Università degli Studi

ROMA TRE

SCUOLA DOTTORALE IN BIOLOGIA (XXII CICLO)

Sezione: SCIENZE BIOMOLECOLARI E CELLULARI

Development of methods for high through-put proteins structure and function prediction through the use of GRID computing infrastructures.

Dottorando: Docente Guida:

Dott. Giovanni Minervini Dott. Fabio Polticelli

ANNO ACCADEMICO 2008-2009

Università degli Studi

ROMA TRE

SCUOLA DOTTORALE IN BIOLOGIA (XXII CICLO)

Sezione: SCIENZE BIOMOLECOLARI E CELLULARI

Sviluppo di metodi per la predizione massiva di struttura e funzione di proteine attraverso l'utilizzo di infrastrutture di calcolo condiviso GRID.

Dottorando:	Docente Guida:
Dott. Giovanni Minervini	Dott. Fabio Polticelli

ANNO ACCADEMICO 2008-2009

BACKGROUND

The number of natural proteins represents a small fraction of all the possible protein sequences and there is an enormous number of proteins never sampled by nature, the so called "never born proteins" (NBPs). A fundamental question in this regard is if the ensemble of natural proteins possesses peculiar chemical and physical properties or if it is just the product of contingency coupled to functional selection. A key feature of natural proteins is their ability to form a well defined 3D structure. Thus, the structural study of NBPs can help to understand if natural protein sequences were selected for their peculiar properties or if they are just one of the possible stable and functional ensembles.

TECHNICAL IUSSUES

The structural characterization of a huge number of random proteins cannot be approached experimentally, thus the problem has been tackled using a computational approach. A large random protein sequences library (2×10^4 sequences) was generated, discarding amino acid sequences with significant similarity to natural proteins, and the corresponding structures were predicted using Rosetta. Given the highly computational demanding problem, Rosetta was ported in grid and a user friendly job submission environment was developed within the GENIUS Grid Portal. Protein structures generated were analysed in terms of net charge, secondary structure content, surface/volume ratio, hydrophobic core composition, etc..

RESULTS

The vast majority of NBPs, according to the Rosetta model, are characterized by a compact three-dimensional structure with a high secondary structure content. Structure compactness and surface polarity are comparable to those of natural proteins, suggesting similar stability and solubility. Deviations are observed in α helix-β sheet relative content and in hydrophobic core composition, as NBPs appear to be richer in helical structure and aromatic amino acids with respect to natural proteins.

CONCLUSIONS

The results obtained suggest that the ability to form a compact, ordered and water-soluble structure is an intrinsic property of polypeptides. The tendency of random sequences to adopt α helical folds indicate that all-α proteins may have emerged early in prebiotic evolution. Further, the lower percentage of aromatic residues observed in natural proteins has important evolutionary implications as far as tolerance to mutations is concerned.

INTRODUZIONE

Il numero di proteine naturali rappresenta una piccola frazione di tutte le possibili sequenze immaginabili. Esiste quindi un numero enorme di proteine potenzialmente non esplorate dalla natura, le cosiddette "Never Born Proteins" (NBPs). Una questione fondamentale, a questo riguardo, è chiarire se l'insieme delle proteine naturali possieda proprietà chimiche e caratteristiche fisiche peculiari, le quali in qualche modo rendano le proteine oggi osservabili in natura, un qualcosa di "irripetibile" o se, meno prosaicamente, esse possano essere solo il prodotto della contingenza unita ad una forte pressione evolutiva operata dalla selezione naturale. Una caratteristica fondamentale delle proteine naturali è la loro capacità di dar luogo ad una struttura tridimensionale ben definita. Lo studio della struttura fine di NBPs può aiutare a capire se le sequenze di proteine naturali siano state scelte per le loro proprietà particolari o se esse stesse rappresentino solo una delle possibili soluzioni in grado di dare luogo ad una struttura stabile e allo stesso tempo funzionale per la vita. Le NBPs proprio per la vastità numerica che ne caratterizza la loro stessa concezione rappresentano una interessante sfida per le attuali tecniche di studio.

Studiarne la struttura significa infatti escogitare un nuovo approccio che permetta di sondare "in maniera massiva" un così ampio bacino di informazioni. Le metodiche sperimentali che si basano su di un approccio computazionale possono aiutare questa esplorazione e fornire allo stesso tempo nuovi mezzi di indagine per lo studio di problemi di complessità analoga, per numero di elementi e qualità della ricerca, come lo studio massivo di interi proteomi.

APPROCCIO METODOLOGICO

La caratterizzazione strutturale di un gran numero di proteine a sequenza casuale non può essere affrontata sperimentalmente, quindi il problema è stato affrontato con un approccio computazionale. E' stata quindi generata una grande libreria di sequenze proteiche casuali (pari a 2×10^4 sequenze), scartando le sequenze amminoacidiche omologhe a proteine note. Le strutture tridimensionali corrispondenti sono stati predette attraverso l'utilizzo del software Rosetta *abinitio*. Data la grande potenza di calcolo richiesta per questo progetto, Rosetta *abinitio* è stato adattato per essere usato in un sistema di calcolo parallelo condiviso: le cosidette GRID. Il porting di tale software è stata anche l'occasione per la creazione, attraverso i tool di sviluppo offetti dal GENIUS Grid Portal, di un'interfaccia grafica

che ne rendesse più familiare l'utilizzo in ambiente GRID. Le strutture proteiche generate sono state analizzate in termini di carica netta, nel contenuto di struttura secondaria, nel rapporto superficie/volume, composizione del nucleo idrofobico, ecc.
Si è inoltre proceduto a sviluppare una metodica statistica che partendo dai risultati ottenuti dall'analisi delle proprietà strutturali, permetta di comprendere quali proprietà rendono una proteina ascrivibile al grande gruppo che forma le proteine naturali. Una siffatta infrastruttura è stata infine usata per affrontare problemi di complessità analoga come la predizione di un gran numero di proteine a funzione ignota da patogeni umani.

RISULTATI

La maggioranza delle NBPs, secondo il modello proprio di Rosetta, sono caratterizzate da una struttura tridimensionale compatta e da un peculiare alto contenuto di struttura secondaria. La struttura generale, la compattezza e la polarità di superficie sono paragonabili a quelli delle proteine naturali, suggerendone pertanto una stabilità e solubilità simili. Deviazioni significative si osservano nel contenuto relativo in α elica e struttura di tipo β e nella composizione del nucleo idrofobico. In generale è possibile affermare che le NBPs sembrano essere più ricche di struttura elicoidale e di residui aromatici rispetto alle proteine naturali. Non meno interessante e' l'osservazione che un campione di NBPs, seppur ristretto, sembra essere caratterizzato dall'avere strutture simili a quelle di proteine esistenti. I risultati finora ottenuti applicando i protocolli sviluppati nella ricerca alle proteine putative del patogeno umano *Streptococcus mutans* evidenziano da un lato la maturità dell'infrastruttura stessa, dall'altro il grande vantaggio che un simile approccio può costituire nello studio di patogeni.

CONCLUSIONI

I risultati riportati in questo lavoro evidenziano come lo studio computazionale di "NBPs", anche se di natura prettamente predittiva, può fornire una panoramica utile sulle proprietà strutturali di base dei polipeptidi e sulle proprietà specifiche delle proteine naturali. Le NBPs appaiono strutturalmente molto simili alle proteine naturali, suggerendo che l'enorme spazio di sequenze NBPs possa essere sfruttato a fini biotecnologici. Una differenza importante tra NBPs e proteine naturali risiede nel diverso rapporto tra aminoacidi aromatici ed alifatici, in particolare nel minor contenuto di amminoacidi aromatici osservati nelle proteine naturali. Queste informazioni possono essere molto utili nella progettazione di esperimenti di evoluzione diretta ed in generale negli studi di ingegneria proteica. Infine, questo studio dimostra che lo sfruttamento di infrastrutture GRID, per progetti di predizione massiva di strutture proteiche è fattibile; da cui è facile immaginare un suo eventuale utilizzo per la predizione della struttura del proteoma di batteri patogeni, per la selezione di nuovi target farmaceutici e negli studi di progettazione di nuove molecole di interesse biomedico.

INDEX

BACKGROUND

The first crystal structure of a macromolecule was solved by Kendrew in 1958, the myoglobin (Kendrew., et al. 1958). This event was a crucial step in biology: in that day, researchers, for the first time, were able to observe a protein in its real form. Fifty years later, however, the easy determination of a protein's structure is considered an issue far from resolved. X-ray crystallography represents a very powerful way of investigation, however it has a very strong limitation: it requires a protein crystal. The techniques used for the crystallization of proteins often require significant economic and technical efforts. To obtain quality crystals apt to yield a high resolution structure it is required the availability of a highly concentrated solution of the protein of interest with a very high degree of purity. It must be also considered that crystallization experiments require long processing times, in other words is quite common that a modern crystallography laboratory spends on average months to solve a single protein structure. Considering that the number of protein coding sequences available in RefSeq Release 38 database (November 11, 2009, www.ncbi.nlm.nih.gov/RefSeq/) is equal to 9.325.214 from 9.166 different organisms and that the number of already solved protein structures is ~ 40.000 (NCBI MMDB - www.ncbi.nlm.nih.gov/sites/entrez?db=structure), it is easy to imagine that the progress of the knowledge about the structure of proteins needs new methods to obtain comparable results in a significantly shorter time. A response to this need is provided by software tools for the prediction of proteins three-dimensional structure. In the last fifteen years in various international laboratories many different software packages have been developed whose clear objective is to allow the prediction of the three-dimensional structure of a protein directly from its amino acid sequence without any prior knowledge of the evolutionary relationships linking the protein of interest to homologous with know structure. The main representatives of this family of programs are the "Rosetta *abinitio* software" (Bonneau et al., 2002, Rohl et al., 2004), the "i-Tasser web server" (Yang, 2008) and "BHAGEERATH web server" (Jayaram et al., 2006). These programs, albeit with limitations, are able to predict the protein fold in a much shorter time if they are compared to crystallography. From this last statement, a reader may be led to think that the structural study of proteins has finally found a fast solution. This is true only in part: the usefulness of these software tools is severely limited by their computational demanding requirements and, more complicated to solve, by the still limited scientific knowledge about the forces that govern the protein folding process. The predictive algorithms approach the simulation of the folding process essentially by simplifying the process through an approximation of the chemical-physical laws that govern protein folding. Moreover some of this algorithms, as the one implemented in Rosetta *abinitio*, combine this simulation procedures with an active search of structural properties on massive databases containing information derived from known protein structures. Computing power, storage space, and archiving of information and results represent a huge challenge even for modern computers. Historically, the first solution adopted to increase the computing power has been to create clusters of computers dedicated to the execution of large computations. A computer cluster is a group of linked computers, working together closely so that in many respects they form a single computer. The components of a cluster are commonly, but not always, connected to each other through fast local area networks. "Clusters are usually deployed to improve performance and/or availability over that of a single computer, while typically being much more cost-effective than single computers of comparable speed or availability" (Baker et al., 2001). The increase in computing power obtained with clusters, however, highlights the intrinsic limits of prediction software. The reliability of this software is good for predictions of proteins with amino acid sequence limited to 200-300 amino acids. Beyond this value the CPU time required for computation makes improper the use of *ab initio* software (several days of computation for each prediction). Nevertheless the protein structure predictions obtained in compliance with this limit clearly show the advantages of this "*in silico"* approach. The recent successes obtained by software packages such as Rosetta in CASP-5 (Critical Assessment of Structure Predictions) competitions confirm however that this kind of software tools can achieve a very high level of reliability (Chivian et al., 2003).

Ab initio protein structure prediction methods allow to undertake the study of large protein sequence datasets with no homology with known natural proteins which can shed light on the nature of the protein folding process. In fact, simple calculations show that the number of known natural proteins is just a tiny fraction of all the theoretically possible sequences. On the other hand, considering random polypeptides of just 100 amino acids in length (the average length of natural proteins being 367 amino acids [Bairoch et al., 2004]), with the 20 natural amino acids co-monomers it is possible to obtain 100^{20} chemically different proteins. This is an astronomically large number which leads to the consideration that there is a tremendous amount of information hidden in these unknown proteins. Previous studies in the field of the origin of life, have identified these sequences as "Never born proteins" or NBPs (Chiarabelli et al., 2006). These studies arised the fundamental question if the set of known natural proteins have particular features which make them eligible for selection, in terms, for example, of particular thermodynamic, kinetic or functional properties. One of the key features of natural protein sequences is their ability to fold and form a stable and well defined three-dimensional structure which in turn dictates their specific biological function (Branden et al., 1999). From this viewpoint, the study of the structural features of NBPs can help to answer the question if the natural protein sequences were selected during molecular evolution because they have unique properties and which are such properties (for instance a peculiar amino acid composition, hydrophobic/hydrophilic amino acids ratio, etc.). Attempts have been made in this direction, often using innovative experimental techniques such as phage display (Chiarabelli et al., 2006). These tests indicate that a computational approach can be extremely useful: software tools as Rosetta, if properly supported by adequate computing power, would allow to produce results otherwise obtainable only through years of experimental activity. In addition, a computational approach allows to evenly sample the protein sequences space in different regions far away enough from the ensemble of natural proteins. However, to obtain statistically significant results the size of the sequence data base to be analyzed must be sufficiently large (at least 10^4 to 10^5 sequences). This is a highly demanding problem from a computational viewpoint. In fact on a single CPU it would require years of computing time to predict the structure of such a large number of protein sequences. Fortunately, from a computational viewpoint this is an embarassingly parallel problem in that the same computation (*i.e.* the prediction of the three-dimensional structure of a protein sequence) must be repeated several times (*i.e.* on a large number of protein sequences). The use of a cluster of computers might provide a solution. However, the execution of thousands of calculations, even considering a scenario in which the code used has been heavily optimized for parallelization, requires years of computation; in addition in this same period the cluster would not be available to run other jobs. The ideal solution for this kind of problems is to use “grid computing” (GRID).

