Journal of
Proteomics & Bioinformatics

Porto et al., J Proteomics Bioinform 2014, 7:7
DOI: 10.4172/0974-276X.1000321

Research Article Open Access

Screening and Functional Prediction of Conserved Hypothetical Proteins from *Escherichia coli*

William F Porto[1], Simone Maria-Neto[1], Diego O Nolasco[1,2] and Octavio L Franco[1,3]*

[1]*Centre for Proteomic and Biochemical Analysis, Graduate Studies in Biotechnology and Genomic Sciences, Catholic University of Brasilia, Brasilia–DF, 70790-160, Brazil*
[2]*Course of Physics, Catholic University of Brasilia, Brasilia-DF, 70966-700, Brazil*
[3]*Post-graduation in biotechnology, Dom Bosco Catholic University, Campo Grande, MS, Brazil*

Abstract

Protein structures can provide some functional evidences. Therefore structural genomics efforts to identify the functions of hypothetical proteins have brought advances in our understanding of biological systems. To this end, a new strategy to mine protein databases in the search for candidates for function prediction was here described. The strategy was applied to *Escherichia coli* proteins deposited in the NCBI's non-redundant database. Briefly, data mining selects small conserved hypothetical proteins without significant templates on Protein Data Bank, without transmembrane regions and with similarity to Eukaryote proteins. Through this strategy, 12 protein sequences were selected for molecular modelling, from a total of 13,306 *E. coli's* conserved hypothetical sequences. From these, only three sequences could be modelled. GI 488361128 model was similar to cupredoxins, GI 281178323 model was similar to β-barrel proteins and GI 227886634 model showed structural similarities to lipid binding proteins. However, only the GI 227886634 seems to have a function related to the similar structures, since it was the unique structure that kept the fold during the molecular dynamics simulation. The method here described can be relevant to select hypothetical sequences that can be targets for *in vitro* and/or *in vivo* functional characterization.

Keywords: Structural genomics; Function prediction; Molecular modelling; Cupredoxins; β-Barrel proteins; Lipid binding proteins

Introduction

In the post-genomic era, several protein sequences have become available through the conceptual translation of putative open reading frames (ORFs). These proteins have been annotated without a detailed structural analysis or functional evaluation. Most of them are not experimentally characterized [1]. Moreover, several problems have been related to data integrity, such as sequencing errors or sample contaminations [2]. Protein function can be predicted by many methods based on sequence similarity. Therefore, assuming that similar sequences have similar functions, such approaches have created prospects for rapid progress in molecular biology [3]. Nevertheless, *in vitro* and/or *in vivo* validations are essential to provide a more accurate prediction, avoiding false results [4].

Commonly, when a protein sequence has no alignment matches or the alignments are non-significant, these sequences are deposited in the databases as hypothetical, unknown or unnamed proteins. When a protein sequence has a significant alignment to a hypothetical (unknown or unnamed) protein, it is deposited as a conserved hypothetical protein [5]. In order to understand the biological functions of these proteins it seems to be essential to learn their role in cell metabolism, as well as their suitability for post translational modifications or as a potential drug target [6]. Nonetheless, identifying the function of a protein is not a trivial task, especially if there is no *in vitro* characterization or evidences of its expression, e.g. the proteins generated by conceptual translation [7]. Nevertheless, it is easier when the unsolved protein sequences possess known domains, as then the functional annotation may be transferred from that domain model to the protein sequence [8]. Indeed, several unsolved proteins have been their functions inferred from the similarities to a structure from the known proteins [9-13]. Therefore, full understanding of the biological functions of such proteins requires structural knowledge, which can yield direct insight into its molecular mechanism, since conserved sequence regions can be important to functional sites [1,7,14]. Bearing this in mind, the structural alignment can be more relevant, since structural alignments are more sensitive than sequence alignments. When identity is below 40%, structural methods produce better alignments than sequence alignments [9,15]. In some cases, the sequence can differ from another sequence which performs the same or similar function [7,16]. Even if the identities are below 20% of residues, the coupling of conserved structures could be preserved during evolution, particularly in active sites [17]. Probably, the functional requirement to form and maintain the active site structure exerts pressure on the protein to adopt the functional fold [18].

Therefore, molecular modelling has become an important tool to reveal these similarities and to predict protein function, mainly for the hypothetical proteins, one of the most challenging problems in structural genomics [11]. The success of computational methods for structure prediction is a product of two factors: structures of small proteins are easier to predict than those of larger ones; and homologous structures can be explored to predict other protein structures [19]. Nevertheless, molecular modelling often needs a semi-automated approach, including manual curation on alignments, choice of templates and analysis of the models; computational power is also needed to perform a molecular dynamics simulation. This makes molecular modelling a costly process, mainly in cases where there are no available templates. Therefore, how can one choose a protein target to make a model for function prediction? We propose the selection of proteins with similarity to proteins from organisms without near phylogenetic relationships. Functional sites are generally conserved

***Corresponding author:** Octavio L Franco, Centre for Proteomic and Biochemical Analysis, Graduate Studies in Biotechnology and Genomic Sciences, Catholic University of Brasilia, Brasilia – DF, 70790-160, Brazil, Tel: (61) 3448-7220; E-mail: ocfranco@gmail.com

Received May 22, 2014; **Accepted** July 03, 2014; **Published** July 08, 2014

Citation: Porto WF, Maria-Neto S, Nolasco DO, Franco OL (2014) Screening and Functional Prediction of Conserved Hypothetical Proteins from *Escherichia coli*. J Proteomics Bioinform 7: 203-213. doi:10.4172/0974-276X.1000321

and it is believed that they are favoured evolutionarily; having the same or similar hydrophobic effect, and they may have the same or similar functions [20]. Therefore, this work describes a novel strategy to mine the NCBI's non-redundant protein database (NR), looking for such proteins. The strategy was based on sequence local alignment searches to find candidates for structural characterization and further function prediction. This strategy was applied to small conserved hypothetical proteins from *Escherichia coli*, aiming to predict their functions and facilitate functional annotation.

Material and Methods

Data mining

The overview of the proposed data mining method is shown in Figure 1. Starting from the NCBI's non-redundant protein database (NR–downloaded at June, 2011), the proteins annotated as conserved hypothetical proteins ranging from 30 to 200 amino acid residues from *E. coli* derived from mRNA, genomic DNA or experimental observation were extracted, composing the initial data set. From this data set, the redundant proteins were removed through JalView [21] with a cut off of 80% of identity. Following that, the remaining sequences were submitted to Phobius [22] to predict transmembrane topology and signal peptide sequences. The predicted transmembrane proteins were discarded and the signal sequences were removed from the translated sequence, generating a data set of mature and non-transmembrane sequences. A new size selection was done again, selecting only the small sequences, ranging from 30 to 100 amino acid residues. Then, two steps of local alignments using BLAST [23] were done. Firstly, all sequences with similarity higher than 30% of identity to any sequence deposited in the Protein Data Bank (PDB) were discarded. Thereby, the remaining sequences with identity higher than 80% to any protein and identity below 40% to proteins from Eukaryotes deposited in the SwissProt Database [24] were discarded. The remaining sequences were submitted to RPS-BLAST search against the Conserved Domain Database (CDD) [8], and the sequences without conserved domains were selected. Then, the sequences predicted to be expressed were selected through Glimmer 3.0 [25]. The last step was used to remove the proteins predicted to have an unstable folding by means of PrDOS [26].

