Nucleic Acids Research Advance Access published June 12, 2012

Nucleic Acids Research, 2012, 1–6
doi:10.1093/nar/gks555

BioShell Threader: protein homology detection based on sequence profiles and secondary structure profiles

Dominik Gront*, Maciej Blaszczyk, Piotr Wojciechowski and Andrzej Kolinski

University of Warsaw, Faculty of Chemistry, Pasteura 1, 02-093 Warsaw, Poland

Received March 22, 2012; Revised May 9, 2012; Accepted May 16, 2012

ABSTRACT

The BioShell package has recently been extended with a web server for protein homology detection based on profile-to-profile alignment (known as 1D threading). Its aim is to assign structural templates to each domain of the query. The server uses sequence profiles that describe observed sequence variability and secondary structure profiles providing expected probability for a certain secondary structure type at a given position in a protein. Three independent predictors are used to increase the rate of successful predictions. Careful evaluation shows that there is nearly 80% chance that the query sequence belongs to the same SCOP family as the top scoring template. The Bioshell Threader server is freely available at: http://www.bioshell.pl/threader/.

INTRODUCTION

The formation of a 3D shape from a linear amino acid chain in its native environment usually occurs easily and quickly. However, this presumably simple and very common process has imposed a very difficult computational challenge. Indeed, the general problem of computing the native structure of a protein based solely on its amino acid sequence still remains unsolved. It is therefore very tempting to bypass the problem by using some of the wealth of information on already known protein structures. It is commonly accepted that the number of distinct protein folds is limited (despite large disagreement as to their exact number) (1). One may therefore expect that a structure of a protein homologous to the one in question (the query) has been already solved. The question then is to find that structure (referred further as a template) and to align its sequence with the sequence of the query. Subsequently, the template or templates and respective alignments are used to create a model of the query protein.

Historically, at least two different methods were proposed for template searching: alignment of query and template sequence-derived data (2,3) and threading of a query sequence through template structures (4–7). Although both methods have continued to develop, to some extent they blended with each other (8). In particular, the sequence alignment methods became equipped with sequence and structural profiles and evolved into so-called 1D threaders. A sequence profile, introduced by Gribskov (2), describes per-position sequence variability within a given sequence family, whereas a structural profile encodes some of the structural properties of a protein as per-residue propensities. One-dimensional methods usually compute query-to-template alignment by a dynamic programming approach that costs $O(N^2)$ when an affine gap penalty is used. To the contrary, ''true'' 3D threading is a NP-hard problem (5), which can be solved only by a heuristic approach. A number of 1D methods have been proposed in the literature. Some of them use Hidden Markov Models to infer the probabilities that comprise a sequence profile: HHpred (9), HMMER (10), SAM-T02 (11). Others calculate per-residue statistics for each residue type: FFAS3 (12), ORFeus (13), FUGUE (14), MUSTER (15). They also differ in the variety of structure-related information employed. The most popular include secondary structure, solvent accessibility, dihedral torsion angles, and structure-dependent gap penalties. Despite their simplicity, profile to profile aligning methods are among the best-performing present-day methods for fold recognition, as can be seen from the results of the blind, automated structure prediction contests CASP (16) and LIVEBENCH (17,18).

In this contribution, we present a new fully automated 1D threading server. Its unique element is the combination of three different secondary structure prediction methods in the scoring system. In addition, our server has been designed to detect template structures for separate protein domains rather than for a whole chain sequence. This decision has been motivated by the fact that for methods commonly used for actual model building, it is

*To whom correspondence should be addressed. Tel: +48 22 8220211 (extn. 310); Fax: +48 22 8220211 (extn. 310); Email: dgront@chem.uw.edu.pl

much easier to build a single domain structure than to calculate a model of a multidomain protein. Moreover, per domain alignment scores are more sensitive than respective values calculated for multidomain sequences. Effective template search is related to the quality of the query-to-template alignment. These two aims, however, often become two opposite goals for method optimization. In this case, we focused solely on template detection. Top-scoring templates will be subsequently used by a recently proposed modeling method (19), which does not require any prior alignment as a modeling input.