GRID was developed in 1999 by Foster with the invention of the "Globus Kit" (a collection of libraries for writing highly parallel software). (Foster et al., 1999). Scientists active in the Physics of particles field had highlighted the need of an innovative approach to computation. In fact physicists needed to run simple calculations (execution time of the order of a few minutes), but in numbers equal to or greater than millions (an example being the thousands of trajectories which have to be calculated to describe the movement of individual sub-atomic particles in hadron collider experiments) (Brodkin., 2008). It is easy to understand that a single computer, however powerful, is not a good choice to tackle this problem. The GRID is based on thousands of computers connected together via Internet and managed by a dedicated software infrastructure: the middleware (Berman et al., 2003, Parvin et al., 2005). GRID is mainly formed by groups of computers called "computing elements" grouped in working nodes. The working nodes are the entities which physically perform the calculations. Data retention is handled by "storage elements" (computer with low computing power, but with very large capacity hard drives). Everything is managed by a "virtual organizations system" which underpin the management of GRID users, data traffic control and ensure privacy and safety (Chau-Hsu., 2006) (Figure 1).

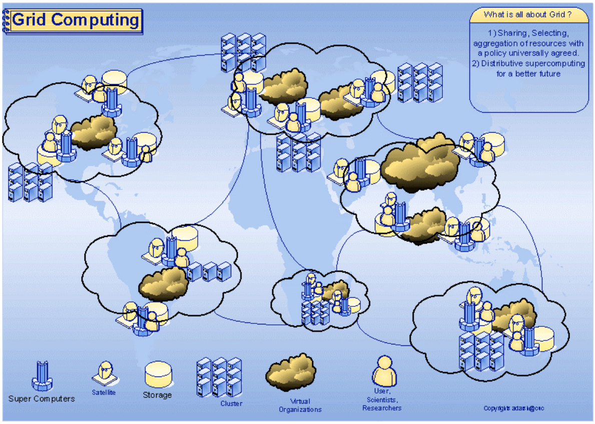

Grid Computing
What is all about Grid ?
1) Sharing, Selecting, aggregation of resources with a policy universally agreed.
2) Distributive supercomputing for a better future
Super Computers
Satellite
Storage
Cluster
Virtual Organizations
User, Scientists, Researchers
Copyright: adash@cic

Figure 1. Grid computing World Wide (from INFM web site)

Although the GRID offers numerous benefits, it requires deep modifications of the software that has to be used on it. A shared environment on global scale is affected by inevitable data transmissions delays. For a software which is designed for processing sequential information this inconvenience must be intended as a lack of execution and computation failure. The software which does not have been natively written for operating in a GRID environment must be adapted through the use of special developing tools. This process is called *gridification* (Mateos et al., 2007). With this aim several programming languages have been written; one of the most famous is the “Job Description Language” (JDL) (Tierney et al., 2001, Burke et al., 2008). The JDL is a scripting language, developed specifically for GRID, characterized by dedicated instructions for monitoring of computing resources, for allowing the jobs submission, for their identification and, not least, for checking the operator's credentials (Burke et al., 2008). JDL is however a very complex language, very powerful but not very easy to be handled by users who don’t have a strong informatics background. Important progress has been made with the introduction of appropriately programmed web interfaces, as the “Genius GRID Portal” (https://genius.ct.infn.it/, http://www.infn.it/, http://www.nice-italy.com/). These web portals can be seen as friendly graphical user interfaces for the utilization of GRID. The prediction of the three-dimensional structure of non natural proteins using Rosetta on a GRID environment, though yields only predictive results, allows to evenly sample the protein sequences space in different regions far away enough from the ensemble of natural proteins. The GRID infrastructure applied to the study of proteins may allow us to understand if natural protein sequences were selected during molecular evolution because they have unique physico-chemical properties or else they just represent a contingent subset of all the possible proteins with a stable and well defined fold. If the latter hypothesis were true, this would mean that the protein realm could be exploited to search for novel folds and functions of potential biotechnological and/or biomedical interest. The same protocols developed to predict the structure of an enormous number of random proteins can be very useful for the study of whole proteomes of biomedical interest. The genomes of several pathogens of medical interest are available in databases such as NCBI (NCBI genome - www.ncbi.nlm.nih.gov/projects/genome/). Frequently many pathogens proteins display little or no homology with better known proteins. In other cases the three-dimensional structure of pathogens proteins has been determined by structural genomics initiatives, but there is no information about their function. Very often there is no other information available apart form the amino acid sequence. An example is the genome of the bacterium *Streptococcus mutans* (*S. mutans*): it contains 1.963 ORFs (open reading frames), 63% of which have been assigned putative functions based on sequences homology (Figure 2). *S. mutans* is the principal causative agent of human dental caries (tooth decay). Dental caries is one of the most common infectious diseases afflicting humans, and tends to remain untreated in many underdeveloped areas, leading to considerable suffering that is often alleviated only by the loss or extraction of the infected tooth (World Health Organization, Geneva 1994). Although 200–300 bacterial species have been found associated with dental plaque, only the presence of *S. mutans* has been consistently linked with the formation of human dental caries (Loesche., 1986). Additionally, *S. mutans* is occasionally associated with non-oral infections, principally sub-acute bacterial endocarditis (Hezenberg., 2000)" (Ajdić et al., 2002). Since 2002, year of the sequencing of *S. mutans* genome, up to now, out of the 63% of proteins with putative function a large part still retains this status (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Protein+Table&list_uids=263). Cases like this one highlight the need to have alternative methods for massive protein structures analyses.

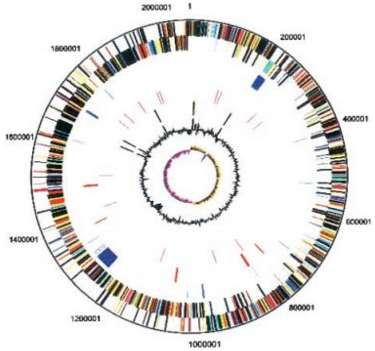


Figure 2. Circular representation of the *S. mutans*, strain UA159 genome. The outer two circles show the position of the probable ORFs on the complementary DNA strands. TheORFs have been color coded by functional category (from NCBI web site)

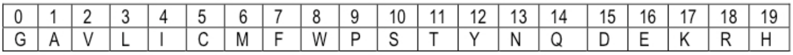
AIM OF THE WORK

Knowledge of the three-dimensional structure of proteins is a fundamental prerequisite for the understanding of biological systems and for the development of new molecules of biological interest. Classical methods such as X-ray crystallography or NMR, while very powerful, require large investments of time and money. On the other hand, the development of software packages for *ab initio* prediction of protein structures has revolutionized the current scientific landscape. These software tools while very promising, however, show their limitations in speed of computation and consequently in their applicability, especially in scenarios involving high flow of data. The shared computing seems to be a good opportunity to enhance the capability of these specific software tools. Unexplored systems such as random protein sequences represent a huge pool of potential compounds for the development of new (macro)molecules, but the huge number of sequences that characterizes NBPs, however, does not allow to approach their study with existing techniques. The first objective of this research is to adapt existing software packages for the prediction of protein structure on shared computing environments. At the same time our aim is to make computing grids accessible to a public not specialized in computer sciences, through the creation of workplace-based graphical user interfaces. The realization of a platform able to handle a large flow of information finds its natural application in studies with high scientific impact such as the NBPs problem or the prediction of the entire proteome of bacterial pathogens.

TECHNICAL ISSUES

Amino acid sequences library generation

Random amino acid sequences (70 amino acids long) were generated using the utility RandomBLAST developed during this project (Evangelista et al., 2007). Briefly, RandomBLAST consists of two main modules: a pseudo random sequence generation module and a BLAST software (Altschul et al., 1990) interface module. The first module is a pseudo-random number generation wich uses the Mersenne Twister 19973 algorithm (Matsumoto et al., 1998) to generate pseudo-random numbers between 0 and 19. Random numbers are translated in single character amino acid code using the conversion matrix show in Table 1 and then concatenated to reach the desired sequence length.



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G	A	V	L	I	C	M	F	W	P	S	T	Y	N	Q	D	E	K	R	H

Table 1. RandomBLAST random number to amino acid type conversion table

Each sequence generated is then given in input to the second RandomBLAST module, an interface to the blastall program which invokes the following command:

blastall -m 8 -p BLASTp -d database -b 1;

where **database** in our case stands for the NR database, and the parameters **–m 8** and **–b 1** indicate the alignment format (tabular form) and the number of sequences to be returned (just the first hit), respectively. BLASTall output is then retrieved by RandomBLAST and the *Evalue* extracted from it. If the *Evalue* is greater than or equals the threshold chosen by the user, the sequence is added to the output file. Note that in this case only the sequences that do not display significant similarity to any protein sequence present in the database are considered valid, so that, contrary to the normal BLAST usage, valid sequences are those displaying an *Evalue* higher than the threshold, set to a value of 1 (Karlin et al., 1990) (Figure 3). The total number of NBPs sequences generated was 20496.

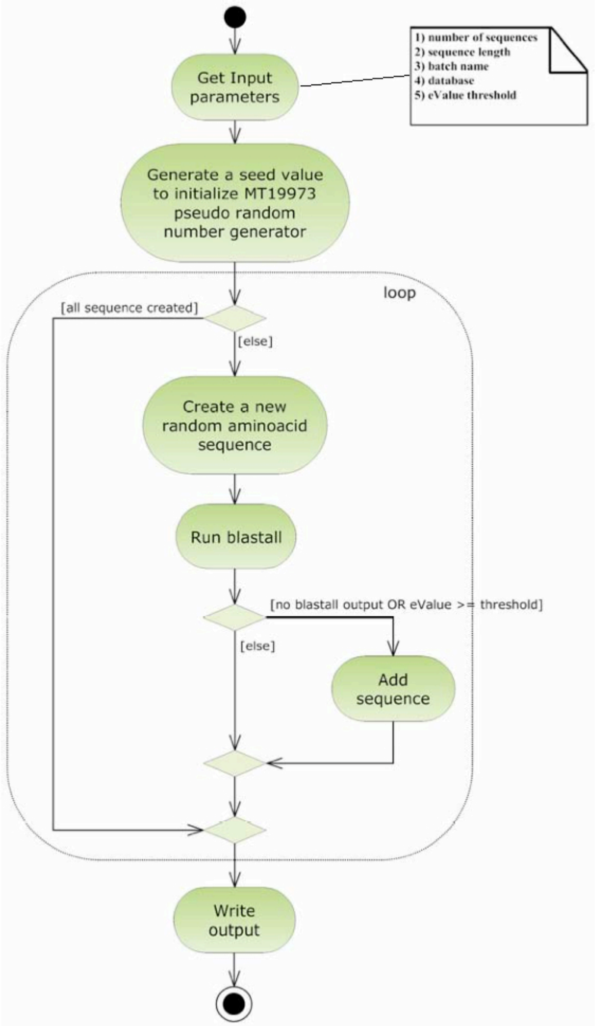


Figure 3. Activity diagram showing the RandomBLAST workflow. The inset details the RandomBLAST input parameters (from Evangelista et al., 2007).

NBPs three-dimensional structure prediction

NBPs three-dimensional structures have been predicted using Rosetta *abinitio* software. Rosetta *abinitio* is an *ab initio* protein structure prediction software which is based on the assumption that in a polypeptide chain local interactions bias the conformation of sequence fragments, while global interactions determine the three-dimensional structure with minimal energy which is also compatible with the local biases (Rohl et al., 2004). To derive the local sequence-structure relationships for a given amino acid sequence (the query sequence) Rosetta *abinitio* uses the Protein Data Bank (Berman et al., 2002) to extract the distribution of conformations adopted by short segments in known structures. The latter is taken as an approximation of the distribution adopted by the query sequence segments during the folding process (Rhol et al., 2004).

