OPEN ACCESS Freely available online

PLoS one

PSPP: A Protein Structure Prediction Pipeline for Computing Clusters

Michael S. Lee[1,2,3], Rajkumar Bondugula[1], Valmik Desai[1], Nela Zavaljevski[1], In-Chul Yeh[1], Anders Wallqvist[1], Jaques Reifman[1]*

1 Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland, United States of America, **2** Computational and Information Sciences Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland, United States of America, **3** Department of Cell Biology and Biochemistry, U.S. Army Medical Research Institute of Infectious Diseases, Fort Detrick, Maryland, United States of America

Abstract

Background: Protein structures are critical for understanding the mechanisms of biological systems and, subsequently, for drug and vaccine design. Unfortunately, protein sequence data exceed structural data by a factor of more than 200 to 1. This gap can be partially filled by using computational protein structure prediction. While structure prediction Web servers are a notable option, they often restrict the number of sequence queries and/or provide a limited set of prediction methodologies. Therefore, we present a standalone protein structure prediction software package suitable for high-throughput structural genomic applications that performs all three classes of prediction methodologies: comparative modeling, fold recognition, and *ab initio*. This software can be deployed on a user's own high-performance computing cluster.

Methodology/Principal Findings: The pipeline consists of a Perl core that integrates more than 20 individual software packages and databases, most of which are freely available from other research laboratories. The query protein sequences are first divided into domains either by domain boundary recognition or Bayesian statistics. The structures of the individual domains are then predicted using template-based modeling or *ab initio* modeling. The predicted models are scored with a statistical potential and an all-atom force field. The top-scoring *ab initio* models are annotated by structural comparison against the Structural Classification of Proteins (SCOP) fold database. Furthermore, secondary structure, solvent accessibility, transmembrane helices, and structural disorder are predicted. The results are generated in text, tab-delimited, and hypertext markup language (HTML) formats. So far, the pipeline has been used to study viral and bacterial proteomes.

Conclusions: The standalone pipeline that we introduce here, unlike protein structure prediction Web servers, allows users to devote their own computing assets to process a potentially unlimited number of queries as well as perform resource-intensive *ab initio* structure prediction.

Citation: Lee MS, Bondugula R, Desai V, Zavaljevski N, Yeh I-C, et al. (2009) PSPP: A Protein Structure Prediction Pipeline for Computing Clusters. PLoS ONE 4(7): e6254. doi:10.1371/journal.pone.0006254

Editor: Eshel Ben-Jacob, Tel Aviv University, Israel

Received March 24, 2009; **Accepted** June 5, 2009; **Published** July 16, 2009

This is an open-access article distributed under the terms of the Creative Commons Public Domain declaration which stipulates that, once placed in the public domain, this work may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose.

Funding: This work was sponsored by the U.S. Department of Defense High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes initiative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jaques.reifman@us.army.mil

Introduction

Three-dimensional (3-D) protein structures are critical for the understanding of molecular mechanisms of living systems. Traditionally, X-ray crystallography or nuclear magnetic resonance methods are used to determine the structures of proteins experimentally. In the post-genomic era, where many new complete genomes are available every year and the number of sequences total in the millions, it is impossible to rely on experimental methods alone for structural characterization. Therefore, computational prediction of protein structures is an essential complement. Predicted protein structures help researchers in several ways. First, fold prediction is an important tool for the functional annotation of proteins at the genomic scale [1–3]. Moreover, fold and structure predictions can be used to infer binding interfaces [4], potential binding partners [5], and catalytic active sites [6]. In addition, *in silico* drug screening can be performed on close homologues of proteins with known structures [7,8].

The quality of protein structure predictions is directly correlated to the similarity of a query sequence to known protein structures [9]. Procedurally, as shown in Fig. 1, query protein sequences are first divided into manageable chunks. Optimally, domain boundaries are used, but these are often experimentally unknown and must be inferred computationally. Each domain sequence is then compared for similarity against a database of known protein structures, i.e., the Protein Data Bank (PDB) [10], which, to date, consists of over 50,000 entries. If no matches can be detected, fold recognition is instantiated, whereby various characteristics of the domain sequence are predicted, including secondary structure,

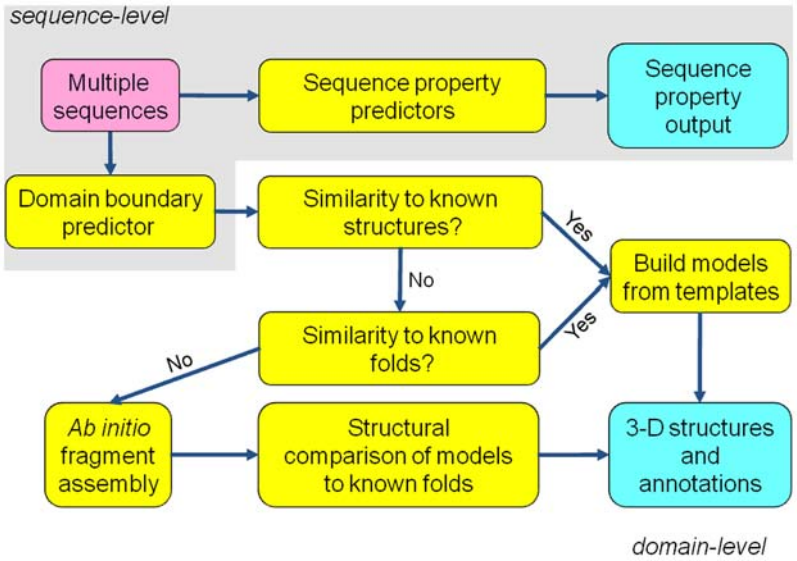

Figure 1. Workflow for the protein structure prediction pipeline given a single query sequence.
doi:10.1371/journal.pone.0006254.g001

solvent accessibility, and sequence-specific substitution matrices. These properties are pairwise aligned against the properties of several thousand known protein folds (e.g., the Structural Classification of Protein (SCOP) database [11]). Finally, if no matches are made in this search, the 3-D atomic structure of the protein domain must be built *ab initio*, i.e., the structure must be assembled using energy functions and filters to guide the packing of residues and multi-residue fragments. Because this is a combinatorial process that does not guarantee a globally optimal solution, thousands of models must be proposed. A high-resolution energy function is then applied to ascertain which models might be closest to the native protein [12,13].

