BMC Bioinformatics

Research article

Open Access

# Structural characterization of genomes by large scale sequence-structure threading

Artem Cherkasov*[1,2] and Steven JM Jones[1]

Address: [1]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada and [2]Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada

Email: Artem Cherkasov* - artc@interchange.ubc.ca; Steven JM Jones - sjones@bcgsc.ca

* Corresponding author

Published: 03 April 2004

*BMC Bioinformatics* 2004, **5**:37

Received: 09 October 2003

Accepted: 03 April 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/37

## Abstract

**Background:** Using sequence-structure threading we have conducted structural characterization of complete proteomes of 37 archaeal, bacterial and eukaryotic organisms (including worm, fly, mouse and human) totaling 167,888 genes.

**Results:** The reported data represent first rather general evaluation of performance of full sequence-structure threading on multiple genomes providing opportunity to evaluate its general applicability for large scale studies.

According to the estimated results the sequence-structure threading has assigned protein folds to more then 60% of eukaryotic, 68% of archaeal and 70% of bacterial proteomes.

The repertoires of protein classes, architectures, topologies and homologous superfamilies (according to the CATH 2.4 classification) have been established for distant organisms and superkingdoms. It has been found that the average abundance of CATH classes decreases from "alpha and beta" to "mainly beta", followed by "mainly alpha" and "few secondary structures".

3-Layer (aba) Sandwich has been characterized as the most abundant protein architecture and Rossman fold as the most common topology.

**Conclusion:** The analysis of genomic occurrences of CATH 2.4 protein homologous superfamilies and topologies has revealed the power-law character of their distributions. The corresponding double logarithmic "frequency – genomic occurrence" dependences characteristic of scale-free systems have been established for individual organisms and for three superkingdoms.

Supplementary materials to this works are available at [1].

## Background

Recent world-wide progress in sequencing projects led to the exponential growth of genomic information and launched a race in the area of structural genomics. The development of bioinformatics tools of gene prediction and methods of sequence matching and threading allowed structural evaluation of sizable portions of complete proteomes, and gave a boost to the emerging areas of comparative structural genomics and proteomics.

The investigation of repertoires of protein structures and functions employed by species at different taxonomy levels led to numerous important discoveries in life science. Thus, the analysis of patterns of folds distribution across

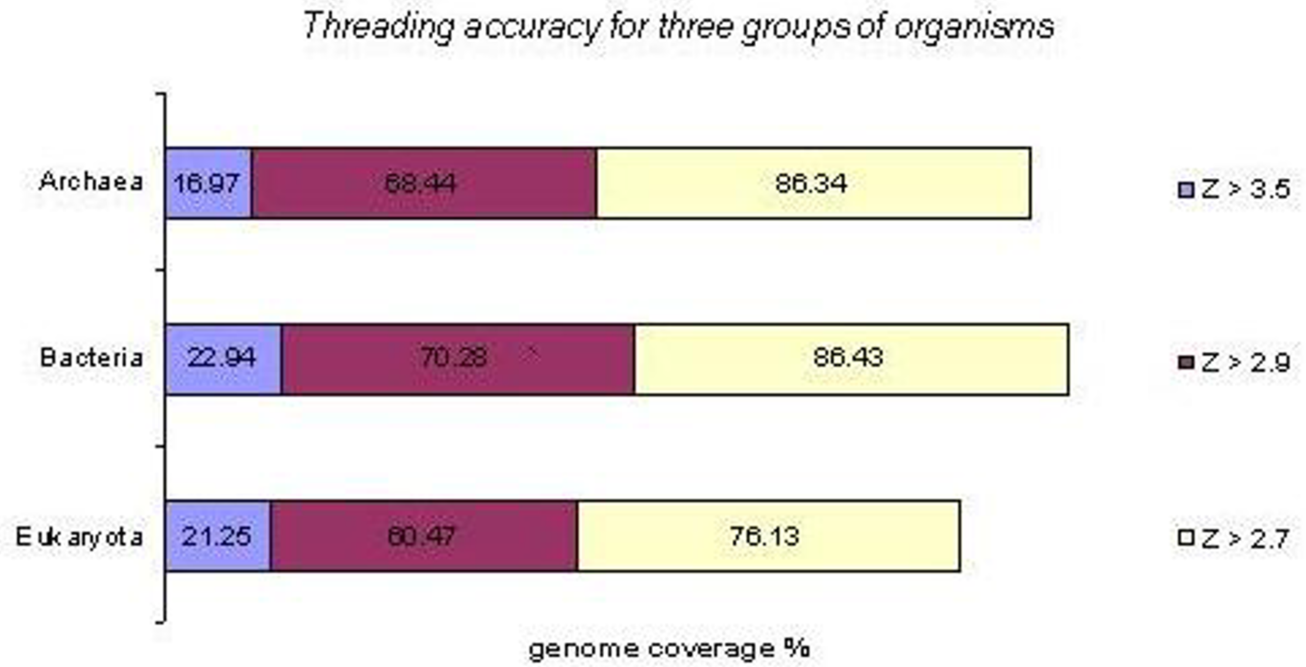

**Figure 1**
The estimated coverage of superkingdoms by medium-, high- and very high quality threading predictions.

three domains of life helped to identify instances of intra- and inter-species lateral gene transfer [2,3]; the characterization of common and unique protein folds provided valuable information about potential drug targets; and the discovery of the scale-free character of gene propagation allowed the estimation of a finite number of basic protein shapes [4].

There is little doubt about the great prospects within the area of comparative proteomics, but there is no shortage of challenges too. An adequate yet general estimation of the three-dimensional structure of a protein from its amino acid sequence remains the biggest obstacle in the field. There are two broad approaches to protein structure prediction: *ab initio* modeling and fold recognition. The former relies on well-understood principles guiding the folding of isolated amino-acid sequences into energetically favorable three-dimensional conformations. Although they are physically justified, these methods do not yet possess useful accuracy, speed and reliability suitable for large-scale proteome studies.

Fold recognition techniques utilize the wealth of experimentally determined protein structures accumulated in the protein databank [5]. Protein scientists thoughtfully analyzed these structures to determine a redundant repertoire of structural motifs used by nature to build proteins. The motifs have been catalogued into numerous standard libraries of protein folds (such as SCOP, CATH, FSSP, MMDB, LPFC, VAST, ASTRAL, SUPERFAMILY) in which protein "building parts" are classified at several hierarchy levels [6-14].

The fold recognition approaches can be divided into two broad classes by the ways in which they utilize libraries of standard folds. The first group of fold recognition methods, called profile-based approaches, represents structural information in linear form, called a profile. A profile reflects the statistically derived probabilities of the occurrence of residues in a particular structure [15-22]. The profile-based fold recognition methods use conventional sequence alignment tools (such as BLAST, FASTA, PSI-BLAST, Hidden Markov Model) to find matches between a probe sequence of unknown structure and the appropriate library entity. The profile-based approaches are very rapid – the modern text alignment algorithms and computer hardware make it a routine operation to process several medium-sized proteomes on a single CPU in a day. In the same time, an unknown protein can only be characterized by a profile if it has reasonable sequence similarity with protein(s) with known structure.