In detail, Rosetta workflow can be divided into two modules:

Module I - Input generation - The query sequence is divided in fragments of 3 and 9 amino acids. The software extracts from the data base of protein structures the distribution of three-dimensional structures adopted by these fragments, based on their specific sequence. For each query sequence a fragments data base is derived which contains all the possible local structures adopted by each fragment of the entire sequence. The procedure for input generation is rather complex due to the many dependencies of module I. In fact, to be executed the first Rosetta *abinitio* module needs the output generated by the programs BLAST (Altschul et al., 1990) and PSIPRED (McGuffin et al., 2000) in addition to the non redundant NCBI protein sequence database (Wheeler et al., 2005). On the other hand this procedure is computationally inexpensive (10 min of CPU time on a Pentium IV 3,2 GHz). Thus it has been chosen to generate the fragments database locally with a *Perl* script that automatizes the procedure for a large dataset of query sequences. The script retrieves query sequences from the random sequence database in FASTA format and executes Rosetta *abinitio* module I creating an input folder with all the files needed for the execution of Rosetta *abinitio* module II.

Module II - *Ab initio protein structure prediction* – Using the derived fragments database and the PSIPRED secondary structure prediction generated by module I for each query sequence, the sets of fragments are assembled in a high number of different combinations by a Monte Carlo procedure by Rosetta *abinitio* module II. The resulting structures are then subjected to an energy minimization procedure using a semi-empirical force field (Rohl et al., 2004). The principal non-local interactions considered by the software are hydrophobic interactions, electrostatic interactions, main chain hydrogen bonds and excluded volume. The structures compatible with both local biases and non-local interactions are ranked according to their total energy resulting from the minimization procedure. A single run with just the lowest energy structure as output takes approx. 10-40 min of CPU time, for a 70 amino acids long NBP and depending on the degree of refinement of the structure.

Integration of Rosetta software on GRID enviroment

Any software to be used in a GRID environment must be prepared through a process called "Gridification”. The porting from a local working environment to a shared system requires an optimization process to operate in such an environment. Operations such as managing system calls to runtime libraries, management of I/O and memory usage must be adapted to the particolar requirements of a shared enviroment (processing unit and storage unit tipically located in different geographic areas). In a system GRID all these issues are solved by one elegant way: the program “believes” to be in a local system and all the necessary management calls are entrusted by a virtual operating system that is responsible for coordination (shared mainframe). All these aspects are managed through the use of a special programming language: Job Description Language (JDL) *(www.grid.org.tr/servisler/dokumanlar/DataGrid-JDL-HowTo.pdf). Using ad hoc* JDL files that describe necessary inputs, outputs are then generated on the resources required by the software. A typical example of a JDL file is as follows:

```
Executable = "/ bin / echo";
Arguments = "Hello World";
StdOutput = "message.txt";
StdError = "stderror";
OutputSandbox = ( "message.txt", "stderror");
Requirements = other.LRMS;
Type == "BPU";
Rank = other.FreeCPUs;
```

where “Executable” is the name of the binary file of the software ported on GRID sistem, “Arguments” is the name of eventual flags required by the binary for its execution, “StdOutput” is the name of the output file, “StdError” is an optional file where are written the error informations, “OutputSandbox” is a list with the name and the number of output files produced at the end of one computation (generally one output file and one error log file), “Requirements” is a string for the hardware requirements (RAM memory amount, Hard Drive space, ecc), “BPU” is the type of CPU used by the binary file and “Rank” specifies the number of CPU necessary for job submission. In the example, the program requires a site with a generic Basic Processing Unit (BPU) tipically a x86 CPU. In this way the job will be submitted only on resources that satisfy this condition. If there is more than one resource satisfying the required conditions, then the rank is used to determine which is the most appropriate and then the one to which the job is submitted (higher values correspond to more resources suitable). JDL is based on Condor libraries (www.bo.infn.it). To check the results of the matching resource without actually submitting a job, the edg-job-list-match command can be used. This command returns an ordered list of resources, in which the most appropriate one appears at the top. A JDL file was created to run Rosetta *abinitio* Module II executable and proceed with workflow execution on the EUChinaGRID (www.euchinagrid.org/) GRID working nodes which use the gLite middleware (http://glite.web.cern.ch/glite/).

Three-dimensional structures analysis

The analysis of the physico-chemical properties of the predicted protein structures was carried out using a collection of different tools. Given the high number of NBPs to be analysed, for each tool the analysis procedure was automated using *ad hoc Perl* scripts. In detail, the programs used were MSMS (Scanner et al., 1996), for molecular volume calculation, SURFace Algorithms (Sridharan et al., 1992), for surface properties analysis (overall molecular surface, per residue solvent accessibility), Freqaa (Tekaia et al., 2002), for amino acid composition analysis and DSSP (Kabsch et al., 1983) for secondary structure content analysis. Surface hydrophobicity was calculated as the ratio between the solvent exposed surface of hydrophobic amino acids and the total solvent exposed surface, both calculated using SURFace Algorithms (Sridharan et al., 1992). To compare the properties of NBPs structures to those of naturally occurring proteins structures, a dataset of natural proteins of length comparable to that of NBPs (55 to 95 amino acids long sequences as compared to NBPs 70 amino acids long sequences) was derived from the Protein Data Bank (Berman et al., 2000). The dataset was cleaned up eliminating protein fragments and proteins whose fold is determined by macromolecular complexes formation (tipically ribosomal proteins). The final natural proteins dataset was formed by 866 proteins.

Statistical analysis of the data

A first exploratory data analysis has been developed to see if there were any significant differences in the structure observed in the two data-sets. Initially few outliers in the data that could affect the analyses were removed, generating a dataset of 18465 NBPs and a dataset of 839 natural proteins. For these sets measures of location, index of dispersions, correlations matrix were derived, and box-plots and scatterplots were built to compare the two data sets. This study was performed on different structure-related variables, which include: volume, surface, surface/volume ratio, net charge, secondary structure content, and surface hydrophobicity. Tests on the Gaussian distribution of the variables led to reject the hypothesis of Gaussianity for the majority of the variables investigated. With a test significance level of 0.05 almost all the variables result with statistically different mean and variance for the two data-sets. The analysis has been also conducted on smaller data-sets of comparable size: a random sample of 1000 observations has been drawn from NBPs dataset and comparisons have been developed. The two analyses generated similar conclusions, presented in the Results section. The statistical software used to analyse the data was R (http://www.R-project.org).

Reliability of three-dimensional protein structure prediction software

In a context of international cooperation related to the European Commission funded project EUChinaGRID and with the aim of testing the reliability of three-dimensional protein structure prediction software when used in an "exotic" context as the structures determination of random proteins, a close cooperation has been established with the research group of Prof. Irena Roterman of the Jagellonian University of Krakow (Poland). The structure of the random sequences library generated in the first phase of the research were predicted through an independent software and based on an alternative approach to that of Rosetta *abinitio*, currently in development by their group. The three-dimensional structures of NBPs were predicted with their “Fuzzy oil drop” method (Twostage program) (Brylinsk et al., 2006; Brylinski et al., 2006b). The resultant predictions were compared with the Rosetta *abinitio* output. The RMSD was calculated only for the backbone atoms. Couples with RMSD less than 6.80 Å were subjected to energy minimization using the software CHARMM (Brooks et al., 1083) in order to reduce the possible clashes between atoms introduced by the prediction software. The RMSD was recalculated for the resulting structure. The structure alignment and its graphic representation were obtained through the use of the CE protein structures alignment software (Shindyalov et al., 1998).

Discrimination Tree

In order to compare properties and functionalities of the random proteins with respect to natural proteins a classification tree was generated. The analysis has been conducted comparing 10 different variables: net charge, volume, surface, β turn, α helix, β strand, % α helix, % β turn, % of secondary structure content and surface hydrophobicity. A first description of the data by evaluating measures of location, dispersion indices, correlations matrices, and box-plots and scatterplots in order to evaluate differences and similarities of the two data sets was developed. With a significance level of 0.05 almost all the variables involved exhibited statistically different means and variances, and tests on the Gaussian distribution of the variables lead to reject the hypothesis of Gaussianity for the majority of the variables investigated. This exploratory data analysis shows significant differences in the structural components of the two data sets. The difference in mean and variance in the two groups of proteins, if highlighting the diversity of the two groups, on the other way prevents the study of the differences through the classical statistical methods. Especially the lack of Gaussianity makes complicated and methodologically incorrect the application of the common linear regression models which are conventionally applied to the study of biological systems. In order to discover and understand the basis of these differences a classification procedure was built by evolving populations of feedforward neural networks. We chose a class of multi-layers neural networks (Haykin et al., 2009, De March et al., 2009) and built a genetic algorithm (De Jong 2006) to evolve the best classification rule.

More specifically, a first random population of networks was generated with the topology of a 2-hidden layers neural networks. This population was formally described as a set of sequences with dichotomic variables (each sequence was a vector of zeros - ones values). Each element of the sequence describes the presence or the absence of a particular structural component of a protein. The topology of these networks, involving different variable compositions, was selected in a random way (first generation of networks), and the response of each network was derived with a two class structure: natural and synthetic proteins. Then, a genetic algorithm was built to evolve the population of networks in a number of generations with the aim of identifying a precise classification rule. Then the response of each network was evaluated deriving a net misclassification rate by 10-fold cross validation procedure: the sequences with smaller values were identified as the more promising solutions. Then the classical genetic operators, such as natural selection, crossover and mutation was applied to the network population, in order to achieve the next generation of promising sequences. These genetic operators are needed to allow the proper learning, by algorithm, of the variability of the variables studied with the neural network. This process is repeated for a fixed number of times or until convergence is reached (end of the learning phase).

Streptococcus mutans **proteins three-dimensional structure prediction**

To evaluate the usefulness of the structure prediction computing platform developed in the study of bacterial proteomes, a feasibility study was carried out on the proteome of the human pathogen *S. mutans*. *S. mutans* protein sequences were retrieved from the NCBI web site, project genome section (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Protein+Table&list_uids=263). The entire *S. mutans* proteome as of 20 march 2009 has approximately 800 protein noted as putative proteins for which do not exist three-dimensional structures or information of homology with known proteins. From this subset of putative proteins a reduced library was extracted consisting of proteins with amino acid sequence length between 25 and 200. During the development of this work a new version of the Rosetta software (Rosetta 2.0) has been released. This version is basically a rewrite of the original Fortran code in C++ programming language. Rosetta 2.0 is characterized by having a better performance and a greater ease of use. The new version differs from the first also for the integration of a module for clustering of the predicted structures. Through this procedure it is possible to isolate from a large set of intermediate models, the prediction that best describes all the predictions produced (in other words "the most similar between the group"). It was therefore decided to use the new version to conduct the prediction of the three dimensional structure of *S. mutans* proteins. To maximize the reliability of the used protocol, different calculations were carried out on samples of natural proteins whose structure was already present in the NCBI database. These preliminary tests showed that to obtain a reliable prediction (global variation compared to crystal structure with average RMSD value less than 3Å) about 990 intermediate structures are required. This value was chosen as the best compromise between computing time and similarity of the proposed model. The protocol used for the prediction of *S. mutans* proteins therefore takes the following form:

1) generation of fragments via the Rosetta fragment module;
2) prediction of 990 intermediate models;
3) clustering of the structures and extraction of the ideal model;
4) refinement of the model.

The computational time needed for the prediction of a single structure with this protocol is of the order of about a week per structure.

Prediction of putative function

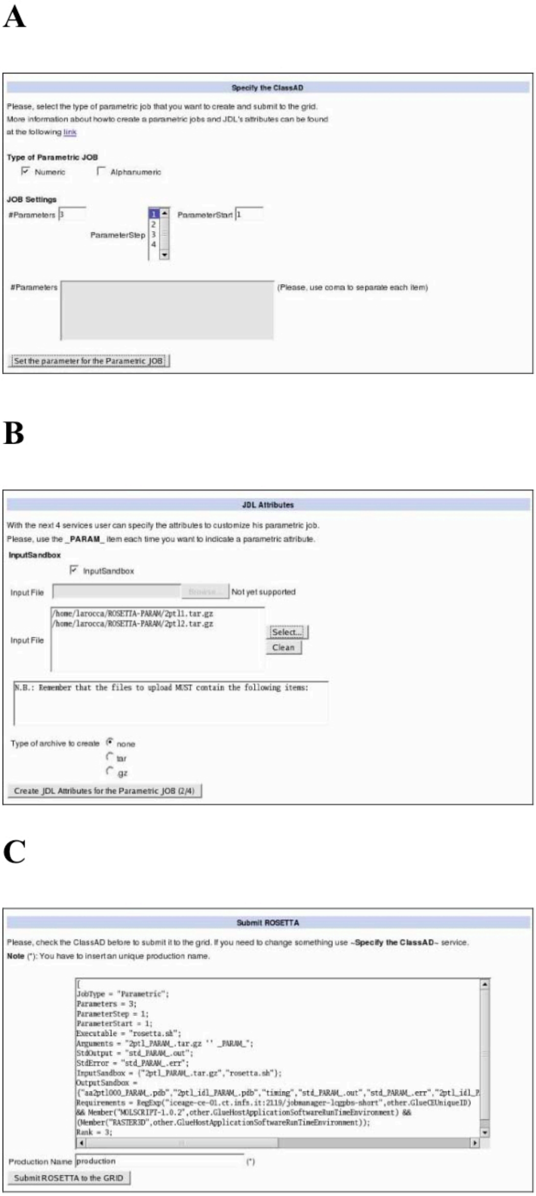
The assignment of a putative function was conducted through the use of the Dali web server, which allows the comparison of the predicted structures with all the protein structures present in the Protein Data Bank and the retrieval of protein structures which display significant structural similarity, and hence similar function, with the input structure (Holm et al., 2008).