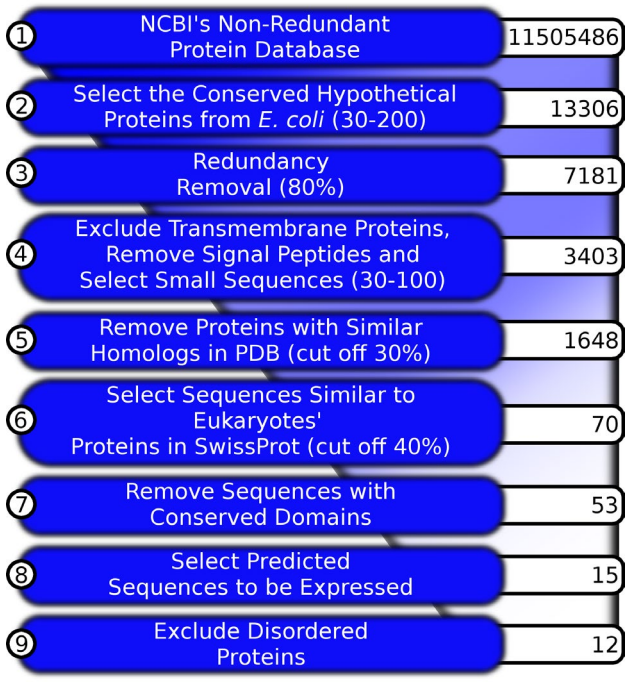

Figure 1: The flowchart of the proposed data mining.

Numbers on the left indicate the step number; numbers on the right indicate the number of sequences in each step.

Molecular modelling

The LOMETS server [27] was used to find the best template for comparative modelling. LOMETS is a Meta threading server which collects the information of other nine threading methods, and then rank the information about the templates. The best templates were selected after manual curation of all alignments generated by LOMETS, taking into account the coverage and the sequence identity. Therefore, one hundred theoretical three-dimensional models were constructed through Modeller 9.9 [28]. The models were constructed using the methods of automodel and environ classes. Furthermore in order to import the ligands, the property io.hetatm from class environ was set as true. The final model was selected according to the discrete optimized protein energy (DOPE) scores. This score assesses the energy of the model and indicates the most probable structures. Then, the sequences without an adequate template or without an inadequate model were submitted to the QUARK *ab initio* modelling server [29], the best ranked algorithm in CASP 9 and 10 in free modelling category. The best models were evaluated through Verify 3D [30], ProSA II [31] and PROCHECK [32]. PROCHECK checks the stereochemical quality of a protein structure, through the Ramachandran plot, where good quality models are expected to have more than 90% of amino acid residues in most favoured and additional allowed regions, while ProSA II indicates the fold quality; additionally, Verify 3D analyses the compatibility of an atomic model (3D) with its own amino acid sequence (1D). Structure visualization was done in PyMOL (The PyMOL Molecular Graphics System, Version 1.4.1, Schrödinger, LLC).

Structural alignments and function predictions

Structural alignments were performed in two ways, through Dali Server [33] and COFACTOR server [34]. On Dali Server, the assessment of structural alignments was done through the Z-Score, where an structural alignment with Z-Score higher than 2 is significant. COFACTOR uses the TM-align structure alignment program [35] to search against the PDB and then examines the binding pockets, predicts the binding pose of ligands into the target structure or model and constructs the protein-ligand complexes. Therefore, this approach allows the identification of the binding position of ligands without docking experiments.

Molecular dynamics

The molecular dynamics simulations (MD) of the protein-ligand complexes were carried out in water environment, using the Single Point Charge water model [36]. The analyses were performed by using the GROMOS96 43A1 force field and computational package GROMACS 4 [37]. The dynamics utilized the three-dimensional models of the protein-ligand complexes as initial structures, immersed in water molecules in cubic boxes with a minimum distance of 0.7 nm between the complexes and the boxes' frontiers. Chlorine ions were also inserted in the complexes with positive charges in order to neutralize the system charge. Geometry of water molecules was constrained by using the SETTLE algorithm [38]. All atom bond lengths were linked by using the LINCS algorithm [39]. Electrostatic corrections were made

by Particle Mesh Ewald algorithm [40], with a cut off radius of 1.4 nm in order to minimize the computational time. The same cut off radius was also used for van der Waals interactions. The list of neighbours of each atom was updated every 10 simulation steps of 2 fs. The conjugate gradient and the steepest descent algorithms (50,000 steps each) were implemented for energy minimization. After that, the system temperature was normalized to 300 K during 2 ns, using the Berendsen thermostat (NVT ensamble). Furthermore the system pressure was normalized to 1 bar during 2 ns, using the Berendsen barostat (NPT ensamble). The systems with minimized energy, balanced temperature and pressure were simulated during 50 ns by using the leap-frog algorithm. The simulations were analysed through RMSD and DSSP. The initial and the final structures were compared through the TM-Score [41], where structures with TM-Scores above 0.5 indicate that the structures share the same fold [42].

Results and Discussion

Data mining

The explosive growth in database data requires techniques and tools to transform the amount of data into useful information and knowledge. Data mining, which is also referred to as knowledge discovery in databases, has clearly increased in importance. Mining information from databases has been recognized by many researchers as a hot spot in different areas [43]. Data mining processes have been carried out to discover novel biological data, such as putative ancestral genes of circular proteins in plants [44], antimicrobial peptides [45,46] or candidates for novel translatable sequences [4]. This work describes a novel data mining approach for selecting hypothetical proteins for functional prediction.

The data mining aims to find small conserved hypothetical proteins (30-100 amino acid residues) from *E. coli,* without significant templates on Protein Data Bank, without transmembrane regions and with similarity to Eukaryote proteins. The data mining process was applied to *E. coli* proteins deposited in the NR. *E. coli* was chosen due to the minor genome complexity when compared to a Eukaryote genome. Moreover, *E. coli* has great importance in experimental, medical and industrial fields and has more than 200 genomes sequenced.

