletion conferred to USA300 an
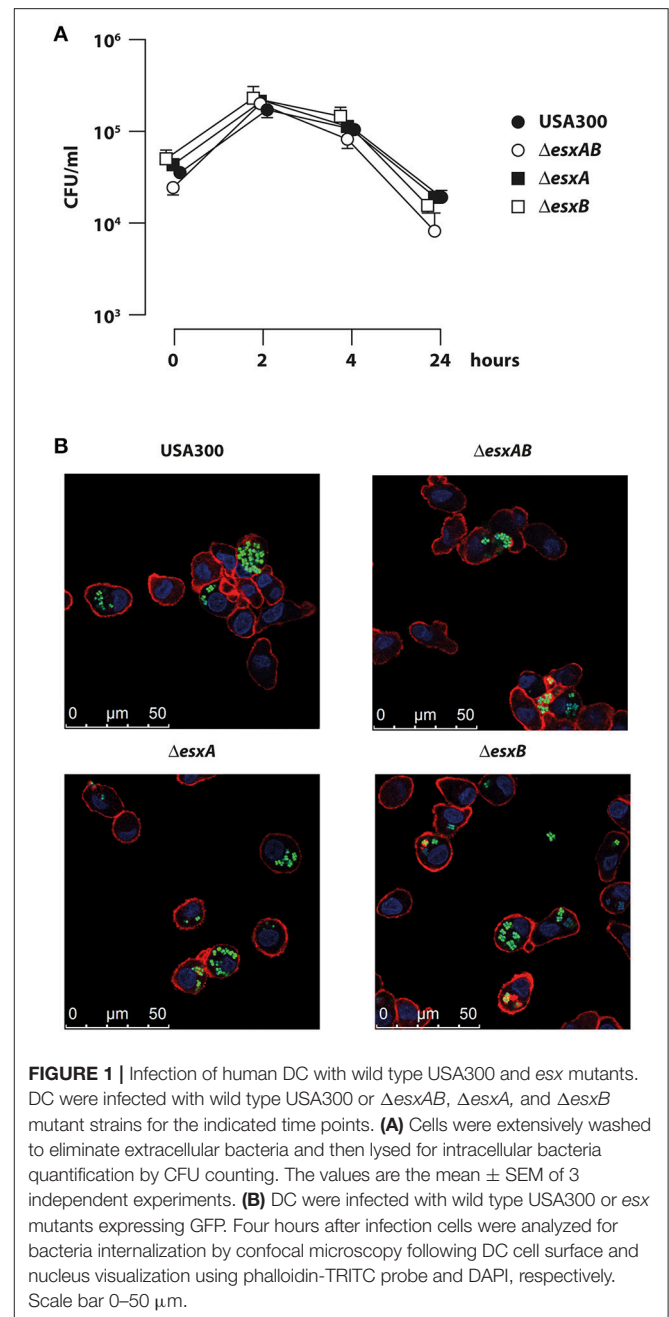
FIGURE 1 | Infection of human DC with wild type USA300 and *esx* mutants. DC were infected with wild type USA300 or ΔesxAB, ΔesxA, and ΔesxB mutant strains for the indicated time points. **(A)** Cells were extensively washed to eliminate extracellular bacteria and then lysed for intracellular bacteria quantification by CFU counting. The values are the mean ± SEM of 3 independent experiments. **(B)** DC were infected with wild type USA300 or *esx* mutants expressing GFP. Four hours after infection cells were analyzed for bacteria internalization by confocal microscopy following DC cell surface and nucleus visualization using phalloidin-TRITC probe and DAPI, respectively. Scale bar 0–50 μm.

increased capacity to promote the release of IL-12, TNF-α, IL-6, and IL-1β (**Figures 4B,E–G**).

Given the differences observed in cytokine production, we investigated the signaling pathways involved in the transcriptional regulation of these factors in response to the infection by wt USA300 and *esx* deletion mutants. In particular, MAPK and NF-kB activation was explored in DC challenged with either wt USA300 or *esx* mutants. In accordance with the observed modulation of DC cytokine milieu, phosphorylation of the MAPK p38 was induced 1 h after infection with both wt USA300 and mutants, although to a different extent. Indeed, while ΔesxAB recombinant poorly
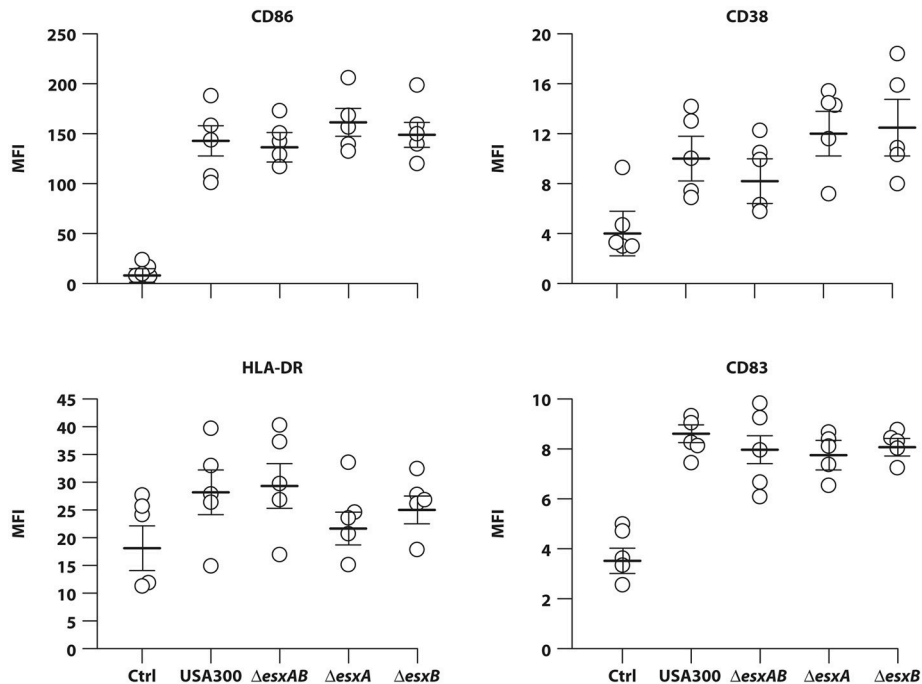
FIGURE 2 | DC maturation in response to infection with wild type USA300 or ΔesxAB, ΔesxA, and ΔesxB mutants. DC were left untreated (Ctrl) or infected for 24 h with wild type USA300 or with USA300 mutants deleted in the *esx* factors. Surface expression of CD86, CD38, HLA-DR, and CD83 was examined by cytometric analysis in 5 independent experiments and graphed by calculating the mean fluorescence intensity (MFI) after the subtraction of the isotype Ab controls. MFI values for each individual experiment and mean MFI ± SEM are shown.

enhanced p38 phosphorylation, ΔesxB mutant resulted to be the strongest inducer as compared to wt USA300 and the other recombinant strains (**Figure 5A**). The p38 activation profile correlates with the down-stream phosphorylation of the NF-kB p65 subunit. Indeed, while ΔesxAB mutant weakly stimulated p65 activation, both ΔesxA and ΔesxB recombinants induced a robust phosphorylation of this NF-kB subunit (**Figure 5A**).

At 3 h post-infection both p38 and p65 activation was lowered in DC infected with the ΔesxB strain as compared to the other mutants, who displayed a sustained phosphorylation of both molecules (**Figure 5B**).

Finally, no difference was found in p42/44 (ERK1/2) phosphorylation 1 and 3 h after the infection with all analyzed *S. aureus* strains, although a slight increase of p42 phosphorylation was observed in ΔesxA-infected DC (**Figure 5**).

Despite the observed differences are not statistically significant, wt USA300 and *esx* mutants may stimulate intracellular signaling activation and, consequently, cytokine production in DC with different kinetics and to a different extent.

Influence of wt USA300 and *esx* Mutant-Induced Cytokine Milieu on IFN-γ and IL-17 Release from CD4[+] T Cells

Finally, to investigate the impact of cytokine milieu released by infected DC on the expansion of IFN-γ and IL-17 producing T cells, we stimulated CD4[+] T cells with anti-CD3/CD28 beads

in presence of supernatants obtained from heterologous DC cultures infected for 24 h with either wt USA300 or ΔesxAB, ΔesxA, and ΔesxB mutants.

In accordance with the high level of IL-12 produced in response to both wt USA300 and *esx* mutant infection (see **Figure 4B**), IFN-γ was strongly released by T cells in all the analyzed experimental conditions (**Figure 6A**). Furthermore, IFN-γ secretion was significantly higher when T cells were stimulated with supernatants derived from ΔesxB-infected DC than what induced by the other supernatants (**Figure 6A**). A similar trend was observed for Th17 expansion (**Figure 6B**), since the highest IL-17 production was observed in CD4[+] T cell culture conditioned by supernatants of ΔesxB-infected DC.

Collectively, these data indicate that *esxB* deletion confers to USA300 a stronger capacity to promote in DC a cytokine production optimal for the expansion of potentially protective Th1/Th17 response as compared to both wt and the other deletion mutants.

