BMC Bioinformatics

BioMed Central

Research article

Open Access

An approach to large scale identification of non-obvious structural similarities between proteins

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

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

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

* Corresponding author

Published: 17 May 2004

BMC Bioinformatics 2004, **5**:61

Received: 02 February 2004

Accepted: 17 May 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/61

Abstract

Background: A new sequence independent bioinformatics approach allowing genome-wide search for proteins with similar three dimensional structures has been developed. By utilizing the numerical output of the sequence threading it establishes putative non-obvious structural similarities between proteins. When applied to the testing set of proteins with known three dimensional structures the developed approach was able to recognize structurally similar proteins with high accuracy.

Results: The method has been developed to identify pathogenic proteins with low sequence identity and high structural similarity to host analogues. Such protein structure relationships would be hypothesized to arise through convergent evolution or through ancient horizontal gene transfer events, now undetectable using current sequence alignment techniques. The pathogen proteins, which could mimic or interfere with host activities, would represent candidate virulence factors.

The developed approach utilizes the numerical outputs from the sequence-structure threading. It identifies the potential structural similarity between a pair of proteins by correlating the threading scores of the corresponding two primary sequences against the library of the standard folds. This approach allowed up to 64% sensitivity and 99.9% specificity in distinguishing protein pairs with high structural similarity.

Conclusion: Preliminary results obtained by comparison of the genomes of *Homo sapiens* and several strains of *Chlamydia trachomatis* have demonstrated the potential usefulness of the method in the identification of bacterial proteins with known or potential roles in virulence.

Background

Pathogen proteins often manipulate host cellular functions by mimicking host activities. In some cases, mimicry is achieved through virulence factors that are direct homologues of host proteins that have been incorporated into the genome of the pathogen through horizontal gene transfer (HGT) [1,2]. In others, convergent evolution has produced new effectors that, although having no obvious amino acid sequence similarity to host factors, mimic them at the structural level [3].

Our recent research was conducted on the discovery of novel bacterial virulence factors through identification of pathogen genes that share a higher degree of sequence similarity to host genes than would otherwise be expected based on their phylogeny suggesting their likely acquisition by HGT [4]. To achieve this objective we developed novel bioinformatics tools to identify genes in complete bacterial genomes, which may be cases of HGT from eukaryotes. Based on a combined analysis of 136,195 genes from 36 bacterial and eukaryotic genome sequences [4], we identified no definitive cases of "recent" (defined as approximately since the divergence of mammals from other amniotes) HGT between bacteria and multicellular eukaryotes, including human genes recently sequenced in the Human Genome Project [5].

We have established that within the limitations of the dataset used, there was a notable lack of genes in the human and other genomes of multicellular eukaryotes that were highly similar to genes from any bacterial species examined. While this analysis did show that bacterial pathogens do contain "host-like" genes that may function as mimics, for the most part these appear to be primarily cases of either maintenance of an orthologous gene that was lost in other lineages, or ancient HGT [4].

It has yet to be established the extent to which convergent evolution events have played a role in the evolution of pathogens, involving alternative mechanisms than then HGT by which pathogens acquire host-mimicking virulence factors. Such genes and their corresponding proteins would usually have distinct sequences from those of the molecule they mimic, but would typically have evolved to imitate, at least in part, the shape and critical chemical groups on the surface of the functional homologues.

In the present work we describe our new efforts to identify pathogen genes, encoding proteins that have low sequence identity but potentially high structural similarity with host proteins. The hypothesis is that under selective pressure, pathogen genes have evolved to encode proteins that functionally mimic host proteins independently of significant primary sequence similarity. Such bacterial proteins mimicking host's functions can, therefore, be considered as potential virulence factors.

Results and Discussion

Genome – wide search for pathogen proteins that have low sequence similarity but significant structural resemblance with host proteins represent an opportunity for new insight into infectious agents, host cell biology and the mechanisms of pathogenesis. Theoretically, such a task should require a comprehensive modeling of three-dimensional (3D) structures of proteins, what is not yet achievable with useful accuracy nor is it computationally feasible on the scale of thousands of sequences.

The existing methods of fold recognition can broadly be divided into two types. In the first, the information is represented in linear form, called a profile, which is based on empirically derived scores for the expected occurrence of residues in a particular structure [6-13]. This type of approach is relatively rapid, but an unknown protein can only be characterized if it has reasonable sequence similarity with protein(s) with known structure. The second strategy is threading, which involves using pair potentials that score the likelihood of two residues being at a certain distance. This approach is based upon the assumption that nature has made certain economic decisions wherein countless different proteins fold into a limited number of shapes (estimated to approximate 4000 [14] and that nearly all natural protein structures can be described based upon these shapes. Threading attempts to assign folds for a protein of unknown structure by sampling it onto each member of a library of known folds using pseudo-energy as a measure of fit [15-20]. Threading approaches have been shown to make accurate predictions even in a "twilight zone" of <25% sequence identity, where sequence-based approaches normally fail [21]. Presently, however, neither profile-based nor threading-based approaches are capable of direct identification of structurally similar proteins from two different sources (such as distinct organismal protein datasets).

The method

In order to allow for structural comparisons to be made across genomes (where limited primary sequence identity is the case) we have adopted an indirect approach to identify potential protein structural similarities, based upon a broader utilization of the numerical outputs from derived from threading applications. For each raw sequence threaded onto the available 1893 model folds using the THREADER2 package [18] we derived Z scores representing the weighted sum of pair wise and solvation energies. Our hypothesis is that each sequence will have its own unique threading profile against a library of model folds and therefore, by correlating these "fingerprints" for two linear sequences it is possible to estimate the degree of potential 3D structural similarity. To support this argument, we have examined if complete sets of threading scores should indeed correlate for two structurally similar proteins. A dataset of 866,631 selected pair wise alignments of protein chains with "possible biologically interesting similarities" and covering whole 0–100% range of sequence similarity has been used for the study.