RESULTS

Integration of Rosetta Module II on the GILDA GRID infrastructure

Single job execution on Grid Infrastructure - A single run of Rosetta *abinitio* Module II consists of two different phases. In the first phase an initial model of the protein structure is generated using the fragment libraries and the PSIPRED secondary structure prediction. The initial model is then used as input for the second phase in which it will be idealised. A shell script has been prepared which registers the program executable (pFold.lnx) and the required input files (fragment libraries and secondary structure prediction file) on the LFC catalog, calls the Rosetta *abinitio* Module II executable and proceeds with workflow execution. A JDL file was created to run the application on the EuChinaGrid working nodes which use the gLite middleware (gLite middleware - http://glite.web.cern.ch/glite/).

Integration on the GENIUS web portal - A key issue to attract the biology community towards the exploiting of the GRID paradigm is to overcome the difficulties connected with the use of the grid middleware by users without a strong background in informatics. This is the main goal that has to be achieved in order to disseminate the use of GRID services by biology applications. To achieve this goal and allow a wide biologists community to run the software using a user friendly interface, Rosetta *abinitio* application has been integrated on the GENIUS (**G**rid **E**nabled web e**N**vironment for site **I**ndependent **U**ser job **S**ubmission) GRID Portal (GENIUS Portal – https://genius.ct.infn.it/), a portal developed by a collaboration between the italian INFN Grid Project (INFN Grid Project – http://www.infn.it/) and the italian web technology company Nice (Nice – http://www.nice-italy.com/). Thanks to this GRID portal, non-expert users can access a GRID infrastructure, execute and monitor Rosetta *abinitio* application only using a conventional web browser. All the complexity of the underlying GRID infrastructure is in fact completely hidden by GENIUS to the end user. In our context, given the huge number of NBPs sequences to be simulated, an automatic procedure for the generation of parametric JDL files has been set up on the GENIUS Grid Portal. With this procedure, exploiting the features introduced by the last release of the gLite middleware, users can create and submit parametric jobs to the GRID. Each submitted job independently performs a prediction of the protein structure. Hereafter is described in detail the workflow adopted to run Rosetta *abinitio* application on GENIUS. After the user has correctly initialized his personal credentials on a MyProxy Server (security check), he can connect to the GENIUS portal and start to set up the attributes of the parametric JDL file that will be created "on the fly" and then submitted to the GRID. First the user specifies the number of runs, equivalent to the number of amino acid sequences to be simulated (Figure 4A). Then, the user specifies the working directory, the name of the shell script (Rosetta *abinitio* executable) to be executed on a GRID resource, loads a .tar.gz input file for each query sequence (containing the fragment libraries and the PSIPRED output file) and specifies the output files (initial and refined model coordinates) in parametric form (Figure 4B). The parametric JDL file is then automatically generated and visualised in order to be inspected by the user and submitted (Figure 4C).



Specify the ClassAD

Please, select the type of parametric job that you want to create and submit to the grid.
More information about howto create a parametric jobs and JDL's attributes can be found
at the following link

Type of Parametric JOB

☑ Numeric ☐ Alphanumeric

JOB Settings

#Parameters 3

ParameterStep 1

ParameterStart 1

#Parameters (Please, use coma to separate each item)

JDL Attributes

With the next 4 services user can specify the attributes to customize his parametric job.
Please, use the **_PARAM_** item each time you want to indicate a parametric attribute.

InputSandbox

☑ InputSandbox

Input File Not yet supported

Input File

```
/home/larocca/ROSETTA-PARAM/2ptl1.tar.gz
/home/larocca/ROSETTA-PARAM/2ptl2.tar.gz
```

```
N.B.: Remember that the files to upload MUST contain the following items:
```

Type of archive to create ● none
○ tar
○ gz

Submit ROSETTA

Please, check the ClassAD before to submit it to the grid. If you need to change something use ~**Specify the ClassAD**~ service.
Note (*): You have to insert an unique production name.

```
[
JobType = "Parametric";
Parameters = 3;
ParameterStep = 1;
ParameterStart = 1;
Executable = "rosetta.sh";
Arguments = "2ptl_PARAM_.tar.gz '' _PARAM_";
StdOutput = "std_PARAM_.out";
StdError = "std_PARAM_.err";
InputSandbox = {"2ptl_PARAM_.tar.gz","rosetta.sh"};
OutputSandbox =
{"aa2ptl000_PARAM_.pdb","2ptl_idl_PARAM_.pdb","timing","std_PARAM_.out","std_PARAM_.err","2ptl_idl_P
Requirements = RegExp("iceage-ce-01.ct.infn.it:2119/jobmanager-lcgpbs-short",other.GlueCEUniqueID)
&& Member("MOLSCRIPT-1.0.2",other.GlueHostApplicationSoftwareRunTimeEnvironment) &&
(Member("RASTER3D",other.GlueHostApplicationSoftwareRunTimeEnvironment));
Rank = 3;
```

Production Name production (*)

Figure 4. Screenshots of the GENIUS GRID portal showing services for the specification of the number of structure predictions to run (top panel), of the input and output files (middle panel) and for the inspection of the parametric JDL file (bottom panel) (from Minervini et al., 2007).

The status of the parametric job as well as the status of individual runs of the same job can be also checked from within the GENIUS portal. When the prediction is done it is also possible, using the portal, to inspect the output produced in graphics form. In addition, in order to allow the user to analyse the predicted NBP structural model, the JMOL Java applet (http://www.jmol.org/) has been embedded into the GENIUS portal. Figure 5 shows some, randomly chosen, examples of the approx. 2×10^4 NBPs three-dimensional structures predicted using Rosetta on grid (Rohl et al., 2004).

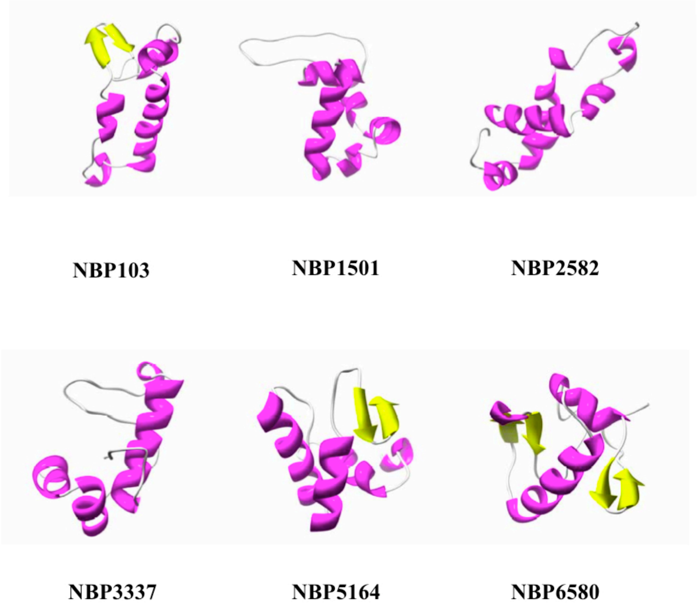
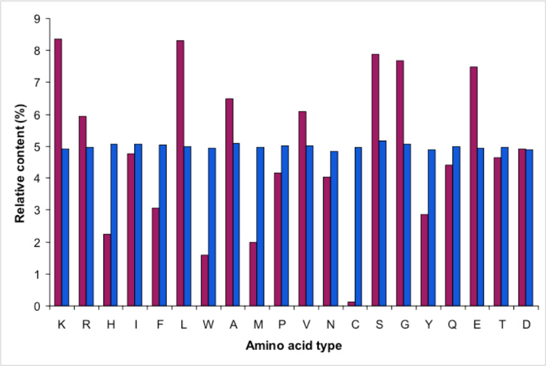


Figure 5. Schematic representation of the three-dimensional structure of randomly chosen NBPs, α helices are coloured in magenta, β strands in yellow (from Minervini et al., 2009).

Amino acid composition analysis

Figure 6 shows the amino acid composition of natural and random protein sequences datasets. For random proteins dataset, each amino acid relative content is about 5%, an obvious consequence of the random nature of these amino acid sequences. Nonetheless, interesting differences are observed when random sequences composition is compared to that of natural proteins. In detail, for the natural proteins dataset five amino acids are significantly over represented: Lys (8.35%), Leu (8.30%), Ser (7.88%), Gly (7.69%) and Glu (7.48%), while three amino acids are under represented: Trp (1,58%), Met (1,98%) and Cys (0,13%) (Figure 6). These differences are not a peculiar characteristic of the subset of natural proteins chosen in this work. In fact similar amino acid composition is observed for the UniProt dataset containing all the known protein sequences (Bairoch et al., 2004) (Figure 6), indicating that the natural proteins subset used in this work is representative of all natural proteins, at least as far as amino acid composition is concerned. Notable differences are observed for Lys and Cys residues which display the highest and lowest percentage in the natural dataset. The former finding can be explained considering that a consistent number of natural proteins with a length in the 55-95 residues range display nucleic acid binding activity and thus a basic character. Cys residues percentage is less straightforward to explain even though it is probably connected to the high reactivity of this amino acid whose presence in natural proteins is tightly evolutionary controlled. Turning to the random proteins dataset, these display a significant excess of aromatic amino acids and a strikingly lower content in Leu, as also evidenced by the low ratio between aliphatics and aromatics as compared to the natural dataset (Table 2).

A



B

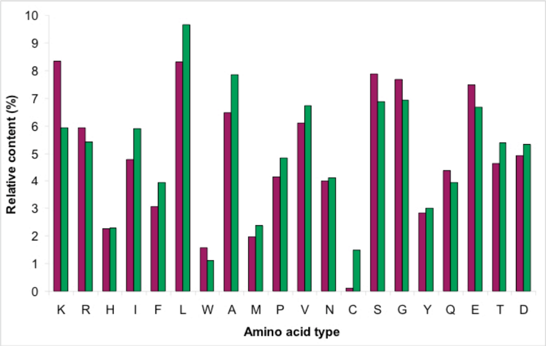
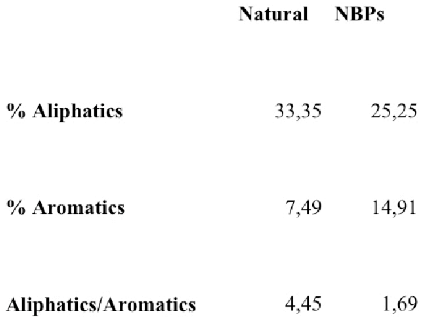


Figure 6. Amino acid composition of NBPs and natural proteins A) Comparison between the amino acid composition of NBPs (blue bars) and that of selected natural proteins of comparable length (purple bars; see Methods). B) Comparison between the amino acid composition of the natural proteins subset used in this work (purple bars) and that of the whole UniProt dataset (Bairoch et al., 2004) (green bars) (from Minervini et al., 2009).

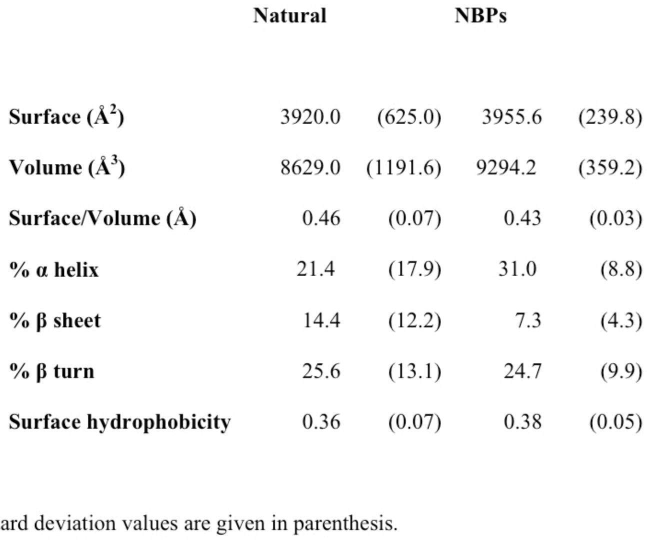


	Natural	NBPs
% Aliphatics	33,35	25,25
% Aromatics	7,49	14,91
Aliphatics/Aromatics	4,45	1,69

Table 2. Hydrophobic amino acids relative content of natural proteins and NBPs (from Minervini et al., 2009).