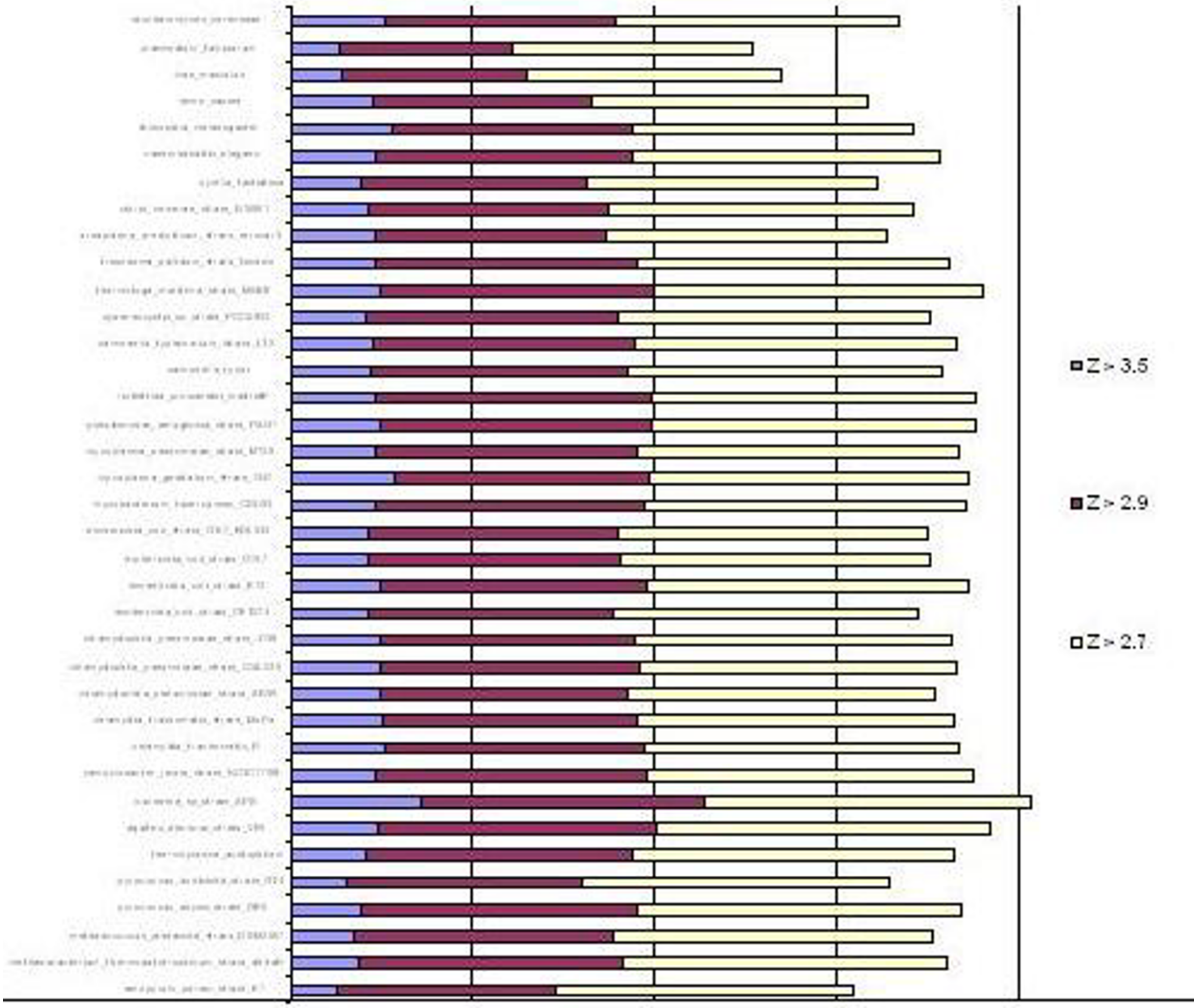

**Figure 2**
The estimated coverage of proteomes by medium-, high- and very high quality threading predictions.

The second strategy of usage of folds libraries is sequence threading. The threading utilizes empirical pair potentials, scoring the likelihood of two residues being at a certain distance in a space. This approach is based upon the assumption that seemingly countless different proteins fold into a limited number of shapes (estimates vary from 4,000 to 10,000+) [4,23] and that nearly all protein structures can be described based upon these shapes. Threading attempts to assign folds for a protein sequence by sampling it onto each member of a folds library using pseudo-energy as a measure of fit [24-29]. Independent from the sequence information, threading has been shown to make accurate predictions even in a "twilight zone" of <25% sequence identity, where sequence-based approaches normally fail. When being benchmarked with a set of proteins with known 3D structures, sequence – structure threading demonstrated accurate performance even well below 25% sequence identity level [30]. However, in contrast to the profile-based methods, the threading-based approaches are too slow to be widely applicable for large-scale structural genomics.

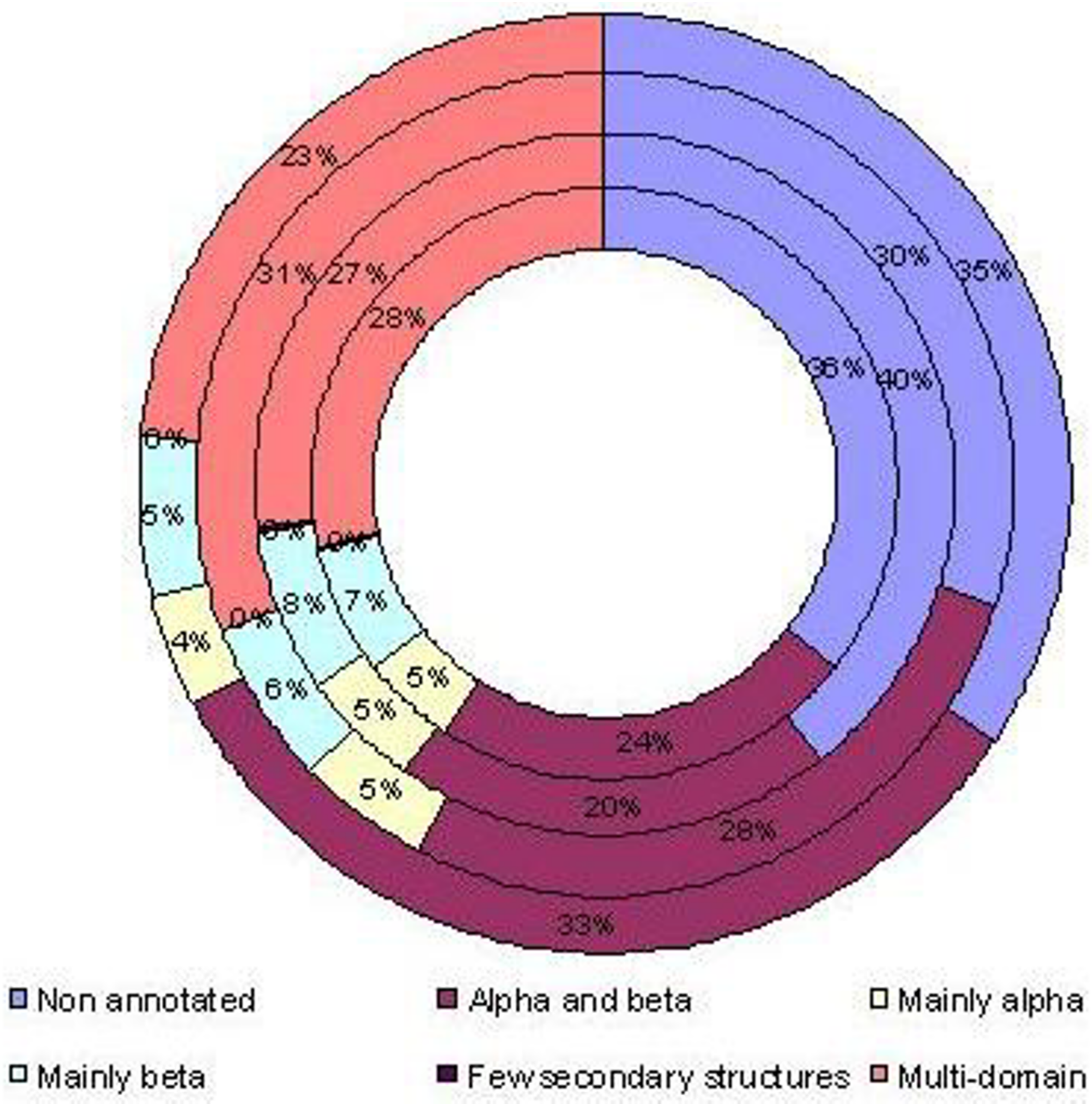

**Figure 3**
Pie charts of total and superkingdom-specific distributions of protein classes.

Another original approach called GenTHREADER using sequence-sequence threading within one day could assign known CATH folds to 46 percents of *Myciplasma genitalium* proteome (containing 468 ORFs) [31]. This fast and accurate approach utilizes some features of classic threading techniques but also largely relies on traditional sequence alignment.

Numerous comparative structural genomic studies have been reported to date describing dozens of structurally characterized proteomes [3,14,32-35]. Large collections of protein folds predictions across genomes are currently available on-line [36-38]. In should be stressed, however, that all these structural predictions have been performed by various automated sequence-sequence matching techniques. Up until now no comparative studies capitalizing on sequence-structure full threading (viewed by some as most accurate and comprehensive) have yet been reported.

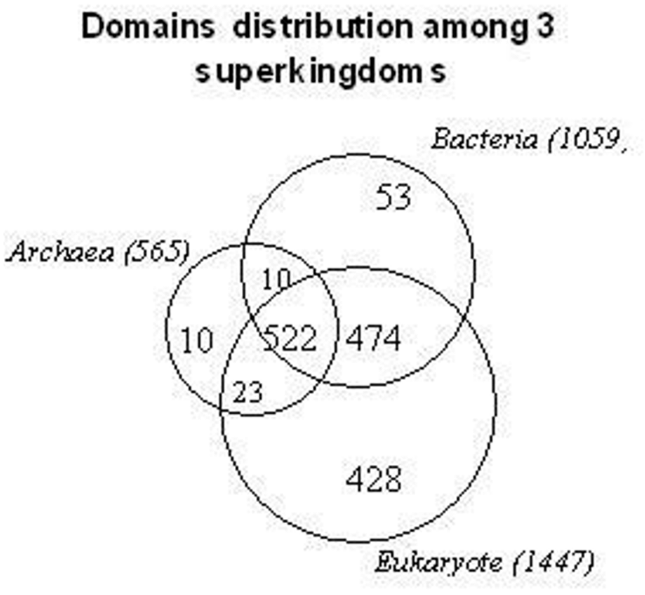

**Figure 4**
Venn diagrams of the distribution of distinct CATH domains shared by species from three domains of life.

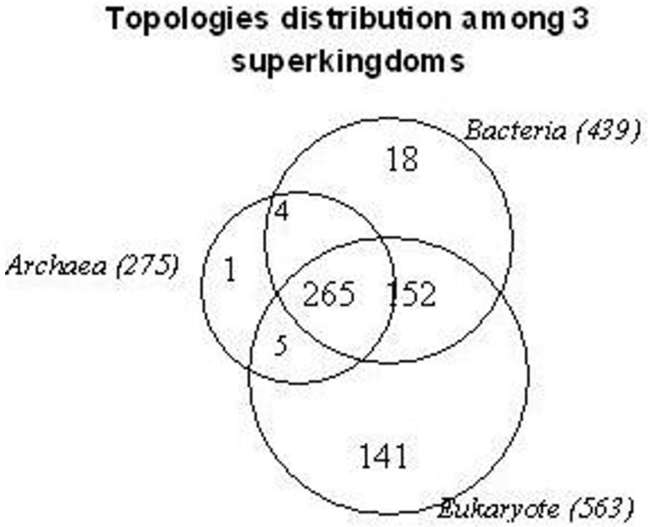

**Figure 5**
Venn diagrams of the distribution of distinct CATH topologies shared by species from three domains of life.

In our recent work, we have used the large-scale full sequence-structure threading to discover novel bacterial virulence factors mimicking host functions. The hypothesis was that, under selective pressure, pathogen genes have evolved to encode proteins that functionally mimic host proteins independently of significant primary sequence similarity. We suggested that such bacterial "mimickers" could be considered as potential virulence factors and, thus, the objective was to identify pathogen genes encoding proteins with low sequence identity but high structural similarity with host counterparts. Since the threading remains the only reliable alternative for comparison of proteins with limited sequence identity, we have adopted the THREADER program [28] which we have customized for large-scale distributed processing.