This dataset has been generated by an "all against all" comparison of protein chains in PDB by the authors of a combinatorial extension (CE) algorithm [22,23]. A CE approach has been shown to accurately identify the similarity of protein structures using a dynamic programming algorithm determining the RMSD and a significance score "Z" for optimal structural alignment. The publicly available CE dataset includes the alignments of proteins (with length difference no more then 30%) corresponding to Z

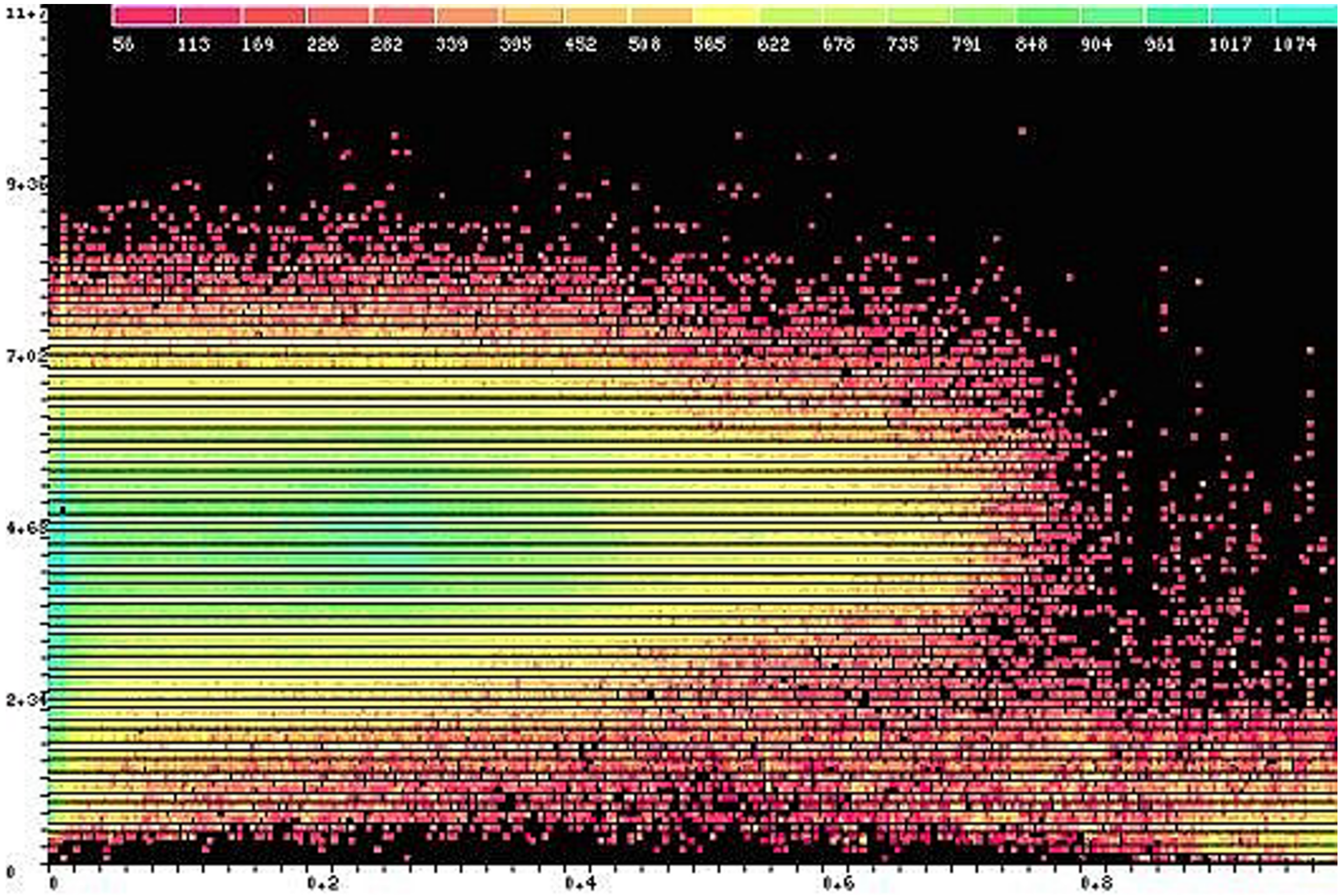

Figure 1
RMSD values of pair wise alignments of representative protein chains versus the corresponding parameters R.

value above 3.5 threshold. Notably, 783,841 or more then 90% of the sequence alignments in the set have sequence identity below 20%. This circumstance makes it very favorable to use it for testing the threading-based approach. We have downloaded CE dataset from [24] and processed the protein chains with a THREADER2 package to produce their threading profiles against 1893 library folds.

For every aligned pair of proteins from the dataset we have estimated a correlation between their threading scores. If a set of threading scores against a standard library is indeed sequence specific, then for proteins with known structures we should be able to observe a defined relationship between coefficients of threading scores correlation and parameters of protein 3D structural similarity.

On a Figure 1 (color-coded according to the plots density) the estimated squared coefficients of correlations between threading scores are plotted against RMSD values for 846,534 selected CE – pair wise alignments of proteins with known structure. As it can readily be seen, the meaningful correlations between threading scores (squared correlation coefficients R^2 above 0.7) correspond to higher quality structural alignments with lower RMSD.

A RMSD value of 2Å is normally considered as a threshold value distinguishing pairs of structurally similar proteins. Thus, as it was anticipated, the thresholds R^2~0.7 and RMSD ~2Å clearly separate two most populated areas which correspond to pairs of proteins with low and high structural similarity ($R^2 < 0.7$; RMSD > 2Å) and ($R^2 > 0.7$; RMSD < 2Å).

A quantitative assessment of protein structural similarity is rather an ambiguous task; thus some structural alignments produce the alignment score Z along with excessively high RMSD parameters, or instead there are few well-superimposed protein pairs in the CE dataset with rather low Z. Therefore, to further support the previous observations, we have introduced an additional