Starting from the NR, 13,306 small and conserved hypothetical proteins from *E. coli* were extracted from a total of 11,505,486 non-redundant sequences (Figure 1). The NR was chosen to select protein sequences derived from diverse sources, such as genomic and transcriptomic data. Another advantage of the NR is its previous removal of redundancy, which permits working with a smaller amount of data. In addition, starting from NR is essential to our workflow, since it summarizes the headers of redundant sequences, for example, the sequence GI 115513274 has two entries into NR, one annotated as hypothetical protein (GI 117624150) and other annotated as conserved hypothetical protein (GI 115513274). This information is useful to solve the problem of hypothetical proteins which are conserved, but remains annotated only as hypothetical. If they are conserved, there must be at least second entry annotated as such.

For a simple genome such *E. coli*'s, 13,306 sequences is a large number of proteins, especially considering that only proteins with range between 30 and 200 amino acid residues were selected. For only one *E. coli* genome, approximately 4,000 proteins were expected, while in the whole genome of *E. coli* K-12 4,288 protein-coding genes were found [47]. In fact, once our starting point was the NR, this subset includes proteins sequences from all *E. coli* strains in the data base. Furthermore, the majority of these proteins (13,306) were given by conceptual translation and, consequently, some of them are not accurate. A number of these were determined by programs, such as ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/), which can identify several ORFs in a putative gene. In this way, a single gene generates two or more protein sequences by conceptual translation (e.g. the sequences GI 193061927 and GI 260858280 were generated by the same predicted gene). Therefore, in order to ensure data quality, several data mining steps was rigorous, avoiding such possible errors.

Our data mining proposes to select proteins with maximum 100 amino acid residues length, once smaller sequences requires less computational power for three-dimensional structure prediction with reliable quality. It also reduces the high computational power dependence to perform the molecular dynamics simulations. However, in the first selection, the maximum length was 200 amino acid residues. This extension in the amino acid number was done in order to avoid some biases. The first one is the precursor proteins, which can have, for example, 120 residues, 30 of them being part of the signal peptide. Then this protein would be excluded if the maximum cut off was set to 100. Another bias is the presence of fragmented sequences, if a protein with 150 residues has a fragment of 80 residues deposited on NR, this fragment must be removed from the analysis, which is done at the next step, the redundancy removal.

JalView was used for redundancy removal, with a cut off of 80%. This step was important firstly for removing fragments of larger proteins and also to cluster the sequences, reducing the amount of data. This step reduced the number of sequences by almost half, leaving 7,181 from 13,306 sequences.

In the next step, the predicted transmembrane proteins were discarded (Figure 1), since interactions between protein and lipid bilayers could cause protein structure modifications [48]. Thus, it may generate molecular models quite different from the native protein structure. In addition, simulations of transmembrane proteins require high computational costs. Following that, the signal peptides were removed from sequences, because the signal peptide is absent from the mature peptide. Then, a novel size cut off was applied, selecting sequences ranging from 30 to 100 amino acid residues. This novel cut off was important for yield models with reliable quality (mainly through *ab initio* molecular modelling) and with low computational costs for performing a molecular dynamics simulation. Moreover, the prediction of a larger protein through *ab initio* modelling is extremely difficult, since by increasing the length of sequence, the conformational phase space of sampling sharply increases, and this results in a loss of accuracy [49,50]. However, predicting the structure of smaller sequences is not useful in predicting their functions through structural alignment, since this method requires at least 30 amino acid residues [33]. After this step, 3,403 sequences remained.

Then, all sequences with identities higher than 30% to PDB structures were discarded, remaining 1,648 sequences (Figure 1). When similarity is below 30%, the detection of structural homology has low statistical significance [51]. In this way, if simple comparative modelling was applied onto these proteins the result would be models with poor quality. In these cases, threading and/or *ab initio* modelling are more useful. In other words, removing the sequences with hits with sequences from PDB, we are discarding the easy and medium cases, which comparative modelling could solve without great problems. Thus, threading or *ab initio* modelling was used to predict the structures and then, structural alignments were performed in order to find similar structures without similar sequences.

At this point, we have non-redundant (at maximum 80% of identity) small (30 to 100 amino acid residues) conserved hypothetical proteins without transmembrane portions and without clear homolog structures in PDB. Therefore, we need to remove proteins that may have their annotations derived from proteins already characterized. To do that, sequences with similarity higher than 80% of identity to proteins deposited in SwissProt were removed (Figure 1). Consequently, at this point, the sequence identity cannot reveal any indications about the function of these proteins. To acquire some clues, knowledge of their structures is crucial. However, selection of appropriate candidates is needed, since molecular modelling has some costs. In order to select suitable candidates, the sequences with identities below 40% to eukaryote protein were discarded. The selection of sequences with some degree of similarity in divergent groups, such as Bacteria and Eukaryote, may indicate selective pressure to maintain certain characteristics with importance for biological processes [52,53]. Therefore, solving the function of these proteins provides a better understanding of biological processes. Then, after removing the proteins with identities higher than 80% to any annotated protein and identities below 40% to eukaryote proteins, 70 protein sequences remained in our data set. Some of these sequences could be resulted from horizontal transfer from *E. coli* to eukaryotes or *vice-versa*. These proteins could have important functions since some regions were conserved, even in divergent life kingdoms.

Nevertheless, there is still an alternative for identifying some clues to a probable function, the identification of a conserved domain, so that when a protein has a conserved domain, functional molecular modelling is unnecessary, because the domain is sufficient to predict the protein's function. We then use the RPS-BLAST for domain identification. The search was done against the CDD database, a collection of well-annotated models for ancient domains and full-length proteins [8]. Therefore, the sequences with identified domains were removed, remaining 53 sequences.

All data mining steps to reach this point were applied with the aim of selecting sequences without direct evidence of a probable function. However, the previous steps do not take into account if the sequences are actually expressed. Therefore, since some of these 53 proteins can be generated by conceptual translation of pseudogenes, or by random ORFs, their respective DNA sequences were submitted to Glimmer 3.0, in order to select only the sequences which may really be expressed by *E. coli*. In this way only 15 sequences were predicted to be expressed, and the remaining sequences were discarded.

Finally, in the last data mining step, PrDOS was used for predicting disordered regions in protein structures. The disordered regions are important for functions of diverse proteins [26,54]. Nevertheless, the structure of such regions could not be solved rightly by experimental methods [26] and by computational methods they do not converge to a consensus structure [55,56]. Therefore, the sequences with more than 80% of residues in predicted disordered regions were removed, since these chaotic structures cannot be adequately modelled, remaining 12 protein sequences (Table 1). These 12 sequences were modelled through comparative or *ab initio* modelling. Nevertheless only three models were approved by the validation methods (data not shown).

GI	Annotation	Number of Amino Acid Residues
115513274	Conserved Hypothetical Protein	97
306907928	Conserved Hypothetical Protein	50
188493836	Conserved Hypothetical Protein	58
331042441	Conserved Hypothetical Protein	47
227883817	Conserved Hypothetical Protein	59
227886634	Conserved Hypothetical Protein	90
253721170	Conserved Hypothetical Protein	95
488361128	Conserved Hypothetical Protein	60
281178323	Conserved Hypothetical Protein	87
300815338	Conserved Hypothetical Protein	61
309705558	Conserved Hypothetical Protein	98
260751868	Conserved Hypothetical Protein	76

Table 1: Sequences retrieved by the data mining process.