To achieve the described objectives we have aimed to process a sufficient number of complete genomes from all domains of live, covering a range of parasitic and free living species. Thus, we have performed sequence-structure threading for more than 30 complete proteomes of organisms from Bacteria, Archaea and Eukaryote superkingdoms. Specific aspects of the discovery of bacterial virulence factors by full threading will be discussed in the separate report. In the present work we utilize the generated information in a conventional form of comparative structural genomic analysis and discuss the applicability of classical full threading for large scale analysis.

## Results

The THREADER fold recognition program uses the CATH folds library, which has four hierarchical levels of classification of proteins: by classes, architectures, topologies and homologous superfamilies [27]. The CATH classes are determined by protein secondary structure composition, the architectures reflect the overall shape of the protein domain, protein topologies depend on both the overall shape and connectivity of the domain, and the homologous superfamily level groups domains with significant sequence similarity [39].

In contrast to profile-based approaches, the threading cannot readily specify several distinct structural domains for one sequence; rather, it tends to associate the entire sequence with a particular CATH entity. The THREADER samples a raw sequence into domains from the CATH library to produce multiple scores quantifying different aspects of the threading. One of them, a Z score of the weighted sum of threading and solvation energies, is regarded as the characteristics of overall goodness of fit between probe sequence and library fold. The results of the threading can be considered at three levels of prediction accuracy. When the Z threading score is above 3.5 then the match between the fold and probe sequence is considered as very significant. The hits with Z > 2.9 are regarded as significant, and Z values between 2.7 and 2.9 represent possibly correct threading prediction [28].

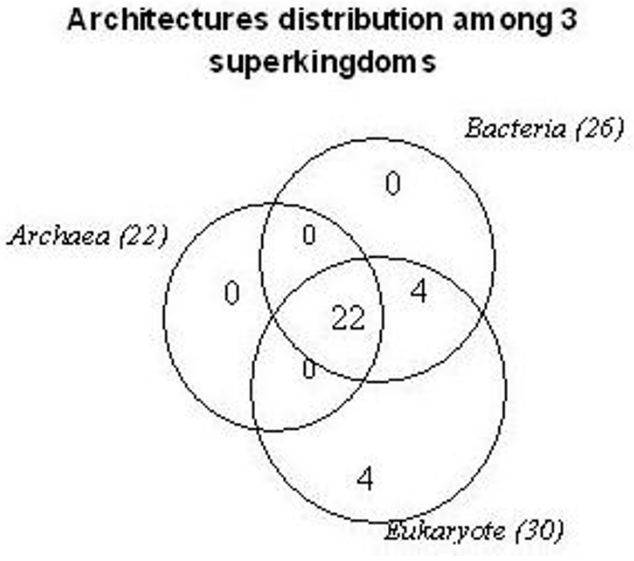

**Figure 6**
Venn diagrams of the distribution of distinct CATH and architectures shared by species from three domains of life.

The threading is a computationally-intense procedure that requires several CPU hours to process a single protein sequence. To allow for a large-scale threading of complete proteomes, we have implemented an automated parallel protocol for distributed processing of THREADER on Beowulf cluster. The distributed processing made it possible to perform the threading-based structural characterization of large sets of genomic information, including complete proteomes of human and major bacterial pathogens (the overall processing took several dozens years of a single CPU time).

We have anticipated that such threading will provide wider genome coverage and more accurate structural predictions than traditional sequence-based approaches and thus will provide a valuable insight into folds composition of the proteomes studied according to the CATH classification. It has also been expected that results will allow general evaluating of accuracy, genomic coverage and CPU usage by classical full threading in order to access feasibility to trade its lower processing speed for higher quality of predictions in large scale studies.

### *The performance of large-scale threading*

Using the sequence-structure threading we have processed complete proteomes of 37 organisms (including 25 bacteria, 6 eukaryotes and 6 archaea) totaling 167,888 sequences. The threading has produced 36,240 predictions with the Z scores above 3.5 threshold that correspond to 21.58 percent of the processed sequences. The fractions of protein structure predictions with high ($Z > 2.9$) and acceptable ($Z > 2.7$) accuracy averaged 64.7 and 80.7 percent respectively. The estimated figures of genome coverage by the threading for the studied organisms are listed in Table 1 (see additional file 1) for three levels of prediction accuracy.

These parameters are plotted on Figures 1 and 2 for the studied organisms and three superkingdoms.

It should be noted, that the produced structural predictions provide first comprehensive enough evaluation of classic sequence-structure threading in large scale structural genomics studies. Data from table 1 (see additional file 1) demonstrate that the average accuracy of protein structure prediction is slightly better for microbial organisms: protein folds have been confidently assigned to more then 60% of eukaryotic genomes, about 68% of archaeal genomes and 70% of bacterial genomes (here and later a protein is considered to be assigned to particular CATH homologous superfamilies if the corresponding Z threading parameter is above 2.9). The better coverage for bacteria may perhaps be explained by the fact that bacterial proteins underwent more extensive experimental characterization and, thus, the Protein Data Bank is heavily biased with bacterial data.

It is difficult to compare the estimated genome coverage by the threading with performance of the sequence-based methods directly, as they do not grade their predictions by levels of accuracy. It is known, however, that gene coverage of profile-based approaches varies between 10 and 45 percent [3,14,33,40-45]. Some newer automated genome annotation techniques could assign up to 62% of certain genomes [46]. Therefore, the estimated results of full sequence-structure threading can be viewed as generally 6–8% better or comparable to those obtained in similar studies. One can argue, however, that when using a less strict cutoff of $Z = 2.7$ (corresponding to "possibly correct" predictions according to the THEADER) the coverage of genomes goes to up to 90% (see Table 1 as an additional file 1). On another hand, as it can be seen from table 2 (see additional file 2), a sizable fraction of the estimated predictions corresponds to multi-domain protein folds according to the CATH 2.4 which was the default library for the THREADER. Unlike SCOP, 3D-PSSM or other similar databases created with a great deal of human insight and curation, the CATH collection is based on automated classification protocols [7]. Apparently, by that time when the CATH 2.4 was created, those protocols could not sufficiently distinguish individual domains in the most complex multi-fold entries. However, in the very latest 2.5 release of the CATH (became available on-line on Dec. 01, 2003) all multi-domain entries have been split into sim-