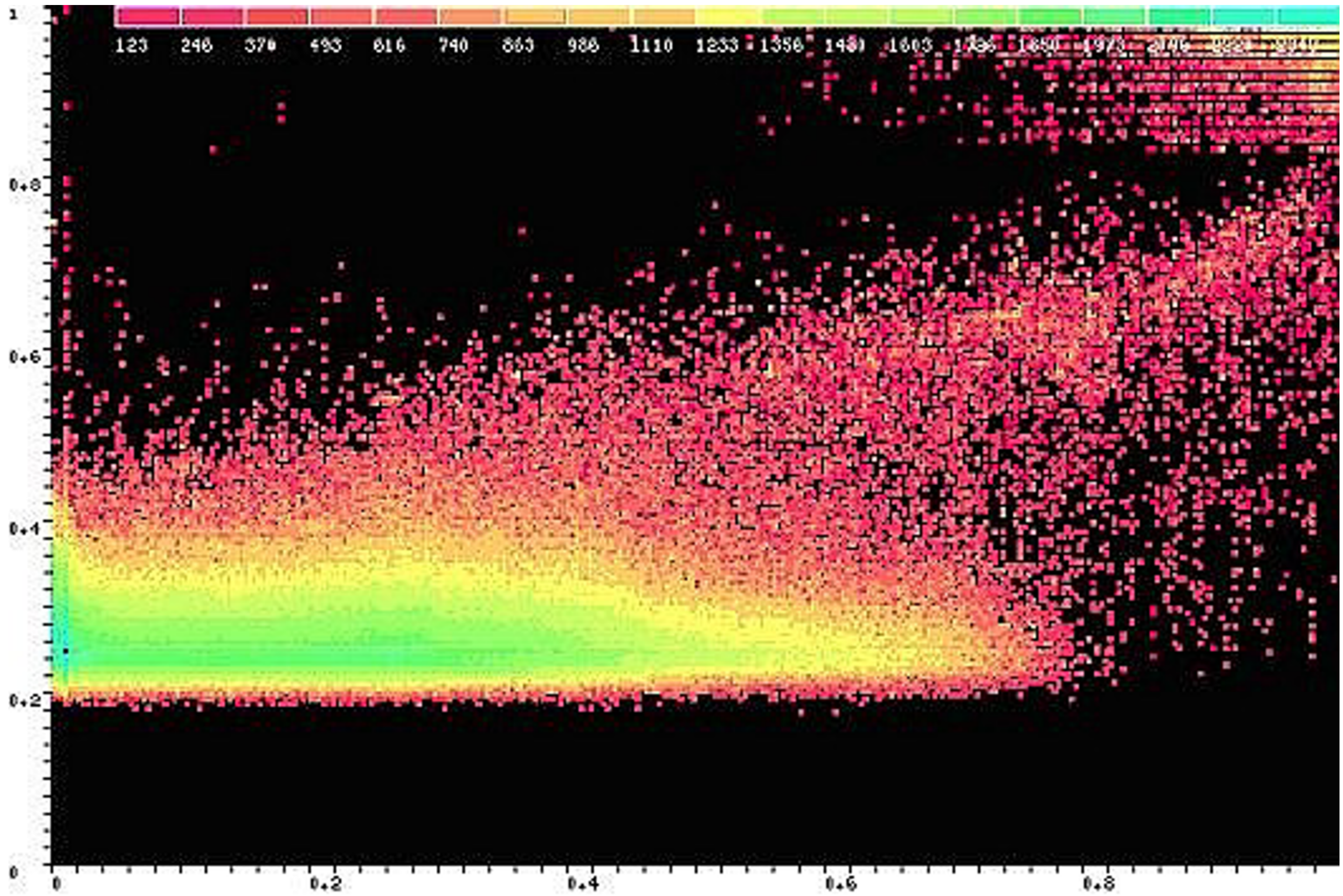

Figure 2
Structure similarity scores (SSS) values for pair wise alignments of representative protein chains versus the corresponding parameters R.

parameter (we called structure similarity score – SSS) calculated as the structural alignment score Z (ranging from 0 to 10) divided by the sum of the corresponding RMSD value and a factor of 10: SSS = Z/(10+RMSD). Thus, SSS is normalized to [0–1] range, where 1 corresponds to a pair of completely similar protein structures superimposed with Z = 10 and RMSD = 0.

In Figure 2, the SSS parameters calculated for 846,534 alignments are plotted against the coefficients of correlations between threading scores. The plots on a graph can be conventionally divided into three major areas of low (SSS < 0.4), medium (0.4 < SSS < 0.6) and high (SSS > 0.6) structural similarity. The graph indicates that the vast majority of protein pairs with correlated threading scores ($R^2 > 0.7$) fail into areas of medium and high structural similarity.

To assess a distinguishing power of R^2 cutoff we have estimated the sensitivity and specificity of the approach in recognizing pairs of superimposed proteins with RMSD < 2Å. The calculated sensitivity and specificity parameters are plotted on a Figure 3 for entire [0–1] range of the R^2.

According to the estimated numbers (true negative predictions (TN): 790630, true positive predictions (TP): 18060, false negative predictions (FN): 28623, false positive predictions (FP): 9221) the approach allows to achieve 99% specificity TN/(FP+TN) in distinguishing protein pairs with RMSD < 2Å when threshold $R^2 = 0.68$ is used. The corresponding sensitivity TP/(TP+FN) of the method reaches 38.7%. The predictive value positive TP/(TP+FP) and the predictive value negative TN/(TN+FN) are 66.2% and 96.5% respectively. Similar evaluation of protein pairs with medium and high degree of similarity (or SSS > 0.4) can also relate 99% specificity level with $R^2 = 0.68$ threshold, while the corresponding sensitivity could be

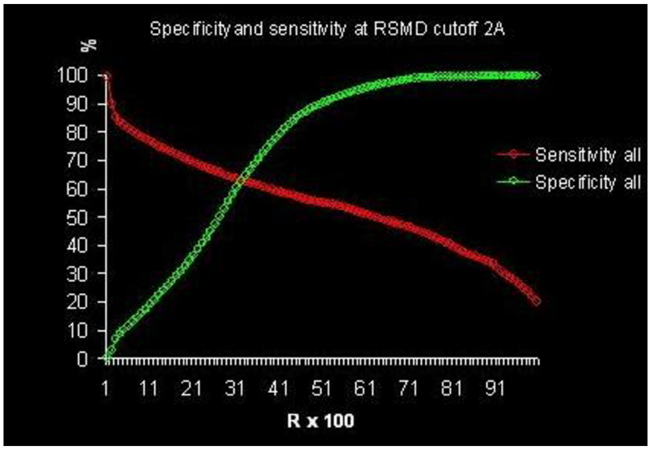

Figure 3
Sensitivity and selectivity of the developed approach in distinguishing meaningful (RMSD < 2A) protein structure alignments

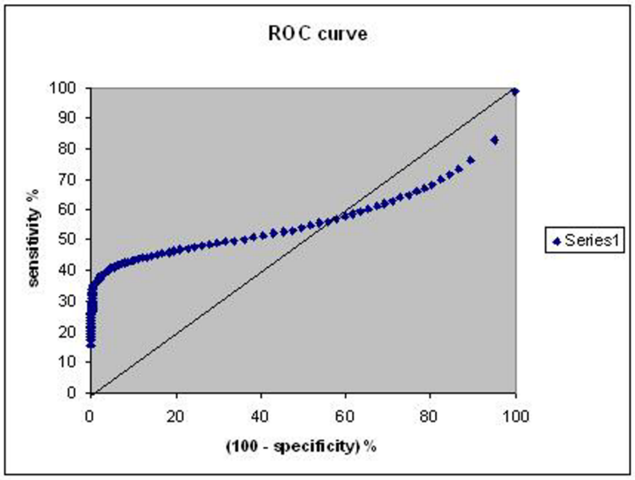

Figure 5
ROC plot for describing the ability of the approach to distinguish superimposed protein pairs with RMSD below 2A threshold.