Because computational protein structure prediction is a complex, multi-step process, it requires many diverse tools, often developed by multiple research laboratories, and the expertise to use them. Many Web servers are available for predicting the structure of a given protein sequence [14–16]. However, depending on publicly available Web servers is not practical for several reasons. First, they are a shared resource, and one may be limited to a small number of sequence submissions in a given amount of time. Conversely, servers that pre-process entire genomes of protein sequences may be limited to offering only comparative modeling results (e.g., ModBase [17]). Second, the confidentiality of data cannot be guaranteed, i.e., the submitted data and predicted results are often publicly viewable on the server. Third, in some cases, one cannot be assured that the servers are properly maintained and use the most recent databases. Fourth, Web servers that require heavy computational processing to perform *ab initio* fragment assembly (e.g., Robetta [15]) may have query limits or long queues. Finally, servers are often discontinued when the grant that establishes them terminates [18].

Given the sheer amount of genomic sequence data, a standalone pipeline is necessary to process thousands of sequences at a time. Moreover, access to a pipeline's source code allows end users to add or replace components as new techniques, software, or databases become available. Standalone protein structure prediction requires the integration of several tools, which have been generously disseminated by various independent research laboratories. If software is to be distributed over multiple nodes (or cores) in a cluster environment, it is often cost effective to rely on freely available software. While open-source software is desirable, it is not possible in some cases. Fortunately, the x86-based Linux operating system is a common standard among computational laboratories, and pre-compiled binaries tend to perform reliably.

In this work, we introduce a Perl-based software pipeline that integrates multiple free software packages to predict protein structures and structural properties [19]. It is composed of sequence-level and domain-level modules (Fig. 1). Beyond what has been described previously [19], the sequence-level module predicts protein domain boundaries and properties, such as secondary structure, solvent accessibility, transmembrane helices, and structural disorder. The domain-level module produces 3-D atomic protein models and structural annotations via three strategies: homology, fold recognition, and *ab initio* fragment assembly. In addition, multiple sequences can be handled simultaneously via parallelization over numerous processing cores with a message passing interface (MPI)-based job scheduling tool.

Methods

The pipeline consists of Perl software modules, C-shell scripts, freely available third-party software (albeit many with license agreements), and an in-house implementation of an MPI job scheduler, Pipeman [2]. The main Perl program, *seq_router.pl*, processes command line parameters and calls sequence analysis, domain boundary detection, and domain-processing modules for individual protein sequences. This program can be run on a single processing core or can run on multiple cores on a single computing node using the multithreading capabilities of PSI-BLAST (sequence searching) and PROSPECT II (fold recognition/threading). A second program, *mpi_seq_router.pl*, performs multiple sequence processing in parallel. This program reads a multiple-sequence FASTA file, writes individually labeled sequences into separate FASTA files, and then dispatches individual *seq_router.pl* jobs via Pipeman. Each component is explained in more detail below. Table 1 lists the third-party software and databases that were integrated into this package.

Table 1. Third-party and in-house software and databases.

Software or Database	Originating Laboratory	Function	Article	Website
BLAST/PSI-BLAST	NCBI	Sequence search	[20]	[52]
CE	Shindyalov & Bourne	Structural similarity search	[44]	[53]
CHARMM[1]	Karplus	Molecular minimization and scoring	[37]	[54]
DISPrO, SSPrO, and ACCPro	Baldi	Disorder, secondary structure, and solvent accessibility prediction	[27]	[55]
Jackal	Honig	Homology modeler	[30]	[56]
MMTSB	Brooks and Feig	CHARMM front-end/structural analysis	[36]	[57]
MUPRED	Xu	Secondary structure and solvent accessibility prediction	[23]	Bundled
NR/PDBAA	NCBI	Sequence databases	[29]	[58]
PDB	RCSB	Database of biological macromolecular structures	[10]	[59]
PROSPECT II	ORNL	Fold recognition/threading	[34]	[60]
PSIPRED	Jones	Secondary structure prediction	[21]	[61]
Rosetta	Baker	*Ab initio* folder	[41]	[62]
SCOP/ASTRAL	Chothia and Murzin	Database of protein folds	[11]	[63]
SCWRL3	Dunbrack	Side chain placement	[42]	[64]
TMHMM	Viklund	Transmembrane helix prediction	[25]	[65]
DFIRE-AA	In house	Atomic scoring function	[35]	Bundled
FIEFDom	In house	Domain boundary prediction	[28]	Bundled
Pipeman	In house	MPI job distribution tool	[2]	Bundled
PROSPECT II templates	In house	Templates for SCOP 1.73 folds	This work	Bundled
PSPP	In house	Core software for the pipeline	This work	Bundled

[1]Optional (requires paid academic or commercial license).
doi:10.1371/journal.pone.0006254.t001

Protein-level predictions

Before proteins are delineated into separate domains, several properties can be predicted for each query protein, including secondary structure, solvent accessibility, disorder, and the presence of transmembrane helices. Most of these programs require a position-specific substitution matrix (PSSM; a.k.a. ''profile'') generated by PSI-BLAST [20] using the *nr* database. For efficiency, we generate the PSI-BLAST profile once and use it for all protein-level predictions. This may slightly degrade the accuracy of certain individual programs, since they are often tuned with specific PSI-BLAST options.

Knowledge of the secondary structure of proteins is helpful in protein classification, understanding protein folding, tertiary structure prediction, and increasing the accuracy of multiple sequence alignments. Although a finer categorization is possible, protein secondary structures are generally classified into three states: helix, strand, and coil. We incorporated three secondary structure prediction tools into the pipeline: PSIPRED [21], SSPro [22], and MUPRED [23].