GI 488361128

The first sequence was GI 488361128. This sequence has 60 amino acid residues and has no signal peptide. This sequence could have its structure predicted by threading algorithms. Inspecting all alignments returned by LOMETS, the structure of umecyanin from *Armoracia rusticana* (PDB 1X9R) [57] was chosen as template. Inspecting the alignment, between sequence and template, a similar copper binding site was observed. Therefore, the copper was imported to the model, which was constructed as a dimeric structure, since its template was also a dimeric structure. Each monomer of the three-dimensional model was composed of two β-sheets and also by a short α-helix (Figure 2). The copper coordination site in the model is composed by residues Gln^{53}, His^{58} and Gln^{60} (Figure 2). The model assessment is summarized in Table 2. Structural alignment shows similarities to other cupredoxins (Table 3), suggesting that GI 488361128 could also be linked to this functional group (Figure 2). Comparing GI 488361128 with cupredoxins, four major differences can be observed: the size of sequences, the lack of a disulphide bridge and two mutations in the copper binding site (His-to-Leu and Cys-to-Gln) (Figure 2). The two mutations may completely inhibit copper coordination; although the glutamine residue could coordinate the copper ion, its side chain is longer, compared to the cysteine residue (Figure 2). In addition, the mutation His-to-Leu lacks a coordination group. However, the copper binding site was changed after the two thousand cycles of energy minimization. The position of Leu^{12} is occupied by Asp^{13} and the Gln^{53} side chain gets closer to copper. In addition, other atoms get closer to copper, such as the carbonyl oxygen of Leu^{12}, Gln^{52} and Gln^{55}, however the side chain of Gln^{60} gets distant from the copper ion (data not shown). In fact the residue Gln^{60} may not be necessary for copper binding, since according to Karlsson et al. [58], the mutants of azurin from *Pseudomonas aeruginosa* with the residue Met^{121} mutated to any natural amino acid or even to a stop codon are able to maintain the copper site. Met^{121} is the equivalent to Gln^{95} in the umecyanin and to Gln^{60} in GI 488361128 (data not shown).

Therefore, the dimeric structure of GI 488361128 in complex with the copper ion was evaluated through MD, where in the chain A, the copper ion was maintained in the site during 20 ns of simulation and then released, and in the chain B, it was released after 5 ns of simulation. The copper releasing indicates that, in fact, the mutations in the copper site may generate a loss of function. In addition, a RMSD variation of 5.2 Å and 3.9 Å were observed respectively for chain A and B (Figure 3a), indicating significant structural changes. This changes are confirmed by the TM-Scores of each monomer indicates that the folding after the 50 ns of simulation are not the same at the beginning, with values of 0.4444 and 0.4324 for chain A and B, respectively. In fact, this small protein undergoes a little secondary structure loss in the first 2 nanoseconds, where an α-helix-to-coil transition was observed (Figure 3c). The DSSP analysis shows that the region comprised by the residues 20 to 40 is very dynamic, since several kinds of secondary structures were observed, varying from coils to 3_{10}-helix (Figure 3c). However, the maintenance of the dimer and the two β-sheets of each

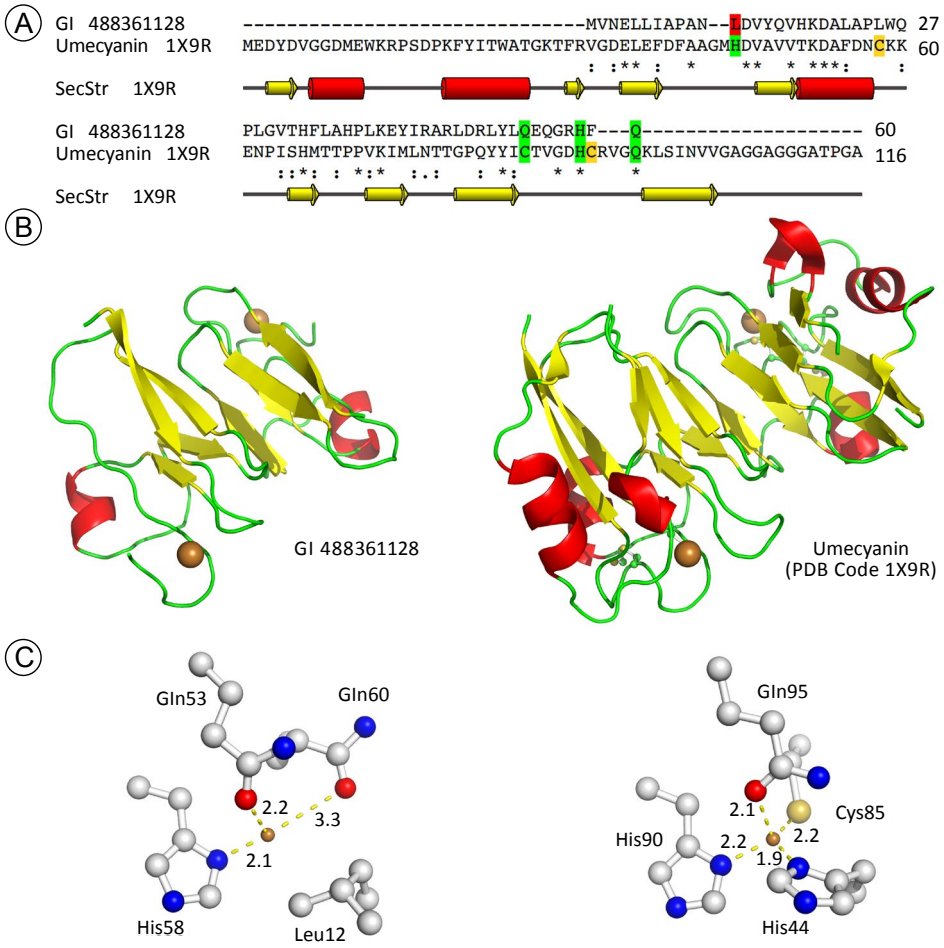

Figure 2: Modelling of GI 488361128.

(A) The sequence alignment used to construct the GI 488361128 3D model. Conserved residues are marked with a star, the residues involved in copper coordination are highlighted in green, and cysteines involved in disulfide bridges are in yellow. The secondary structure of 1X9R is indicated below the alignment (arrows for β-strands, cylinders for α-helices and grey lines for loops). (B) The 3D model of GI 488361128 (left) and the X-ray structure of 1X9R (right). (C) The putative copper coordination site of GI 488361128 (left) and the copper coordination of 1X9R.