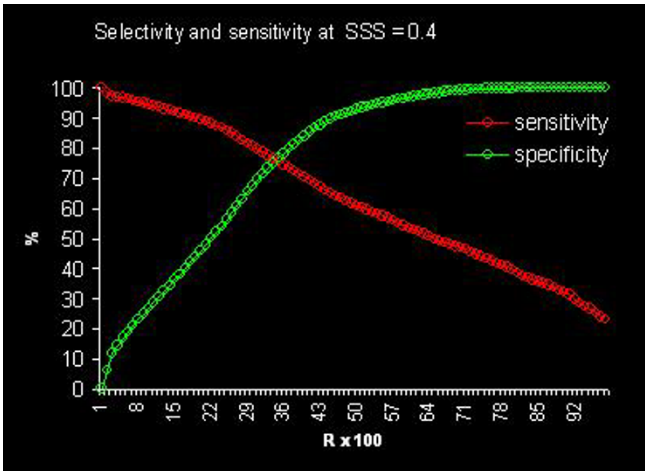

Figure 4
Sensitivity and selectivity of the developed approach in distinguishing meaningful (SSS > 0.4) protein structure alignments

estimated as 47% (Figure 4). The sensitivity of the developed approach improves with increase of the structural similarity criteria. For highly similar proteins (alignments producing SSS > 0.6) from the CE dataset sensitivity reaches 92%.

Thus, it is feasible to conclude, that the coefficient of correlation between 1893 threading scores for 2 raw sequences can adequately indicate putative three-dimensional similarity between the corresponding protein structures. When the threshold $R^2 = 0.68$ is used the general accuracy of the developed approach (TP+TN)/(TP+TN+FP+FN) is 95.1%.

The training set has also been used to estimate the receiver operating characteristic plots (ROC). The ROC plot is obtained by plotting all sensitivity values (true positive fraction) on *y* axis against their equivalent (1-specificity) values (false positive fraction) for all available thresholds on the *x* axis. The area under the ROC curve (AUC) is usually taken as an important single measure of overall accuracy of approach that is not dependent on the particular threshold.

We have plotted the specificity versus selectivity parameters of the developed approach on 0–1.0 range of the R threshold with the step of 0.01. The calculation has been conducted on the training set of low sequence similarity alignments. The resulting ROC curve is presented on a Figure 5. As it can be seen, the resulting ROC covers more then a half of the chart's area reflecting the fact, that the approach gives better then a chance performance. Therefore, the developed approach has a valid general predictive power and, thus, provides an opportunity to investigate potential structural similarity between two proteins without actual modeling of their structures.

In addition, it should also be outlined, that the developed approach does not impose requirement for high quality assignment of sequence to particular fold(s) by threading.