Solvent accessibility prediction helps in the understanding of protein tertiary structure, antigenic determinants, protein stability analysis, protein structure classification, and protein interaction analysis. We include ACCPro [22] and MUPRED [24] for solvent accessibility prediction in our pipeline. Both programs predict relative solvent accessibility and can be used for classifying residues as exposed or buried using a threshold value.

In addition, transmembrane proteins are an important class of proteins crucial to all multi-cellular organisms. They play a vital role in signal transduction, ion transport, and other significant functions. TMHMM [25] is incorporated in the pipeline to designate different segments of a given protein sequence as intracellular, extracellular, or transmembrane.

Moreover, intrinsically disordered proteins are often responsible for molecular recognition, molecular assembly, protein modification, and entropic chain activities in organisms [26]. In the pipeline, disordered regions in proteins are predicted using DISPro [27]. For each residue, its profile along with the predicted secondary structure (using SSPro) and predicted solvent accessibility (using ACCPro) are input to an artificial neural network that outputs a residue level index from 0 to 9 (where 0 = fully ordered and 9 = fully disordered) [27].

Finally, the query protein sequences are delineated into separate domains using FIEFDom [28], a novel domain prediction method that we have developed. Briefly, FIEFDom performs a PSI-BLAST search of the full protein sequence against a database of known multiple domain structures. A consensus identification of domain boundary regions is accumulated from profile-sequence matches with known structures. If FIEFDom predicts one or more domains longer than 250 residues, which is often a result of failed domain recognition, the user is provided with an option to use Bayesian statistics to break the sequence into smaller blocks.

Domain-level predictions

After delineation of the query sequence into domains, each domain sequence is routed to homology modeling, fold recognition, and *ab initio* fragment assembly (Fig. 1). If homology modeling is successful, i.e., at least one template is found above a user-specified sequence similarity threshold, the domain module proceeds directly to all-atom scoring. Otherwise, fold recognition is initiated. A fold confidence above a user-specified threshold will trigger model building. After the generation of template-based models by homology and/or fold recognition, all-atom scoring on the models is performed as described below. Finally, if no models

are built, *ab initio* fragment assembly (if it has been selected as an option) is instantiated.

In homology modeling, PSI-BLAST [20] is used to find sequences of PDB structures that align to the query sequence. First, the domain sequence is compared against the *nr* database to generate a profile using three iterations of PSI-BLAST. The sequence and profile are then compared against all sequences in the PDB (i.e., the *pdbaa* database [29]), and the top hits are ranked by sequence identity (i.e., the number of exactly matched residues in the alignment divided by the length of the query sequence). Finally, the program Nest, which is part of the Jackal suite of protein modeling programs from the Honig laboratory [30], is used to build homology models from the most promising alignments. Regions of the model that do not align to the template are treated as loops, which must be predicted "*ab initio*." For this reason, loop regions are often the largest sources of structural errors in homology models. At present, the pipeline only supports one template per comparative model.

The next prediction at the domain level is based on fold recognition. While many good programs exist for this function [31–33], few are freely available for download. One of the best options is the free, but closed-source, program PROSPECT II from the Oak Ridge National Laboratory [34]. Fold recognition involves profile-profile alignment of the query sequence to a template database of known folds. In addition to profile, other features, such as secondary structure and solvent accessibility, are evaluated in the alignment procedure. We built PROSPECT II-compatible templates from the SCOP 1.73 database of protein structures [11] (95% sequence similarity filter: $N \sim 15{,}000$). The pipeline performs three PROSPECT II passes. In the first pass, a search is performed against all of the SCOP templates. In the second pass, the top-ranked templates (based on a support vector machine-estimated score) are threaded using a more definitive, but costlier, Z-score procedure. In the final pass, templates with the top-ranked Z-scores are threaded using the pairwise interaction option. Finally, Nest is used to build models from the alignments that pass a certain threshold of fold confidence. Fold confidence is computed via an analytical fit to the data points in the table of the original PROSPECT II article [34]. Even with solid confidence scores, 3-D structural models generated by fold recognition will often have quality issues because of slight errors in the template alignment in addition to the loop region problem discussed earlier.

The template-based models resulting from the comparative modeling and fold recognition modules are scored with two procedures. The first scoring program is an in-house implementation of the DFIRE-AA all-atom statistical potential [35]. This potential is derived from an analysis of the inter-atomic distances between pairs of atom types in a large set of known protein structures. The second scoring module first minimizes the model with MMTSB [36] and CHARMM [37] using the PARAM22 [38] all-atom force field and a distance-scaled electrostatic potential with a dielectric constant equal to 4 [39]. The minimized structure is then scored using the PARAM22 force field with the GBMV2 implicit solvent potential [13,40] and a surface area-based non-polar term. The CHARMM-based scoring module is only available with an academic CHARMM (or commercial CHARMm) license and thus is an optional, albeit valuable, component. The DFIRE-AA and PARAM22/GBMV2 (GB22) scores are output as the raw score divided by the number of residues in the model. This formula is a simple, though imperfect, way to compensate for different-sized models.

If a domain level sequence is too distant from known folds, template-based modeling is no longer a viable option. In this case, the pipeline calls *ab initio* folding, which uses the popular RosettaAbInitio program from the Baker laboratory [41]. The RosettaAbInitio procedure begins by constructing a library of three- and nine-residue fragments from PDB structures with similar sequences, secondary structures, and profiles as stretches of the query sequence. Rosetta assembles these fragments into full-sized protein backbone models using various energy terms and filters. Because the internal united-residue energy function is often unable to discriminate near-native models, the models must be ranked via a post-process. While newer versions of Rosetta offer side chain packing, minimization, and scoring, our pipeline uses its own post-processing algorithm. First, all-atom models are generated by building the side chains onto each backbone model using SCWRL3 [42]. Next, the all-atom models are scored by our in-house implementation of DFIRE-AA [35]. Finally, the top DFIRE-scoring models are minimized and scored using CHARMM, as described above.