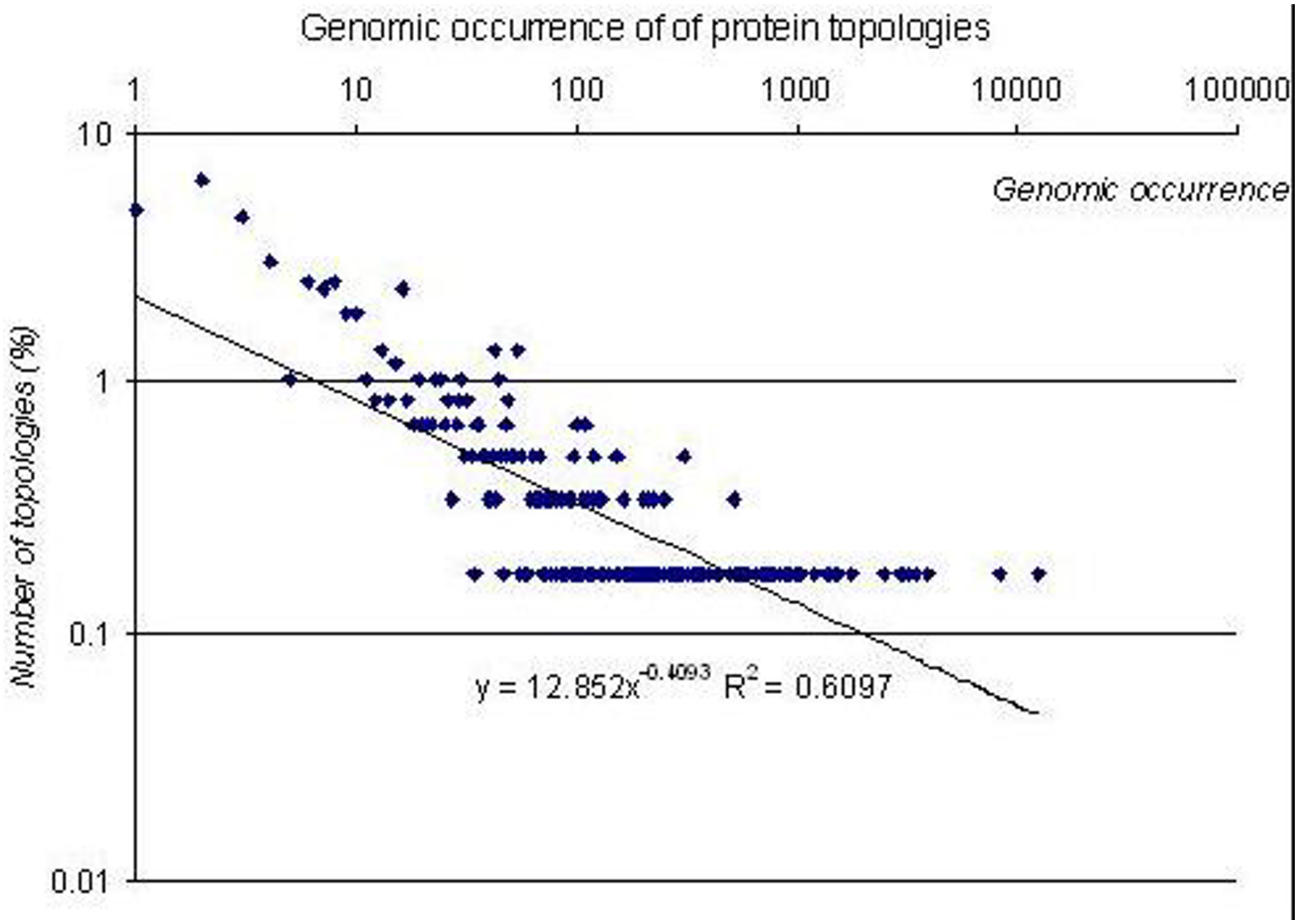

**Figure 7**
General genomic occurrence of frequencies of protein topologies (in % of totals).

pler α, β and (α+β) components. Thus, in view of these recent changes, our multi-domains predictions can also be considered as not successful for assigning defined classes to the corresponding proteins. It will reduce the C-level genome coverage by sequence-structure threading to 12%, 38% and 45% for very significant-, significant – and possibly correct predictions respectively. In the same time it should be stressed, that although the multi-domain predictions do not directly contribute to the knowledge about representation of α, β, α+β and "few secondary structures" elements in complete proteomes, they do provide meaningful insight on three dimensional structures for the target sequences.

Thus, it would be possible to summarize, that in our experience the achieved precision and genome coverage by the full threading may not compensate for more then 3 hours of CPU time we had to spend to process a single sequence with the THREADER package. In the same time, it cannot be underestimated that the reported data represent first broad application of full *sequence-structure* threading to multiple genomes with all its pluses and minuses. There is no doubt in our mind that sequence-structure threading currently remains the only suitable instrument for protein structure predictions when no sequence homology information is available. The full sequence-structure threading can be used as powerful complimentary approach for structural genomic studies and the reported results (independent from any sequence homology information) can be viewed as very complementary to the existing structural genomic databases. Thus, we have developed the web-based interface: [1] for open access to our results. It should also be stressed that the accuracy of the threading could probably be improved further by using pre-computed protein secondary structures and domain boundaries. However, these approaches have not been used, since we

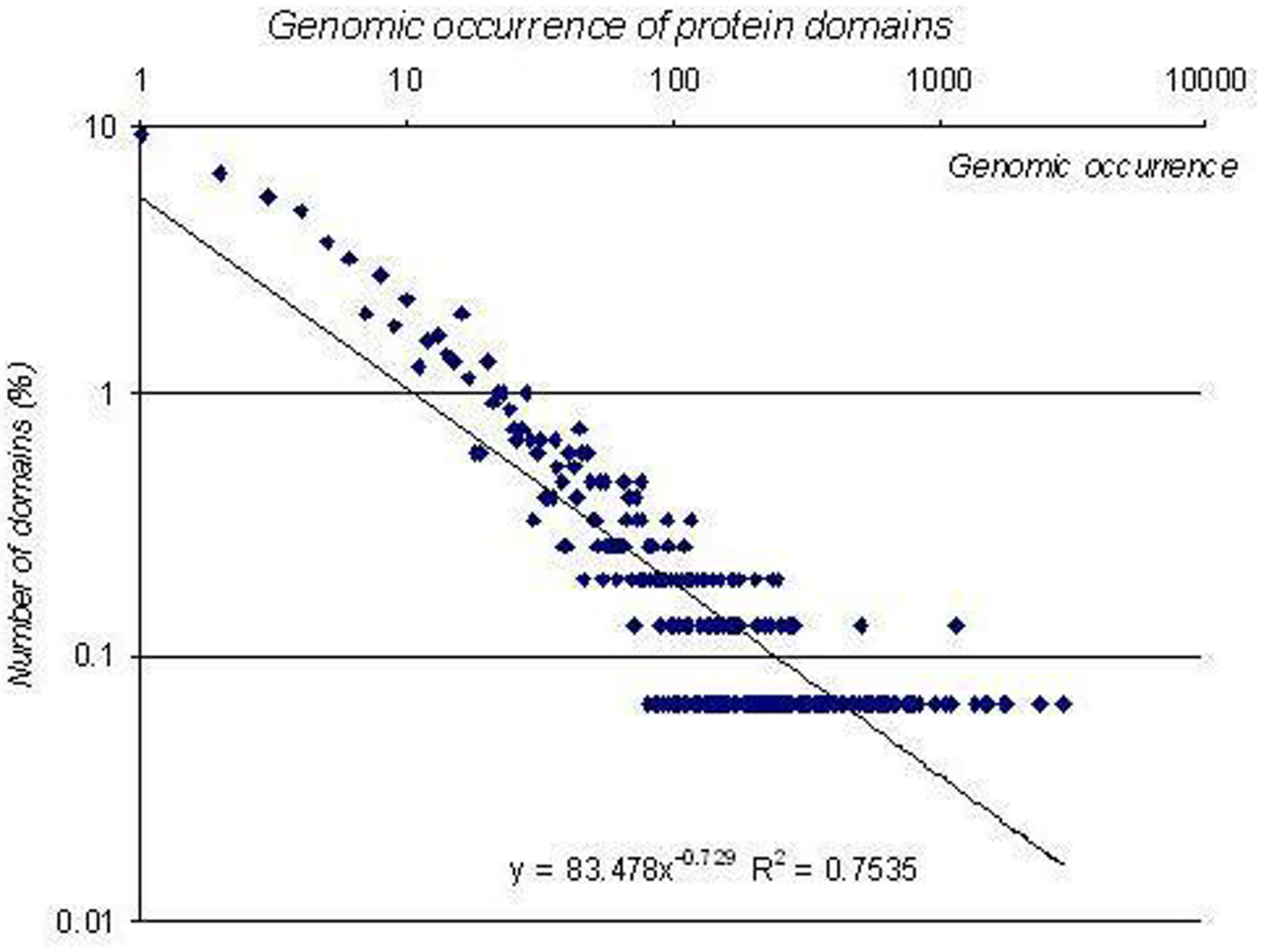

**Figure 8**
General genomic occurrence of frequencies of protein domains (in % of totals).

tried to minimize human intervention and maximize the automation of large-scale protein structure prediction.

## Discussion

### *Fold repertoires of eukaryotes, bacteria and archaea*

The statistics of protein folds distribution represent one of the most important aspects of structural genomics. Information about the most abundant and unique folds can provide valuable insight into evolutionary relations and can serve as an important source of drug target information [4,34]. Previous studies have produced several very similar lists of the most abundant protein folds according to the SCOP classification. Thus, Wolf et al. have identified P-loop NTP-ase as the most abundant fold in all three superkingdoms. The next most common protein structures in bacteria and archaea have been characterized as ferredoxin-like fold, TIM barrel and methyltransferase, whereas in eukaryotic proteomes the most common fold was followed by protein kinase, β-propeller and TIM barrel. Muller et al. have also named P-loop NTP-ase as the most common fold, while other members of their "top five" lists varied. Hegyi et al. have developed a fold ranking system which produced similar folds abundance order: P-loop NTP-ase, ferrodoxin, TIM barrel. It has also been generally agreed that the significant fraction of known protein folds can be found in all three major groups of organisms [47] and that the distribution of folds within organisms can differ significantly [33,47].