MATERIALS AND METHODS

BioShell Threader: input data and overall workflow

The only data required as an input are a protein sequence in the FASTA format. The sequence is used as a starting point for a procedure that consists of the following steps:

(i) sequence profile is built by PSI-BLAST (20) with the following settings: five iterations with 1e-5 e-value threshold, BLOSUM62 with gap parameters: −10,-2 (opening and extending, respectively); NCBI-nr database is used for the search.
(ii) secondary structure (SS) profiles are computed with PsiPred (21), Porter (22) and SAM (23), all of them with their default settings. Such an SS profile holds three probability values per residue in a query sequence and describes the expected chance for finding the residue in a Helix, Loop or Extended conformation.
(iii) aforementioned four profiles for the query sequence are aligned against an in-house database of corresponding profiles created for SCOP (24) domains.

Profile similarity is assessed by the Picasso3 (25) score, whereas secondary structure similarity is measured with L1 metrics. Although the state-of-the-art SS predictors may feature 80% or higher success rate, the remaining 20% mispredictions are not evenly distributed over the query protein. Conversely, predictors often miss or even mispredict a whole secondary structure element, which may lead to wrong assignment of a protein family. The use of three independent prediction programs has been introduced to minimize the effect of such errors. Such an approach (also using PsiPred, Porter and SAM) has been previously applied for de novo protein structure prediction (26).

Template databases

To be able to detect templates for domains rather than whole protein chains, the database of templates for BioShell Threader was based on the most recent structural classification of proteins (SCOP) (24) 1.75 classification. For each SCOP domain, a separate database entry was created, which holds the domain's 3D coordinates, sequence profile (computed with PsiBlast in the same manner as for query sequences) and secondary structure assigned by dictionary of secondary structure of proteins (DSSP) (27). Unfortunately the most recent SCOP 1.75 edition that has been released in 2009 covers only a half of today's protein data bank (PDB) content. Therefore, the SCOP-based set of templates has been extended by the PDB chain entries, which are not in the SCOP database yet.

Output

Finally, the user obtains a number of the best scoring templates, both in the form of alignments with the query sequence and as PDB-formatted 3D data. The format of the alignments displayed by the server (Edinburgh format) was chosen considering its transparency. User can also download the results both in Edinburgh and FASTA formats. Conversion to other commonly used formats can be easily done with simple BioShell utility scripts.

The PDB files may be directly used to create a comparative model for the query. The server also provides secondary structure prediction for the query. Additionally, the result page provides summary statistics for the top-scoring templates such as Z-score and SCOP family assignment. A sample result page is presented in Figure 1.

Server architecture

The server physically comprises two independent computers: a front-end server and a computing cluster (see Figure 2). The front-end is a WWW server with a PostgreSQL database used to store data about submitted jobs and their results. The computing host periodically checks the server's queue, retrieves new jobs and uploads the results. The front-end server never solicits any computations. It is always the computing server that initiates communication, performed via an HTTP protocol. Such an approach facilitates rearrangements of the computing server(s) (e.g. substituting one by another during maintenance or adding more computing hosts if required). All the data processing operations and alignment calculations are performed by BioShell 28,29) package modules.

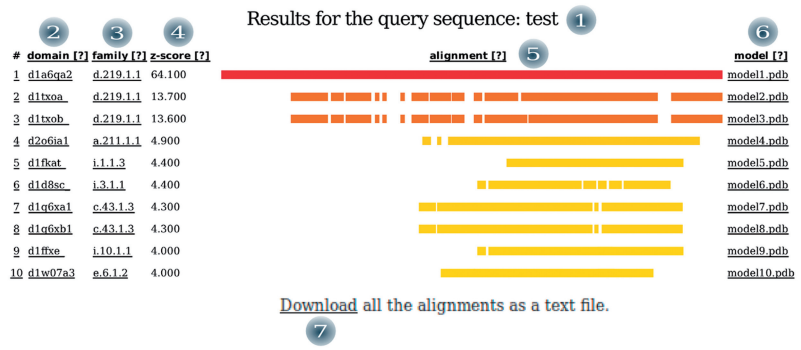
Algorithm and its validation

Our one dimensional threading presented in this article implements a pair-wise sequence alignment algorithm with affine gap penalty. As mentioned earlier, we defined the score value for a match between the i^{th} position in a query and the j^{th} position in a template as:

$$S(i,j) = \text{Picasso3}(i,j) + w_{PSIPRED} L_1(i,j) + w_{PORTER} L_1(i,j) + w_{SAM} L_1(i,j) \quad (1)$$

The method, therefore, employs two similarity functions that have to be chosen: one to compare sequence profiles (Picasso3 in Formula 1) and the other for secondary structure profiles (L_1 in Formula 1). Alignments may be computed according to the global or to the local variant. There are also six independent parameters: three weights, bias (only for local alignment) and two gap parameters. The optimization and validation of all these settings were performed on a carefully selected subset (30) of the SCOP database. From all the SCOP families, only those were selected that contained at least four protein domains, similar in no more than 30% to one another. When a family provided more than four such domains, only four

BioShell Threader is a fully automated server for protein homology detection based on 1D threading which employs profile-profile alignments for both sequence and secondary structure. Below you can see an example of the result page and explanation of its functionality.