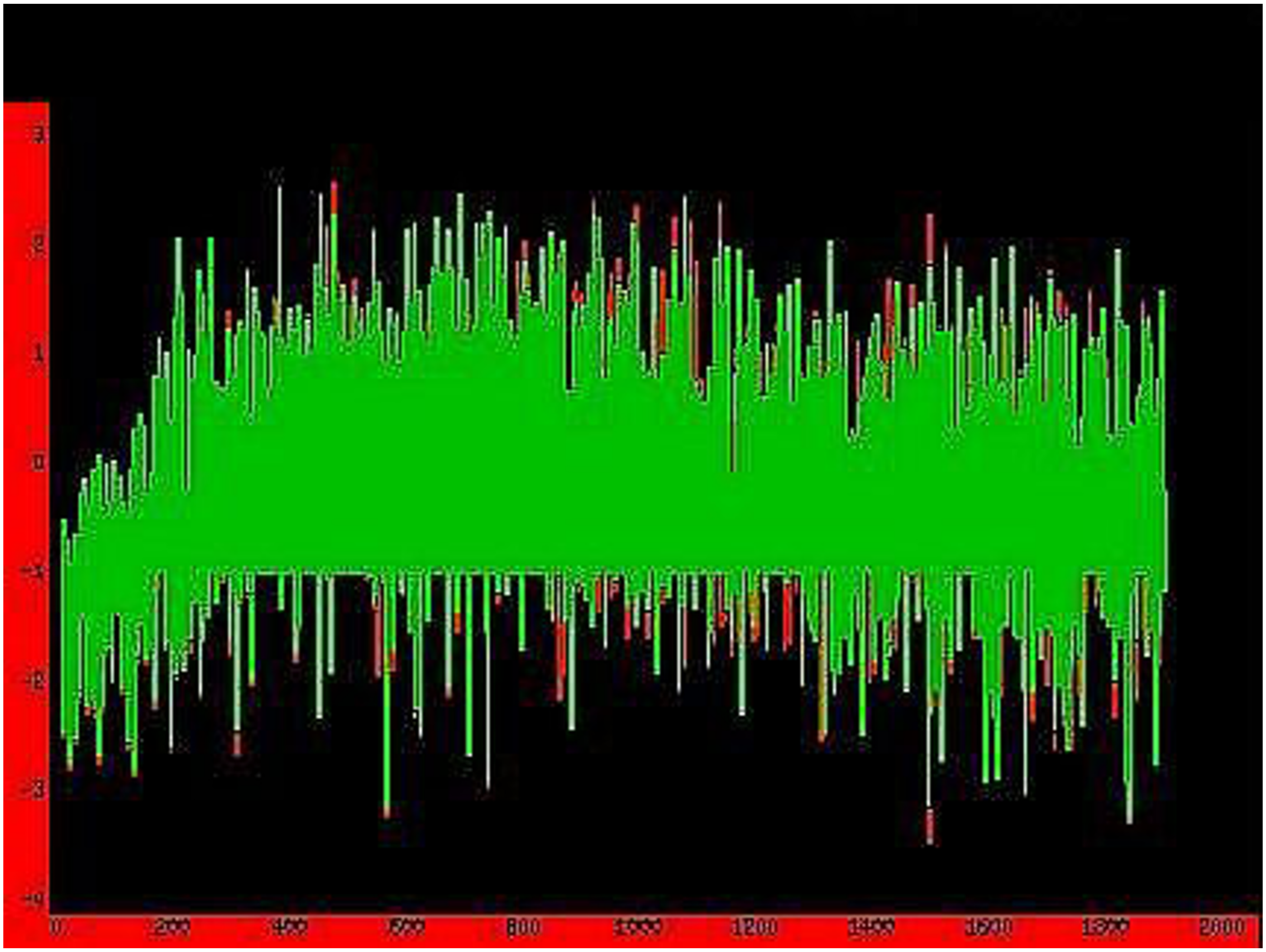

Figure 6
Z values of pseudo energies of threading of protein chains 2BUI:C and 2SAK through 1893 CATH model folds.

In fact, one or even both proteins may not be assigned by THREADER2 to any known folds, but their resulting threading profiles can be used to reveal the existing structural similarity.

To illustrate this point, the threading scores for sequences of staphylokinases 1BUI:C and 2SAK are presented on a Figure 6 as histograms. As it can readily be seen, neither of these sequences could be assigned to certain fold, as a minimal threshold for reliable fold recognition is 3.5 (displayed as a horizontal bar on a Figure 6). At the same time, structures of proteins 1BUI:C and 2SAK can be superimposed with RMSD = 1.3Å and therefore are very similar.

In spite the fact, that neither 1BUI:C or 2SAK could not be assigned to certain fold, their threading possess a correlation coefficient, R, of 0.83 clearly indicating high structural similarity between 1BUI:C and 2SAK.

The developed approach has further been applied to a random set of proteins with known structure. Using CE programs have been superimposed 1800 randomly selected proteins from Protein Databank (PDB) on "all against all" basis to generate 3,240,000 redundant structural alignments (small fraction of alignments could not be produced). In the same time, all 1700 corresponding sequences have been processed with the THREADER2 package and the generated threading scores datasets have been cross-correlated to produce 3,222,731 correlation coefficients.

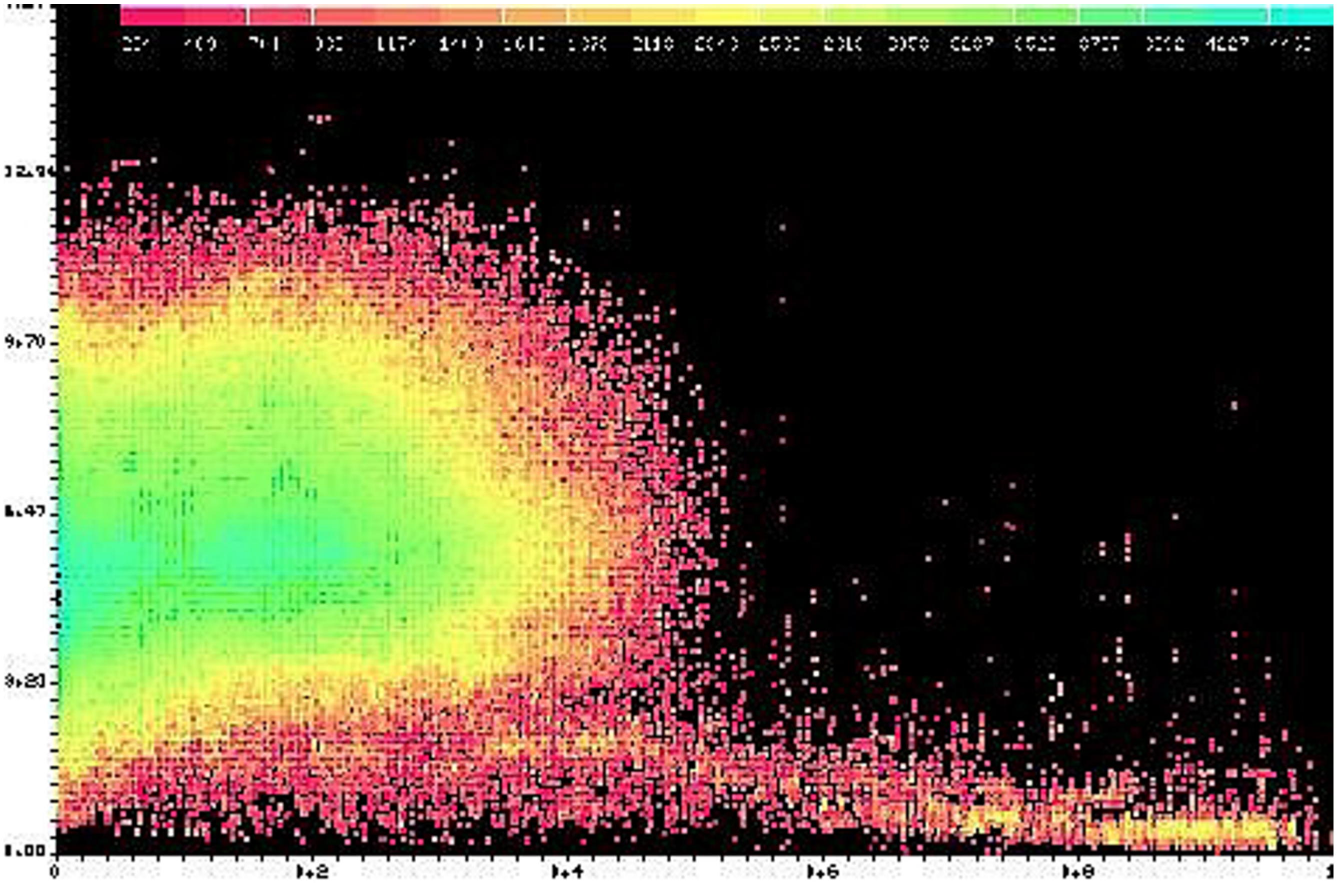

Figure 7
RMSD values of pair wise alignments of randomly selected protein chains versus the corresponding parameters R.