Sequence ID	Z-Score (PROSA II)	Verify 3D (1D-3D Average)		Ramachandran Plot (%)		
		Minimum	Maximum	Favoured Regions	Allowed Regions	Generously Allowed
GI 488361128	-1.28	-0.02	0.36	81.7	14.4	3.8
GI 281178323	-4.16	-0.03	0.51	53.4	34.2	6.8
GI 227886634	-4.96	0.09	0.51	82.9	12.9	2.9

Table 2: Summary of molecular modelling validation assessments.

Molecule	PDB Code	Z-Score	RMSD (Å)	Reference
Umecyanin	1X9R	9.3	0.5	[60]
Mavicyanin	1WS7	8.6	0.8	[61]
Plantacyanin	1F56	7.4	1.3	[62]
Phytocyanin	2CBP	7.3	1.2	[63]
Stellacyanin	1JER	6.0	1.1	[58]

Table 3: Structural alignment search of GI 488361128 against PDB.

monomer were also observed (Figure 3b). Indeed, this protein seems not to be related to cupredoxins, however, it is important to highlight that the structural prediction was not completely wrong. Probably there is no copper coordination site, but we could observe the dimer and the β-sheets maintenance.

GI 281178323

The second sequence was GI 281178323. This sequence has 87 amino acid residues and has no signal peptide. For this sequence, the predicted structure was yielded by QUARK. The three-dimensional model was composed of 9-stranded β-barrel with simple up-down topology (Figure 4). The model assessments are showed in Table 2. Structural alignment shows similarities to several β-barrel proteins (Table 4), being the similarity restricted to the tertiary structure (data not shown). Through the COFACTOR server, no complexes with significant BS-Scores were obtained.

Therefore, the predicted three-dimensional structure was evaluated alone by molecular dynamics, where a RMSD of about 5.4Å was observed (Figure 5a). Despite the final structure maintains the β-strands (Figure 5b), the TM-Score of 0.3769 indicates that the final structure does not share the same fold with the initial one. The DSSP analysis (Figure 5c) indicated that there are changes in some β-strand segments and also in the loops connecting the β-strands, where bend-to-turn and turn-bend transitions could be observed. Once the final structure is not in the same fold as the initial, the actual function of GI 281178323 seems to be unrelated to the proteins identified in the structural alignments

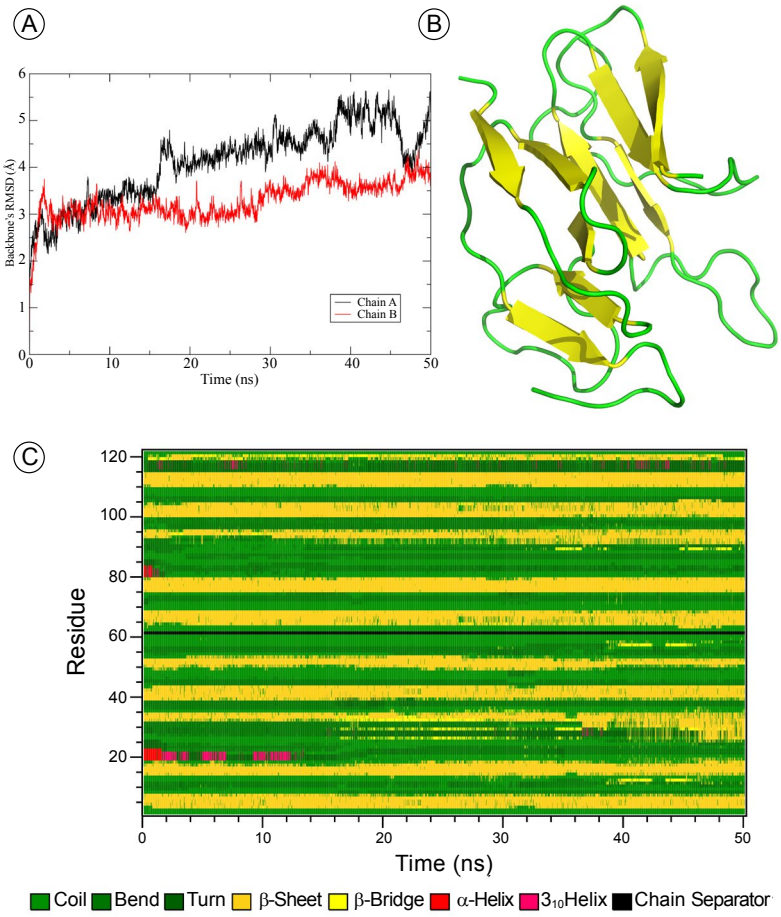

Figure 3: Molecular dynamics simulation of GI 488361128.

(A) The Backbone's RMSD variation during 50ns of simulations. (B) The GI 488361128 model after 50 ns of simulation. (C) The DSSP analysis of secondary structure during the simulation.

Molecule	PDB Code	Z-Score	RMSD (Å)	Reference
Coagulase	2X4M	6.1	3.0	[64]
Fatty Acid-Bidding Protein	2JU3	5.8	2.2	[65]
Avidin	1LDQ	5.8	2.7	[66]
Avidin	1NQN	5.8	2.7	[67]
Streptavidin	1RXJ	5.8	2.8	[68]

Table 4: Structural alignment search of GI 281178323 against PDB.

Molecule	PDB Code	Z-Score	RMSD (Å)	Reference
Myotoxin II	1CLP	2.8	3.0	[69]
Prophospholipase A2	1HN4	2.7	3.1	[70]
Phospholipase A2	1AYP	2.7	2.8	[71]
Phospholipase A2	1RGB	2.7	2.8	[72]
Phospholipase A2	1POD	2.7	3.0	[73]

Table 5: Structural alignment search of GI 227886634 against PDB.

(Table 4), since the initial folding is not maintained.

GI 227886634

The third sequence here selected and evaluated was GI 227886634. This sequence has 90 amino acid residues length and has no signal peptide. Its structure was predicted by QUARK. The three-dimensional model was composed of two α-helixes and several loops (Figure 6). The validation parameters are summarized in table 2. Structural alignment shows similarities to several phospholipases (Table 5). Nevertheless there is no sequence similarity; the similarity is restricted to the tertiary structure (data not shown). The COFACTOR server also indicates that the predicted folding of this sequence is related to lipid binding proteins. Therefore, the complex between the model of GI 227886634 and the lipid 3,7,11,15-tetramethyl-hexadecan-1-ol (ARC) was generated by COFACTOR with a BS-Score of 1.11.