Using the generated threading results we have estimated the distribution of CATH classifications among three major groups of organisms. The populations of "all-

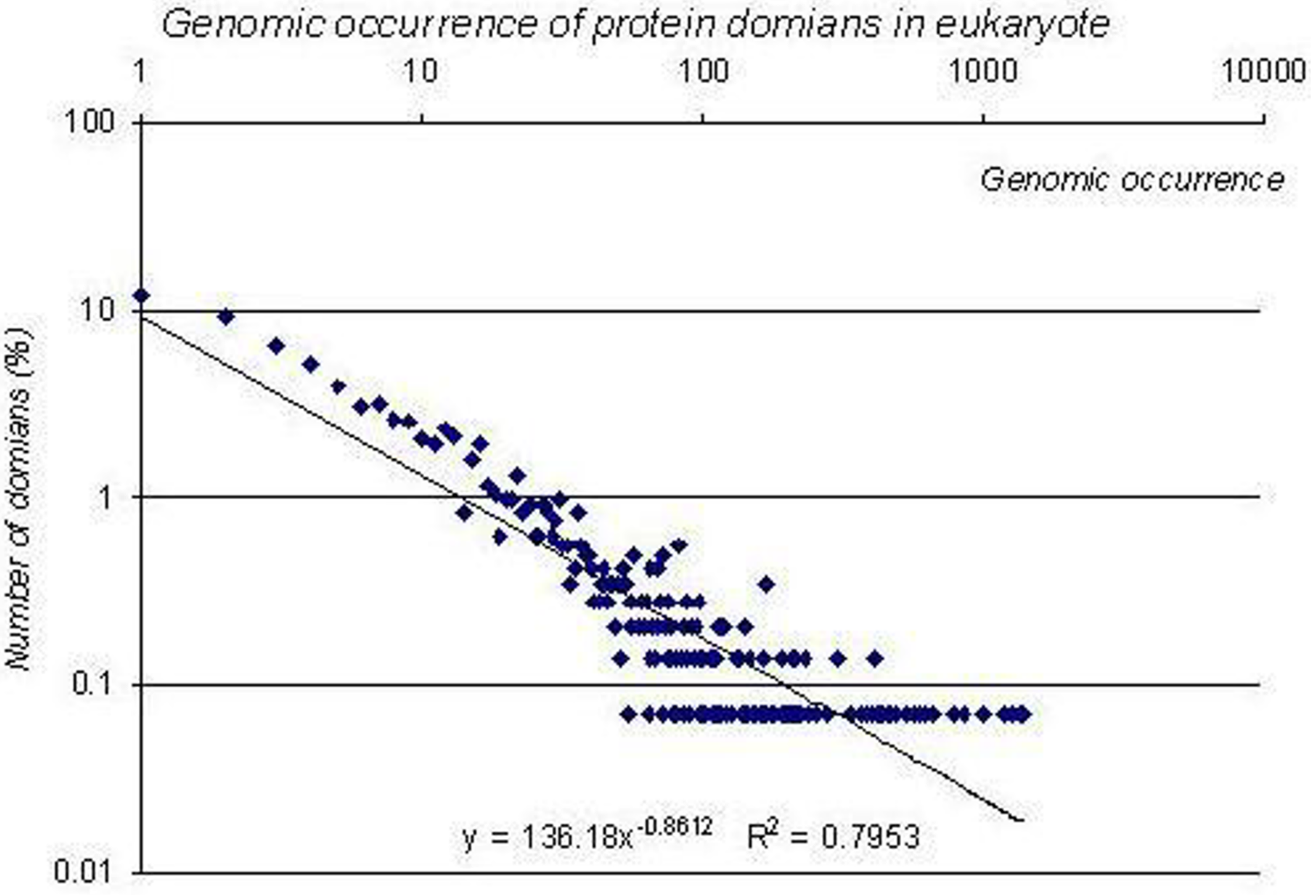

**Figure 9**
Genomic occurrence of frequencies of protein domains in eukaryote.

alpha", "all-beta", "alpha and beta", "few secondary structures" classes are presented in Figure 3 for eukaryotes, bacteria and archaea.

According to the numerical data from Table 2 (see additional file 2), apart from multi-domain predictions accounting for more then 44% of the threading results (and requiring further separation into conventional classes), the most abundant protein class is "alpha and beta" (36.6%), followed by "mainly beta" (10.8%), "mainly alpha" (8.4%) and "few secondary structures" (0.2%). The difference in the abundance of classes when three major groups of organisms are considered is noticeable. More than half of the archaeal proteins belong to "alpha and beta" class while no proteins with "few secondary structures" have been detected. The distributions of eukaryotic and bacterial protein classes are similar. The only difference is that the proportions of secondary structures are more homogenous for eukaryotes: their "mainly alpha" and "mainly beta" proteins have higher occurrences, and "alpha and beta" have lower occurrences, when compared to bacteria. The estimated higher proportion of multi-domain predictions agrees with the results of global studies by Teichmann with co-authors [48,49] assigning larger parts of prokaryote and particularly eukaryote proteomes to multi-domain folds.

From the total of 1893 default CATH homologous superfamilies used by the THREADER, we have identified 1520 as being present in the organisms studied. The data generated suggests that eukaryotic species contain the largest fraction of the established H-classifications – 1447, while 1059 distant homologous superfamilies have been found in bacteria and 565 in archaea.

The 3bct00 CATH fold corresponding to "Armadillo repeat" topology, "horseshoe" architecture and "mainly alpha" class has the highest total count (2865) within the studied proteomes. Two other folds – 1gdoA0 (class: alpha and beta, architecture: 4-layer sandwich, topology: glutamine phosphoribosylpyrophosphate) and 1gotB0 (mainly beta, 7 propellor, methylamine dehydrogenase) – have been counted 2728 and 2283 times respectively. This abundance order is changed to 1gdoA0, 3bct00, 1gotB0 if the fold abundance is calculated as a sum of fractions of distant protein fold in individual proteomes. The repertoires of protein functions in the studied organisms can be illustrated by distributions of protein topologies. In total, 588 out of 703 CATH topologies have been identified in the studied proteomes. It has been found that eukaryotic organisms contain the largest variety of protein topologies – 563. Bacterial species are constituted with 439 topologies, archaeal with 275.

Two CATH topologies – Rossman fold and TIM barrel – produce the highest frequency of occurrence. These two topologies are the most abundant for all the studied organisms except *Plasmodium falciparum* (containing an unusually large fraction of hydrolases). Rossman fold and TIM barrel account for 14 to 28 percent of topology compositions of the individual proteomes, what, likely, is determined by known multi-functionality of these folds. Several other highly abundant topologies also been identified within the studied organisms. These folds mostly associated with transport and metabolism functions include hydrolase, oxidoreductase, neuraminidase, transferase, Armadillo repeat, glutamine phosphoribosyl-pyrophosphate, methylamine dehydrogenase, and isomerase/synthase bifunctional proteins. Some species contain large fractions of phosphotransferases, binding proteins, sugar transport proteins, isomerazes also related to transport and metabolism. Various toxins folds also have high abundance, particularly in the bacterial genomes. The list of top 10 common topologies identified by the current study is presented in Table 3 (see additional file 3).

The ranking of protein topologies has been conducted in three different ways: based on topology count for the entire dataset, using a sum of fractions of distinct topologies within individual genomes, and by numerical ranking of topologies within organisms. All three approaches have produced very similar abundance orders identifying Rossman fold and TIM barrel as the top-ranked protein topologies.

The threading results have also illustrated the fact there are very few protein topologies having a high frequency of genome occurrence; the overwhelming majority of protein topologies occur quite rarely (the uneven character of the distribution of protein folds will be discussed in greater detail later). This makes it difficult to produce a very precise topologies abundance ranking and to compare the results of different structural genomics studies (also biased by the choice of the particular organisms studied). Nonetheless, it is fair to conclude that the estimated CATH T-ranking generally agrees with the previous studies mentioned above, which have identified P-loop NTPase, TIM barrel and Rossman fold as the most abundant folds according to the SCOP classification [3,33,47]. The similarity is even more noticeable considering that the organizations of SCOP and CATH 2.4 libraries are quite distinct.

According to the threading results, the most abundant protein architectures can be placed in the following order: 3-layer (aba) sandwich, barrel, 2-layer sandwich. The top three are followed by non-bundle, 4-layer sandwich, horseshoe and 6-bladed propeller. As it can be noted from the list (and previously mentioned by other authors), the most abundant protein folds possess rather high symmetry. Perhaps, the corresponding symmetric protein compositions correspond to energetically more favorable configurations and, hence, have some evolutionary advantage.

30 out of 31 distinct known protein architectures have been identified in the studied organisms. Eukaryotic proteomes contain all 30 identified architectures, 26 architectures have been identified in the studied bacteria and 22 in archaea. The distribution of architectures can also be characterized as highly uneven. Remarkably, one protein architecture has been identified at only one occasion (a super-fold) and several others could be found fewer than four or five times. Not surprisingly, all these architectures also appeared to be superkingdom-specific. Figure 4 represents a Venn diagram of the established distribution of protein architectures in three superkingdoms.

Figures 4, 5, 6 demonstrate that bacterial and archaeal proteins do not form any superkingdom-specific architectures and do not share any architectures that would also be absent from eukaryotes. Eukaryotes, however, exclusively possess four protein architectures not present elsewhere. Already mentioned eukaryotic super-fold is associated with bactericidal function and involved into eukaryotic host defense. Besides, there are 5-stranded propeller, orthogonal prism and aligned prism architectures which can only be found in eukaryotes. The corresponding configurations have been adopted by various membrane associated proteins in eukaryotes. Likely, these eukaryotic protein architectures evolved to accommodate specifics of eukaryotic cell wall composition. Orthogonal prism configuration is related to lactins and mannose-binding function. Based on the observation that the corresponding architecture has been observed only in mouse, it is feasible to speculate it may be related to certain mammalian – specific features. Four other architectures are exclusively shared between eukaryotes and bacteria: rib-