Rosetta has been successfully used in remote fold recognition and annotation for genome- scale applications [1,43]. To classify and annotate the folds of the models that result from the Rosetta code, the structures of the top few models are compared against ASTRAL PDB-style coordinates of the SCOP 1.73 fold library [11] using CE [44]. The top CE matches ranked by Z-score are listed in the output along with the SCOP annotations. If CHARMM is not present on the computer system, the top models as scored by DFIRE are selected instead. The pipeline can parallelize this module over multiple processing cores using the MPI compilation of Rosetta and the Pipeman job distribution tool for the post-processing steps.

Output formats

As the software evolved, several output options in different formats were developed. The first format is a text-based human-readable output. Hypertext markup language (HTML) output is also available and incorporates a query-template sequence alignment view as well as DFIRE and GB22 scores. Web pages are organized by a hyperlinked directory tree. In addition, tab-delimited output, containing much of the same information as in the HTML output, is generated so that users can import annotation results into spreadsheet applications.

Results and Discussion

In this section, we demonstrate an application of our pipeline for large-scale protein structure prediction. Then, we show the value of the scoring schemes implemented in our pipeline. Next, we discuss the performance of the pipeline in the CASP7 competition. In addition, we discuss ongoing biological applications using the pipeline. Finally, we discuss computational time and scaling issues.

To demonstrate the use of the pipeline for large-scale processing, we performed structural annotation of the *variola* (smallpox) virus genome [45–48], which consists of 197 protein-coding genes. A summary of the results for this run is presented in Table 2. For roughly 10% of the proteins, homology models were produced that might be suitable for drug design (i.e., >50% sequence similarity to a known protein structure). We present the results for three *variola* proteins, which are all labeled "hypothetical," to show the various outcomes of the prediction workflow (Fig. 1). Figure 2 shows the HTML output for NP_042212.1, one of the *variola* proteins for which a reasonable comparative model could be built. The top model has a 44.4% sequence identity to the known structure of mouse protein guanylate kinase (PDB ID: 1LVG). The GB22 and DFIRE energy scores of the top model were also the highest in rank versus the other comparative models

Table 2. Summary statistics for template-based structure prediction of the proteins encoded by the *variola* (smallpox) genome using the pipeline.

	No. of Proteins	No. of Domains
Total queries	197	355
Sequence similarity to a PDB structure		
>90%	12 (6%)	20 (6%)
Between 50% and 90%	8 (4%)	14 (4%)
Between 30% and 50%	11 (6%)	15 (4%)
Fold recognition (<30% sequence similarity to PDB)		
>90% confidence	32 (16%)	39 (11%)
Between 50% and 90% confidence	21 (11%)	29 (8%)
<50% confidence	113 (57%)	238 (67%)

doi:10.1371/journal.pone.0006254.t002

generated. Note that in the HTML output, the index column is hyperlinked to the comparative model in PDB format, if a model is predicted.

For the protein domains in the *variola* genome that did not have homologous PDB structures, fold recognition was automatically called. An example of a positive fold result is shown in Fig. 3. *Variola* protein NP_042071.1 can be aligned to two SCOP fold templates with confidence scores >50%. Thus, for these two templates, structural models were produced.

Perhaps typical of viral genomes, less than half of the proteins encoded by the *variola* genome could be structurally characterized using either homology modeling or fold recognition. Consequent-

Comparative Modeling (PDB-blast)

#	PDB	Hom	N(match)	Ident	Pos	Gap	N(model)	DFIRE	GB22	Description
1	1LVG_A	44.4%	149	67	101	0	151	-66.49	-39.70	Chain A, Crystal Structure Of Mouse Guanylate Kinase ...
2	1Z6G_A	31.1%	141	47	81	3	151	-61.98	-38.69	Chain A, Crystal Structure Of Guanylate Kinase From P...
3	2QOR_A	29.8%	141	45	83	3	-	-	-	Chain A, Crystal Structure Of Plasmodium Vivax Guanyl...
4	2AN9_A	26.5%	149	40	73	4	-	-	-	Chain A, Crystal Structure Of Oligomeric E.Coli Guany...
5	1EX6_A	25.8%	143	39	79	2	-	-	-	Chain A, Crystal Structure Of Unliganded Form Of Guan...
6	1S96_A	25.2%	149	38	71	4	-	-	-	Chain A, The 2.0 A X-Ray Structure Of Guanylate Kinas...
7	1S4Q_A	25.2%	141	38	70	5	-	-	-	Chain A, Crystal Structure Of Guanylate Kinase From M...
8	1ZNW_A	25.2%	141	38	70	5	-	-	-	Chain A, Crystal Structure Of Unliganded Form Of Myco...
9	2J41_A	24.5%	148	37	77	4	-	-	-	Chain A, Crystal Structure Of Staphylococcus Aureus G...
10	1Z8F_A	24.5%	141	37	69	5	-	-	-	Chain A, Guanylate Kinase Double Mutant A58c, T157c F...

Comparative Modeling Alignments

```
                1       2       3       4       5       6
        ....:....0....:....0....:....0....:....0....:....0....:....0
Query Start MEREGVDYHYVNREAIWKGIATGNFLEHTEFLGNIYGTSKTAVNTAAINNRICVMDLNID
1LVG   47   -EEDGKDYYFVTREMMQRDIAAGDFIEHAEFSGNLYGTSKEAVRAVQAMNRICVLDVDLQ
1Z6G   66   -EKEGVDYYFIDKTIFEDKLKNEDFLEYDNYANNFYGTLKSEYDKAKEQNKICLFEMNIN

                7       8       9       10      11      12
        ....:....0....:....0....:....0....:....0....:....0....:....0
Query       GVRSLKNTYLMPYSVYIKPTSLKMVETKLRRRNTEADDKIHRRAMLAKTDMDEANEAGLF
1LVG        GVRSIKKTDLCPIYIFVQPPSLDVLEQRLRLRNTETEESLAKRLAAARTDMESSKEPGLF
1Z6G        GVKQLKKSIKNALYIFIKPPSTDVLLSRLLTRNTENQEQIQKRMEQLNIELHEAN-LLNF

                13      14      15
        ....:....0....:....0....:....0.
Query       NTTIIEDNVNLAYSKLIQILQDRIRMYFNTN                          Stop
1LVG        DLVIINDDLDKAYATLKQALSEEIKKAQGT-                          195
1Z6G        NLSIINDDLTLTYQQLKNYL----------                            205
```