The complex of GI 227886634-ARC was evaluated through molecular dynamics, where the maintenance of the complex was observed during the whole simulation. However, we observed a RMSD variation of about 5.5Å (Figure 7a), which could be related to a large displacement suffered by the N-Terminal loop, since it is responsible for stabilizing the protein-lipid complex together with the C-Terminal α-helix (Figure 7b). In addition, there is a gain in secondary structure presented during the simulation (Figure 7c). The region comprised by the residues His^{15} and Thr^{19} forms a short α-helix, and the region comprised by the residues Pro^{35} and Gln^{38} forms a 3_{10}-helix (Figure 7c). However, the TM-Score of 0.5535 indicates that the initial and the final structure share the same fold. Indeed, the RMSD variation of about 5.5 Å is due to the N-Terminal loop displacement, since removing the first

Figure 4: Modelling of GI 281178323.

(A) The structural prediction of GI 281178323; the predicted secondary structure is indicated below the sequence (arrows for β-strands and grey lines for loops). (B) The 3D model of GI 281178323.

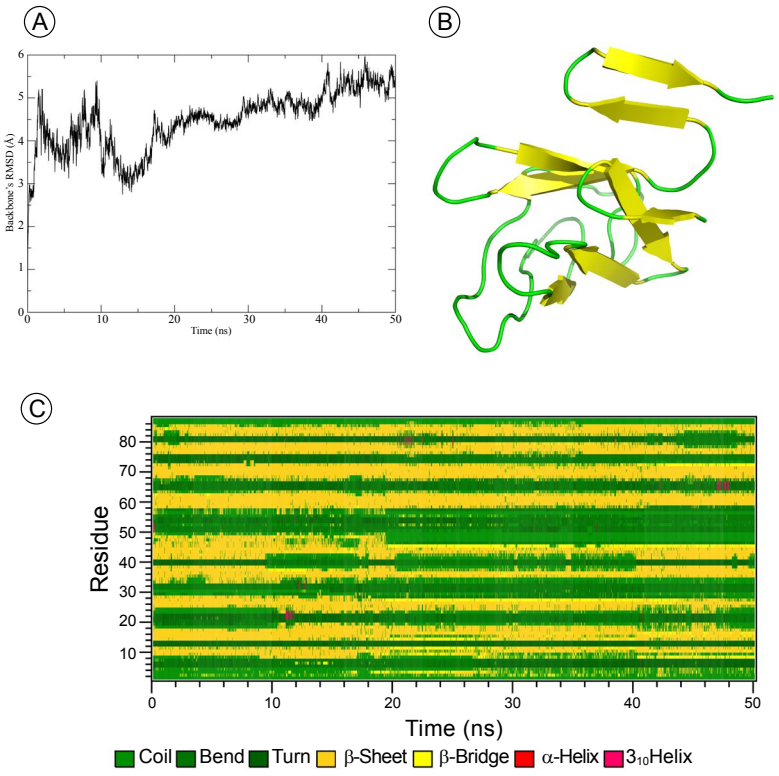

Figure 5: Molecular dynamics simulation of GI 281178323.

(A) The Backbone's RMSD variation during 50 ns of simulations. (B) The GI 281178323 model after 50 ns of simulation. (C) The DSSP analysis of secondary structure during the simulation.

17 residues from the RMSD calculation, we could observe a RMSD of about 3 Å (Figure 7a).

In fact, this is an intriguing sequence, since there is no sequence similarity, the functional annotation of a phospholipase cannot be transferred to it. However, this hypothetical protein can be annotated as a putative lipid binding protein, even with its actual function remaining unclear.

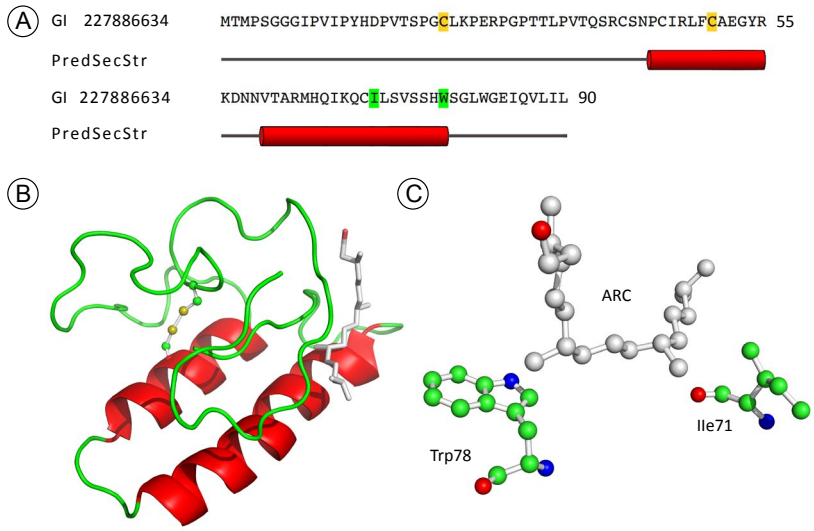

Figure 6: Modelling of GI 227886634.

(A) The structural prediction of GI 227886634; the predicted secondary structure is indicated below the sequence (cylinders for α-helices and gray lines for loops), the cysteine residues involved in disulphide bonds are highlighted in yellow, and the residues with hydrophobic interactions to ARC are highlighted in green. (B) The 3D model of GI 227886634 in complex with ARC. (C) The amino acid residues with hydrophobic interactions to ARC.

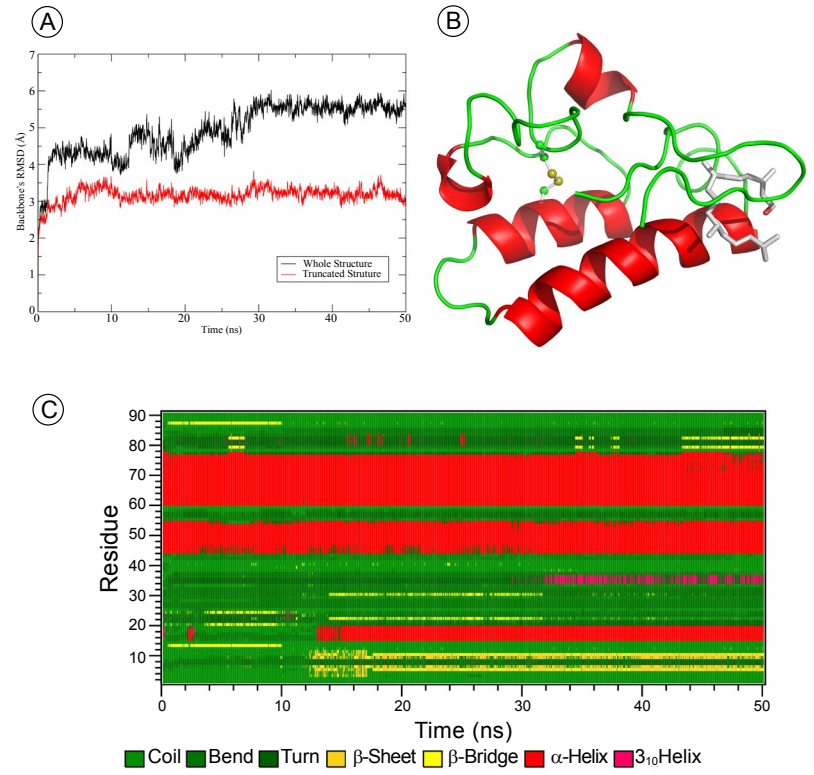

Figure 7: Molecular dynamics simulation of GI 227886634.