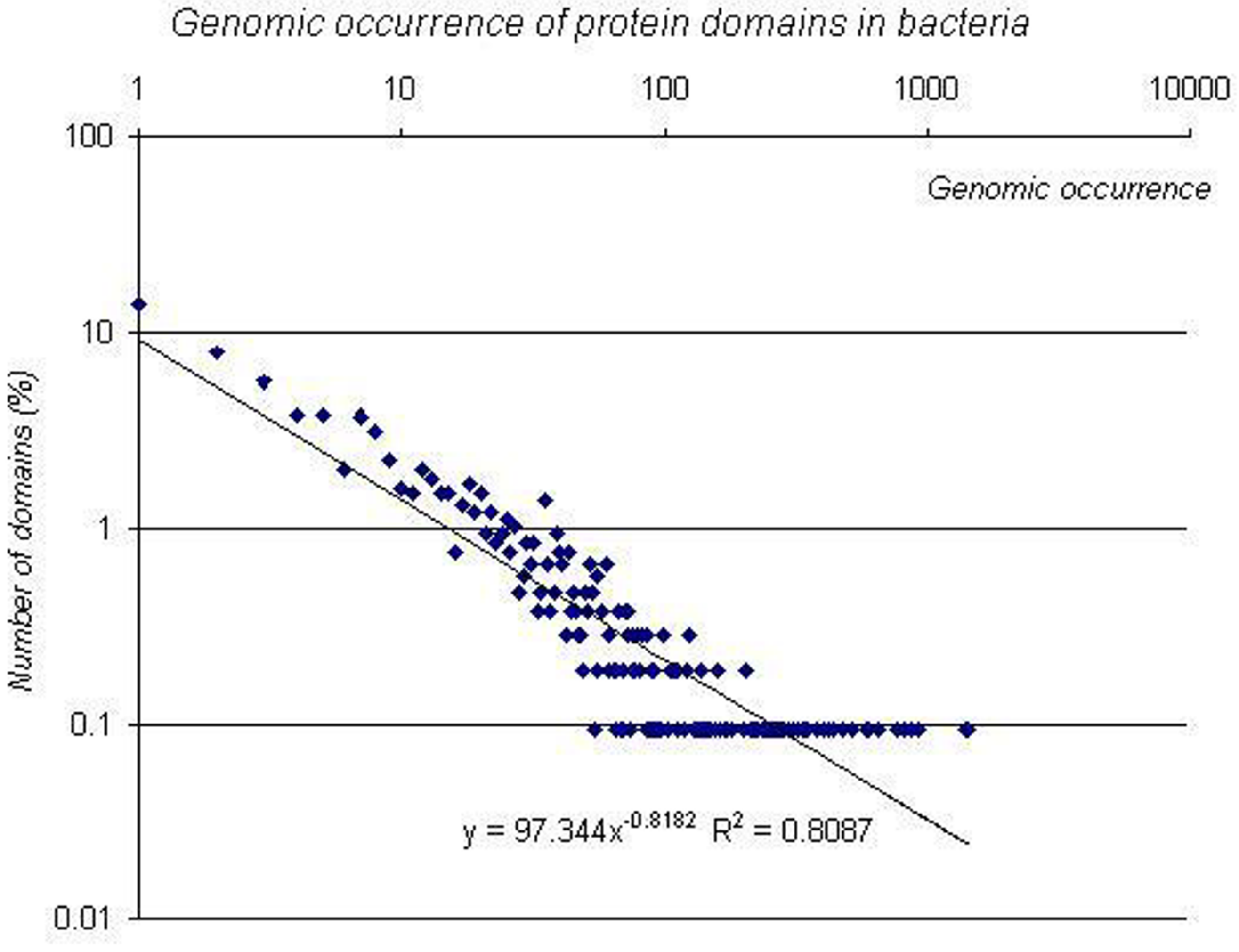

**Figure 10**
Genomic occurrence of frequencies of protein domains in bacteria.

bon, 3-layer sandwich, distorted sandwich and irregular architecture. It seems to be a general observation that low complexity and irregular structures are absent from archaea; species of this group also lack the entire "few secondary structures" protein class. It has been previously outlined by Gerstein and Levitt that small folds prevail in eukaryotic proteomes as they are mostly involved into intercellular communication and regulation in vertebrates [46]. This may explain why the representatives of the smaller "few secondary structures" class have not been found in archea, while 42 of them have been identified in human- and 82 in mouse proteomes.

The distribution of CATH topologies between three superkingdoms follows similar trends. 588 out of the total of 703 CATH protein topologies have been identified in the studied organisms. As can be seen from Figure 5, most of them are at least present in two superkingdoms. 265 protein topologies can be found in all three species groups. These compose almost the entire topology repertoire of the archaea, which exclusively share only 4 topologies with bacteria and 5 with eukaryotes. Only one merely archaeal topology could be found thus far. This observation agrees with the previous studies that have pointed to the near absence of archaea-specific SCOP folds [33,47]. These findings provide another example of uniqueness of archaeal fold repertoire (in addition to the previously indicated absence of superkingdom-specific architectures, unusually high fraction of "alpha and beta" proteins and lowered "low-complexity" content). The low

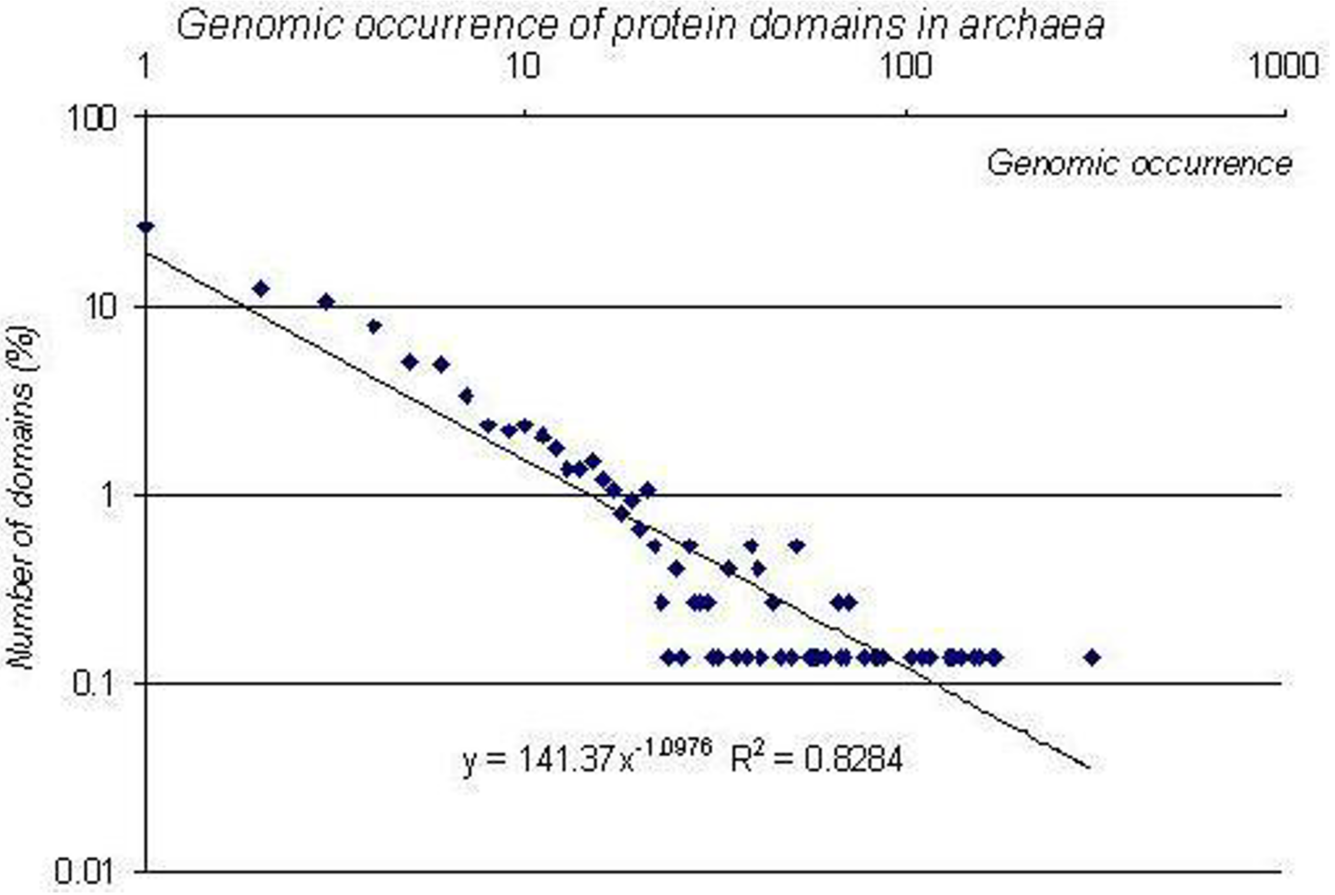

**Figure 11**
Genomic occurrence of frequencies of protein domains archaea.

proportion of accurate structural predictions for archaeal proteins can also be viewed as supporting this idea (see Table 1 in additional file 1).

Another important observation of the work coincides with the previous findings that bacteria and eukaryotes share a significant fraction of their folds repertoires. The threading results indicate that out of 563 CATH topologies found in eukaryotes and 439 in bacteria, 152 are exclusively shared between these two superkingdoms. This figure composes more then 25 percent of the entire pool of identified protein topologies. This is in good agreement with results of Wolf et al. who indicated that more than 20 percent of SCOP folds can be shared between bacteria and eukaryotes. The fraction of protein homologous superfamilies exclusively shared between eukaryotic and bacterial organisms is even greater. Our results demonstrate that bacteria share almost 45 percent of their homologous superfamilies with eukaryotes (474 out of 1059), while the H-level sharing between bacteria and archaea is very limited (see Figure 4). The situation with the estimated superkingdom-specific homologous superfamilies is very similar to the picture of distribution of protein topologies: only 10 protein homologous superfamilies can be exclusively found in archaea and 53 in bacteria while the eukaryotes contain 428 unique H-representatives (out of the total of 1447).