DISCUSSION

The specialized type VII secretion system found in Gram-positive *Actinobacteria* and *Firmicutes*, plays a variety of roles in bacterial physiology and pathogenesis (Green and Mecsas, 2016). The effector molecules transported by this secretion system are often involved in a wide range of important functions, such as survival in different environment, manipulation of the host response and
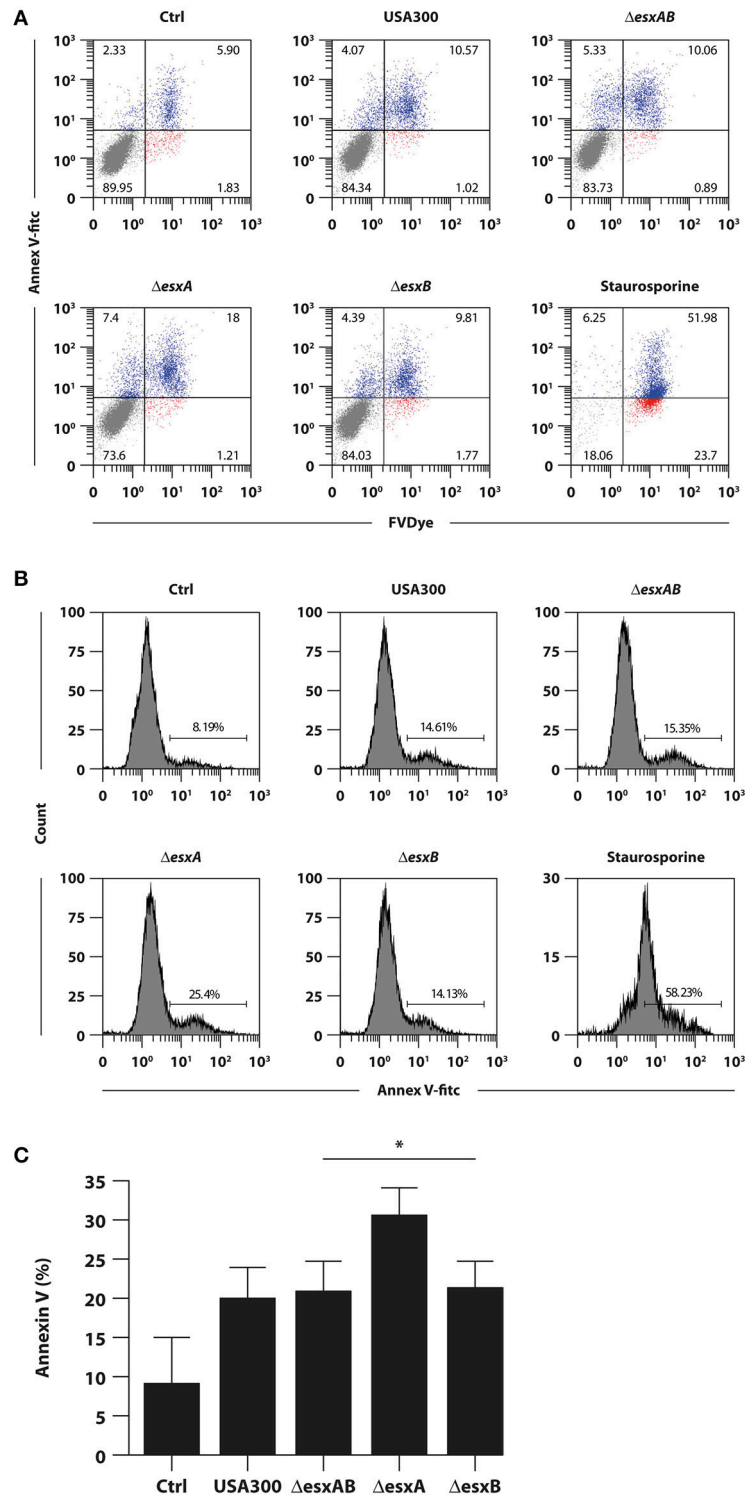
FIGURE 3 | Impact of wild type USA300 and *esx* mutants on DC apoptosis. DC were left untreated (Ctrl) or infected for 24 h with wild type USA300 or ΔesxAB, ΔesxA, and ΔesxB recombinants. **(A)** Apoptosis was assessed by cytofluorimetric analysis through Annexin-V and FvDye staining. As internal control DC were treated with 1 μM staurosporine for 16 h. A representative experiment, out of 3 independent experiments performed that yielded similar results, is shown. Numbers in the dot plots correspond to the percentage of cells in each quadrant. **(B)** The expression of Annexin-V for the different experimental conditions was reported as the histogram of fluorescence intensity. Numbers in each panel correspond to the percentage of total Annexin-V positive cells. **(C)** Histogram of the means of Annexin-V positive cell percentage ± SEM of 3 independent experiments. Statistical significance was calculated by Student's *t*-test for paired data. *$P = 0.004$ (wild type USA300- vs. ΔesxA recombinant-infected DC). Statistics refers to strains at the extremity of the horizontal lines.
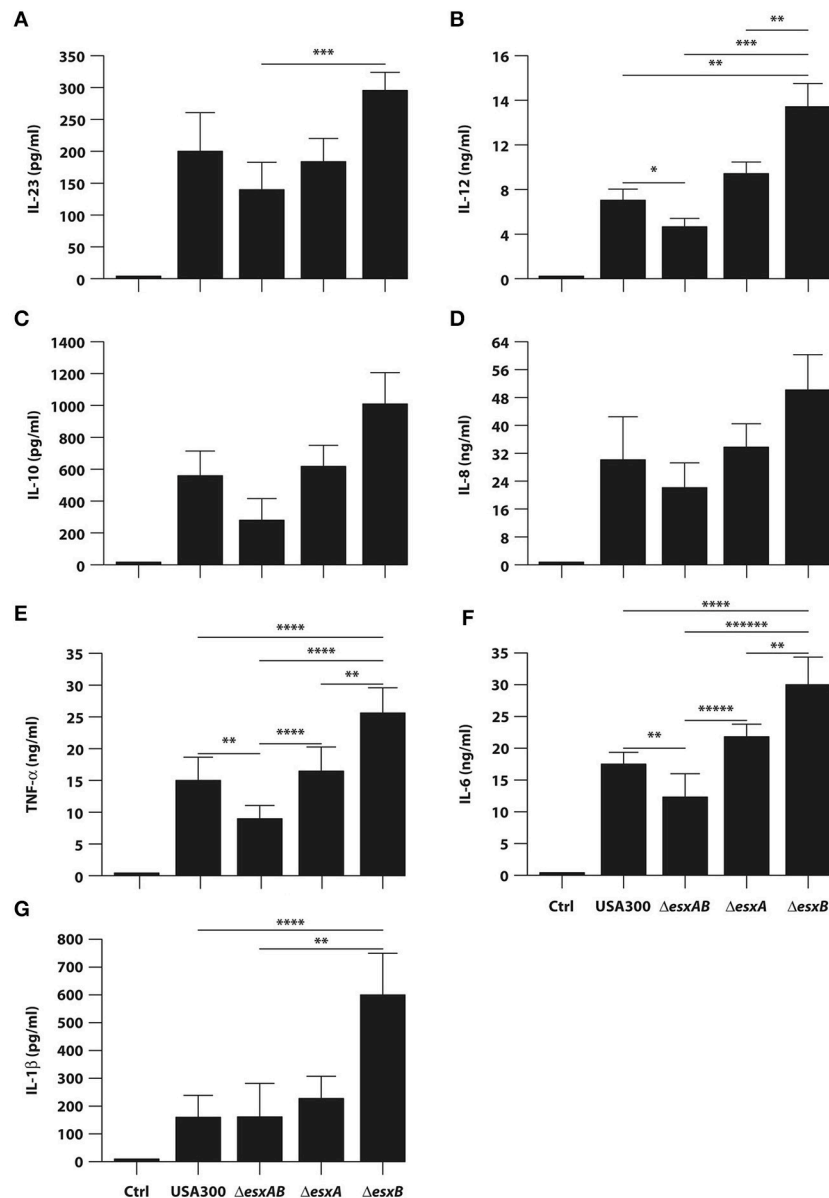
FIGURE 4 | Cytokine production by DC infected with wild type USA300 or *esx* mutants. DC were left untreated (Ctrl) or infected for 24 h with wild type USA300 or *ΔesxAB*, *ΔesxA*, and *ΔesxB* mutants. **(A–G)** IL-23 production was measured in DC culture supernatants by ELISA while, the secretion of IL-12, IL-10, IL-8, TNF-α, IL-6, and IL-1β was instead evaluated by Inflammatory Cytokine array kit. The results represent means ± SEM of 6 independent experiments. Statistical Significance was calculated by Student's *t*-test for paired data. $^{*}P = 0.003$; $^{**}P < 0.05$; $^{***}P < 0.01$; $^{****}P < 0.02$; $^{*****}P = 0.03$; $^{******}P = 0.006$. Statistics refers to strains at the extremity of the horizontal lines.

establishment of a replicative niche (Groschel et al., 2016; Bottai et al., 2017).

The first identified and the best characterized type VII machinery is the mycobacterial ESX-1 system (Bottai et al., 2017). In the context of *M. tuberculosis* infection, several cellular events have been demonstrated to be manipulated by the activity of a functional ESX-1 system: inhibition of phagosomal maturation and acidification (Deretic et al., 2009), cytosolic access, autophagy block (Romagnoli et al., 2012), host cell-death

(Derrick and Morris, 2007), modulation of the inflammatory (Stanley et al., 2003), and adaptive immune response (Etna et al., 2015), indicating that this secretion apparatus constitutes a major virulence factor (Green and Mecsas, 2016). Less defined is the role played by the *S. aureus* Ess in the modulation of host immune response. Notwithstanding the type VII-like Ess of *S. aureus* and the mycobacterial ESX-1 system are phylogenetically distant, these systems share two types of conserved elements: the membrane-bound ATPase of the FtsK/SpoIIIE that forms
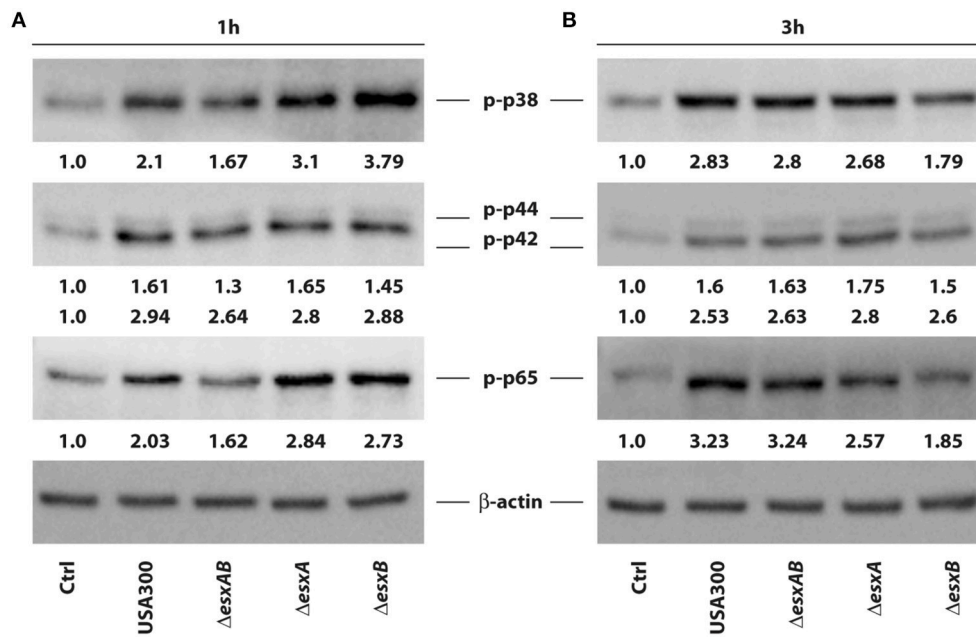
FIGURE 5 | Activation of p38, p44/42 MAPK and NF-kB p65 in DC infected with wild type USA300 or *esx* mutants. The activation of intracellular pathway involving p38, p42/44 MAPK, and NF-kB p65 subunit was investigated by western blotting analysis of the phosphorylated isoforms in whole cell extracts prepared from DC infected for 1 **(A)** and 3 h **(B)** with wild type USA300 or *esx* mutants. β-actin levels were analyzed as control for protein loading. Fold changes of each analyzed protein, indicated at the bottom of each immunoblot, were calculated by dividing values obtained in infected conditions with those of the uninfected counterpart. A representative experiment, out of 3 independent experiments performed that yielded similar results, is shown.
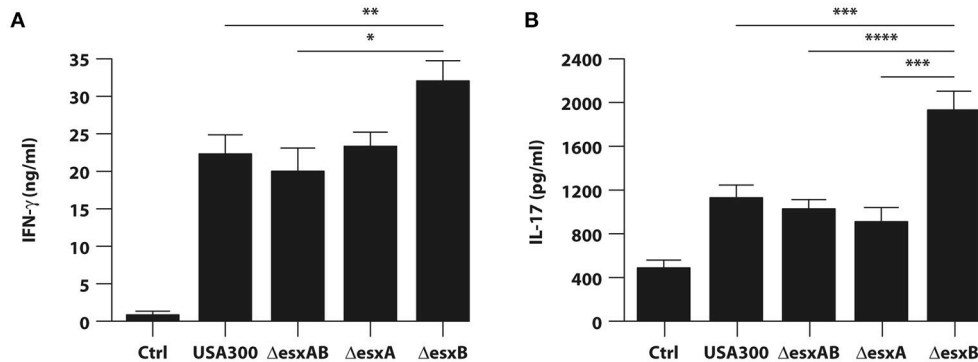
FIGURE 6 | Expansion of IFN-γ or IL-17 producing T cells by cytokine milieu released from DC infected with wild type USA300 and *esx* mutants. Total CD4[+] T cells were stimulated with anti-CD3/CD28 magnetic beads in presence of supernatants derived from DC cultures untreated or infected for 24 h with wild type USA300 or *esx* mutants. IFN-γ **(A)** and IL-17 **(B)** production was evaluated in harvested supernatants by ELISA. The results shown represent means ± SEM of 4 independent experiments. Significance was calculated by Student's *t*-test. [*]$P = 0.006$; [**]$P = 0.01$; [***]$P = 0.02$; [****]$P = 0.05$.

hexameric membrane-spanning pores, and the secreted WXG100 proteins (Jager et al., 2016).