1 job name

2 SCOP domain identification code

3 SCOP family identification code

4 alignment z-score - evaluation of a quality of the alignment

5 alignment - query coverage by a template. Click on the bar to move to the text representation of the alignment. 8 Bar color corresponds to the alignment z-score. 4

6 click to download thread (partial model) based on an alignment between the query sequence and a template structure. Go to 9 to see visualization of the model.

7 click to download all alignments in a single text file.

8 text representation of the alignment with information about the secondary structure of the query (consensus derived from three prediction methods: PSI-PRED, PORTER and SAM) and the template (recognized using DSSP).

Figure 1. Sample Bioshell Threader result page describes functionalities of the server.

of them were selected. Such a selection procedure resulted in a set of 423 SCOP families. Two random out of each four domains were moved to a '*train*' set used for parameter optimization. The other pair of domains was moved to a "*test*" set, necessary for final validation. Each of these two sets, therefore, comprises the same number of SCOP domains, always two domains from the same family.

The optimization goal was to maximize the chance for finding the right family member for a query. Each domain from the 'train' set was aligned with every other domain in

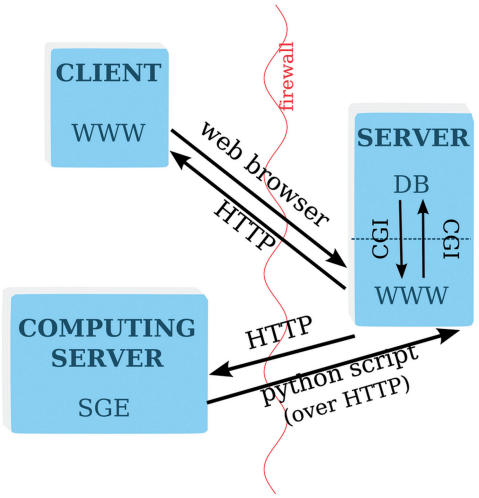


Figure 2. Interaction between the computing cluster(s), the font-end server and the user.

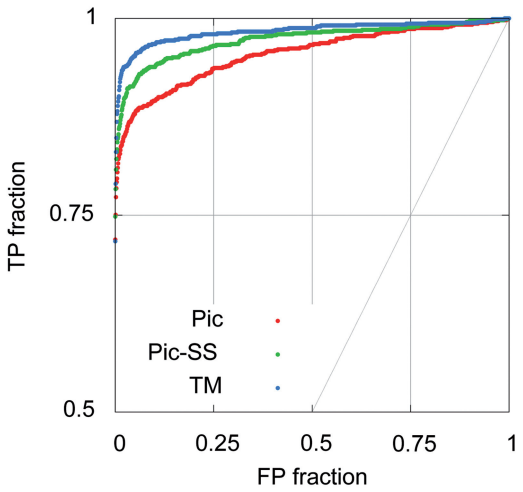


Figure 3. ROC curves for SCOP Family assignment obtained by three methods: sequence profile alignment (Pic), alignment with secondary structure profiles (Pic-SS) and structure alignment (TM). Area under ROC curve: 0.955, 0.957 and 0.984, respectively.

the set, which required 423 * (423 −1) alignments. From these results we computed how often the top-scoring template shared the same family as the query, a parameter further referred to as HIT_RATE. Such a procedure was repeated for each reasonable parameter combination, for a number of substitution matrices as well as for several profile–profile scoring schemes, with and without secondary structure profiles, both for local and global alignment. In total, more than 5000 CPU hours were devoted to test several thousand combinations. The best results (HIT_RATE = 0.80) were obtained with the Picasso3 scoring scheme used in local alignment of sequence profiles combined with the L1 score used to assess SS similarity. The optimal gap parameters were: −1.7, −0.5 (gap open and gap extend, respectively). The optimal weights for scaling secondary structure similarity were 0.4, 0.2 and 0.2 for PsiPred, Porter and SAM, respectively. Finally, these settings were validated on the 'test' benchmark subset yielding HIT_RATE = 0.788.