Comparative structural analysis

Several structure-related parameters (volume, surface, surface/volume ratio, secondary structure content, and surface hydrophobicity) have been computed for the two datasets in order to compare their statistical and structural properties (Table 3). As a general consideration, the average value of the analysed structural parameters and the corresponding standard deviation values are statistically different between NBPs and natural proteins with a significance level of 0.05. In particular natural proteins are characterised by a higher standard deviation whereas NBPs seem to be narrowly distributed around the experimental average. Despite the differences observed in amino acid composition between the two datasets, the structural analysis of NBPs reveals that these are in most cases characterized by a well ordered structure. In fact, secondary structure content of NBPs appears to be comparable to that of natural proteins, with an average total secondary structure content (including α helix, β sheet and β turn) slightly higher than 60% for both datasets (Table 3). However, NBPs appear to be less compact than natural proteins, as evidenced by the higher average volume (9294.0 Å and 8630.0 Å for NBPs and natural proteins, respectively, Table 3) and lower surface/volume ratio (0.43 Å and 0.46 Å for NBPs and natural proteins, respectively, Table 3) (Figure 7A). Interestingly, NBPs display a significantly higher α helix content, and a corresponding lower β sheet content, with respect to natural proteins (Figure 7B), in the absence of any clear correlation between amino acid composition and amino acids propensities for formation of a specific secondary structure. Surface hydrophobicity of the two datasets has also been calculated and results to be very similar (Figure 8A), indicating a predicted water solubility of NBPs structures comparable to that of natural proteins. Interestingly, comparison of the amino acid composition of the two datasets with data relative to solvent accessibility of different amino acids types highlights how aromatic amino acids are more represented in the hydrophobic core of NBPs with respect to natural proteins. As an example, Trp residues are approx. three times more frequent in NBPs with respect to natural proteins (Figure 6A). However, solvent accessible surface of NBPs Trp residues is, on average, only twice that of natural proteins Trp residues (Figure 8B), indicating that a higher proportion of Trp residues is buried within the hydrophobic core of NBPs. Similar considerations apply to Phe and Tyr residues (compare Figures 6A and 8B), leading to the conclusion that aromatic residues contribute to NBPs hydrophobic core formation to a higher degree than in natural proteins.

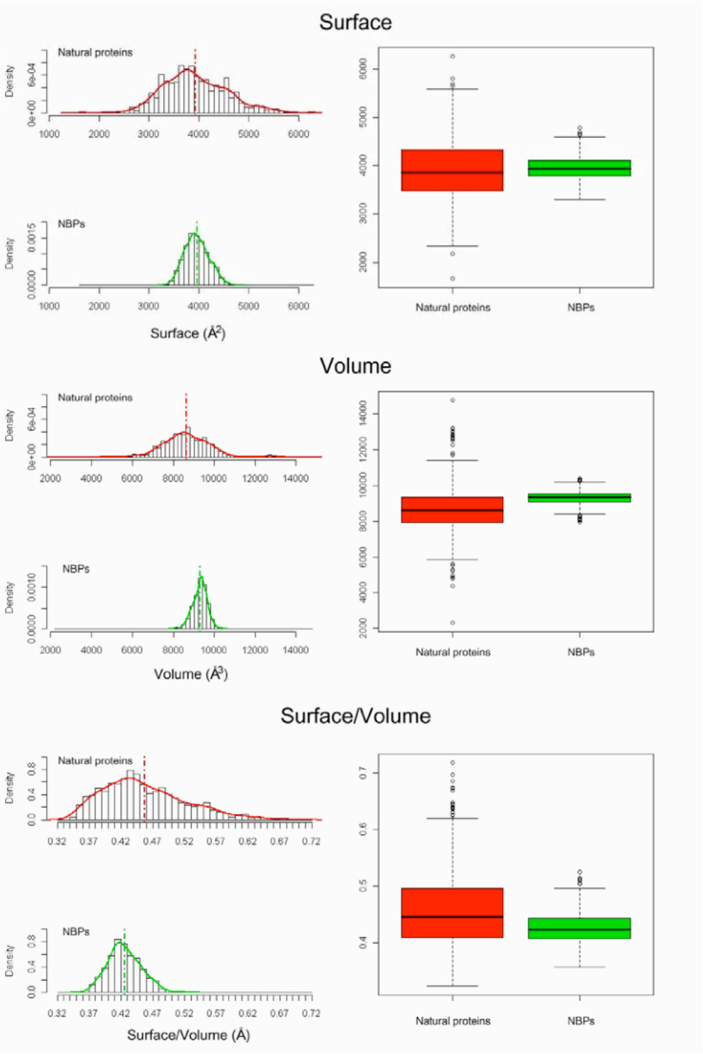


	Natural		**NBPs**	
Surface (Å^2)	3920.0	(625.0)	3955.6	(239.8)
Volume (Å^3)	8629.0	(1191.6)	9294.2	(359.2)
Surface/Volume (Å)	0.46	(0.07)	0.43	(0.03)
% α helix	21.4	(17.9)	31.0	(8.8)
% β sheet	14.4	(12.2)	7.3	(4.3)
% β turn	25.6	(13.1)	24.7	(9.9)
Surface hydrophobicity	0.36	(0.07)	0.38	(0.05)

[a]Standard deviation values are given in parenthesis.

Table 3. Average values of the structure-related parameters calculated for natural proteins and NBPs (from Minervini et al., 2009).

A


Surface
Natural proteins
Density
0e+00
6e-04
1000
2000
3000
4000
5000
6000
NBPs
0.0000
0.0015
Surface (Å²)
2000
3000
4000
5000
6000
Natural proteins
NBPs
Volume
Natural proteins
2000
4000
6000
8000
10000
12000
14000
NBPs
0.0000
0.0010
Volume (Å³)
Surface/Volume
Natural proteins
0.0
0.4
0.8
0.32
0.37
0.42
0.47
0.52
0.57
0.62
0.67
0.72
NBPs
Surface/Volume (Å)
0.4
0.5
0.6
0.7

B

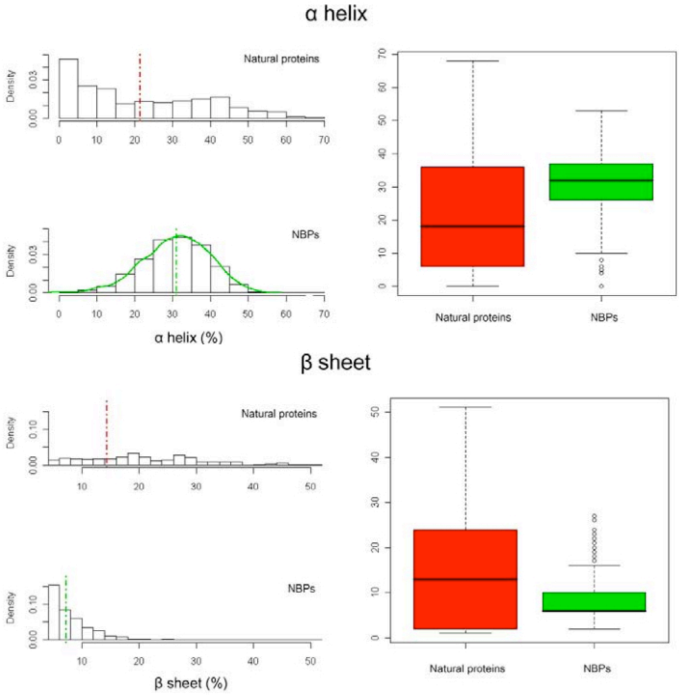


Figure 7. Structural properties of NBPs and natural proteins A) Surface, volume and surface/volume ratio distribution for NBPs and natural proteins; B) Secondary structure content of NBPs and natural proteins. In this and in the following figure, boxplots are shown in the right panels (from Minervini et al., 2009)

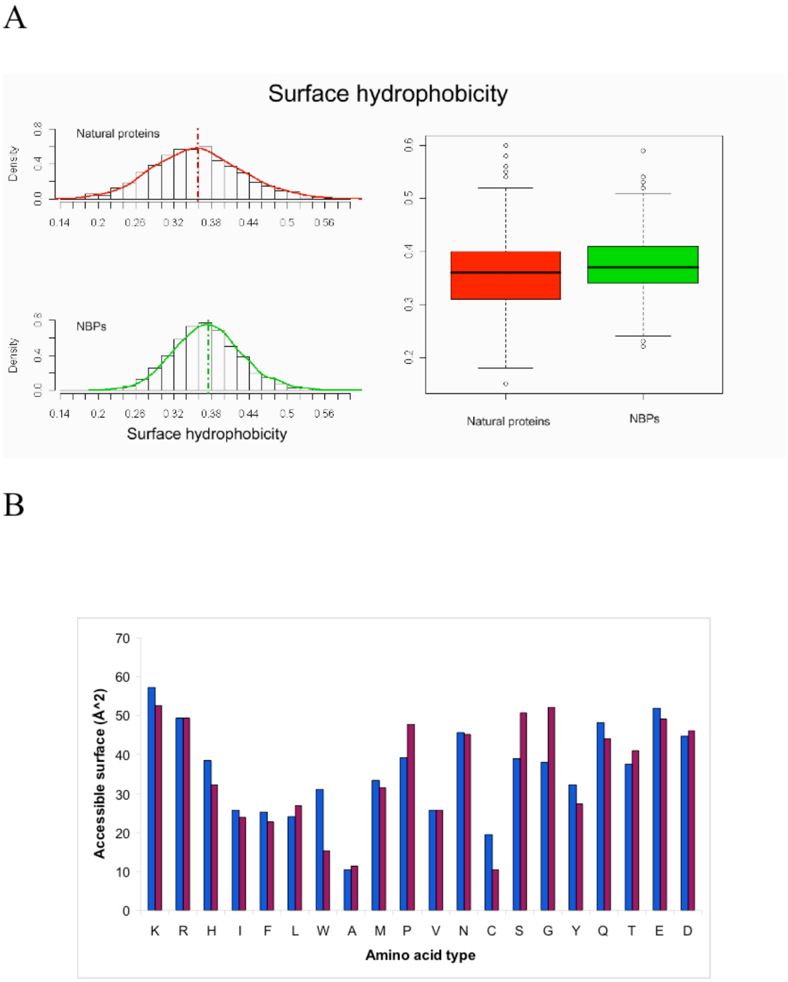
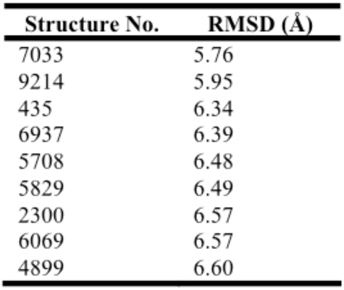


Figure 8. Surface properties of NBPs and natural proteins. A) Surface hydrophobicity of NBPs and natural proteins; B) Amino acids solvent accessibility for NBPs (blue) and natural proteins (purple) (from Minervini et al., 2009).

Comparison of Rosetta *abinitio* and Twostage results

In the framework of an international collaboration within the EU funded EUChinaGRID project (http://www.euchinagrid.org), NBPs structures predicted with Rosetta *abinitio* were compared to those predicted with a different software based on the "fuzzy oil drop" method (Twostage program) (Brylinski et al., 2006; Brylinski et al., 2006b). Table 4 clearly shows that RMSD values obtained are never lower than 5.86 Å which represents the best value produced by the pair of predictions for the NBPs sequence 7033. Although this is not an exciting result, it is important in that highlights the challenging conditions in which the two software packages were tested. For both methods, the predicted structure is formed largely by alpha helices (Figure 9, F1d and Figure 9, F1e). The related topology maps (Figure 9, F1a and Figure 9, F1b) show that both predictions are characterized by the presence in the C-terminus region of two segments of alpha helix (Rosetta, residues 43-50 and 51-61; Twostage, 43-49 and 62-58), the first of which can be viewed as the central axis of the molecule. Around this central axis are located two other alpha-helical regions interspersed by short loops, which seem to embrace the center with the shape of a ring (Rosetta, 20-28 and 29-36; Twostage, 21-29 and 35-39). In total disagreement however, is the prediction for the N-terminus region for which Rosetta predicts the presence of a short beta-sheet while Twostage identifies a further stretch of alpha helix (Rosetta, 5-2 and 8-12; Twostage, 9 -14). The overlap of the predicted structures relative to sequence number 7033 (Figure 9, F1f), illustrates some agreement between the two predictions. The structural alignment (Figure 9, F1c) clearly shows the presence of a good structural similarity and if the region of disagreement at the N-terminus is excluded, the calculated RMSD value decreases to ~4.39 Å, reinforcing the agreement between the two methods.



Structure No.	RMSD (Å)
7033	5.76
9214	5.95
435	6.34
6937	6.39
5708	6.48
5829	6.49
2300	6.57
6069	6.57
4899	6.60

Table 4: List of the first nine structures arranged for better RMS-D value (from Minervini et al 2008).