Based on this evidence and having found that mycobacterial ESX-1 system profoundly impacts on human DC functions (Romagnoli et al., 2012; Etna et al., 2015), here we sought to investigate how and whether *S. aureus* Ess is able to play a similar role.

In the context of *S. aureus*-DC interplay, some analogies were observed with the mycobacterial counterpart with respect to the ability to modulate the apoptotic process (Derrick and

Morris, 2007; Choi et al., 2010; Aguilo et al., 2013; Etna et al., 2015). Indeed, the *S. aureus* mutant lacking the expression of EsxA factor significantly promotes apoptosis in DC with respect to wt USA300 strain **(Figure 3)** in agreement to previous data by Korea and collaborators indicating an inhibitory role of EsxA protein on the apoptotic process in human epithelial cells (Korea et al., 2014). Nevertheless, unexpected data were obtained when Δ*esxAB* double mutant was used to infect DC since the apoptotic process was not impaired but even induced at a level similar to that present in wt USA300-infected DC **(Figure 3)**

likely suggesting a not yet characterized role of EsxB factor in controlling the apoptotic process. This result was also in contrast with the increased apoptosis observed by Korea and collaborators in Δ*esxAB* mutant-infected epithelial cells (Korea et al., 2014). A possible explanation for this opposing data may rely on the different cell type in which the apoptotic process was analyzed, suggesting that the Esx factors could play a different role depending on the infected cell.

Another important finding of our study relies on the differential expression of regulatory and inflammatory cytokines (**Figure 4**), in spite of the absence of significant differences in the induction of DC maturation in response to the infection with wt USA300 and *esx* mutant strains (**Figure 2**). Indeed, all strains induced a robust expression of the analyzed maturation markers as compared to the uninfected counterpart irrespective of the presence of Esx-encoded factors suggesting that Esx proteins do not interfere with this process (**Figure 2**). Instead, our comparative analysis of cytokine production revealed that the induction of both regulatory (IL-12) and pro-inflammatory (TNF-α, IL-6, IL-1β) cytokines was higher in DC infected with the isogenic mutant lacking *esxB* expression respect to wt USA300 or the other recombinant strains (**Figure 4**). Surprisingly, this cytokine induction was reduced in DC cultures infected with Δ*esxAB* mutant (**Figure 4**). This unexpected and apparently counterintuitive result may be explained by the reported capacity of *S. aureus* to express other Esx factors (Anderson et al., 2013). In particular, Anderson and collaborators have described that, although the deletion of either *esxA* or *esxB* blocks the secretion of all Esx factors in USA300, different effects on Esx production were observed: the absence of *esxA* does not affect the expression of EsxB, EsxC, and EsxD while the deletion of *esxB* impairs only the production of EsxD (Anderson et al., 2013).

Moving from Anderson's data, we could speculate that Esx proteins could be released by dying bacteria and, once in the cell cytosol, they might interact with host factors/sensors. Thus, the different combination of multimeric complexes or single Esx subunits, which are present in knockout mutants, might contribute to the DC phenotype observed in this manuscript. Therefore, by tuning Esx balance, *S. aureus* could modulate the host cytokine production and programmed cell death (Fraunholz and Sinha, 2012), complex events at the crossroad of host's defensive responses and the pathogen's virulence and evasion mechanisms. This is an intriguing hypothesis whose demonstration requires additional analysis, including the complementation of the deleted target genes to formally demonstrate the specific role of EsxA and EsxB factors in controlling DC immune-regulatory functions.

Interestingly, our data on Esx factor-mediated modulation of cytokine expression are in line with findings from Anderson and collaborators, who demonstrated that, in a mouse model of *S. aureus* bloodstream infection, the EssE-mediated secretion of effectors proteins via the Ess pathway contributes to the manipulation of host immune response by affecting the production or the suppression of specific cytokines (Anderson et al., 2016). These findings are also important in light of the effect of cytokine production on anti-staphylococcal immunity.

In this context, it is conceivable that the enhanced release of IL-12, TNF-α, IL-1β, and IL-6 by Δ*esxB* infection likely promotes a robust Th1 and Th17 response (**Figure 6**), whose crucial role in the immunity against *S. aureus* has been demonstrated by several studies (Spellberg et al., 2008; Lin et al., 2009; Krishna and Miller, 2012; Zielinski et al., 2012; Bagnoli et al., 2015; Kolata et al., 2015). In addition, in spite of several differences in the experimental setting, such as infection with live bacterium in our human DC-based model vs. the immunization with the recombinant staphylococcal proteins of the murine model, our data well correlate with the results by Zhang and collaborators showing that in mice EsxA and EsxB recombinant proteins promote both Th1 and Th17 responses (Zhang et al., 2015).

We also investigated the mechanisms underlying the cytokine production by studying kinetics of activation of MAPK p38 and NF-kB p65 pathways (**Figure 5**), known regulators of *S. aureus*-induced immune response (Krishna and Miller, 2012; Armbruster et al., 2016). We observed that wt USA300 and *esx* mutants stimulate intracellular MAPK and NF-kB activation with different kinetics and to a different extent. Although, the observed differences in MAPK and NF-kB activation were not statistically significant, these results suggest that *S. aureus esx* mutants might differentially control intracellular signaling involved in the regulation of cytokine expression.

In this regard, structural bioinformatics analysis revealed that both EsxA and EsxB factors share significant structural homology with the human proteins inhibitor growth factor 4 (ING4) and charged multivesicular body protein 4B (CHMP4B). Interestingly, ING4 is a factor involved in the regulation of NF-kB and p53 activity, which mediate different cellular functions including apoptotic process and cytokine production (Coles et al., 2010; Mathema and Koh, 2012) supporting our findings about the involvement of EsxA and EsxB in the modulation of DC apoptosis and cytokine secretion. Moreover, EsxA and EsxB homology with CHMP4B, a component of ESCRT-III (endosomal sorting complex required for transport III) required for degradation of surface receptor proteins and formation of endocytic multivesicular bodies (Ariumi et al., 2011), suggests a possible interference with the endocytic pathway, which could favor *S. aureus* escape from phagocytes. This hypothesis well correlates with data from Korea and collaborators demonstrating that both EsxA and EsxB are required for the release of *S. aureus* from the infected epithelial cells (Korea et al., 2014). In line with this view, it has been recently demonstrated that also the mycobacterial EsxH factor, secreted by the ESX-3 type VII secretion system, is able either to impair phagosome maturation or to undermine MHC class II presentation by disrupting ESCRT function (Mehra et al., 2013; Portal-Celhay et al., 2016) thus, supporting the idea on the interfering role of Esx factors with host response.

Although, our bioinformatics analysis predicts that EsxA and EsxB may mimic or interact with host-specific proteins, the mechanism by which EsxA and/or EsxB exert their function on DC immune response needs to be further investigated and experimentally validated in future study.

In the attempt to investigate EsxA and EsxB heterodimeric and/or multimeric protein complex formation, *in silico*

bioinformatics analysis was performed indicating that both the heterodimeric and the heterotetrameric assembly of EsxA with EsxB is stereochemically feasible (Supplementary Figures 1, 2). The physical interaction between EsxA and EsxB factors has been hypothesized since their discovery (Burts et al., 2005), mainly on the basis of the similarity with the related mycobacterial proteins, although empirical demonstration of such a complex has not yet been obtained (Sundaramoorthy et al., 2008; Anderson et al., 2013). The experimental validation of the predicted EsxA-EsxB heterodimer/heterotetramer and the definition of the features of this interaction might help to explain the different behavior of the double mutant ΔesxAB with respect to both the single ΔesxA and ΔesxB mutants in the modulation of DC immune functions.

In conclusion, our findings indicate that staphylococcal Esx factors may influence DC functions and, in turn, the adaptive immune response fate providing important information, which could be exploited for developing innovative preventive or therapeutic interventions against *S. aureus*.

AUTHOR CONTRIBUTIONS

MC: Conducted DC infection, sample preparation for confocal microscopy and FACS analysis, protein extraction, western blotting, T cell polarization experiments, predictive analysis of *S. aureus* Esx factor interaction, analyzed the results and wrote the manuscript; ME: Conducted sample preparation for FACS analysis, T cell polarization experiments, predictive analysis of *S. aureus* Esx factor interaction, analyzed the results and wrote the manuscript; RC and SS: Conducted *S. aureus* strain preparations and CFU counting; EG: Conducted DC preparation, analyzed the results and discussed the data; MS: Conducted FACS experiments, analyzed the results and discussed the data; FR: Conducted FACS experiments and analyzed the results; ZP: Conducted confocal microscopy experiments and analyzed the results; VB: Conducted predictive modeling and interaction analysis; FP: Conducted predictive modeling, interaction analysis and discussed the data; EA and AP: Analysis and discussion of the results; FB: Analysis, interpretation, discussion of the results and revised the manuscript critically; EC: Guided the idea generation, experimental work, and manuscript preparation. All authors reviewed the manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcimb.2017.00330/full#supplementary-material

REFERENCES

Aguilo, J. I., Alonso, H., Uranga, S., Marinova, D., Arbues, A., de Martino, A., et al. (2013). ESX-1-induced apoptosis is involved in cell-to-cell spread of *Mycobacterium tuberculosis*. *Cell. Microbiol.* 15, 1994–2005. doi: 10.1111/cmi.12169

Anderson, M., Aly, K. A., Chen, Y. H., and Missiakas, D. (2013). Secretion of atypical protein substrates by the ESAT-6 secretion system of *Staphylococcus aureus*. *Mol. Microbiol.* 90, 734–743. doi: 10.1111/mmi.12395

Anderson, M., Ohr, R. J., Aly, K. A., Nocadello, S., Kim, H. K., Schneewind, C. E., et al. (2016). EssE promotes *Staphylococcus aureus* ESS-dependent protein secretion to modify host immune responses during infection. *J. Bacteriol.* 199, e00527–16. doi: 10.1128/JB.00527-16

Ariumi, Y., Kuroki, M., Maki, M., Ikeda, M., Dansako, H., Wakita, T., et al. (2011). The ESCRT system is required for hepatitis C virus production. *PLoS ONE* 6:e14517. doi: 10.1371/journal.pone.0014517

Armbruster, N. S., Richardson, J. R., Schreiner, J., Klenk, J., Gunter, M., Kretschmer, D., et al. (2016). PSM peptides of *Staphylococcus aureus* activate the p38-creb pathway in dendritic cells, thereby modulating cytokine production and t cell priming. *J. Immunol.* 196, 1284–1292. doi: 10.4049/jimmunol.1502232