Figure 2. Screenshot of the pipeline-rendered comparative modeling results in hypertext markup language (HTML) format for *variola* protein NP_042212.1. Color-coding for the amino acid letters is as follows: red, acidic; blue, basic; green, polar; and black, apolar.
doi:10.1371/journal.pone.0006254.g002

Fold Recognition (PROSPECT II)

#	Template	Conf %	Hom %	SCOP ID	N(model)	DFIRE	GB22	Classification
1	d2flux1	100.0	18.8	b.68.11.1	179	-87.79	-20.38	Kelch motif Kelch-like ECH-associated protein 1, KEAP1
2	d1k3ia3	100.0	13.3	b.69.1.1	179	-81.89	-18.34	Galactose oxidase, central domain Galactose oxidase, central domain
3	d1kb0a2	47.9	7.9	b.70.1.1	-	-	-	Quinoprotein alcohol dehydrogenase-like Quinoprotein alcohol dehydrogenase, N-terminal domain
4	d1mdah_	45.3	4.9	b.69.2.1	-	-	-	Methylamine dehydrogenase, H-chain Methylamine dehydrogenase, H-chain
5	d1sqja1	44.2	6.2	b.69.13.1	-	-	-	Oligoxyloglucan reducing end-specific cellobiohydrolase Oligoxyloglucan reducing end-specific cellobiohydrolase
6	d2ad6a1	44.2	6.9	b.70.1.1	-	-	-	Quinoprotein alcohol dehydrogenase-like Methanol dehydrogenase, heavy chain
7	d1sqja2	41.3	7.2	b.69.13.1	-	-	-	Oligoxyloglucan reducing end-specific cellobiohydrolase Oligoxyloglucan reducing end-specific cellobiohydrolase
8	d1ijqa1	39.8	6.3	b.68.5.1	-	-	-	YWTD domain Low density lipoprotein (LDL) receptor
9	d1n1ta2	34.6	8.4	b.68.1.1	-	-	-	Sialidases (neuraminidases) Trypanosoma sialidase
10	d1l0qa2	33.6	9.9	b.69.2.3	-	-	-	YVTN repeat Surface layer protein

Fold Recognition Alignments

```
                 1         2         3         4         5         6
                 ....:....0....:....0....:....0....:....0....:....0....:....0
QueryStart       MNNEIHNNAIAVNYISNNWIPIPPMNSPKLYASGIPANNKLYVVGGLPNPTSVERWFHGN
2flu   385 -DNTDSSALDCYNPMTNQWSPCAPMSVPRNRIGVGVIDGHIYAVGGSHGHNSVERYEPER
1k3i   197 -SPGGITLTSSWDPSTGIVSDRTVTVTKHDMFPGISMNGQIVVTGG-NDAKKTSLYDSSS
1kb0   361 ASGYDKHGKPIGIDGSKPQDAVPGPYGAHNWHPMSFNPQTLVYLPAQNVPVNLMDDKKWE
1mda    91 -KGKRTDYVEVFDPVPIADIELPDAPRFSPRVHIIGNSACLLFFLFGSS--AAAGLSVPG
1sqj    86 YVGDEWAAFYVSEDRGQSFTIYENDMGRNNGERLAVNPSNEVWMGTRTEG--IWKSSDRA
2ad6   369 TRMD--HKGTNICP---------SAMGFHNQGVDSYDPESRTLYAGLNHICMDWE-PFMLP
1sqj   459 LLSGIGDISMKHDDLTKPQKMFGAPQFSN--LDSIDAPNVVVRAGSSGHCARGAYATDGG
1ijq   415 VASNRIRMICSTQLYDTVI---RDIQAPD---GLAVDHSNIYWTDS--VLGTVSVADTKG
1n1t   117 FNKQHRDATISWGKPVSLKPLFPAEFDGILTKEFVGASNLVYPVQIAGRVFTKIMYEDDG
1l0q    39 ISPDGTNDVSIIDTATNNVIATVPAGS---SPQGVAVGKQVYVTNMASS--TLSVIDTTS
```

Figure 3. Screenshot of the pipeline-rendered fold recognition results in HTML format for *variola* protein NP_042071.1. This picture truncates the alignment after the first block of 60 residues.
doi:10.1371/journal.pone.0006254.g003

ly, comprehensive analysis of this genome requires that a large number of the domains be processed via the computationally intensive *ab initio* method. For example, fold recognition on *variola* protein NP_042054.1 did not identify any template with more than 50% confidence. Figure 4 shows the partial HTML output of an *ab initio* run on this protein. Z-scores, computed via CE, that are >5 are considered to be matches at the superfamily level [44]. The top-scoring *ab initio* model (energy rank = 1) structurally aligns with a membrane-bound chloride channel and several all-α-helical protein folds. Sequence property predictions for this protein sequence are shown in Fig. 5 (truncated at 120 residues for display purposes). The output includes predictions of transmembrane helices, disorder, secondary structure, and solvent accessibility aligned with the query sequence. Most notably, a membrane-bound fold can tentatively be ruled out because TMHMM did not predict any regions of transmembrane helices.