The generated RMSD values for random pair wise structural alignments are plotted against the corresponding R^2 parameters on a Figure 7. The shape of the graph resembles previously obtained well-like "RMSD *vs* R^2" dependence for selected CE dataset. Major areas of true positive and true negative observations for the random dataset can be separated by R^2 threshold value of 0.5. Similar cutoff level can be observed on a Figure 8. representing a relationship between SSS and R^2 parameters for 3,222,731 random structural alignments under consideration. Areas of protein alignments with low (SSS < 0.4) and medium (0.4 < SSS < 0.6) structural similarity are clearly separated by the R^2~0.5 threshold. There are very few highly similar proteins (with SSS > 0.6) have been observed within a random dataset.

Thus, when applied to the random dataset of protein with known 3D structures, the developed approach (operating R^2 = 0.68 cutoff value) has a sensitivity of 50% for the alignments with RMSD < 2Å (651 FP, 23,432 FN, 23,765 TP and 3,174,891 TN) and 64% if SSS = 0.4 is used as a criteria of similar structures. The specificity in both cases remains at 99.9% level. On a random dataset of aligned protein structures the R^2 = 0.68 threshold value identifies protein pairs with SSS > 0.6 with the sensitivity of 97 %. When SSS value reaches >= 0.7, both sensitivity and specificity of the developed approach stay around 99% (Figure 9).

The results obtained on selected and random datasets of proteins with known structures allow concluding, that the developed approach is enable to identify with reasonable accuracy proteins with medium and high levels of structural similarity. To address the question whether the developed approach is sequence dependent, we have estimated its sensitivity and selectivity in distinguishing structurally similar proteins (with SSS > 0.6) at 0 – 20%, 20 – 40% and 40 – 60 % levels of sequence identity. The corresponding results presented on a Figure 10 illustrate that the predictive power of the developed approach varies

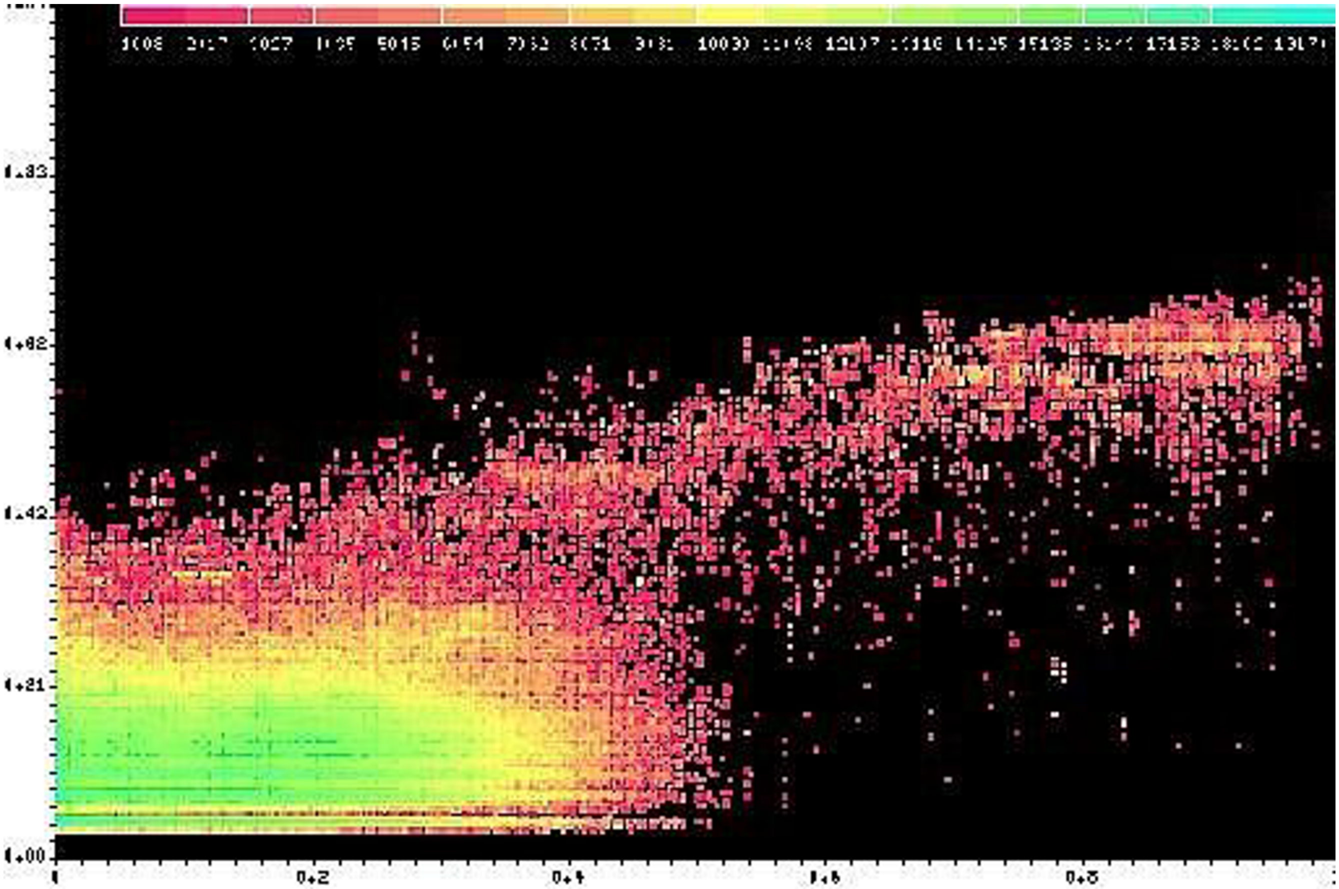

Figure 8
Structure similarity scores (SSS) for pair wise alignments of randomly selected protein chains versus the corresponding parameters R.

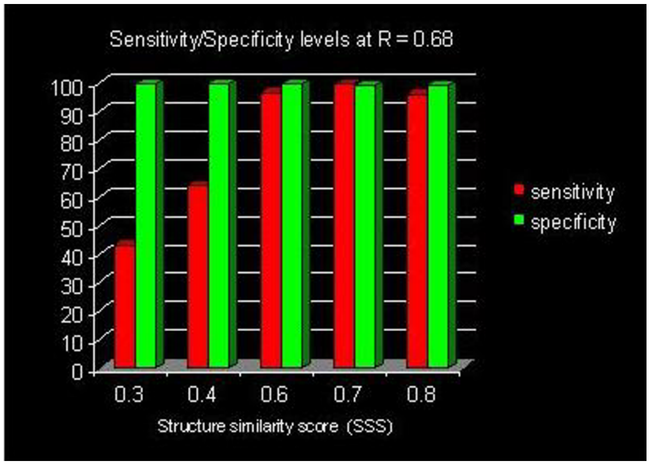

Figure 9
Distinguishing power of the developed approach at different levels of protein structural similarity.

upon different levels of protein sequence similarity. Apparently, the threading profiles of structurally similar proteins become less resembling as their sequence identity drops. This observation is somewhat contradictory since threading is considered to be independent from sequence identity information.