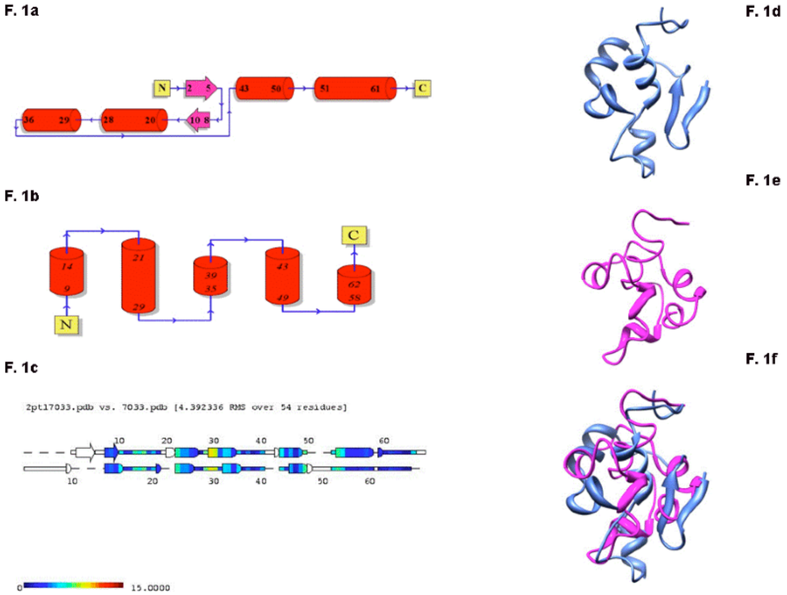


Figure 9. Schematic representation of topology maps, structural alignment and graphics representation of predicted structures for sequence n° 7033. F1a and F1b, topology maps; F1c, structural alignment between structures predicted with two different methods; F1d and F1e, structures obtained with Rosetta and Twostage, respectively; F1f, superimposition of the two proposed predictions (from Minervini et al., 2008).

Evolutionary Neural Network Approach

The application of an Evolutionary Neural Network Approach (ENNA) to classify random versus natural protein sequences yielded interesting results. In figure 10 is reported, for 10 generations, the best sequence response value (black line) and mean sequence response value (blue line) in terms of misclassification rate, and the number of times in which each sequence element is selected in the last generation. More specifically in Figure 10, is showed how the error rate of discrimination between the two groups decreases dramatically at the tenth generation. This indicates that the algorithm developed is able, after ten cycles of learning, to distinguish the two groups with good confidence. Figure 10 also shows the average rate of correct classification obtained in comparison with a single sequence of known source which is used as benchmarck. The results show that when using a particular set of structural components, the method used is able to classify the proteins in a quite precise way. The best set of selected variables provides a classification accuracy equal to 89%: by evaluating the properties related to these variables it is thus possible to identify a natural protein from a random one and predict which protein structures potentially display specific and relevant functionalities. Furthermore a high degree of reliability of the components identified is observed, because of the high probability with which they are selected (as shown in figures 10b and 11). The small number of the selected structural components is also a relevant result and can be a very useful insight for protein engineering. Comparison with the classification tree approach (Hastie et al., 2009) shows the better performance of the evolutionary approach. The decision tree was built by the C4.5 algorithm using 10-fold cross validation and fixing to 10 the minimum number of observations per leaf. The decision tree has a classification accuracy of 90%.

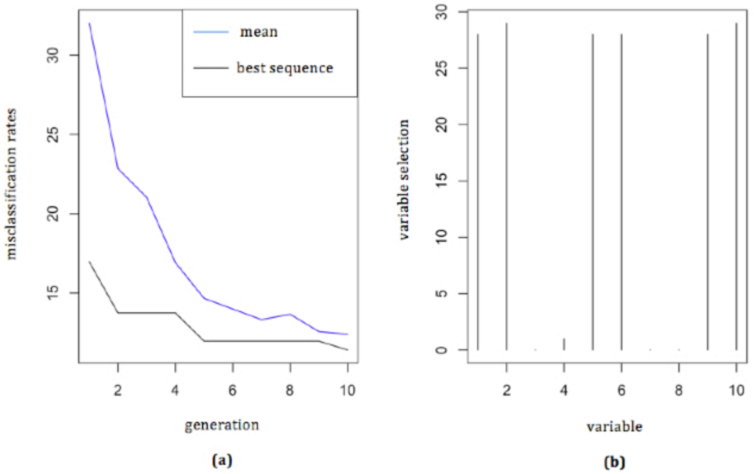


Figure 10. a) Best and mean misclassification rates in 10 generations; b) Number of times in which each variable is selected in the last generation.

The classification tree has confirmed that is possible to distinguish natural proteins from NBP. Analysis of the variables that best discriminate the two groups shows that the set of natural proteins is significantly different for the variables volume, surface, helix and strand (Figure 11). It is very interesting to note that a small group of NBPs (approximately 300 proteins) has characteristics consistent with the average properties of natural proteins.

This group whose members were identified as "pseudo-natural proteins" is composed of random proteins which are identified by the classification algorithm as natural proteins. On this restricted set of pseudo-natural proteins the protocols developed for the study of *S. mutans* proteins are currently being applied. The first encouraging results (under validation) show that these particular NBPs not only exhibit chemical-physical properties comparable to those of natural proteins, but also general fold similar to proteins that actually exist in nature as is the case of protein_263 showed in Figure 12.

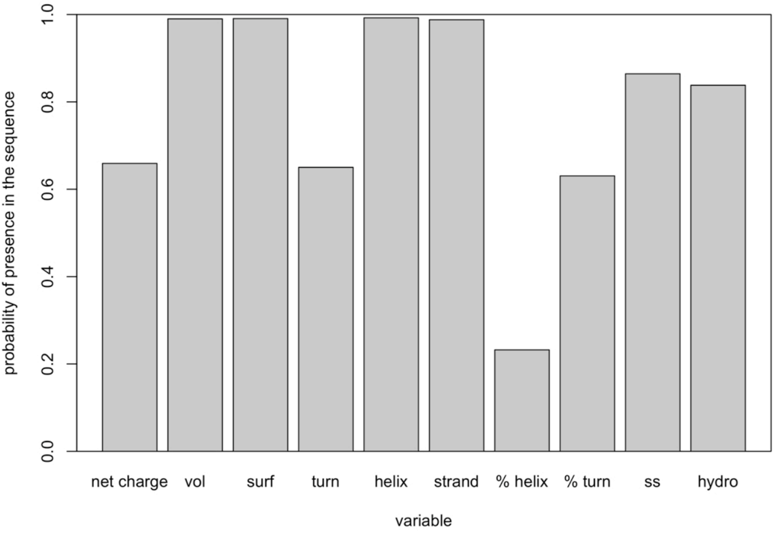


Figure 11. Variables that best discriminate the NBPs group from the natural proteins group; vol stands for volume, surf for surface, ss for secondary structure content and hydro for hydrophobicity.

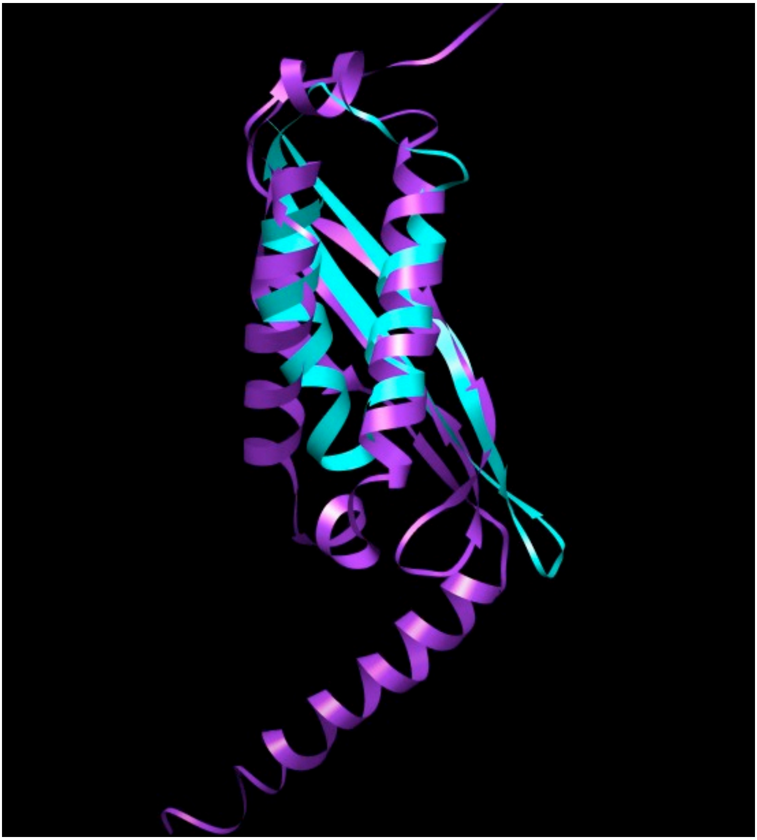


Figure 12. Structural superimposition between protein_263 (turquoise) and cristal structure of a core domain of stomatin from *Pyrococcus horikoshii* (purple, PDB code 3BK6).

Implementation of the protocols to the case of real proteomes: preliminary results

The usefulness of the structure prediction computing platform developed in this thesis for the study of bacterial proteomes was tested using the human pathogen *S. mutans* as a test case. The three dimensional structure of about 200 *S. mutans* proteins was predicted, which represent 10% of the entire proteome of the bacterium. This value takes on a greater significance when we consider that it represents about 30% of the proteins annotated as putative proteins in the bacterial genome. All predicted models are currently under verification and the assignment of putative functions to the proteins is currently underway. To verify the reliability of the described approach, the structure of 5 additional proteins of *S. mutans*, for which is available the experimentally determined three-dimensional structure, have been predicted. All the predicted models show a good degree of agreement with the relative crystal structure and the average RMSD value is less than 3Å An example of this preparatory predictions is showed in Figure 13. The model obtained for the DnaD-like replication protein (PDB code 2ZC2, RMSD 2,20 Å) well explains the level of accuracy obtainable.

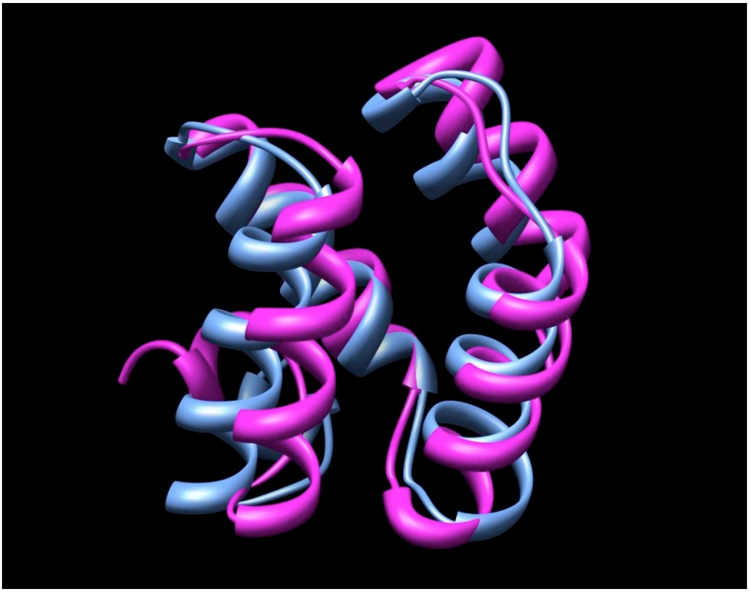
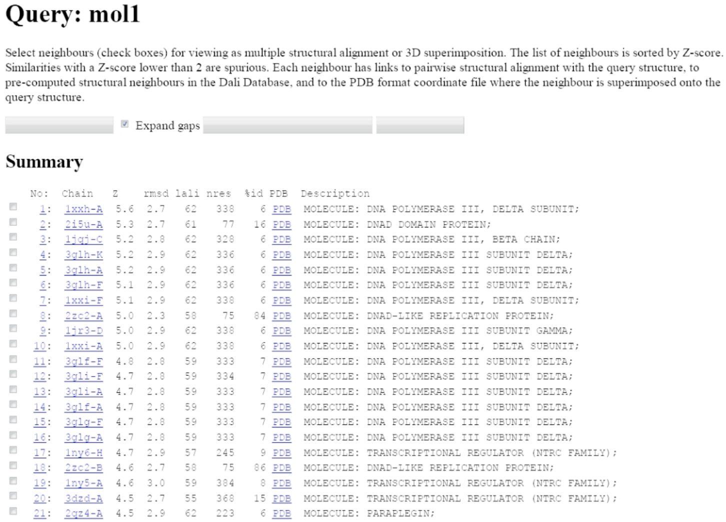


Figure 13. Superimposition between the crystal structure of the DNAD-like replication protein from Streptococcus mutans UA159 (cyan, PDB code 2ZC2) and the model obtained with Rosetta *abinitio* (RMSD 2,20 Å).