(A) The Backbone's RMSD variation during 50 ns of simulations. (B) The GI 227886634 model after 50 ns of simulation. (C) The DSSP analysis of secondary structure during the simulation.

Advantages and Limitations

From more than 11 million sequences in the NR, through the data mining process 12 conserved hypothetical sequences (Table 1) were selected for further structure and function predictions. However, unfortunately, nine sequences could not have their structures correctly predicted (data not shown) and consequently neither their functions.

Thereby, only three models (25%) could be used for function prediction. In a similar work [56], 400 domains of unknown functions (DUFs) ranging from 30 to 100 amino acid residues, without transmembrane portions were modelled using ab initio procedures. From these, 85 validated models were obtained (21.25%) [56]. This work had a similar performance, in proportion taking into account only the validation reports.

Comparing with 400 sequences from previous work [56], the outcome of 12 sequences is extremely small. Nevertheless, 12 sequences compose an adequate set for performing structure predictions and further molecular dynamics, since these procedures require manual curation.

Here, from the three cases in which the tertiary structure could be predicted, only the GI 227886634 seems to have a function related to the similar structures, since it was the unique structure that kept the folding during the molecular dynamics simulation. It is important to highlight that the molecular dynamics simulation has a pivotal role in such predictions, once through the simulations, the structure stabilization and/or ligand accommodation could be observed (or not). In this context, the molecular dynamics simulation data aggregate reliability to the predicted functions.

In this context, two limitations are clearly evident, the first one is the *in silico* structure prediction methods dependence and the second one is the size of target sequences (30 to 100 amino acid residues), which is induced by the *in silico* methods. However, by simply increasing the size cut-off, the method can be easily adapted to select larger proteins for structure determination by using an experimental approach such as NMR spectroscopy or X-ray diffraction. In fact, the data mining method could be used for selecting sequences for novel rounds of structural genomics projects or even novel CASP experiments.

Conclusions

Indeed, the functions of other hypothetical proteins (including those which could not be predicted here) may be adequately predicted through NMR spectroscopy or X-ray diffraction, as observed for the hypothetical protein MJ0577 from *Methanococcus jannaschii*, which was solved by X-ray diffraction [9]. This protein structure was found to bind to ATP, since the ATP molecule was crystallized together with the protein, so that MJ0577 was predicted as ATPase or ATP-mediated molecular switch. It is clear that in the case of GI 488361128, GI 281178323 and GI 227886634 more experiments are needed to confirm the *in silico* predictions. The hypothesis that the predictions could be wrong cannot be ruled out, but now we have a starting point from which to study these three hypothetical proteins.

Efforts to identify the functions of hypothetical proteins could bring novel advances in our understanding of biological systems. In 2004, Roberts [59] proposed that novel systems or pipelines should be developed to predict the function of hypothetical proteins, with these approaches being applied initially to prokaryotes [59]. Despite the *in silico* structure prediction methods become more advanced and reach more accurate results, this kind of prediction is not a trivial task, especially if the predictions will be made through *in silico* methods, as this report clearly shows. Although the *in silico* structure predictions dependence, the data mining method described in this paper can be applied for mining databases and/or genomes, looking for hypothetical sequences that can be targets for *in vitro* and/or *in vivo* functional characterization. Through this method, novel advances in structural genomics could be reached.

Acknowledgement

The authors are grateful to Center for Scientific Computing (NCC/GridUNESP) of the São Paulo State University (UNESP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Universidade Católica de Brasília (UCB) for the support.

References

1. Brenner SE (2001) A tour of structural genomics. Nat Rev Genet 2: 801-809.
2. Carter K, Oka A, Tamiya G, Bellgard MI (2001) Bioinformatics issues for automating the annotation of genomic sequences. Genome Inform 12: 204-211.
3. Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17: 429-431.
4. Desler C, Suravajhala P, Sanderhoff M, Rasmussen M, Rasmussen LJ (2009) *In Silico* screening for functional candidates amongst hypothetical proteins. BMC Bioinformatics 10: 289.
5. Sivashankari S, Shanmughavel P (2006) Functional annotation of hypothetical proteins - A review. Bioinformation 1: 335-338.
6. Gabow AP, Leach SM, Baumgartner WA, Hunter LE, Goldberg DS (2008) Improving protein function prediction methods with integrated literature data. BMC Bioinformatics 9: 198.
7. Kinoshita K, Nakamura H (2003) Protein informatics towards function identification. Curr Opin Struct Biol 13: 396-400.
8. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37: D205-210.
9. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, et al. (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc Natl Acad Sci U S A 95: 15189-15193.
10. Das K, Xiong X, Yang H, Westland CE, Gibbs CS, et al. (2001) Molecular modeling and biochemical characterization reveal the mechanism of hepatitis B virus polymerase resistance to lamivudine (3TC) and emtricitabine (FTC). J Virol 75: 4771-4779.
11. Martinez-Cruz LA, Dreyer MK, Boisvert DC, Yokota H, Martinez-Chantar ML, et al. (2002) Crystal structure of MJ1247 protein from *M. jannaschii* at 2.0 A resolution infers a molecular function of 3-hexulose-6-phosphate isomerase. Structure 10: 195-204.
12. Keller JP, Smith PM, Benach J, Christendat D, deTitta GT, et al. (2002) The crystal structure of MT0146/CbiT suggests that the putative precorrin-8w decarboxylase is a methyltransferase. Structure 10: 1475-1487.
13. Nan J, Brostromer E, Liu XY, Kristensen O, Su XD (2009) Bioinformatics and structural characterization of a hypothetical protein from *Streptococcus* mutans: implication of antibiotic resistance. PLoS One 4: e7245.
14. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294: 93-96.
15. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8: 995-1005.
16. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. Nature 405: 823-826.
17. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823-826.
18. Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 16: 399-408.
19. Zhang Y, Skolnick J (2004) Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys J 87: 2647-2655.
20. Chakrabarti S, Lanczycki CJ (2007) Analysis and prediction of functionally important sites in proteins. Protein Sci 16: 4-13.
21. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.
22. Käll L, Krogh A, Sonnhammer EL (2007) Advantages of combined

transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res 35: W429-432.

23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

24. UniProt Consortium1 (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71-75.

25. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673-679.

26. Ishida T, Kinoshita K (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 35: W460-464.

27. Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35: 3375-3382.

28. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci Chapter 2: Unit 2.

29. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80: 1715-1735.

30. Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356: 83-85.

31. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35: W407-410.

32. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26: 283-291.