Considering the specific need of the approach to recognize structurally similar proteins with low sequence identity and taking into account a large number of protein alignments with SSS > 0.6 (negative counts) in the investigated CE database, we have compiled an additional training set. The set only included protein pairs with low sequence identity (>20%) and had equal representation of proteins alignments with SSS below and above 0.6 threshold.

Thus, 244 pair wise sequence alignments with low similarity have been extracted from the CE set. This comprised

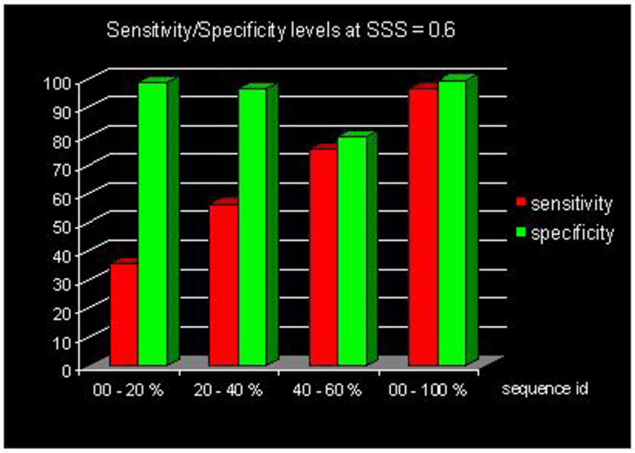

Figure 10
Distinguishing power of the developed approach at different levels of protein sequence identity.

all 122 alignments with SSS above 0.6 threshold and 122 randomly selected ones with SSS < 0.6. The use of R = 0.68 threshold has yielded the following predictions: TP: 44, TN: 122, FP: 0, FN: 78. These correspond to 36 % sensitivity TP/(TP+FN), 100% specificity TN/(FP+TN), 100 % predictive value positive TP/(TP+FP) and 61% predictive value negative TN/(TN+FN).

The estimated numbers allow to conclude that the developed approach utilizing quantitative outputs of threading possesses useful sensitivity and specificity in recognizing proteins with low sequence identity (below 20%) and high structural similarity (SSS > 0.6). This makes it suitable for genome scaled studies.

Identification of structural homologues between H. sapien and C. trachomatis proteins

The developed approach has been used to test the hypothesis that pathogenicity of microorganisms can dependent on number of proteins in their genomes mimicking structures of host analogues. In order to evaluate this assumption we have examined human structural homologues among proteins from *Chlamydia trachomatis* organism.

First of all, currently available proteomes of *Homo sapien* and *Chlamydia trachomatis* strain D (30585 and 894 entries respectively) have been processed with the THREADER2.

The generated threading profiles of human and Chlamydia proteins then have then been compared using the developed approach, aiming to produce 30585 * 894 = 27,342,990 parameters R. All proteins from two genomes have also been compared on "all-against-all" manner for sequence similarity to identify those pairs with no sequence homology but similar threading profiles (high R scores).

Overall, we were able to produce 25,649,384 pair wise comparisons of threading profiles of human and Chlamydia proteins (some short sequences have been rejected by the threading). Out of these, only 636 protein alignments produced sequence similarity value above 20%. Among 636 pairs of similar proteins, 86 (or 13.5 percent) have R parameter above 0.68 threshold. The fraction of structurally similar proteins among those with low sequence identity (<20%) is 5.5 percent: 1,409,914 out of 25,648,748 alignments. This is 1.5 folds higher then the proportion of potentially similar proteins found in the training set of the CE sequence alignments (27,281 (FP+TP) out of the total of 846,534, or 3.2 percent). This is an interesting finding, considering that the CE set of "possible biologically interesting similarities" is already heavily enriched with structurally similar proteins. On another hand, these finding demonstrate the CE – training set we have used for the threshold estimation, can be used as rather adequate representation of bacterial genome.

We have also compared the estimated positive count ratio of 5.5 percent with the corresponding number for randomly sampled PDB – chains alignments. In this case we have found much more significant difference of 8 folds: 5.5 versus 0.7 percent (the later can be calculated as a sum of 23,765 true positive and 651 false positive predictions for 3,222,731 random sequence alignments). Such rather elevated occurrence of *Chlamydia* proteins with potential structural similarity with human counterparts may illustrate the importance of the factors of convergent evolution.

From the pool of human and *Chlamydia* protein alignments we have identified 40 pairs of single domain proteins with no detected sequence similarity (at E = 0.00001) and the highest R scores which are presented in Table 1. Multi-domain proteins have not been considered in the study to simplify the exercise.

If our assumptions about the conversion nature of bacterial virulent mimicry were correct, then we would expect some chlamydial virulence factors to be found in the table. Evaluation of the table indicates that a large part of the presented *Chlamydia* proteins has indeed been already previously identified as potential virulence factors. Thus, five *Chlamydia trachomatis* putative out-membrane proteins (F, H, I, E and A types corresponding to the table entries 2, 15, 17, 25 and 27 respectively) have been detected by the developed approach as top virulence candidates. An important role of these proteins in *Chlamydia* antigenic polymorphism has been previously underlined in [25].

Among other proteins from the Table 1, gi3328486gbAAC67681.1 (entry 26) is known as a *Chlamydia* virulence factor responsible for pathogen survival in Ca – deficient environment; protein gi3328822gbAAC67993.1 (entry 29) is a heat shock protein – one of potential *Chlamydia* virulence factors; protein YopC (entry 35) is involved in secretion of pathogenic genetic material.

Chlamydia is "energy parasite" [25] importing ATP from host cells. Thus, it came to no surprise, that ATP transport protein gi3328511gbAAC67704.1 (kinase fold) has also been identified as potential virulence factor (entry 10). Two other potential Chlamydia virulence factors perform transport functions: transpeptitase gi3329155gbAAC68296.1 and protein translocase gi3329345gbAAC68468.1 (entries 12 and 3 respectively).