For the same group was also tested the accurancy of the function prediction offered by the utilization of DALI web server when it is used on a theoretical model. In figure 14 is possible to see how the abinitio model of 2ZC2 is well recognized as DNA-D domain.



Query: mol1

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, to pre-computed structural neighbours in the Dali Database, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

☑ Expand gaps

Summary

```
    No:  Chain    Z    rmsd lali nres  %id PDB  Description
□    1:  1xxh-A  5.6  2.7   62   338    6 PDB  MOLECULE: DNA POLYMERASE III, DELTA SUBUNIT;
□    2:  2i5u-A  5.3  2.7   61    77   16 PDB  MOLECULE: DNAD DOMAIN PROTEIN;
□    3:  1jqj-C  5.2  2.8   62   328    6 PDB  MOLECULE: DNA POLYMERASE III, BETA CHAIN;
□    4:  3glh-K  5.2  2.9   62   336    6 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□    5:  3glh-A  5.2  2.9   62   336    6 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□    6:  3glh-F  5.1  2.9   62   336    6 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□    7:  1xxi-F  5.1  2.9   62   338    6 PDB  MOLECULE: DNA POLYMERASE III, DELTA SUBUNIT;
□    8:  2zc2-A  5.0  2.3   58    75   84 PDB  MOLECULE: DNAD-LIKE REPLICATION PROTEIN;
□    9:  1jr3-D  5.0  2.9   62   338    6 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT GAMMA;
□   10:  1xxi-A  5.0  2.9   62   338    6 PDB  MOLECULE: DNA POLYMERASE III, DELTA SUBUNIT;
□   11:  3glf-F  4.8  2.8   59   333    7 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□   12:  3gli-F  4.7  2.8   59   334    7 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□   13:  3gli-A  4.7  2.8   59   333    7 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□   14:  3glf-A  4.7  2.8   59   333    7 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□   15:  3glg-F  4.7  2.8   59   333    7 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□   16:  3glg-A  4.7  2.8   59   333    7 PDB  MOLECULE: DNA POLYMERASE III SUBUNIT DELTA;
□   17:  1ny6-H  4.7  2.9   57   245    9 PDB  MOLECULE: TRANSCRIPTIONAL REGULATOR (NTRC FAMILY);
□   18:  2zc2-B  4.6  2.7   58    75   86 PDB  MOLECULE: DNAD-LIKE REPLICATION PROTEIN;
□   19:  1ny5-A  4.6  3.0   59   384    8 PDB  MOLECULE: TRANSCRIPTIONAL REGULATOR (NTRC FAMILY);
□   20:  3dzd-A  4.5  2.7   55   368   15 PDB  MOLECULE: TRANSCRIPTIONAL REGULATOR (NTRC FAMILY);
□   21:  2qz4-A  4.5  2.9   62   223    6 PDB  MOLECULE: PARAPLEGIN;
```

Figure 14. The results N°1-16 show how the correct function recognition can be obtained with DALI web server even when it is applied to theoretical models

As an example of the results being obtained on *S. mutans* proteins with unknown structure and only putative function, the prediction of the structure of SMU.28 is shown in figure 15. SMU.28 is assigned to be a multidrug ABC transporter on the basis of its sequence homology with the protein Sav1866 from *Staphylococcus aureus* (PDB code 2ONJ). The comparison between the model obtained for SMU.28 and the crystal structure of Sav1866 shows that strong structural correlation exists (Figure 15). In this case, the assignment of function obtained by structural homology by DALI server, confirms the putative function assigned to the hypothetical protein SMU.28 (Figure 16).

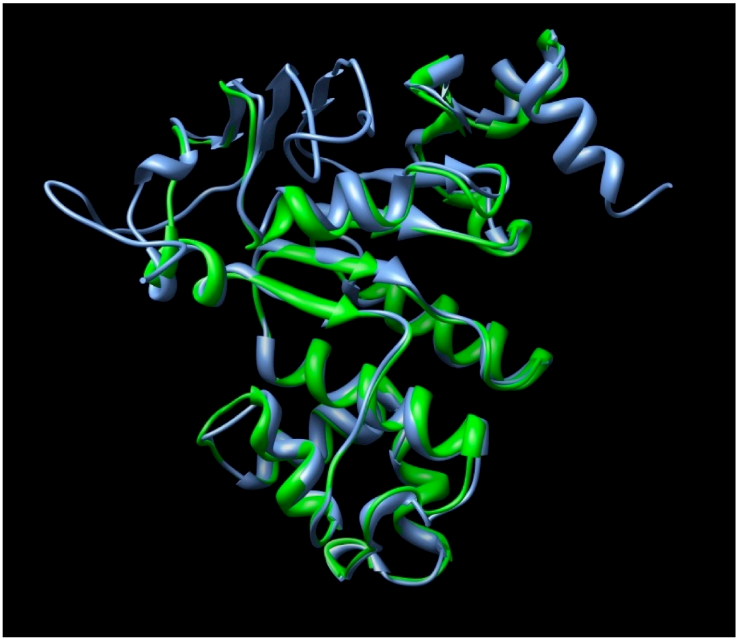
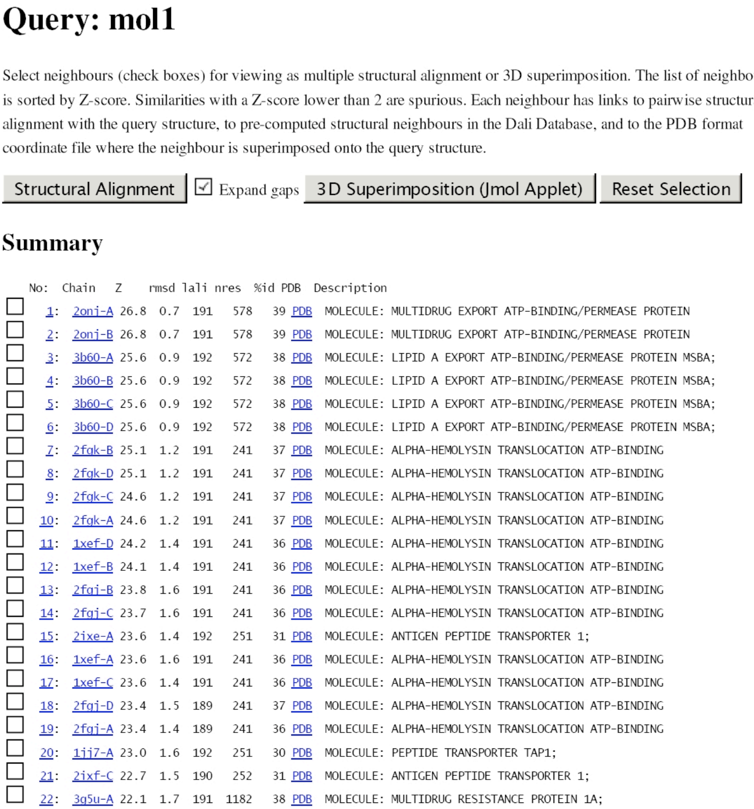


Figure 15. Superimposition between the crystal structure of the multidrug ABC transporter Sav1866 from *Staphylococcus aureus* (cyan, PDB code 2ONJ) and the model obtained for the putative protein SMU.28 from *S. mutans* (green).



Query: mol1

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, to pre-computed structural neighbours in the Dali Database, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Summary

```
    No:  Chain   Z    rmsd lali nres  %id PDB  Description
□    1:  2onj-A 26.8  0.7  191   578   39 PDB  MOLECULE: MULTIDRUG EXPORT ATP-BINDING/PERMEASE PROTEIN
□    2:  2onj-B 26.8  0.7  191   578   39 PDB  MOLECULE: MULTIDRUG EXPORT ATP-BINDING/PERMEASE PROTEIN
□    3:  3b60-A 25.6  0.9  192   572   38 PDB  MOLECULE: LIPID A EXPORT ATP-BINDING/PERMEASE PROTEIN MSBA;
□    4:  3b60-B 25.6  0.9  192   572   38 PDB  MOLECULE: LIPID A EXPORT ATP-BINDING/PERMEASE PROTEIN MSBA;
□    5:  3b60-C 25.6  0.9  192   572   38 PDB  MOLECULE: LIPID A EXPORT ATP-BINDING/PERMEASE PROTEIN MSBA;
□    6:  3b60-D 25.6  0.9  192   572   38 PDB  MOLECULE: LIPID A EXPORT ATP-BINDING/PERMEASE PROTEIN MSBA;
□    7:  2fgk-B 25.1  1.2  191   241   37 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□    8:  2fgk-D 25.1  1.2  191   241   37 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□    9:  2fgk-C 24.6  1.2  191   241   37 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   10:  2fgk-A 24.6  1.2  191   241   37 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   11:  1xef-D 24.2  1.4  191   241   36 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   12:  1xef-B 24.1  1.4  191   241   36 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   13:  2fgj-B 23.8  1.6  191   241   36 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   14:  2fgj-C 23.7  1.6  191   241   36 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   15:  2ixe-A 23.6  1.4  192   251   31 PDB  MOLECULE: ANTIGEN PEPTIDE TRANSPORTER 1;
□   16:  1xef-A 23.6  1.6  191   241   36 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   17:  1xef-C 23.6  1.4  191   241   36 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   18:  2fgj-D 23.4  1.5  189   241   37 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   19:  2fgj-A 23.4  1.4  189   241   36 PDB  MOLECULE: ALPHA-HEMOLYSIN TRANSLOCATION ATP-BINDING
□   20:  1jj7-A 23.0  1.6  192   251   30 PDB  MOLECULE: PEPTIDE TRANSPORTER TAP1;
□   21:  2ixf-C 22.7  1.5  190   252   31 PDB  MOLECULE: ANTIGEN PEPTIDE TRANSPORTER 1;
□   22:  3g5u-A 22.1  1.7  191  1182   38 PDB  MOLECULE: MULTIDRUG RESISTANCE PROTEIN 1A;
```

Figure 16. Functional assignament obtained for hypotetical protein SMU.28 using DALI web server.

An encouraging result for the validation of this approach is provided also by a publication appeared during the development of this research in which it is presented the crystal structure of the putative *S. mutans* protein SMU.440 obtained with standard protocols (Nan et al., 2009). The figures 17A and 17B show the superimposition of the crystal structure 3IJT and the model predicted using the protocols developed in this research

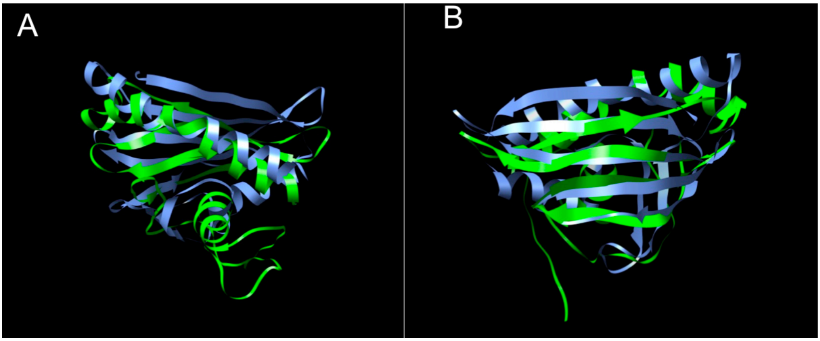


Figure 17. Superimposition between the crystal structure of the putative protein SMU.440 (cyan) from *S. mutans* (PDB code 3IJT) and the model obtained with the protocols developed in this research (green). A) front view; B) rear view.

DISCUSSION

Meaningful interpretation of the results described in the present work rely heavily on the validity of the structure predictions obtained using Rosetta. However, the Rosetta model has been shown to perform fairly well and even yield near-atomic resolution structures in a number of cases (Bradley et al., 2005). Results shown in Figure 18 for a sample of natural proteins confirm that Rosetta predictions are in most cases fairly accurate in terms of overall fold, secondary structure content and topology. In some cases the agreement between the experimental and predicted structures is even surprising, as is the case of the predicted structure of the protein *nusa* (indicated in figure 4 with the PDB code 1UL9) which displays an overall backbone RMSD of only 1.74 Å with respect to the experimentally determined structure.