Figures 4, 5, 6 illustrate similar distribution trends at all three levels of CATH classification. The results demonstrate that bacterial and eukaryotic organisms have similar protein organizations and share a high degree of commonality. The similarities in their protein repertoires may illustrate the relevance of lateral gene transfer between bacteria and eukaryotes as well as an intensity of ancestral relationships between bacteria and eukaryotic organelles [50]. Archaeal organisms demonstrate quite distinct trends in protein organization: they seem to lack superkingdom-specific topologies and architectures and,

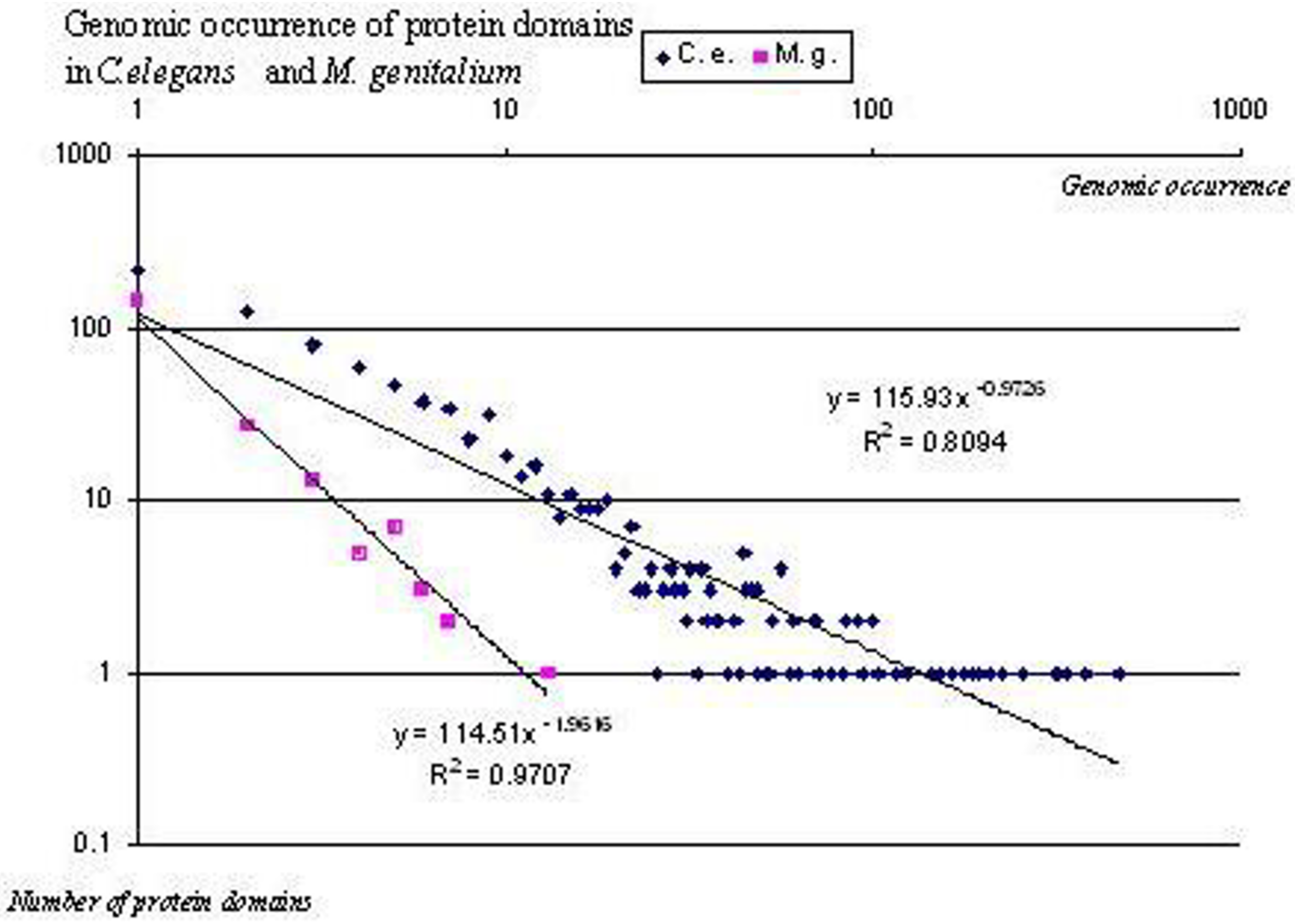

**Figure 12**
Genomic occurrence of frequencies of protein domains in *C. elegans* and *M. genitalium*.

as well, they do not contain proteins with irregular or low complexity structures.

At the moment, it is difficult to speculate whether it is possible to identify any species-specific protein topologies, as we have processed only a limited number of proteomes. At the same time, the structural characterization of complete proteomes by threading is in continuous progress and new conclusions may emerge.

### *Folds occurrences and power-law behavior of fold distributions*

It has been recently demonstrated by several independent studies that protein fold distribution is drastically uneven: there are few extremely common protein structures while most protein folds occur very infrequently. It has been previously shown that the occurrence of SCOP protein families, superfamilies and folds follow asymptotic power laws. The double logarithmic linear plots could be established for distribution of protein folds by the number of families, distribution of families by the number of domains, etc [4,34,35,51,52]. These findings laid the foundation for characterizing the evolution of the protein universe in terms of a growing scale-free system in which individual genes are represented as the nodes of a propagating network. The estimated scale-free character of such a network indicates a preference to duplicate genes encoding for already common protein folds [35].

We have analyzed structural predictions generated by the threading for the frequency of occurrence of particular CATH classes, architectures, topologies and homologous superfamilies. The frequency distributions have been established for the studied organisms and for superkingdoms. The double logarithmic plots estimated for frequencies of total genomic occurrence of protein homologous superfamilies, topologies and architectures for all proteins combined are presented in Figures 7, 8.

It can readily be seen from the figures that these dependences can be described by a power-law f($i$)~$i^{-b}$ function relating occurrences $i$ of CATH classifications with their corresponding frequencies of occurrence f($i$). The total distribution of protein homologous superfamilies determined for all the studied proteins has produced an exponent $b = 0.729$. The power factor for the distribution of frequencies of genomic occurrence of protein topologies has been established as $b = 0.409$.

The estimated values of the $b$ exponent appeared to be out of typical range (1.5÷3), characteristic for scale-free systems. Figures 7 and 8 demonstrate that the fitted $f(i) = ai^{-b}$ double logarithmic functions clearly deviate from the upper end of the distribution trends favoring the lowered magnitudes of $b$. The obvious reason for such behavior is in the fact that the majority of points in Figures 7, 8 are concentrated at the lower parts of the distributions (the areas corresponding to higher genomic occurrence).

Such "tailing" puts a significant statistical weight to the lower part of the funnel-like dependence, so the trend line tends to fit the majority of the data points at the bottom of a graph. Similar deviations of power-law trend lines can also be recognized in Figures 9, 10, 11, illustrating the genome occurrences of protein homologous superfamilies within three superkingdoms.

However, it should be noted that double logarithmic dependences estimated for superkingdoms have more profound scale-free character. The corresponding power factors $b$ for distribution of protein homologous superfamilies increase from -1.0976 for archaeal, through -0.8612 for eukaryotic, to -0.8182 for bacterial proteomes.

Parameters $a$ and $b$ of the power law dependences $f(i) = ai^{-b}$ relating the genomic occurrences $i$ with the frequencies of protein homologous superfamilies and topologies $f(i)$ have also been calculated for the distinct organisms and are presented in Table 4 (see additional file 4).

The numbers in the table illustrate that the magnitude of $b$ can vary significantly among the species. Thus, Figure 12 plots the genomic occurrences of homologous superfamilies frequencies within the genomes of *C. elegans* and *M. genitalium* where the difference in $b$ factors for the two organisms (-1.96 and -0.97 respectively) can readily be recognized.

It is difficult to speculate at this point whether the estimated deviation of the power exponents from the scale free range truly reflect specific aspects of distribution of protein folds in organisms, or merely result from the poor ability of power function to describe funnel-like dependences.

It is clear, however, that the estimated results illustrate the need for development of new statistical functions describing protein fold distributions in a more accurate way than the conventional double logarithmic "frequency – genomic occurrence" dependences. The development of such new statistical functions and tools is currently underway. We expect that new statistical approaches will help us to answer the questions raised by the reported study.

## Conclusions

We have analyzed the results of the large-scale automated threading procedure applied to complete proteomes of 6 eukaryotic, 25 bacterial and 6 archaeal organisms. The coverage and reliability of unmodified full threading procedure have been assessed for large-scale automated protein structure predictions. The sequence-structure threading allowed satisfactory assignment of structures to more than 60% of eukaryotic, 68% of archaeal and 70% of the bacterial proteomes analyzed.

The folds recognition results have also been estimated for very high and lower levels of prediction confidence; the estimated accuracy, genomic coverage and CPU usage by the classical full threading have generally demonstrated that the trade of lower processing speed of the method for its higher quality of predictions may not be justified for large scale studies.

The current work relying on sequence-structure threading has identified the most abundant and unique CATH 2.4 folds in individual species and superkingdoms. 3-layer (aba) sandwich has been characterized as the most abundant protein architecture and Rossman fold as the most common topology.