Bagnoli, F., Fontana, M. R., Soldaini, E., Mishra, R. P., Fiaschi, L., Cartocci, E., et al. (2015). Vaccine composition formulated with a novel TLR7-dependent adjuvant induces high and broad protection against *Staphylococcus aureus*. *Proc. Natl. Acad. Sci. U.S.A.* 112, 3680–3685. doi: 10.1073/pnas.1424924112

Bien, J., Sokolova, O., and Bozko, P. (2011). Characterization of virulence factors of *Staphylococcus aureus*: novel function of known virulence factors that are implicated in activation of airway epithelial proinflammatory response. *J. Pathog.* 2011:601905. doi: 10.4061/2011/601905

Bottai, D., Gröschel, M. I., and Brosch, R. (2017). Type vii secretion systems in gram-positive bacteria. *Curr. Top. Microbiol. Immunol.* 404, 235–265. doi: 10.1007/82_2015_5015

Burts, M. L., DeDent, A. C., and Missiakas, D. M. (2008). EsaC substrate for the ESAT-6 secretion pathway and its role in persistent infections of *Staphylococcus aureus*. *Mol. Microbiol.* 69, 736–746. doi: 10.1111/j.1365-2958.2008.06324.x

Burts, M. L., Williams, W. A., DeBord, K., and Missiakas, D. M. (2005). EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1169–1174. doi: 10.1073/pnas.0405620102

Choi, H. H., Shin, D. M., Kang, G., Kim, K. H., Park, J. B., Hur, G. M., et al. (2010). Endoplasmic reticulum stress response is involved in *Mycobacterium tuberculosis* protein ESAT-6-mediated apoptosis. *FEBS Lett.* 584, 2445–2454. doi: 10.1016/j.febslet.2010.04.050

Coles, A. H., Gannon, H., Cerny, A., Kurt-Jones, E., and Jones, S. N. (2010). Inhibitor of growth-4 promotes IkappaB promoter activation to suppress NF-kappaB signaling and innate immunity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11423–11428. doi: 10.1073/pnas.0912116107

David, M. Z., and Daum, R. S. (2010). Community-associated methicillin-resistant *Staphylococcus aureus*: epidemiology and clinical consequences of an emerging epidemic. *Clin. Microbiol. Rev.* 23, 616–687. doi: 10.1128/CMR.00081-09

Deretic, V., Delgado, M., Vergne, I., Master, S., De Haro, S., Ponpuak, M., et al. (2009). Autophagy in immunity against *Mycobacterium tuberculosis*: a model system to dissect immunological roles of autophagy. *Curr. Top. Microbiol. Immunol.* 335, 169–188. doi: 10.1007/978-3-642-00302-8_8

Derrick, S. C., and Morris, S. L. (2007). The ESAT6 protein of *Mycobacterium tuberculosis* induces apoptosis of macrophages by activating caspase expression. *Cell. Microbiol.* 9, 1547–1555. doi: 10.1111/j.1462-5822.2007.00892.x

Etna, M. P., Giacomini, E., Pardini, M., Severa, M., Bottai, D., Cruciani, M., et al. (2015). Impact of *Mycobacterium tuberculosis* RD1-locus on human primary dendritic cell immune functions. *Sci. Rep.* 5:17078. doi: 10.1038/srep17078

Etna, M. P., Giacomini, E., Severa, M., Pardini, M., Aguilo, N., Martin, C., et al. (2014). A human dendritic cell-based *in vitro* model to assess *Mycobacterium tuberculosis* SO2 vaccine immunogenicity. *ALTEX* 31, 397–406. doi: 10.14573/altex.1311041

Fraunholz, M., and Sinha, B. (2012). Intracellular *Staphylococcus aureus*: live-in and let die. *Front. Cell. Infect. Microbiol.* 2:43. doi: 10.3389/fcimb.2012.00043

Gordon, R. J., and Lowy, F. D. (2008). Pathogenesis of methicillin-resistant *Staphylococcus aureus* infection. *Clin. Infect. Dis.* 46(Suppl. 5), S350–S359. doi: 10.1086/533591

Gould, I. M., David, M. Z., Esposito, S., Garau, J., Lina, G., Mazzei, T., et al. (2012). New insights into meticillin-resistant *Staphylococcus aureus* (MRSA) pathogenesis, treatment and resistance. *Int. J. Antimicrob. Agents* 39, 96–104. doi: 10.1016/j.ijantimicag.2011.09.028

Green, E. R., and Mecsas, J. (2016). Bacterial secretion systems: an overview. *Microbiol. Spectr.* 4. doi: 10.1128/microbiolspec.vmbf-0012-2015

Groschel, M. I., Sayes, F., Simeone, R., Majlessi, L., and Brosch, R. (2016). ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat. Rev. Microbiol.* 14, 677–691. doi: 10.1038/nrmicro.2016.131

Holm, L., and Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Res.* 38, W545–W549. doi: 10.1093/nar/gkq366

Jager, F., Zoltner, M., Kneuper, H., Hunter, W. N., and Palmer, T. (2016). Membrane interactions and self-association of components of the Ess/Type VII secretion system of *Staphylococcus aureus*. *FEBS Lett.* 590, 349–357. doi: 10.1002/1873-3468.12065

Jin, J. O., Zhang, W., Du, J. Y., and Yu, Q. (2014). BDCA1-positive dendritic cells (DCs) represent a unique human myeloid DC subset that induces innate and adaptive immune responses to *Staphylococcus aureus* infection. *Infect. Immun.* 82, 4466–4476. doi: 10.1128/IAI.01851-14

Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., and Meiler, J. (2010). Practically useful: what the rosetta protein modeling suite can do for you. *Biochemistry* 49, 2987–2998. doi: 10.1021/bi902153g

Kneuper, H., Cao, Z. P., Twomey, K. B., Zoltner, M., Jager, F., Cargill, J. S., et al. (2014). Heterogeneity in ess transcriptional organization and variable contribution of the Ess/Type VII protein secretion system to virulence across closely related *Staphyloccccus aureus* strains. *Mol. Microbiol.* 93, 928–943. doi: 10.1111/mmi.12707

Kolata, J. B., Kühbandner, I., Link, C., Normann, N., Vu, C. H., Steil, L., et al. (2015). The fall of a Dogma? Unexpected high t-cell memory response to *Staphylococcus aureus* in humans. *J. Infect. Dis.* 212, 830–838. doi: 10.1093/infdis/jiv128

Korea, C. G., Balsamo, G., Pezzicoli, A., Merakou, C., Tavarini, S., Bagnoli, F., et al. (2014). Staphylococcal Esx proteins modulate apoptosis and release of intracellular *Staphylococcus aureus* during infection in epithelial cells. *Infect. Immun.* 82, 4144–4153. doi: 10.1128/IAI.01576-14

Kosugi, S., Hasebe, M., Tomita, M., and Yanagawa, H. (2009). Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10171–10176. doi: 10.1073/pnas.0900604106

Krishna, S., and Miller, L. S. (2012). Innate and adaptive immune responses against *Staphylococcus aureus* skin infections. *Semin. Immunopathol.* 34, 261–280. doi: 10.1007/s00281-011-0292-6

Lin, L., Ibrahim, A. S., Xu, X., Farber, J. M., Avanesian, V., Baquir, B., et al. (2009). Th1-Th17 cells mediate protective adaptive immunity against *Staphylococcus aureus* and Candida albicans infection in mice. *PLoS Pathog.* 5:e1000703. doi: 10.1371/journal.ppat.1000703

Mathema, V. B., and Koh, Y. S. (2012). Inhibitor of growth-4 mediates chromatin modification and has a suppressive effect on tumorigenesis and innate immunity. *Tumour Biol.* 33, 1–7. doi: 10.1007/s13277-011-0249-3

Mehra, A., Zahra, A., Thompson, V., Sirisaengtaksin, N., Wells, A., Porto, M., et al. (2013). *Mycobacterium tuberculosis* type VII secreted effector EsxH targets host ESCRT to impair trafficking. *PLoS Pathog.* 9:e1003734. doi: 10.1371/journal.ppat.1003734

Missiakas, D., and Schneewind, O. (2016). *Staphylococcus aureus* vaccines: deviating from the carol. *J. Exp. Med.* 213, 1645–1653. doi: 10.1084/jem.20160569

Monaco, M., Pimentel de Araujo, F., Cruciani, M., Coccia, E. M., and Pantosti, A. (2016). Worldwide epidemiology and antibiotic resistance of *Staphylococcus aureus*. *Curr. Top. Microbiol. Immunol.* doi: 10.1007/82_2016_3. [Epub ahead of print].

Ohr, R. J., Anderson, M., Shi, M., Schneewind, O., and Missiakas, D. (2017). EssD, a nuclease effector of the *Staphylococcus aureus* ESS pathway. *J. Bacteriol.* 199:e00528-16. doi: 10.1128/JB.00528-16

Olaniyi, R., Pozzi, C., Grimaldi, L., and Bagnoli, F. (2016). *Staphylococcus aureus*-associated skin and soft tissue infections: anatomical localization, epidemiology, therapy and potential prophylaxis. *Curr. Top. Microbiol. Immunol.* doi: 10.1007/82_2016_32. [Epub ahead of print].

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T., and Weng, Z. (2014). ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* 30, 1771–1773. doi: 10.1093/bioinformatics/btu097

Pierce, B., and Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins* 67, 1078–1086. doi: 10.1002/prot.21373

Portal-Celhay, C., Tufariello, J. M., Srivastava, S., Zahra, A., Klevorn, T., Grace, P. S., et al. (2016). *Mycobacterium tuberculosis* EsxH inhibits ESCRT-dependent CD4$^+$ T-cell activation. *Nat. Microbiol.* 2:16232. doi: 10.1038/nmicrobiol.2016.232

Pozzi, C., Lofano, G., Mancini, F., Soldaini, E., Speziale, P., De Gregorio, E., et al. (2015). Phagocyte subsets and lymphocyte clonal deletion behind ineffective immune response to *Staphylococcus aureus*. *FEMS Microbiol. Rev.* 39, 750–763. doi: 10.1093/femsre/fuv024

Romagnoli, A., Etna, M. P., Giacomini, E., Pardini, M., Remoli, M. E., Corazzari, M., et al. (2012). ESX-1 dependent impairment of autophagic flux by *Mycobacterium tuberculosis* in human dendritic cells. *Autophagy* 8, 1357–1370. doi: 10.4161/auto.20881

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738. doi: 10.1038/nprot.2010.5

Spellberg, B., Ibrahim, A. S., Yeaman, M. R., Lin, L., Fu, Y., Avanesian, V., et al. (2008). The antifungal vaccine derived from the recombinant N terminus of Als3p protects mice against the bacterium *Staphylococcus aureus*. *Infect. Immun.* 76, 4574–4580. doi: 10.1128/IAI.00700-08

Stanley, S. A., Raghavan, S., Hwang, W. W., and Cox, J. S. (2003). Acute infection and macrophage subversion by *Mycobacterium tuberculosis* require a specialized secretion system. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13001–13006. doi: 10.1073/pnas.2235593100