The use of energetic scoring functions, such as DFIRE-AA and GB22, in template-based modeling improves the chances of detecting the most accurate model [49]. As a test case, we predicted 224 all-atom comparative models for the α-spectrin SH3 domain (sequence derived from PDB ID: 1SHG) using only homology modeling. As shown in Fig. 6a, the highest sequence homology hits correctly produced the most accurate Nest-built models. However, suppose that there were no templates with $>90\%$ sequence identity. In this thought experiment, percent identity appears to be a poor determinant of model accuracy as measured by the root mean squared deviation (RMSD) of the C_α-trace between the native X-ray structure and the model (C_α RMSD). On the other hand, the DFIRE-AA and GB22 functions (Fig. 6, b and c) show scoring funnels [13] for this query sequence, i.e., as the score improves, so does the model accuracy. Therefore, if only lower sequence

Domain	Energy Rank	Z-score rank	SCOP ID	Z-score	SCOP class	SCOP family	Description
NP_042054.1_134	1	1	d1otsa_	5.06	f.20.1.1	Clc chloride channel	Clc chloride channel
NP_042054.1_134	1	2	d1gwcal	5.01	a.45.1.1	Glutathione S-transferase (GST), C-terminal domain	Class tau GST
NP_042054.1_134	1	3	d1gwcal	5.01	a.45.1.1	Glutathione S-transferase (GST), C-terminal domain	Class tau GST
NP_042054.1_134	1	4	d1gw5a_	5.01	a.118.1.10	Clathrin adaptor core protein	Adaptin alpha C subunit N-terminal fragment
NP_042054.1_134	1	5	d1u6gc_	4.98	a.118.1.2	HEAT repeat	Cullin-associated NEDD8-dissociated protein 1 (Tip120)
NP_042054.1_134	1	6	d1paqa_	4.90	a.118.1.14	MIF4G domain-like	Translation initiation factor eIF-2b epsilon
NP_042054.1_134	1	7	d1tu7al	4.81	a.45.1.1	Glutathione S-transferase (GST), C-terminal domain	Class pi GST
NP_042054.1_134	1	8	d1tu7al	4.81	a.45.1.1	Glutathione S-transferase (GST), C-terminal domain	Class pi GST
NP_042054.1_134	1	9	d1glqal	4.81	a.45.1.1	Glutathione S-transferase (GST), C-terminal domain	Class pi GST
NP_042054.1_134	1	10	d1kpla_	4.79	f.20.1.1	Clc chloride channel	Clc chloride channel
NP_042054.1_134	2	1	d1u6gc_	4.90	a.118.1.2	HEAT repeat	Cullin-associated NEDD8-dissociated protein 1 (Tip120)
NP_042054.1_134	2	2	d1u6gc_	4.90	a.118.1.2	HEAT repeat	Cullin-associated NEDD8-dissociated protein 1 (Tip120)

Figure 4. Screenshot of the pipeline-rendered SCOP annotations derived from the *ab initio* results in HTML format for *variola* protein NP_042054.1. The full output (not shown) includes a total of 5 models (ranked by GB22 energy) and the top 10 SCOP matches for each model.
doi:10.1371/journal.pone.0006254.g004

Figure 5. Screenshot of the pipeline-rendered HTML output showing predicted sequence properties for *variola* protein NP_042054.1. i, intracellular; D, disordered; H, helix; E, strand; B, buried.
doi:10.1371/journal.pone.0006254.g005

identity templates were available, these two scoring functions could aid in selecting the most accurate model.

We participated in the CASP7 experiment in 2006 using an older version of the pipeline and submitted 408 3-D models for 92 targets. Our overall performance ranked in the middle of the 130 participating groups. The noteworthy successes were that one of our homology model predictions and one of our *ab initio* predictions ranked no. 1 in the "Top 1" model category, as measured by the global distance test [50]. We attribute these two cases to the use of the GB22 and DFIRE-AA scoring functions. Our modest performance could be attributed to our lack of advanced loop modeling capabilities or alignment optimization [49] and our reliance on single-template models. Also, at the time, we did not have the domain recognition algorithm FIEFDom to break larger query sequences into more manageable chunks. Regardless, the performance of our standalone pipeline will only improve as new downloadable technologies are shared by research laboratories with the larger community.

Users of the pipeline are currently applying the *ab initio* component to deduce the function of several proteins encoded in virus genomes, including the VP24 protein of Ebola and Marburg viruses [51]. In addition, they are using the pipeline in proteomic surveys of the *Escherichia coli* and *Yersinia pestis* genomes to determine which protein structures can be built by homology, such that protein-protein interactions can be modeled. In addition, the pipeline is helping researchers infer the functions of proteins that, up to now, have been labeled as "hypothetical."

It is worth discussing the computational effort of the homology modeling and fold recognition run on the *variola* genome. While running in parallel on 64 Xeon 3.0-GHz cores, the pipeline required, on average, nearly 4 CPU-hours per domain when utilizing a shared file system. In contrast, repeating the same calculations using the hard drives of the local nodes instead averaged a much more reasonable ~1 CPU-hour per domain. We believe that most of the performance degradation on the shared file system can be attributed to PROSPECT II, which uses frequent I/O operations of opening and closing ~15,000 template and temporary output files for each domain. One solution we are considering is switching to the newly available open-source fold recognition program OpenProspect [38] and modifying it so that it processes large blocks of templates at a time.

In comparison, an *ab initio* run scales well up to 32 processing cores (results not shown). While the Rosetta-MPI component scales almost linearly up to 64 cores, too many simultaneous instances of the structural CE-based similarity search over a shared file system leads to asymptotic limits in speedup. Similar to the situation with PROSPECT II templates, copying the SCOP fold database to the hard drives of the local computing nodes improves parallel performance, albeit with a trade-off of some wall-clock time for copying the database files from shared to local file systems.

Conclusions

We have introduced a standalone, Perl-based pipeline for protein structure prediction that integrates freely downloadable software components from various academic and government research laboratories. Unlike Web services, which either limit the number of query sequences for processing or perform only a limited subset of prediction techniques, our pipeline allows researchers to harness the power of their own computational resources to perform protein structure predictions at the genomic level. Salient features of our structure prediction software include all-atom scoring, structural annotation of *de novo* models, annotations and sequence alignments in HTML format, and an MPI-parallel framework for large-scale studies.