The results highlight similarities and differences in the protein compositions of eukaryotes, bacteria and archaea. It has been found that eukaryotes share a significant portion of their protein repertoires with bacteria, which illustrates the intensity of their ancestral relationships. The protein composition of archaeal organisms was characterized as being quite distinct and generally missing low complexity and protein structures.

It has been found that protein homologous superfamilies and topologies distributions in the studied organisms and superkingdoms obey the power law dependence characteristic of scale-free systems. The corresponding double logarithmic "frequency – genomic occurrence" dependences characteristic for scale-free systems have been established for individual organisms and for three superkingdoms.

## Methods

Threading has been carried out by the THREADER2 [28] program with default parameters. The CATH v4.2. fold assembly has been used as a library of standard folds.

The large-scale threading has been conducted on Beowulf cluster with 52 dual processor blades (2 × 1 GHz, 1 G RAM). The automated control has been implemented by the PVM-supported Perl scripts.

The threading results have been stored and manipulated within the MySQL database.

## Authors' contributions

SJ has developed the general concept of the work and participated in drawing the conclusions; AC has performed the fold prediction and carried out all the calculations.

## Additional material

### Additional File 1

*Accuracy of structural characterization of distant genomes by threading.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-37-S1.doc]

### Additional File 2

*Distribution of major classes of proteins from distinct organisms.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-37-S2.doc]

### Additional File 3

*The most abundant folds for the studied organisms.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-37-S3.doc]

### Additional File 4

*List of genomic properties characterizing the power law distribution of protein hierarchies.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-37-S4.doc]

## Acknowledgements

The work has been funded by the Vancouver Hospital and Health Sciences Centre research award for AC and by the Functional Pathogenomics of Mucosal Immunity project, funded by Genome Prairie, Genome BC and their industry partners, Inimex Pharmaceuticals and Pyxis Genomics.

SJMJ is a Michael Smith Foundation for Health Research Scholar.

## References

1. [http://www.pathogenomics.bc.ca/cgi-bin/threader/threader_folds.cgi].
2. Yanai I, Wolf Y, Koonin EV: **Evolution of gene fusions: horizontal transfer versus independent events.** *Genome Biology* 2002, **3(5):**1-0024.
3. Hegyi H, Lin J, Greenbaum D, Gerstein M: **Structural genomics analysis: characteristics of atypical, common and horizontally transferred folds.** *Proteins: Struct Funct Genet* 2002, **47:**126-141.
4. Koonin EV, Wolf YI, Karev GP: **The structure of the protein universe and genome evolution.** *Nature* 2002, **420:**218-223.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank.** *Nucl Acids Res* 2000, **28:**235-242.
6. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2002: refinements accommodate structural genomics.** *Nucleic Acids Res* 2002, **30:**264-247.
7. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds.** *Nature* 1994, **372:**61-634.
8. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, **6:**377-385.
9. Holm L, Sander C: **Dali/FSSP classification of three-dimensional protein folds.** *Nucleic Acids Res* 1997, **25:**231-234.
10. Orengo CA, Flores TP, Taylor WR, Thornton JM: **Identification and classification of protein fold families.** *Protein Eng* 1993, **6:**485-500.
11. Schmidt R, Gerstein M, Altman R: **LPFC: an internet library of protein family core structures.** *Protein Sci* 1997, **6:**246-248.
12. Madej T, Gibrat J-F, Bryant SH: **Threading a database of protein cores.** *Proteins* 1995, **23:**356-369.
13. Brenner SE, Koehl P, Levitt M: **The ASTRAL compendium for sequence and structure analysis.** *Nucl Acids Res* 2000, **28:**254-256.
14. Gough J, Karplus K, Hughey R, Chotia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313:**903-919.
15. Russell RB, Saqi MAS, Bates PA, Sayle RA, Sternberg MJE: **Recognition of analogous and homologous protein folds – assessment of prediction success and associated alignment accuracy using empirical substitution matrices.** *Protein Engineering* 1998, **11:**1-9.
16. Bowie JU, Luthy R, Eisenberg G: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, **253:**164-170.
17. Bates A, Jackson RM, Sternberg MJE: *Genomes, Molecular Biology and Drug Discovery* London, Academic Press; 1996.
18. Russell RB, Copley RR, Barton GJ: **Protein fold recognition by mapping predicted secondary structure.** *J Mol Biol* 1996, **259:**349-365.
19. Rice DW, Eisenberg GA: **3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence.** *J Mol Biol* 1997, **267:**1026-1038.
20. Rost B, Schneider R, Sander C: **Protein fold recognition by prediction – based threading.** *J Mol Biol* 1997, **270:**471-480.
21. Defay TR, Cohen FE: **Multiple sequence information for threading algorithms.** *J Mol Biol* 1996, **262:**314-323.
22. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF: **IMPALA: matching a proteins sequence against a collection of PSI-BLAST – constructed position-specific score matrices.** *Bioinformatics* 1999, **15:**1000-1011.
23. Machalek AZ: **Structural genomics: a slice of the proteomics pie.** *ASM News* 2001, **67:**441-446.
24. Godzik A, Skolnick J: **Sequence – structure matching in globular proteins: application to supersecondary and tertiary structure determination.** *Proc Natl Acad Sci* 1992, **89:**12098-12102.
25. Bryant SH, Altschul SF: **Statistics of sequence – structure threading.** *Curr Opin Struct Biol* 1995, **5:**236-244.
26. Murzin AG, Bateman A: **Distant homology recognition using structural classification of proteins.** *Proteins* 1997, **Suppl 1:**105-112.
27. Jones DT, Miller RT, Thornton JM: **Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing.** *Proteins* 1995, **23:**387-397.

28. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition.** *Nature* 1992, **358:**86-89.
29. Taylor WR: **Multiple sequence threading: an analysis of alignment quality and stability.** *J Mol Biol* 1997, **269:**902-943.
30. Levitt M: **Competitive assessment of protein fold recognition and alignment accuracy.** *Proteins* 1997, **Suppl 1:**92-104.
31. Jones DT: **GenTHREADER: an efficient and reliable protein fold recognition method for genome sequences.** *J Mol Biol* 1999, **287:**797-815.
32. Wolf YI, Aravind L, Koonin EV: **Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange.** *Trends Genet* 1999, **15:**173-175.
33. Wolf YI, Brenner SE, Bash PA, Koonin EV: **Distribution of protein folds in the three superkingdoms of life.** *Genome Research* 1999, **9:**17-26.
34. Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M: **The dominance of the population by a selected few: power-law behavior applies to a wide variety of genomic properties.** *Genome Biology* 2002, **3(8):**0040.1-0040.7.
35. Qian J, Luscombe NM, Gerstein M: **Protein family and occurrence in genomes: power-law behavior and evolutionary model.** *J Mol Biol* 2001, **313:**673-681.
36. [http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY].
37. [http://bioinf.cs.ucl.ac.uk/psipred].
38. [http://www.sbg.bio.ic.ac.uk/3dgenomics].
39. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH – A Hierarchic Classification of Protein Domain Structures.** *Structure* 1997, **5:**1093-1108.
40. Peitsch MC: **PROMOD and SWISS-MODEL – Internet-based tools for automated comparative protein modeling.** *Biochem Soc Trans* 1996, **24:**274-279.
41. Peitsch MC, Wilkins MR, Tonella L, Sanchez JC, Appel RD, Hochstrasser DF: **Large scale protein modeling and integration with the SWISS-PROT and SWISS-2DPAGE databases: the example of Escherichia coli.** *Electrophoresis* 1997, **18:**498-501.
42. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A: **Comparative protein structure modeling of genes and genomes.** *Ann Rev Biophys Biomol Struct* 2000, **29:**291-325.
43. Sanchez R, Sali A: **Large-scale protein structure prediction of the *Saccharomyces cerevisiae* genome.** *Proc Natl Acad Sci USA* 1998, **95:**13597-13602.
44. Fisher D, Eisenberg D: **Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium.** *Proc Natl Acad Sci USA* 1997, **94:**11929-11934.
45. Muller A, MacCallum RM, Sternberg MJE: **Structural characterization of human proteome.** *Genome Research* 2002, **12:**1625-1641.
46. Iliopoulos I, Tsoka S, Andrade MA, Janssen P, Audit B, Tramontano A, Valencia A, Leroy C, Sander C, Ouzonis CA: **Genome sequences and great expectations.** *Genome Biology* 2001, **2:**0001. INTERACTIONS Epub 2000
47. Gerstein M, Levitt M: **A structural census of the current population of protein sequences.** *Proc Natl Acad Sci* 1997, **94:**11911-11916.
48. Apic G, Huber W, Teichmann SA: **Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination.** *J Struct Funct Genomics* 2003, **4:**67-78.
49. Teichmann SA, Park J, Chotia C: **Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements.** *Proc Natl Acad Sci USA* 1998, **95:**14658-14663.
50. Brinkman FSL, Blanchard JL, Cherkasov A, Av-Gay Y, Brunham RC, Fernandez RC, Finlay BB, Otto SP, Oullette BF, Keeling PJ, Rose AM, Hankock REW, Jones SJM: **Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria and the chloroplast.** *Genome Research* 2002, **12:**1159-1167.
51. Rzhetski A, Gomez SM: **Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome.** *Bioinformatics* 2001, **17:**988-996.
52. Yanai I, Camacho CJ, DeLisi C: **Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification.** *Phys Rev Lett* 2000, **85:**2641-2644.