Sundaramoorthy, R., Fyfe, P. K., and Hunter, W. N. (2008). Structure of *Staphylococcus aureus* EsxA suggests a contribution to virulence by action as a transport chaperone and/or adaptor protein. *J. Mol. Biol.* 383, 603–614. doi: 10.1016/j.jmb.2008.08.047

Warne, B., Harkins, C. P., Harris, S. R., Vatsiou, A., Stanley-Wall, N., Parkhill, J., et al. (2016). The Ess/Type VII secretion system of *Staphylococcus aureus* shows unexpected genetic diversity. *BMC Genomics* 17:222. doi: 10.1186/s12864-016-2426-7

Zhang, B. Z., Hua, Y. H., Yu, B., Lau, C. C., Cai, J. P., Zheng, S. Y., et al. (2015). Recombinant ESAT-6-like proteins provoke protective immune responses against invasive *Staphylococcus aureus* disease in a murine model. *Infect. Immun.* 83, 339–345. doi: 10.1128/IAI.02498-14

Zhang, Q. C., Petrey, D., Garzon, J. I., Deng, L., and Honig, B. (2013). PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.* 41, D828–D833. doi: 10.1093/nar/gks1231

Zielinski, C. E., Mele, F., Aschenbrenner, D., Jarrossay, D., Ronchi, F., Gattorno, M., et al. (2012). Pathogen-induced human TH17 cells produce IFN-gamma or IL-10 and are regulated by IL-1beta. *Nature* 484, 514–518. doi: 10.1038/nature 10957

OXFORD

Structural bioinformatics

LIBRA-WA: a web application for ligand binding site detection and protein function recognition

Daniele Toti[1,†], Le Viet Hung[2,†], Valentina Tortosa[1], Valentina Brandi[1] and Fabio Polticelli[1,3,*]

[1]Department of Sciences, University of Roma Tre, 00146 Rome, Italy, [2]Department of Science and Technology, Nguyen Tat Thanh University, Ho chi Minh City, Vietnam and [3]National Institute of Nuclear Physics, Roma Tre Section, 00146 Rome, Italy

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Summary: Recently, LIBRA, a tool for active/ligand binding site prediction, was described. LIBRA's effectiveness was comparable to similar state-of-the-art tools; however, its scoring scheme, output presentation, dependence on local resources and overall convenience were amenable to improvements. To solve these issues, LIBRA-WA, a web application based on an improved LIBRA engine, has been developed, featuring a novel scoring scheme consistently improving LIBRA's performance, and a refined algorithm that can identify binding sites hosted at the interface between different subunits. LIBRA-WA also sports additional functionalities like ligand clustering and a completely redesigned interface for an easier analysis of the output. Extensive tests on 373 apoprotein structures indicate that LIBRA-WA is able to identify the biologically relevant ligand/ligand binding site in 357 cases ($\sim$96%), with the correct prediction ranking first in 349 cases ($\sim$98% of the latter, $\sim$94% of the total). The earlier stand-alone tool has also been updated and dubbed LIBRA+, by integrating LIBRA-WA's improved engine for cross-compatibility purposes.

Availability and implementation: LIBRA-WA and LIBRA+ are available at: http://www.computationalbiology.it/software.html.

Contact: polticel@uniroma3.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, structure-based protein function recognition has gained renewed interest due to the availability of a large number of experimental protein structures, determined within the context of structural genomics initiatives, whose function is unknown (Grabowski *et al.*, 2016; Petrey *et al.*, 2015). In this framework, we recently developed and described LIBRA, a graph theory-based software tool that, given a protein's structural model, predicts the presence and identity of active sites and/or small molecule ligand binding sites (Viet Hung *et al.*, 2015). Extensive tests carried out on the LigaSite (Dessailly *et al.*, 2008) set of approximately 400 apoproteins indicated that LIBRA was able to identify the correct

binding/active site in $\sim$90% of the cases analyzed, outperforming other structure-based function recognition software such as SiteSeer (Laskowski *et al.*, 2005a,b), EF-Seek (Murakami *et al.*, 2013) and ASSIST (previously developed in our lab; Caprari *et al.*, 2014), while displaying a performance comparable to ProFunc, which employs a combined sequence/structure approach (Laskowski *et al.*, 2005 b). However, the identified correct site ranked first only in 80% of the cases, a suboptimal performance that needed to be improved for LIBRA to be able to handle the most challenging cases. For this purpose, an improved version of LIBRA featuring a novel scoring system has been developed both as a web application, LIBRA-WA and a standalone tool, LIBRA+. The new system also features an

improved algorithm to deal with binding sites located at the interface of different protein subunits and clustering of identified ligands according to their chemical similarity. Tests carried out on the same set of apoproteins earlier used on LIBRA demonstrate a significant improvement of the performance, as LIBRA-WA is able to identify the correct binding site in ∼96% of the cases, with the correct site ranking first in ∼94% of the cases. Comparative tests demonstrate that LIBRA-WA has a performance comparable to the state-of-the-art COACH meta-server.

2 Materials and methods

2.1 LIBRA-WA's improved engine and features

LIBRA-WA features an improved active/ligand binding site detection engine and a number of additional features, including a redesigned GUI freely accessible online. The core improvement of the engine lies in a novel scoring system, which takes advantage of a clustering process carried out on more than 17 000 unique small molecule ligands stored in the application's database, based on their SMILES representation. LIBRA-WA, for each alignment record, now provides a score obtained by combining the contributions given by the aligned binding site's clique size (number of matching residues between the input protein and the target binding site), RMSD value, and the relative size of the cluster containing the ligand. A detailed description of the calculation of this combined score is provided in the Supplementary Material. Besides, the detection algorithm has been further refined by allowing the identification of binding sites hosted at the interface between different subunits. Recognition jobs can be launched against two pre-compiled databases: a ligand binding sites database, including more than 173 000 entries, and a database of active sites derived from the Catalytic Site Atlas (Furnham, 2014) (∼1000 entries) that can be used for the prediction of the catalytic activity of an input protein. For a detailed description of the procedure used to build the two databases, see Viet Hung et al. (2015). Aside from that, as a web application, sharing the same architectural framework of (Atzeni et al., 2011a, b; Toti et al., 2012), LIBRA-WA is freely accessible by any web user, who can schedule multiple recognition jobs. Optionally, LIBRA-WA also enables users to create a personal workspace and access their results at a later time, by notifying the users once the jobs' executions have terminated. Results can be also graphically displayed in three-dimensions via the Jmol HTML5 plug-in (Hanson, 2010). Furthermore, the LIBRA desktop application has been updated by incorporating the new detection engine and the information about the ligand clusters: this new version, which has been dubbed LIBRA+, can read the results exported from LIBRA-WA and is backward-compatible with the output files produced with the original version of LIBRA. A more thorough description of LIBRA-WA's additional functionalities is reported in the Supplementary Material.

3 Results

The effectiveness of LIBRA-WA has been tested on the LigaSite set. A detailed analysis of the results is reported in Supplementary Table S1. As shown in the table, LIBRA-WA finds the biologically relevant ligand/binding site in ∼96% of the cases. More important for the predictive power of the application, the correct ligand/binding site ranks first in ∼94% of all cases. In fact, in 'real life' applications, where no functional information is available on the protein of interest, it is essential that the correct prediction is found in the few first-ranking hits. Even removing from the database the holo-proteins present in the LigaSite set, the application still performs fairly well. In fact, LIBRA-WA still identifies a biologically relevant ligand in 88% of the cases, with the correct ligand ranking first in 80% of the cases (Supplementary Table S4). Particularly striking is the ability of LIBRA-WA to pick out similar ligand binding motifs even in structures that do not display significant sequence/structure similarity. One such example, illustrated in Supplementary Figure S3, is that of the *E.coli* adenylate kinase (apoprotein PDB code 4AKE) which, upon ADP binding undergoes a large conformational change (holo-protein PDB code 2ECK). Therefore, the program does not identify the ADP binding site contained in the database entry 2ECK as a correct match. Nonetheless, it correctly identifies the ADP binding site in the input protein by virtue of the structural similarity with the ADP binding site of the human kinesin-8 motor domain (PDB code 3LRE). As detailed in the Supplementary Material, the *E.coli* adenylate kinase and the human kinesin-8 motor domain do not display a similar fold and share a non-significant 11% sequence identity. A combined execution of LIBRA-WA using both the ligand binding sites and the catalytic sites databases allows a user to obtain information on both the location of the binding site, the identity of the ligand(s) and, in case the input protein is an enzyme, its catalytic activity, and thus assign a function to the input protein with high confidence. For example, on the *E.coli* adenylate kinase and using the ligand binding sites database, LIBRA-WA detects as first hit an ADP binding site similar to that of the kinesin-8 motor domain. However, an execution using the catalytic sites database detects as first hit a catalytic site similar to that of *Bacillus stearothermophilus* adenylate kinase (PDB code 1ZIO). Combining the two information together leads to a highly reliable function prediction for the input protein.

4 Discussion

In this paper, the development of LIBRA-WA, a web application based on an improved LIBRA engine has been described. By employing an enhanced, composite scoring system, in LIBRA-WA both precision and recall are significantly improved with respect to LIBRA, as it can be clearly seen from the results of the extensive tests detailed in Supplementary Table S1. Furthermore, LIBRA-WA outperforms SiteSeer while displaying a performance comparable to that of COACH (Yang et al., 2013), ranked as the best method in the weekly CAMEO ligand Binding Site Prediction Experiments (Haas et al., 2013), even though the latter uses a combination of structure-based and sequence-based algorithms, while LIBRA-WA is purely structure-based (Supplementary Tables S2 and S3).

Conflict of Interest: none declared.

References

Atzeni,P. *et al.* (2011a) A framework for semi-automatic identification, disambiguation and storage of protein-related abbreviations in scientific literature. *Proc. Int. Conf. Data Eng.*, **5767646**, 59–61.

Atzeni,P. *et al.* (2011b) An automatic identification and resolution system for protein-related abbreviations in scientific papers. *Lect. Notes Comput. Sci.*, **6623**, 171–176.

Caprari,S. *et al.* (2014) ASSIST: a fast versatile local structural comparison tool. *Bioinformatics*, **30**, 1022–1024.

Dessailly,B.H. *et al.* (2008) LigASite: a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.*, **36**, 667–673.

Furnham,N. *et al.* (2014) The Catalytic Site Atlas 2. *Nucleic Acids Res.*, **42**, 485–489.

Grabowski,M. *et al.* (2016) The impact of structural genomics: the first quindecennial. *J. Struct. Funct. Genomics*, **17**, 1–16.

Haas,J. *et al.* (2013) The Protein Model Portal–a comprehensive resource for protein structure and model information. *Database (Oxford)*, **2013**, bat031.

Hanson,R.M. (2010) Jmol-a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.*, **43**, 1250–1260.

Laskowski,R.A. *et al.* (2005a) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **3**, 89–93.

Laskowski,R.A. *et al.* (2005b) Protein function prediction using local 3D templates. *J. Mol. Biol.*, **351**, 614–626.

Murakami,Y. *et al.* (2013) Exhaustive comparison and classification of ligand-binding surfaces in proteins. *Protein Sci.*, **22**, 1379–1391.

Petrey,D. *et al.* (2015) Template-based prediction of protein function. *Curr. Opin. Struct. Biol.*, **32**, 33–38.