Availability and Requirements

- Project name: Protein Structure Prediction Pipeline
- Project download page: http://www.bhsai.org/structure2.html
- Operating system: Linux
- Programming languages: Perl5, tcsh, and C++

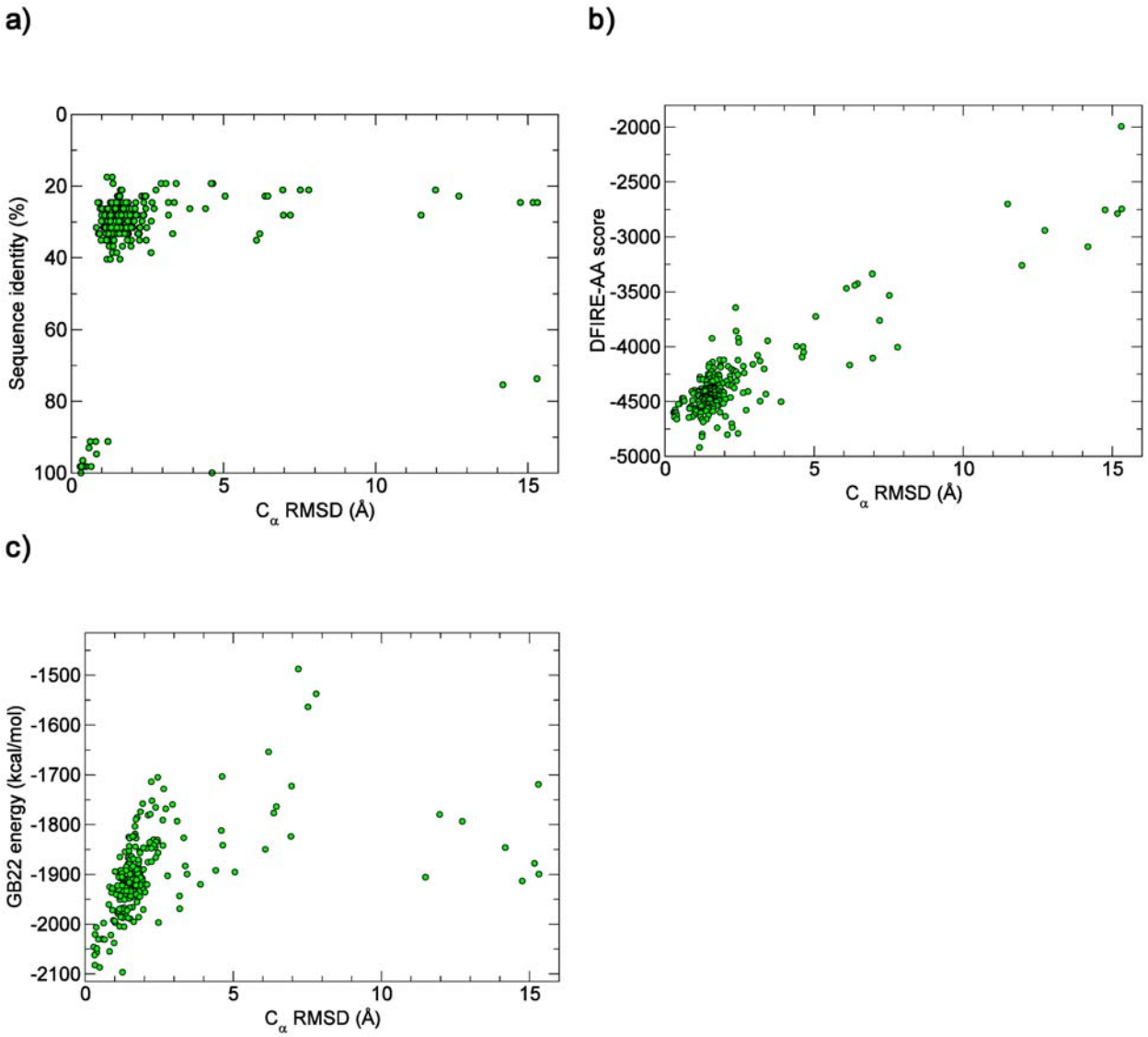

Figure 6. Comparison of various homology model quality criteria against structure accuracy as measured by C_α, root mean squared deviation (RMSD) for the α-spectrin SH3 domain (PDB ID: 1SHG): sequence identity (a), DFIRE-AA score (b), and GB22 energy (c).
doi:10.1371/journal.pone.0006254.g006

Acknowledgments

The authors express their gratitude to the developers of the numerous open-source and freely available programs implemented in the pipeline. We also thank Drs. C. Yu and P. Wilson for their Pipeman software and Dr. M. Olson for helpful discussions. This work was sponsored by the U.S. Department of Defense High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes Initiative.

Disclaimer. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

Author Contributions

Conceived and designed the experiments: MSL. Performed the experiments: MSL. Analyzed the data: MSL. Contributed reagents/materials/analysis tools: MSL RB VD NZ ICY. Wrote the paper: MSL RB AW. Oversaw the project: JR AW. Revised the manuscript: JR AW.

References

1. Bonneau R, Baliga NS, Deutsch EW, Shannon P, Hood L (2004) Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1. Genome Biol 5: R52.
2. Yu C, Zavaljevski N, Desai V, Johnson S, Stevens FJ, et al. (2008) The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. BMC Bioinformatics 9: 52.
3. Yu C, Zavaljevski N, Desai V, Reifman J (2009) Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases. Proteins 74: 449–460.
4. Konc J, Janezic D (2007) Protein-protein binding-sites prediction by protein surface structure conservation. J Chem Inf Model 47: 940–944.
5. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ (2007) SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res 35: D580–589.
6. Torrance JW, Bartlett GJ, Porter CT, Thornton JM (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 347: 565–581.