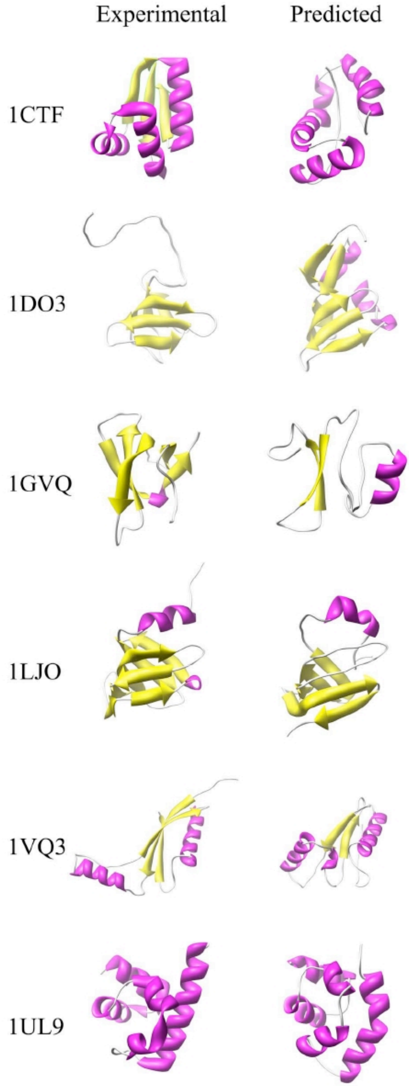


Figure 18. Sample natural proteins predicted three-dimensionalstructures. Schematic representation of the experimental (left panel) and predicted (right panel) three-dimensional structures of randomly chosen natural proteins. α helices are coloured in magenta, β strands in yellow. Protein Data Bank identification codes are indicated on the left for reference. (from Minervini et al., 2009).

Analysis of the structural properties of the predicted NBPs structures yielded several interesting and in some cases counterintuitive results. In fact one would expect that in a large population of random amino acid sequences, a large proportion would be “unfoldable” and thus unstructured. Given the assumption of the Rosetta model, which however has been shown to perform fairly well, our results indicate that this is not the case. Indeed most of the NBPs structures are compact and well ordered, as indicated by the average surface/volume ratio and secondary structure content (Figure 7A and Table 3). Surface polarity is similar to that of natural proteins (Figure 8) suggesting that water solubility is an intrinsic property of random polypeptides. The main differences observed between NBPs and natural proteins are the lower compactness and the higher α helix content of the former. The lower compactness observed for NBPs is probably related to their significantly higher aromatics/aliphatics ratio with respect to natural proteins (Table 2). In fact, a higher proportion of aromatic residues in NBPs results in a hydrophobic core composition more prone to packing “defects”, given the rigid character of aromatic sidechains with respect to branched aliphatic residues such as Leu. Indeed, Leu is largely over represented in natural proteins while the opposite is observed for aromatic residues. The latter finding has important evolutionary implications. In fact a hydrophobic core made up of branched aliphatic amino acids is probably more tolerant to mutations, in that residue substitutions are more easily accomodated by conformational changes of the flexible aliphatic side chains. Regarding secondary structure content, NBPs display a higher α helix content with respect to natural proteins and a very low β sheet content (Figure 7B and Table 3). This could be related to the local nature of the interactions within the α helix. In fact a helical fold can accommodate random sequences by packing together α helical elements interrupted by loops in which bad helix forming residues are located. This is much more difficult in β sheets in which precise pairing of β strands, far away from each other along the amino acid sequence, is required to form a stable structure. From this point of view it can be hypothesized that helical folds are more tolerant to random amino acid sequences. This is a fascinating hypothesis that would be very interesting to test experimentally. In fact in a prebiotic scenario, in which the first polypeptides were probably characterized by random amino acid sequences, α helix could have emerged early as an intrinsic structural property of polypeptides (Minervini et al., 2009).

Based on the differences existent between the two groups is extremely fascinating the possibility to investigate more thoroughly about the properties still hidden in these NBPs. A ranking algorithm capable of learning by its error seems to be a very promising way to identify what are the fundamental characteristics that make a protein "natural". If we stop to think that NBP is just a protein with a random sequence is not surprising the ability to discriminate correctly the natural proteins from random proteins. Nevertheless it is particularly interesting that a subset of NBPs are characterized by natural-like features. The observation that the three-dimensional structure of these pseudo-natual proteins is somehow compatible with what already exists in nature tends to suggest that the variability of forms and folds that we see today in real proteins is largely due to initial conditions in which the first polypeptide chains were formed: an aqueous medium together with the limited availability of certain amino acids. In this light, therefore the proteins commonly found in organisms could be considered the result of a deep biological evolution made by nature on a small number of "keys common structures".

From a technical viewpoint, this work has proved that is possible to use in a shared computing environment applications which were not designed for this purpose. This thesis work has also showed that, to be truly usable by the biologists scientific community, GRID infrastructure must be simplified in its basic commands and features, and in general should adopt a common graphical interface. Another result that emerges between the lines of this project is the goodness of Rosetta *abinitio* software. A semi-empirical approach as the generation of protein-fragments based on existing proteins results dramatically according to the evidence that the protein designed to be "different" from the natural like NBPs, are similar to existing fold. Unlike other methods such as the software "twostage", though promising, clearly bring to light the need to continue to study the principles that govern the protein fold. Considering the results obtained with the two software packages, it is clear that both are capable of simulating the formation of secondary structure fragments (obtaining in this respect also a very good agreement), but the comparison also shows that the program "Twostage" partly fails on the determination of the tertiary structure. Probably the “Twostage” assumption that the amino acid residues of a protein can be approximated with spheres and that the whole protein could be described by a spheroid (Brylinski et al., 2006; Brylinski et al., 2006b) is not always in agreement with the many different protein folds that exist in nature.

The preliminary results obtained predicting proteins of unknown function in *Streptococcus mutans* indicates that the protocols and infrastructure used in the case of NBPs are mature. The structures generated show also that in the case of natural sequences the reliability of the *abinitio* prediction is fairly good. At the same time, however, the large number of intermediate models that are needed to produce a good model makes clear the need of using a GRID infrastructure to face this kind of challenges.

CONCLUSIONS

Results reported in this work highlight how the computational study of “never born proteins”, though predictive in nature, can give a useful insight on the basic structural properties of polypeptides and on the specific properties of natural proteins. NBPs appear to be structurally very similar to natural proteins, suggesting that the enormous sequence space of NBPs could indeed be exploited for biotechnological purposes. An important difference between NBPs and natural proteins resides in the different aromatic/aliphatic amino acids content, in particular in the lower content of aromatic amino acids observed in natural proteins. This information can be very useful in the design of directed evolution and protein engineering studies. Finally, this study demonstrates that exploitation of grid infrastructures for massive structure prediction projects is feasible, possible applications including genome wide protein structure prediction of bacterial pathogens for target selection and drug design studies.

REFERENCES

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., (1990) Basic local alignment search tool, J. Mol. Biol., 215, 403-410

Bairoch A., Boeckmann B., Ferro S., Gasteiger E. (2004) Swiss-Prot: Juggling between evolution and stability, Brief. Bioinform. 5, 39-55

Baker MA., and Apon A., (2001) Middleware, in A White paper on Cluster Computing, International Journal of High Performance Computing, Sage Science Press, Volume 15, number 2, pp 102

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank, Nucleic Acids Res., 28, 235-242

Berman, Fran; Anthony J. G. Hey, Geoffrey C. Fox (2003). Grid Computing: Making The Global Infrastructure a Reality. Wiley

Bradley P, Malmström L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D (2005) Free modeling with Rosetta in CASP6. Proteins, 61(Suppl 7):128-134.

Branden C., Tooze J. (1999) Introduction to protein structure, Garland Publishing, New York

Brodkin, Jon. (2008) "Parallel Internet: Inside the Worldwide LHC computing grid". Techworld

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. (1982) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations J. Comp. Chem., 4: 187

Brylinski M, Kochanczyk M, Konieczny L, Roterman I. (2006) Sequence-structures-function relation characterized in silico. Biomol. Struct. Dyn., 23: 519

Brylinski M, Konieczny L, Roterman I. (2006) Hydrophobic collapse in (in silico) protein folding. J. Comput. Biol. Chem., 30: 255

Chao-Hsu Yao (2006) “Grid Computation – The Fastest Supercomputer in the World”. CSA Discovery Guide

Chiarabelli C., Vrijbloed J.W., De Lucrezia D., Thomas R.M., Stano P., Polticelli F., Ottone T., Papa E., Luisi P.L. (2006) Investigation of de novo totally random biosequences, Part II: Onthe folding frequency in a totally random library of de novo proteins obtained by phage display, Chem. Biodivers., 3, 840-859

Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CEM, Bonneau R, Rohl CA, Baker D. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53 Suppl 6:524-33

Condor - http://www.bo.infn.it/calcolo/condor/high_throu_sys.html

De Jong, K. A.(2006) Evolutionary Computation. MIT, N.Y.

De March, D., Forlin, M., Slanzi, D., Poli I. (2009) An evolutionary predictive approach to design high dimensional experiments. Artificial Life and Evolutionary Computation. World Scientific

Evangelista G, Minervini G, Luisi PL, Polticelli F. (2007) RandomBlast a tool to generate random “never born protein” sequences. Bio-Algorithms and Med-Systems, 3: 27-31

EuChinaGRID - http://www.euchinagrid.org/

Foster I, Kesselman C. (1999) The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers

gLite middleware - http://glite.web.cern.ch/glite/

GENIUS Portal – https://genius.ct.infn.it/

Hastie, T., Tibshirani, R., Friedman, J. (2009) The Elements of Statistical Learning: Data Mining,Inference, and Prediction. Springer Series in Statistics

Haykin, S. (2009) Neural Networks and Learning Machines. Pearson, N.Y.

Holm L, Kaariainen S, Rosenstrom P, Schenkel A. (2008) Searching protein structure databases with DaliLite v.3. Bioinformatics 24, 2780-2781

INFN Grid Project – http://www.infn.it/

Jayaram B., Bhushan K., Shenoy R. S., Narang P., Bose S., Agrawal P., Sahu D., Pandey V. (2006) Bhageerath: an energy based web enabled computer software suite for limitino the search space of tertiary structures of small globular proteins. Nucleic Acids Research, Vol. 34, No. 21 6195-6204

Jmol: An open-source Java viewer for three-dimensional molecular structures. http://www.jmol.org/

Job Description Language -www.grid.org.tr/servisler/dokumanlar/DataGrid-JDL-HowTo.pdf

Kabsch W, Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22: 2577-2637

Karlin S, Altschul SF. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA , 87: 2264-2268

Kendrew JC, Bodo G, Dintzis R, Parrish H, Wyckoff H, Phillips DC. (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181, 662-666

Matsumoto M, Nishimura T. (1998) Mersenne Twister A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation, 8: 3-30

McGuffin L.J., Bryson K., Jones D.T., (2000) The PSIPRED protein structure prediction server, Bioinformatics, 16, 404-405

Minervini G, La Rocca G, Luisi PL, Polticelli F. (2007) High throughput protein structure prediction in a GRID enviroment. Bio-Algorithms and Med-Systems. Vol. 3, No. 5, pp. 39-43

Minervini G, Evangelista G, Polticelli F, Piwowar M, Kochanczyk M, Flis L, Malawski M, Szepieniec T, Wisniowsky Z, Matczynska E, Prymula K, Roterman I. (2008) Never born proteins as a test case fora b inizio protein structures prediction. Bioinformation 3(4): 177-179

Minervini G, Evangelista G, Villanova L, Slanzi D, De Lucrezia D, Poli I, Luisi PL, Polticelli F. (2009) Massive non-natural proteins prediction usig grid technologies. BMC Bioinformatics, 10(Suppl 6):S22

Nan J., Brostomer E., Liu X., Krisyensen O., Su X. (2009) Bioinformatics and Structural Characterization of a Hypothetical Protein from Streptococcus mutans: Implication of Antibiotic Resistance, PlosOne 4(10): e7245

Nice – http://www.nice-italy.com/

NCBI genome - www.ncbi.nlm.nih.gov/projects/genome/

NCBI MMDB - www.ncbi.nlm.nih.gov/sites/entrez?db=structure/

Parvin Asadzadeh, Rajkumar Buyya, Chun Ling Kei, Deepa Nayar, and Srikumar Venugopal. (2005) Global Grids and Software Toolkits: A Study of Four Grid Middleware Technologies, High Performance Computing: Paradigm and Infrastructure, Laurence Yang and Minyi Guo (eds), pp.431-458 (Chapter 22), ISBN: 0-471-65471-X, Wiley Press

R Development Core Team: R (1992) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3- 900051-07-0, URL http://www.R-project.org

RefSeq - http://www.ncbi.nlm.nih.gov/RefSeq/

Rohl C.A., Strauss C.E., Misura K.M., Baker D. (2004) Protein structure prediction using Rosetta. Methods Enzymol. 383, 66-93

Sanner MF, Olson AJ, Spehner JC (1996) Reduced Surface: an efficient way to compute molecule surfaces. Biopolymers, 38: 305-320

Shindyalov N. and Bourne P. E. (1998) structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(*9*) *739*-747

Sridharan S, Nicholls A, Honig B. (1992) A new vertex algorithm to calculate solvent accessible surface areas. Biophys J, 61: A174

Streptococcus mutans genome
www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Protein+Table&list_uids=263

Tekaia F, Yeramian E, Dujon B. (2002): Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene, 297: 51-60

Wheeler D.L., Barrett T., Benson D.A., Bryant, S.H., Canese K., Church D.M. (2005) Database resources of the National Center for Biotechnology Information Nucleic Acids Res., 33, D39-D45

Yang Zhang. (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, vol 9, 40