Toti,D. *et al.* (2012) Automatic protein abbreviations discovery and resolution from full-text scientific papers: the PRAISED framework. *Bio-Algorithms Med-Syst.*, **8**, 13.

Viet Hung,L. *et al.* (2015) LIBRA: LIgand Binding site Recognition Application. *Bioinformatics*, **31**, 4020–4022.

Yang,J. *et al.* (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.

Short Communication

Valentina Tortosa, Maria Carmela Bonaccorsi di Patti, Valentina Brandi, Giovanni Musci
and Fabio Polticelli*

An improved structural model of the human iron exporter ferroportin. Insight into the role of pathogenic mutations in hereditary hemochromatosis type 4

Abstract: Ferroportin (Fpn) is a membrane protein representing the major cellular iron exporter, essential for metal translocation from cells into plasma. Despite its pivotal role in human iron homeostasis, many questions on Fpn structure and biology remain unanswered. In this work, we present two novel and more reliable structural models of human Fpn (hFpn; inward-facing and outward-facing conformations) obtained using as templates the recently solved crystal structures of a bacterial homologue of hFpn, *Bdellovibrio bacteriovorus* Fpn. In the absence of an experimentally solved structure of hFpn, the structural predictions described here allow to analyze the role of pathogenic mutations in the Fpn-linked hereditary hemochromatosis disease and represent a valuable alternative for reliable structure-based functional studies on this human iron exporter.

Keywords: ferroportin; iron transport; major facilitator superfamily; molecular modeling.

Introduction

Iron is the second most abundant metal on Earth's crust and plays an essential biological role in all living organisms [1]. Biological properties of iron are linked to its efficient electron transfer properties, allowing it to accept or donate electrons. This ability enables it not only to participate as a catalyzing cofactor in various biochemical reactions but also to contribute to the formation of potentially damaging free radical species [2]. Thus, iron concentration in body tissues needs to be strongly regulated. In fact, it is well known that the imbalance of iron homeostasis has noticeable effects on human health [1].

Ferroportin (Fpn) is a membrane protein representing the major cellular iron exporter, essential for metal translocation from cells into plasma [3]. It belongs to one of the largest secondary transporter families, the major facilitator superfamily (MFS), and together with hepcidin, a 25-amino-acid peptide that binds to hFpn and drives its degradation, it is a key player in the control of human iron homeostasis [3]. Mutations in the gene coding for hFpn cause type 4 hemochromatosis [4], an autosomal dominant iron overload condition. Mutations are generally classified into two subclasses: "loss-of-function" and "gain-of-function". The majority of pathogenic variants are defined as "loss-of-function," as they lead to decreased iron export from cells. These give rise to type 4A hemochromatosis, also known as Fpn disease [5]. On the other hand, "gain-of-function" mutations, responsible for type 4B hemochromatosis, confer resistance to the action of hepcidin. Hepcidin binding to Fpn leads to protein internalization, ubiquitination, and degradation. Gain-of-function mutations are named this way because they inhibit this mechanism and cause Fpn to permanently stay on the membrane and continue to be active in iron efflux.

Although the importance of Fpn in human iron homeostasis has been widely demonstrated, its exact three-dimensional structure remains unknown. Based on the knowledge of the structure of other members of the MFS family, it was proposed that Fpn operates through an alternating-access mechanism that involves a movement of the two halves of the protein, together with local structural rearrangements. The rotation of the two domains composing the protein leads to three different conformational

*Corresponding author: Fabio Polticelli,** Department of Sciences, Roma Tre University, Viale Guglielmo Marconi 446, 00146 Rome, Italy, Phone: +39-06-57336362, Fax: +39-06-57336321, E-mail: fabio.polticelli@uniroma3.it.
http://orcid.org/0000-0002-7657-2019; and National Institute of Nuclear Physics, Roma Tre Section, 00146 Rome, Italy
Valentina Tortosa and Valentina Brandi: Department of Sciences, Roma Tre University, 00146 Rome, Italy
Maria Carmela Bonaccorsi di Patti: Department of Biochemical Sciences, Sapienza University of Roma, 00185 Rome, Italy
Giovanni Musci: Department of Biosciences and Territory, University of Molise, 86090 Pesche, Italy

states: outward-facing, occluded, and inward-facing [6, 7]. Despite the lack of a crystal structure, different models of hFpn have been proposed. The first available hFpn model [8] can be considered a "low-resolution" one. In fact, the transmembrane helices were predicted and built *de novo* and then manually fitted over the three-dimensional structure of the *Escherichia coli* glycerol-3-phosphate transporter in the inward-facing conformation (PDB code: 1PW4 [9]). A few years later, Le Gac and coworkers [10] described a novel model obtained using the three-dimensional structure of the MFS transporter EmrD (PDB code: 2GFP [11]) in the occluded conformation. However, loops outside of the transmembrane helices were not modeled, and some critical structural details of this model were later shown to be incorrect. For instance, Asp181, a residue affected in loss-of-function mutants, was hypothesized to be exposed toward the membrane while it was later shown to be essential for iron binding and transport and thus exposed within the substrate-binding pocket [12]. The first structural models of hFpn in both the inward- and outward-facing conformations were later obtained by Bonaccorsi di Patti and coworkers [12] on the basis of the three-dimensional structures of the glycerol-3-phosphate transporter (PDB code:1PW4 [9]) and the FucP fucose transporter (PDB code: 3O7Q [13]), respectively. These models allowed to uncover and experimentally validate the identity of residues involved in iron binding and transport. Finally, a possible iron-transport mechanism was hypothesized on the basis of hFpn models in all three mechanistically relevant conformations (inward-facing, occluded, and outward-facing) [7].

Recently, a bacterial homologue of hFpn has been identified in the predatory bacterium *Bdellovibrio bacteriovorus* [14, 15] and the three-dimensional structure of this protein (BdFpn) has been determined in the inward- and outward-facing conformational states. Taking advantage of the availability of these structures, the only ones available for a member of the Fpn family, updated models of hFpn in the inward- and outward-facing conformations have been built and a comprehensive structure-function analysis of known Fpn pathogenic mutations has been carried out.

Methods

Molecular modeling

The structural models of hFpn have been built using the *ab initio*/threading approach implemented in the

I-TASSER pipeline [5] and the homology modeling software packages MODELLER [16] and SwissModel [17]. In the case of I-TASSER and SwissModel, no query/template alignment has been provided in input. In fact, in the case of I-TASSER, the server uses LOMETS (Local Meta-Threading Server) to thread the query sequence through a representative PDB structure library to search for the possible folds compatible with the sequence of the protein of interest [18, 19]. SwissModel instead uses PSI-BLAST [20] to construct the best query-template alignment. In the case of MODELLER, a reliable sequence alignment between hFpn and BdFpn has been given in input. This was obtained using CLUSTAL-omega [21] to align the amino acid sequence (NCBI GI code reported within parentheses) of BdFpn (gi 426403971) to those of Fpn from *Homo sapiens* (gi 7657100), *Bos taurus* (gi 118151032), *Rattus norvegicus* (gi 158635998), *Mus musculus* (gi 7109247), *Anas platyrhynchos* (gi 483522646), *Gallus* (gi 61098366), *Chelonia mydas* (gi 465955074), *Danio rerio* (gi 7109245), *Echinops telfairi* (gi 507641023), *Tetraodon nigroviridis* (gi 47225539), *Branchiostoma floridae* (gi 229294181), *Caenorhabditis elegans* (gi 351061132) and *Ciona intestinalis* (gi 198437857). In all three cases, modeling has been carried out by specifying as templates the three-dimensional structures of BdFpn in the inward-facing (PDB code: 5AYO [14]) and outward-facing (PDB code: 5AYM [14]) conformation. The overall quality of the models has been assessed using PROCHECK [22], and the corresponding G-factor values (the PROCHECK G-factor threshold for good-quality three-dimensional structures being −0.5) are reported in Supplementary Materials Table S1.

Structure-based sequence alignment of the models to the three-dimensional structure of BdFpn in the inward-facing and outward-facing conformations has been obtained using ESPript 3.0 [23].

Docking simulations

Docking simulations between the three-dimensional structure of hepcidin (PDB code: 3H0T [24]) and the structural model of hFpn in the outward-facing conformation have been performed using the protein-protein docking server ZDOCK (version 3.0.2), which is based on a Fast Fourier transform algorithm and a scoring system combining shape complementarity, electrostatics and statistical potential terms [25]. The 2000 complexes predicted by ZDOCK were re-ranked using ZRANK, which uses a more detailed potential combining electrostatics, van der Waals and desolvation terms [26].

Collection of pathogenic mutations

A list of all known hFpn pathogenic mutations has been obtained through a careful manual analysis of the available literature. Most mutations were retrieved from the comprehensive meta-analysis by Mayr et al. [27]; the list has recently been updated [5, 27].

Results and discussion

Analysis of the structural models of hFpn in the inward- and outward-facing conformations

The structural models of hFpn in the inward- and outward-facing conformations obtained with I-TASSER, MODELLER, and SwissModel are shown in Figure 1. To further assess the degree of agreement between the different models at the residue level, a structure-based sequence alignment has been carried out. As can be seen in Supplementary Materials Figures S1 and S2, the three different models display a high degree of consensus in the core structure (transmembrane helices) while, for obvious reasons, large deviations are observed at the level of the intracellular and extracellular surface loops. From this viewpoint, it must be noted that in the models obtained by homology modeling, the surface loops are not structured. On the contrary, the threading/*ab initio* approach implemented in I-TASSER, including replica exchange

molecular dynamics simulations, allows to predict a putative structure of the large loops, which characterize hFpn. In other structurally characterized members of the MFS, these loops are involved in the inward to outward conformational change [28]. Therefore, once assessed, the substantial agreement between the three methods, the subsequent structural analyses were conducted on the hFpn models generated with I-TASSER.

The superposition of BdFpn structures with the hFpn models generated by I-TASSER yields RMSD values of 1.24 Å for the inward-facing state and of 1.06 Å for the outward-facing state, with the most noticeable differences being located at the level of the TM2-TM3 connecting loop. This difference is acceptable, as this loop is considerably longer in hFpn with respect to BdFpn, and has been subjected to an *ab initio* modeling procedure in the I-TASSER workflow.

As already emerged from previous studies [7, 8, 10, 12, 14], the structural model of hFpn consists of 12 transmembrane helices (TM), which are arranged in two domains (N-domain and C-domain), each containing six consecutive transmembrane helices, surrounding a substrate translocation pore, with the N- and C-termini located on the intracellular side. Helices, 1, 2, 4, 5, 7, 8, 10 and 11 shape the central transport cavity, and helices 3, 6, 9 and 12 form the outer domain.