The majority of other *Chlamydia* proteins presented in the Table 10 can be divided into proteases (entries 23, 31, 37 proteases and metalloprotease 34) and proteins related to DNA transcription (entries 6,8 – transcription proteins, 20 – nucleotide transport, 19 – DNA isomerase, 7,9,40 – ribonucleases, riboreductase, 4,13,18,32 – Gly, isoLeu, Ala and Leu tRNA synthetases).

Thus, the preliminary results allow to conclude, that out of 33 top *Chlamydia* hit with assigned functions presented in the Table, up to 11 proteins either have been previously identified as pathogenic virulence factors or possess define virulent characteristics. The predictive value positive (TP/TP+FP) of the approach above the separating threshold $R^2 > 0.9$ is as high as 92.33 % (for 846534 predictions 1052 FP, 34170 FN, 12513 TP and 798799 TN). Therefore, it is expected that the most if not all of 40 Chlamydia proteins presented in Table 1 can be reliably considered as structurally highly similar to their human counterparts.

To assess the actual ability of the developed approach to enrich for proteins attributable to virulence we need to evaluate how many virulence factors can be found by chance in random pool of 33 *Chlamydia* proteins. This is not a trivial task as it requires the knowledge of the total number of virulence genes in *Chlamydia trachomatis* genome. At the moment the exact virulence content of the *Chlamydia trachomatis* genome remains unknown, so we attempted its evaluation using available literature data.

Thus, an indirect justification for this number can be derived from the results of the work of Fields and all 1986 who have experimentally identified 81 genes of *Salmonella typhimurium* responsible for its survival in professional phagocytes [26,27]. Taking similar to the previous guess that the real number of virulent factors is as twice as high, the hypothetical virulence content of *Salmonella typhimurium* genome can be contemplated around 3.6% (162 out of 4451 genes).

Thus, by the analogy, we may expect that about 4 percents of an average bacterial proteome can be assigned to virulence associated proteins. Therefore, there is roughly 4 percent probability of random finding of virulence factors in arbitrary pool of bacterial genes.

Based on that estimate, we may expect that among 33 annotated *Chlamydia trachomatis* proteins presented in Table 1 one or two potential virulence factors could be identified by chance. The fact, that there are about 11 of them demonstrates that the developed approach is indeed capable of 6 – 10 folds enriching for bacterial virulence factors.

Conclusions

Virulence factors candidates from bacteria and viruses having low sequence and high 3D similarity with host proteins can be readily identified by the developed approach. Its sensitivity can future be improved as efforts to complete and organize the inventory of model folds are successful [14] (as it has been mentioned the THREADER2 takes into account only known 2000 model folds that covers only about 50% of 4000 folds predicted).

The developed approach is not only applicable for identification of potential novel virulence factors in pathogen genomes, but may be broadly used for all kinds of protein similarity studies.

Methods

Sequence similarity search has been conducted with BLAST program [28] with E value of 0.00001.

Threading has been carried out by the THREADER2 [18] program with default parameters. The CATH v2.0 (November 2000) fold assembly has been used as a library of standard folds.

Human proteome has been downloaded from ENSEMBL database; the proteome of *Chlamydia trachomatis serovar D* – from NCBI site.

Authors' contributions

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

Additional material

Additional File 1

Protein pairs from H. sapien and C. trachomatis with low sequence identity and high structural similarity.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-61-S1.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. Davies J: **Origins and evolution of antibiotics eresistance.** *Microbiologia* 1996, **12:**9-16.
2. Wolf YI, Aravind L, Koonin EV: **Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange.** *Trends Genet* 1999, **15:**173-175.
3. Stebbins CE, Galan JE: **Maintenance of an unfolded polypeptide by a cognate chaperone in bacterial type III secretion.** *Nature* 2001, **412:**70-81.
4. Brinkman FSL, Blanchard JL, Cherkasov A, Av-Gay , Brunham RC, Fernandez RC, Finlay BB, Otto SP, Oullette BF, Keeling PJ, Hancock REW, Rose AM, 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.
5. International Human Genome Sequencing Consortium: *Nature* 2001, **409:**860.
6. 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.
7. 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.
8. Bates A, Jackson RM, Sternberg MJE: *Genomes, Molecular Biology and Drug Discovery* Academic Press, London; 1996.
9. Russell RB, Copley RR, Barton GJ: **Protein fold recognition by mapping predicted secondary structure.** *J Molec Biol* 1996, **259:**349-365.
10. Rice DW, Eisenberg G: **A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence.** *J Molec Biol* 1997, **267:**1026-1038.
11. Rost B, Schneider R, Sander C: **Protein fold recognition by prediction – based threading.** *J Molec Biol* 1997, **270:**471-480.
12. Defay TR, Cohen FE: **Multiple sequence information for threading algorithms.** *J Mol Biol* 1996, **262:**314-323.
13. 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.
14. Machalek AZ: *ASM News* 2001, **67:**441-447.
15. 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.
16. Bryant SH, Altschul SF: **Statistics of sequence-structure threading.** *Curr Opin Struct Biol* 1995, **5:**236-244.
17. Murzin AG, Bateman A: **Distant homology recognition using structural classification of proteins.** *Proteins (Suppl)* 1997, **1:**105-112.
18. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition.** *Nature* 1992, **358:**86-89.
19. Jones DT, Miller RT, Thornton JM: **Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing.** *Proteins* 1995, **23:**387-397.
20. Taylor WR: **Multiple sequence threading: an analysis of alignment quality and stability.** *J Molec Biol* 1997, **269:**902-943.
21. Levitt M: **Competitive assessment of protein fold recognition and alignment accuracy.** *Proteins (Suppl)* 1997, **1:**92-104.
22. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Engineering* 1998, **11:**739-747.
23. Shindyalov IN, Bourne PE: **A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) alrorithm.** *Nucleic Acids Research* 2001, **29:**228-229.
24. **CE Database** [http://cl.sdsc.edu/ce.html]
25. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW: **Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis.** *Science* 1998, **282:**754-759.
26. Fields PI, Swanson RV, Haidaris CG, Heffron F: **Mutants of Salmonella typhimurium that cannot survive within the macrophage are avirulent.** *Proc Natl Acad Sci* 1986, **86:**5189-5193.
27. Gahring LC, Heffron F, Finlay BB, Falkow S: **Invasion and Replication of Salmonella typhimurium in Animal Cells.** *Infection and Immunity* 1990, **58:**443-448.
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.