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

34. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40: W471-477.

35. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33: 2302-2309.

36. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B (Ed) Intermolecular Force. Dordrecht, Reidel.

37. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4: 435-447.

38. Miyamoto S, Kollman PA (1992) SETTLE. An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J Comp Chem 13: 2134-2144.

39. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS. A linear constraint solver for molecular simulations. J Comp Chem 18: 1463-1472.

40. Darden T, York D, Pedersen L (1993) Particle mesh ewald: an n long(n) method for ewald sums in large systems. J Chem Phys 98: 10089-10092.

41. Zhang Y, Skolnick J (2004) Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys J 87: 2647-2655.

42. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26: 889-895.

43. Chen MS, Han J, Yu PS (1996) Data Mining: An overview from database perspective. IEEE Trans on Knowledge and Data Engineering 8: 866-883.

44. Mulvenna JP, Mylne JS, Bharathi R, Burton RA, Shirley NJ, et al. (2006) Discovery of cyclotide-like protein sequences in graminaceous crop plants: ancestral precursors of circular proteins? Plant Cell 18: 2134-2144.

45. Fernandes FC, Porto WF, Franco OL (2009) A wide antimicrobial peptides search method using fuzzy modeling. Lecture Notes in Computer Science 5676: 147-150.

46. Porto WF, Souza VA, Nolasco DO, Franco OL (2012) *In silico* identification of novel hevein-like peptide precursors. Peptides 38: 127-136.

47. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462.

48. Killian JA, Nyholm TK (2006) Peptides in lipid bilayers: the power of simple models. Curr Opin Struct Biol 16: 473-479.

49. Floudas CA, Fung HK, McAllister SR, Mönnigmann M, Rajgaria R (2006) Advances in protein structure prediction and *de novo* protein design: a review. Chem Eng Sci 61: 966-988.

50. Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 5: 17.

51. Koehl P, Levitt M (2002) Sequence variations within protein families are linearly related to structural variations. J Mol Biol 323: 551-562.

52. Webster G, Genschel J, Curth U, Urbanke C, Kang C, et al. (1997) A common core for binding single-stranded DNA: structural comparison of the single-stranded DNA-binding proteins (SSB) from *E. coli* and human mitochondria. FEBS Lett 411: 313-316.

53. Kelly TJ, Simancek P, Brush GS (1998) Identification and characterization of a single-stranded DNA-binding protein from the archaeon *Methanococcus jannaschii*. Proc Natl Acad Sci U S A 95: 14634-14639.

54. Tavares LS, Rettore JV, Freitas RM, Porto WF, Duque AP, et al. (2012) Antimicrobial activity of recombinant Pg-AMP, a glycine-rich peptide from guava seeds. Peptides 37: 294-300.

55. Porto WF, Nolasco DO, Franco OL (2014) Native and recombinant Pg-AMP1 show different antibacterial activity spectrum but similar folding behavior. Peptides 55: 92-97.

56. Rigden DJ (2011) Ab initio modeling led annotation suggests nucleic acid binding function for many DUFs. OMICS 15: 431-438.

57. Koch M, Velarde M, Harrison MD, Echt S, Fischer M, et al. (2005) Crystal structures of oxidized and reduced stellacyanin from horseradish roots. J Am Chem Soc 127: 158-166.

58. Karlsson BG, Nordling M, Pascher T, Tsai LC, Sjölin L, et al. (1991) Cassette mutagenesis of Met121 in azurin from *Pseudomonas aeruginosa*. Protein Eng 4: 343-349.

59. Roberts RJ (2004) Identifying protein function-a call for community action. PLoS Biol 2: E42.

60. Xie Y, Inoue T, Miyamoto Y, Matsumura H, Kataoka K, et al. (2005) Structural reorganization of the copper binding site involving Thr15 of mavicyanin from *Cucurbita pepo* medullosa (zucchini) upon reduction. J Biochem 137: 455-461.

61. Einsle O, Mehrabian Z, Nalbandyan R, Messerschmidt A (2000) Crystal structure of plantacyanin, a basic blue cupredoxin from spinach. J Biol Inorg Chem 5: 666-672.

62. Guss JM, Merritt EA, Phizackerley RP, Freeman HC (1996) The structure of a phytocyanin, the basic blue protein from cucumber, refined at 1.8 A resolution. J Mol Biol 262: 686-705.

63. Hart PJ, Nersissian AM, Herrmann RG, Nalbandyan RM, Valentine JS, et al. (1996) A missing link in cupredoxins: crystal structure of cucumber stellacyanin at 1.6 A resolution. Protein Sci 5: 2175-2183.

64. Eren E, Murphy M, Goguen J, van den Berg B (2010) An active site water network in the plasminogen activator pla from *Yersinia pestis*. Structure 18: 809-818.

65. He Y, Yang X, Wang H, Estephan R, Francis F, et al. (2007) Solution-state molecular structure of apo and oleate-liganded liver fatty acid-binding protein. Biochemistry 46: 12543-12556.

66. Pazy Y, Kulik T, Bayer EA, Wilchek M, Livnah O (2002) Ligand exchange between proteins. Exchange of biotin and biotin derivatives between avidin and streptavidin. J Biol Chem 277: 30892-30900.

67. Pazy Y, Eisenberg-Domovich Y, Laitinen OH, Kulomaa MS, Bayer EA, et al. (2003) Dimer-tetramer transition between solution and crystalline states of streptavidin and avidin mutants. J Bacteriol 185: 4050-4056.

68. Eisenberg-Domovich Y, Pazy Y, Nir O, Raboy B, Bayer EA, et al. (2004) Structural elements responsible for conversion of streptavidin to a pseudoenzyme. Proc Natl Acad Sci U S A 101: 5916-5921.

69. Arni RK, Ward RJ, Gutierrez JM, Tulinsky A (1995) Structure of a calcium-independent phospholipase-like myotoxic protein from Bothrops asper venom. Acta Crystallogr D Biol Crystallogr 51: 311-317.

70. Epstein TM, Yu BZ, Pan YH, Tutton SP, Maliwal BP, et al. (2001) The basis for k(cat) impairment in prophospholipase A(2) from the anion-assisted dimer structure. Biochemistry 40: 11411-11422.

71. Oh BH (1995) A probe molecule composed of seventeen percent of total diffracting matter gives correct solutions in molecular replacement. Acta Crystallogr D Biol Crystallogr 51: 140-144.

72. Georgieva DN, Rypniewski W, Gabdoulkhakov A, Genov N, Betzel C (2004) Asp49 phospholipase A(2)-elaidoylamide complex: a new mode of inhibition. Biochem Biophys Res Commun 319: 1314-1321.

73. Scott DL, White SP, Browning JL, Rosa JJ, Gelb MH, et al. (1991) Structures of free and inhibited human secretory phospholipase A2 from inflammatory exudate. Science 254: 1007-1010.