Experimental studies carried out on hFpn have identified functional roles for selected residues in iron binding and transport. In agreement with the biochemical characterization of other MFS transporters, the hypothetical substrate-binding site of hFpn is located halfway into the membrane [6], with the TM7 contributing more
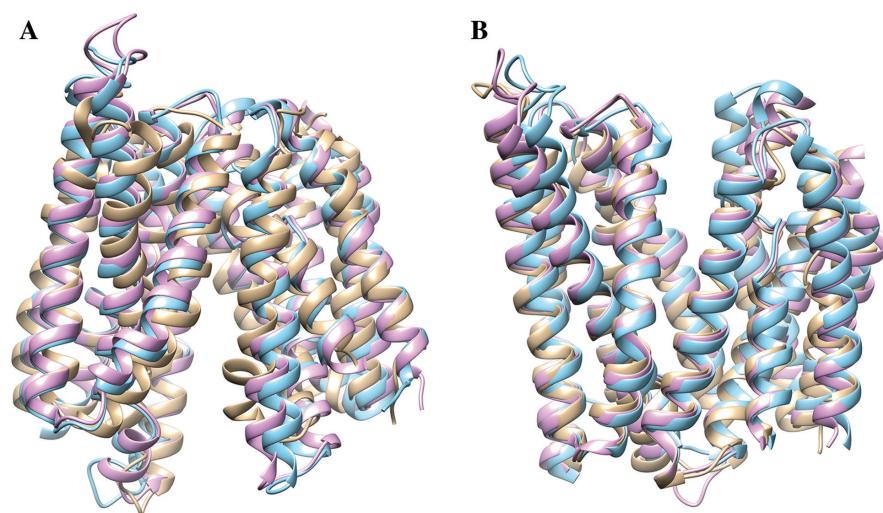

Figure 1: Schematic representation of the structural models of hFpn obtained in this study.
Structural models of hFpn obtained with I-TASSER (tan), MODELLER (cyan), and SwissModel (purple) in the inward-facing (A) and outward-facing (B) conformations.

than half of the interacting residues for ligand coordination. In detail, Asp39, Asp181 and Asp325, whose mutation completely abolishes the iron binding and transport ability [12], form a putative iron binding site (Figure 2);
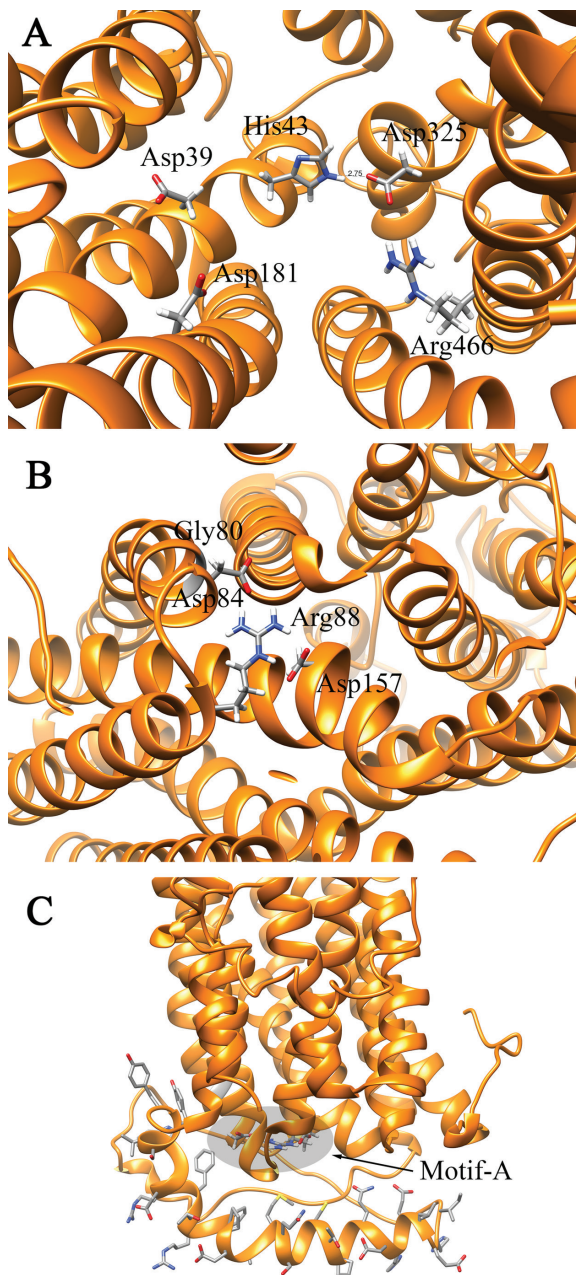

Figure 2: Selected, functionally-relevant structural details of the hFpn structural models.
Schematic representation of (A) the hFpn putative iron binding site viewed from the cytoplasmic side in the inward-facing conformation; (B) Motif-A involving Gly80, Asp84, Arg88 and Asp157 in the outward-facing conformation (Gly80 position indicated by the grey ribbon color); (C) the cytoplasmic kinked, amphipatic α-helix interacting with the cytoplasmic ends of the transmembrane helices in the outward-facing conformation.

Arg466 mutation significantly reduces the iron transport ability of Fpn and has been hypothesized to act as an "electrostatic switch", which, upon iron binding, facilitates the inward to outward transition of Fpn [7]. These residues are conserved in BdFpn, with the exception of hFpn Asp325, which substitutes His261 present in the orthologous position in BdFpn [15]. Unexpectedly, mutation of BdFpn His261 into Asp (mimicking hFpn in this regard) completely abolished iron transport [15]. Analysis of the latest inward-facing model of hFpn provides a likely explanation for this result. In fact, Asp325 is predicted to be located in the vicinity of His43 (Figure 2), and a compensative substitution of hFpn His43 with Asp28 in BdFpn is observed, indicating a functional role of an Asp-His pair (Asp325-His43 in hFpn; Asp28-His261 in BdFpn), which is disrupted in the BdFpn His261Asp mutant. Thus, the two Fpns contain similarly oriented His-Asp pairs, and this observation, together with the reduced iron transport ability observed upon mutation of Asp325 in hFpn [12] and His261 in BdFpn [15], indicates that this residues pair likely plays a role in iron binding and/or translocation. The overall structure of Fpn shows that the transporter consists of irregular, highly kinked transmembrane helices that could be stabilized by the hydrophobicity and/or size of the amino acid side chains. Several "loss-of-function" mutations are located in positions that could induce irregularities in helices [29] indicating that, although none of these residues is required for hFpn activity, the bulkyness of the side chain at these positions could be important for conformational flexibility.

Another important motif identified in this molecule is Motif-A [28] (Figure 2), located at the level of the cytoplasmic loop connecting TM2 and TM3, formed by residues Gly80 (Gly65 in BdFpn), Asp84 (Asp69 in BdFpn) and Arg88 (Arg73 in BdFpn) [7, 15]. This motif, largely conserved in members of the MFS family, does not likely play a direct role in the iron transport mechanism but may only function as a cytoplasmic anchor point probably coordinating the conformational changes of MFS transporters by stabilizing the outward-facing conformation [6, 30].

An interesting feature of the outward-facing model of hFpn is the presence, on the cytoplasmic side, of a kinked amphipatic α-helix seating at the cytoplasm-membrane boundary and interacting with the cytoplasmic ends of the transmembrane α-helices (Figure 2). This helix corresponds to the long loop connecting the N- and C-terminal domains in the inward-facing conformation. Interestingly, in other members of the MFS, the structuring of this loop into an amphipatic α-helix has been suggested to drive the inward- to outward-facing conformation [30].

Mapping of all known mutations on the structural models of hFpn

Table 1 shows a list of all hFpn mutations known up to date, together with the corresponding clinical and biochemical phenotypes. From this viewpoint, it must be mentioned that a structure-function analysis of a restricted subset of these mutations was already reported by Le Gac and coworkers [10] on the basis of the hFpn structural model in the occluded conformation. However, as briefly mentioned in the Introduction, some of the structural details of this model were later shown not to be consistent with experimental data and, given the occluded conformation of the model, the role of mutations conferring hepcidin resistance was difficult to assess.

Here we report a detailed analysis of the putative structural-functional role of the gain-of-function and loss-of-function hFpn mutations based on the novel models obtained. A large part of the mutations is located on trans-membrane helices with the following exceptions: Arg88, Ile152, Asp157, Trp158, Val162, Leu233, Lys240, Asp270, Arg296, Gly323, Cys326 and Ser338. As can be seen from Figure 3, especially in the inward-facing model of hFpn, most of the loss-of function mutations are located on the N-terminal half of the protein where Asp39, His43, and Asp181 are located. This region hosts also the residues building up Motif-A, essential for stabilizing the outward-facing conformation of the protein, allowing the release of the substrate. Indeed loss-of-function mutations affect residues Ala77, Gly80, Asp84 and Arg88 (Table 1). The latter three residues form hFpn Motif-A (Figure 2), and the mutations observed either do not allow the close contact between TM2 and TM11 needed to reach the outward-facing conformation (A77D, G80S/V, D84E) or disrupt the charge interaction between Asp84 and Arg88 needed to stabilize the same conformation (R88G/T). From this viewpoint, the outward-facing model of hFpn clearly shows that Asp157 is part of the charge network involving Asp84 and Arg88 (Figure 2) and indeed loss-of-function mutations affecting Asp157 are observed as well (D157G/Y/N; Table 1).

Concerning gain-of-function mutations, the outward-facing model provides a ready explanation of their effect. Indeed almost all of these mutations are clustered around an area centered on TM2 and TM11 where the hepcidin "docking" site (defined by Cys326 and His507 [28]) is located (Figure 3). These mutations involve residues Tyr64, Ala69, Ser71, Val72, Asn144, Cys326, Tyr501, Asp504 and His507 (Table 1), which define an almost continuous patch of the cavity of hFpn in the outward-facing

Table 1: Classification of hFpn missense variants.

Mutation	Clinical phenotype	Biochemical phenotype
Y64N/H	Non classical	Gain of function
A69T	Non classical	Gain of function
S71F	Non classical	Gain of function
V72F	Non classical	Gain of function
A77D	Variable	Loss of function
G80S/V	Variable	Loss of function
D84E	Classical	Loss of function
R88G/T	Classical	Loss of function
L129P	Classical	Loss of function
N144H/D/T	Variable	Gain of function
I152F	Classical	Loss of function
D157G/Y/N	Classical	Loss of function
W158L/C	Classical	Loss of function
V162del	Classical	Loss of function
L170F	N/A	Loss of function
N174I	Variable	Loss of function
R178Q	Variable	N/A
I180T	–	Neutral
D181V/N	Classical	Loss of function
Q182H	Classical	Loss of function
N185D	Variable	Loss of function
G204S/R	Variable	Gain of function
T230N	–	Neutral
L233P	Variable	Loss of function
K240E	Non classical	Gain of function
Q248H	–	SNP
M266T	–	Neutral
G267D	N/A	N/A
D270V	Non classical	Gain of function
R296Q	Classical	Loss of function
G323V	N/A	Loss of function
F324S	–	SNP
C326S/Y/F	Non classical	Gain of function
S338R	Non classical	Gain of function
G339D	–	SNP
L345F	–	Neutral
I351V	–	Neutral
R371Q	–	Neutral
L384M/V	–	SNP
F405S	–	SNP
M432V	–	SNP
P443L	–	SNP
R489S/K	Variable	Loss of function
G490D/S	Variable	Loss of function
Y501C	Non classical	Gain of function
D504N	Non classical	Gain of function
H507R	Non classical	Gain of function
R561G	–	SNP

SNP, single nucleotide polymorphism; N/A, not available.