7. Kairys V, Fernandes MX, Gilson MK (2006) Screening drug-like compounds by docking to homology models: a systematic study. J Chem Inf Model 46: 365–379.
8. Jiang X, Kumar K, Hu X, Wallqvist A, Reifman J (2008) DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0. Chem Cent J 2: 18.
9. Venclovas C, Zemla A, Fidelis K, Moult J (2003) Assessment of progress over the CASP experiments. Proteins 53 Suppl 6: 585–595.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
11. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
12. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309: 1868–1871.
13. Lee MS, Olson MA (2007) Assessment of Detection and Refinement Strategies for de novo Protein Structures Using Force Field and Statistical Potentials. J Chemical Theory and Computation 3: 312–324.
14. Guo JT, Ellrott K, Chung WJ, Xu D, Passovets S, et al. (2004) PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. Nucleic Acids Res 32: W522–525.
15. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, et al. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53 Suppl 6: 524–533.
16. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) Structure prediction meta server. Bioinformatics 17: 750–751.
17. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, et al. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34: D291–295.
18. Veretnik S, Fink JL, Bourne PE (2008) Computational biology resources lack persistence and usability. PLoS Comput Biol 4: e1000136.
19. Lee MS, Yeh IC, Zavaljevski N, Wilson P, Reifman J (2006) A Software Pipeline for Protein Structure Prediction. In: Parmentola JH, Rajendran AM, eds. Proceedings of the 25th Army Science Conference. Orlando, FL: Tech Science Press. pp 423–430. http://www.bhsai.org/publications.html.
20. Altschul SF, Madden TL, Schaffer 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.
21. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195–202.
22. Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33: W72–76.
23. Bondugula R, Xu D (2007) MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. Proteins 66: 664–670.
24. Bondugula R, Xu D (2006) Combining Sequence and Structural Profiles for Protein Solvent Accessibility Prediction. In: Proceedings of the Computational Systems Bioinformatics, CSB2008 Conference Proceedings; 2008 Aug 26–29. Stanford, California, USA: Imperial College Press. pp 195–202.
25. Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci 13: 1908–1917.
26. Dunker AK, Brown CJ, Obradovic Z (2002) Identification and functions of usefully disordered proteins. Adv Protein Chem 62: 25–49.
27. Cheng J, Sweredoski M, Baldi P (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. Data Mining and Knowledge Discovery 11: 213–222.
28. Bondugula R, Lee MS, Wallqvist A (2008) FIEFDom: A Transparent Domain Prediction System Using a Fuzzy Mean Operator. Nucleic Acids Res 37: 452–462.
29. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–65.
30. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, et al. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins 53 Suppl 6: 430–435.
31. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33: W284–288.
32. Kelley LA, MacCallum RM, Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299: 499–520.
33. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, et al. (2005) SAM-T04: what is new in protein-structure prediction for CASP6. Proteins 61 Suppl 7: 135–142.
34. Kim D, Xu D, Guo JT, Ellrott K, Xu Y (2003) PROSPECT II: protein structure prediction program for genome-scale applications. Protein Eng 16: 641–650.
35. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714–2726.
36. Feig M, Karanicolas J, Brooks CL 3rd (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model 22: 377–395.
37. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminatham S, et al. (1983) CHARMm: A program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 4: 187.
38. Mackerell AD Jr, Bashford D, Bellott DM, Dunbrack Jr RL, Evanseck JD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B 102: 3586–3616.
39. Feig M, Brooks CL 3rd (2002) Evaluating CASP4 predictions with physical energy functions. Proteins 49: 232–245.
40. Lee MS, Feig M, Salsbury FR Jr, Brooks CL 3rd (2003) New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J Comput Chem 24: 1348–1356.
41. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268: 209–225.
42. Bower MJ, Cohen FE, Dunbrack RL Jr (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 267: 1268–1282.
43. Malmstrom L, Riffle M, Strauss CE, Chivian D, Davis TN, et al. (2007) Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS Biol 5: e76.
44. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11: 739–747.
45. Shchelkunov SN, Blinov VM, Resenchuk SM, Totmenin AV, Olenina LV, et al. (1994) Analysis of the nucleotide sequence of 53 kbp from the right terminus of the genome of *variola major* virus strain India-1967. Virus Res 34: 207–236.
46. Shchelkunov SN, Blinov VM, Sandakhchiev LS (1993) Genes of *variola* and *vaccinia* viruses necessary to overcome the host protective mechanisms. FEBS Lett 319: 80–83.
47. Shchelkunov SN, Totmenin AV (1995) Two types of deletions in orthopoxvirus genomes. Virus Genes 9: 231–245.
48. Shchelkunov SN, Totmenin AV, Sandakhchiev LS (1996) Analysis of the nucleotide sequence of 23.8 kbp from the left terminus of the genome of *variola major* virus strain India-1967. Virus Res 40: 169–183.
49. Chivian D, Baker D (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res 34: e112.
50. Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 31: 3370–3374.
51. Lee MS, Lebeda FJ, Olson MA (2009) Fold prediction of VP24 protein of Ebola and Marburg viruses using *de novo* fragment assembly. J Struct Biol, doi: 10.1016/j.jsb.2009.05.001.
52. BLAST ftp://ftp.ncbi.nih.gov/blast/.
53. CE http://cl.sdsc.edu/.
54. CHARMM http://www.charmm.org/.
55. DISPro http://scratch.proteomics.ics.uci.edu/.
56. Jackal http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Jackal.
57. MMTSB http://blue11.bch.msu.edu/mmtsb/Main_Page.
58. NR ftp://ftp.ncbi.nih.gov/blast/.
59. PDB www.rcsb.org/pdb.
60. PROSPECT II http://compbio.ornl.gov/structure/prospect2/.
61. PSIPRED http://bioinf.cs.ucl.ac.uk/psipred/.
62. Rosetta http://www.rosettacommons.org/software/index.html.
63. SCOP http://scop.berkeley.edu/.
64. SCWRL3 http://dunbrack.fccc.edu/SCWRL3.php.
65. TMHMM http://www.pdc.kth.se/~hakanv/prodiv-tmhmm/.