conformation (Figure 4), in line with a homology model based on BdFpn in the outward open conformation [14]. Mutation of each of these residues likely modifies the hepcidin interaction surface on hFpn, resulting in a decreased
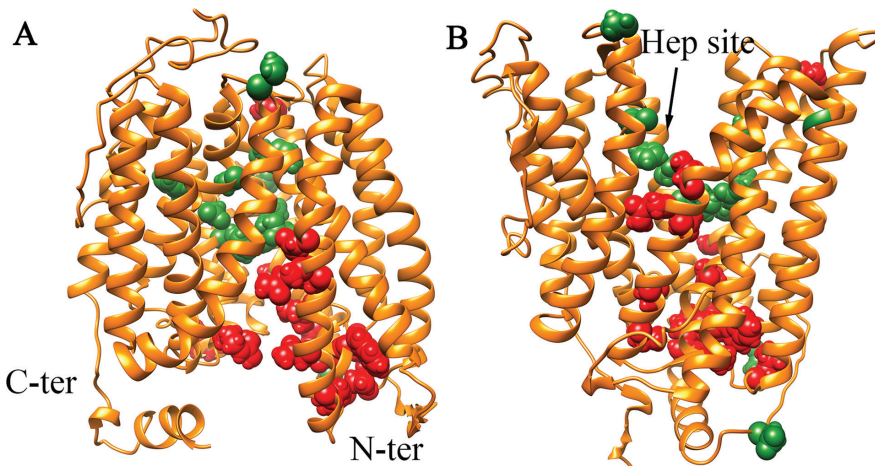
Figure 3: Pathogenic mutations of hFpn mapped onto the structural models.
Location of the residues affected by loss-of-function mutations (red) and gain-of-function mutations (green) in the hFpn inward-facing (A) and outward-facing (B) conformations.
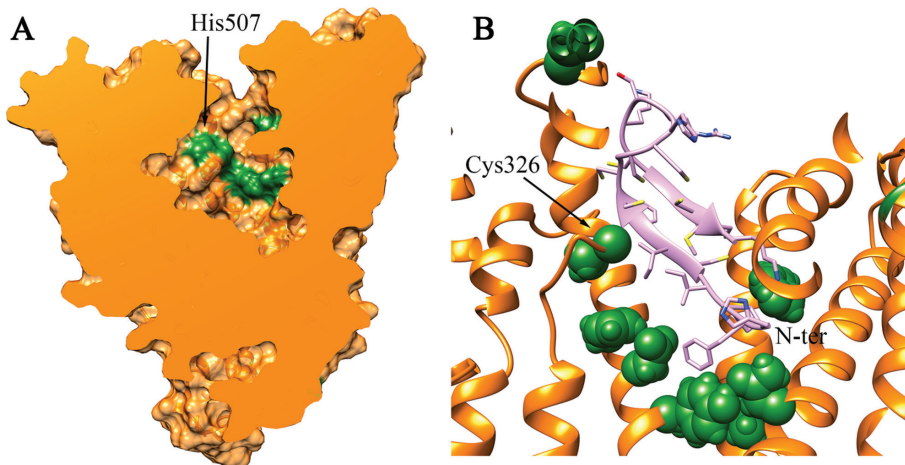

Figure 4: Structural details of the hFpn hepcidin binding site.
(A) Residues affected by gain-of-function mutations (green) located in the hepcidin binding cavity mapped onto the molecular surface of hFpn in the outward-facing conformation. (B) Putative hFpn-hepcidin complex obtained by docking simulations. Note the close proximity between hepcidin N-terminal residues and hFpn residues affected by gain-of-function mutations.

or impaired binding of this hormone and failure to initiate hFpn internalization and degradation. This hypothesis is confirmed by docking simulations of hepcidin onto the outward-facing model of hFpn, carried out using ZDOCK (see Methods section for details). Indeed, in the best ranking complex, obtained by rescoring the docking poses with ZRANK, hepcidin nicely sits within the hFpn cavity, contacting all the residues building up the above mentioned surface patch (Figure 4). Interestingly, in this complex hepcidin N-terminal end contacts, almost all of the residues affected by gain-of-function mutations, in agreement with the experimental evidence that the first nine residues of hepcidin are critical for hFpn binding

and for the disulphide exchange process with Cys326, which triggers hFpn internalization and degradation [31, 32]. Other residues involved in gain-of-function mutations are Gly204, Lys240 and Asp270. Gly204 is positioned at the extracellular end of TM6 near the membrane boundary and the short loop connecting TM5 and TM6, in a position in which the two α-helices are in close contact. Substitution of this residue with a bulkier residue probably does not affect hepcidin binding, but it may destabilize the outward-facing conformation and/or interfere with the internalization process. Lys240 is located in the long intracellular loop connecting TM6 and TM7, and it has been shown to be involved in hepcidin-dependent

ubiquitination [33, 34]. Also, Asp270 is found in this loop and replacement with Val might modify the loop conformation and the accessibility of the Lys residues required for ubiquitination of hFpn.

Conclusion

In this work, novel and improved structural models of the inward- and outward-facing conformations of hFpn, based on the recently solved BdFpn three-dimensional structures, have been presented. The models presented in this manuscript provide a novel insight into human Fpn structure-function relationships with respect to previously published models. In particular, the inward-facing model uncovers a structural role of the His43-Asp325 pair, never hypothesized before, also corroborated by compensative mutations observed in BdFpn in which an Asp28-His261 pair is observed; the outward-facing model highlights the presence also in hFpn of an amphipatic α-helix stabilizing this conformation, as observed in other members of the MFS, never postulated on the basis of previous models. The same model also provides the first view at atomic level of the Asp157 involvement in the Motif-A charge relay, this latter observation being particularly noteworthy as this residue is affected in loss-of-function mutants. Furthermore, the higher reliability of our outward-facing model allowed to obtain for the first time a structural model of the hFpn-hepcidin complex, which is fully in line with the current knowledge of hepcidin mechanism of action and consistent with the location of gain-of-function mutations. The models allowed the analysis at the structural level of the effect of the loss-of-function and gain-of-function mutations leading, respectively, to type 4A and type 4B hemochromatosis. In the absence of an experimental structure, the models here described represent a valuable alternative for structure-based functional studies on hFpn. Particularly interesting, in this respect, is the reliability of the outward-facing model as far as the hepcidin binding site is concerned. This model appears to be a good starting point for the design of hepcidin substitutes useful for the treatment of iron overload disorders.

References

1. Bogdan AR, Miyazawa M, Hashimoto K, Tsuji Y. Regulators of iron homeostasis: new players in metabolism, cell death, and disease. Trends Biochem Sci 2016;41:274–86.
2. Ray PD, Huang BW, Tsuji Y. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. Cell Signal 2012;24:981–90.
3. Drakesmith H, Nemeth E, Ganz T. Ironing out ferroportin. Cell Metab 2015;22:777–87.
4. Montosi G, Donovan A, Totaro A, Garuti C, Pignatti E, Cassanelli S, et al. Autosomal-dominant hemochromatosis is associated with a mutation in the ferroportin (SLC11A3) gene. J Clin Invest 2001;108:619–23.
5. Pietrangelo A. The ferroportin disease: pathogenesis, diagnosis and treatment. Haematologica 2017;102:1972–84.
6. Yan N. Structural biology of the major facilitator superfamily transporters. Annu Rev Biophys 2015;44:257–83.7.
7. Tortosa V, Bonaccorsi di Patti MC, Musci G, Polticelli F. The human iron exporter ferroportin. Insight into the transport mechanism by molecular modeling. Bio-Algorithms Med-Systems 2016;12:1–7.
8. Wallace DF, Harris JM, Subramaniam VN. Functional analysis and theoretical modeling of ferroportin reveals clustering of mutations according to phenotype. Am J Physiol Cell Physiol 2010;298:C75–84.
9. Huang Y, Lemieux MJ, Song J, Auer M, Wang DN. Structure and mechanism of the glycerol-3 phosphate transporter from Escherichia coli. Science 2003;301:616–20.
10. Le Gac G, Ka C, Joubrel R, Gourlaouen I, Lehn P, Mornon JP, et al. Structure-function analysis of the human ferroportin iron exporter (SLC40A1): effect of hemochromatosis type 4 disease mutations and identification of critical residues. Hum Mutat 2013;34:1371–80.
11. Yin Y, He X, Szewczyk P, Nguyen T, Chang G. Structure of the multidrug transporter EmrD from Escherichia coli. Science 2006;312:741–4.
12. Bonaccorsi di Patti MC, Polticelli F, Cece G, Cutone A, Felici F, Persichini T, et al. A structural model of human ferroportin and of its iron binding site. FEBS J 2014;281:2851–60.
13. Dang S, Sun L, Huang Y, Lu F, Liu Y, Gong H, et al. Structure of a fucose transporter in an outward-open conformation. Nature 2010;467:734–8.
14. Taniguchi R, Kato HE, Deshpande CN, Wada M, Ito K, Ishitani R, et al. Outward- and inward-facing structures of a putative bacterial transition-metal transporter with homology to ferroportin. Nat Commun 2015;6:8545.
15. Bonaccorsi di Patti MC, Polticelli F, Tortosa V, Furbetta PA, Musci G. A bacterial homologue of the human iron exporter ferroportin. FEBS Lett 2015;21:3829–35.
16. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–73.
17. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL Workspace: a web-based environment for protein structure homology modelling. Bioinformatics 2006;22:195–201.

18. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: protein structure and function prediction. Nat Methods 2015;12:7–8.

19. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35:3375–82.

20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

21. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, et al. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 2010;38:W695–9.

22. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK – a program to check the stereochemical quality of protein structures. J App Cryst 1993;26:283–91.

23. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res 2014;42:W320–24.

24. Jordan JB, Poppe L, Haniu M, Arvedson T, Syed R, Li V, et al. Hepcidin revisited, disulfide connectivity, dynamics, and structure. J Biol Chem 2009;284:24155–67.

25. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. Bioinformatics 2014;30:1771–3.

26. Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 2007;67:1078–86.

27. Mayr R, Janecke AR, Schranz M, Griffiths WJ, Vogel W, Pietrangelo A, et al. Ferroportin disease: a systematic meta-analysis of clinical and molecular findings. J Hepatol 2010;53:941–49.

28. Jiang D, Zhao Y, Wang X, Fan J, Heng J, Liu X, et al. Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. Proc Natl Acad Sci USA 2013;110:14664–9.

29. Madej MG, Soro SN, Kaback HR. Apo-intermediate in the transport cycle of lactose permease (LacY). Proc Natl Acad Sci USA 2012;109:E2970–8.

30. Jiang D, Zhao Y, Wang X, Fan J, Henga J, Liua X, et al. Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. Proc Natl Acad Sci USA 2013;110:14664–69.

31. Fernandes A, Preza GC, Phung Y, De Domenico I, Kaplan J, Ganz T, et al. The molecular basis of hepcidin-resistant hereditary hemochromatosis. Blood 2009;114:437–43.

32. Preza GC, Ruchala P, Pinon R, Ramos E, Qiao B, Peralta MA, et al. Minihepcidins are rationally designed small peptides that mimic hepcidin activity in mice and may be useful for the treatment of iron overload. J Clin Investigation 2011;121:4880–8.

33. Qiao B, Sugianto P, Fung E, del-Castillo-Rueda A, Moran-Jimenez MJ, Ganz T, et al. Hepcidin-induced endocytosis of ferroportin is dependent on ferroportin ubiquitination. Cell Met 2012;15:918–24.

34. Ross SL, Tran L, Winters A, Lee KJ, Plewa C, Foltz I, et al. Ferroportin internalization requires ferroportin lysines, not tyrosines or JAK-STAT. Cell Met 2012;15:905–17.