RESULTS

Comparison with other methods

Training and validation of the method was based on a custom benchmark set derived from SCOP classification. It is therefore very difficult to compare our approach with other methods, because they often rely on pre-calculated databases or are implemented as web servers. To give a rough estimate of the performance of BioShell Threader, we compared our profile-based protocol with simple sequence alignment and structural alignment. The former of the two methods was optimized on the 'train' benchmark in the same manner as BioShell Threader. After the optimization the BLOcks of amino acid SUbstitution matrix (BLOSUM62) was chosen with gap penalties (open, extend) = (-10,-1). As for the structural alignment, a TM-align (31) method was used. Validation results show that the BioShell Threader method not only outperforms sequence alignment but is even better than the structural alignment method. The results presented in Figure 3 show that the alignment of profiles already enables correct SCOP Family assignment in 74.8% of cases, whereas combined with secondary structure alignment, it yields 78.8% correct predictions. This is already slightly better than the structure alignment method TM-align (75.2% successful predictions). Another assessment measure, area under ROC curves, confirms the high sensitivity of the method (ref. Figure 3), although according to this criterion TM-align is slightly better than our method. Bioshell Threader also yields better results than ORFeus, a similar, previously described method, which also uses secondary structure profiles and it correctly assigned only 583 out of 1713 representatives (34%) (13). This most likely results from the use of dot product (DP), a significantly less effective scoring scheme. On our benchmark, DP yielded approximately 58% successful SCOP family assignments. Moreover ORFeus uses results from only one secondary structure prediction program.

Additionally, we have compared our approach to popular HHpred method (9). It also conducts similarity search but is based on hidden markov model (HMM) profiles. On our benchmark, HHpred yielded 75.6% successful SCOP family assignments when a HMM profile was used as a query and slightly below 71% when just a sequence was used. The difference between the performance of HHpred and BioShell Threader most likely results from the fact that our method uses three independent secondary structure prediction methods. Obviously, the necessity for running four external programs (the tree predictors and PsiBlast for a query sequence profile) increases the computational time. Processing a single query requires 10–30 minutes.

Effect of independent secondary structure predictors

The use of secondary structure prediction in fold recognition certainly helps; nevertheless the improvement is limited. All three SS-related weights sum up to 0.6, whereas the sequence profile score has a weight of 1.0.

Table 1. Comparison of the three secondary structure prediction methods (PsiPred, Porter and SAM) used by the server

Method	Q3	DSSP	PsiPred	Porter	SAM
DSSP		1.00	0.73	0.68	0.64
PsiPred	80.8% (7.7%)	0.73	1.00	0.67	0.68
Porter	77.4% (8.8%)	0.68	0.67	1.00	0.60
SAM	76.8% (8.1%)	0.64	0.64	0.60	1.00

The second column shows Q3 prediction accuracy, whereas the last four columns show mutual correlation between the predictors and DSSP as the secondary structure definition

The dynamic range of the Picasso3 score is nearly twice as high as the range of the L1 score, which means that both the sequence-related and SS-related components are equally important. The impact of secondary structure is hindered by inaccuracy of predictions. The rates of successful prediction (Q3) measured as a percentage of correctly assigned (H, E, L) letters measured on the benchmark set were: 80.8% (7.7%), 77.4% (8.8%) and 76.8% (8.1%) for PsiPred, Porter and SAM, respectively, with standard deviation values in brackets (see Table 1). The three predictors yield similar accuracy, being, however, rather loosely correlated with one another (See Table 1). In an easy case when a lot of sequence homologs can be found for a query sequence of a 'typical' protein, all three methods return essentially correct predictions. For difficult targets all of them make mistakes but the mistakes are differently distributed along the sequence. Therefore, the use of several predictors increases the chance for successful fold recognition.

CONCLUSION

In this contribution, we described a fold recognition server that assigns structural templates for plausible domains in a query sequence. The use of three independent secondary structure predictors significantly increases the rate of successful template assignment. The server, based on a thoroughly tested BioShell software library, will be continuously developed in the future toward a unified environment for protein structure modeling.

FUNDING

Foundation for Polish Science TEAM project [TEAM/2011-7/6 to M.B. and P.W.] co-financed by the European Regional Development Fund operated within the Innovative Economy Operational Programme; Polish National Science Center (NCN) [2011/01/D/NZ2/07683 to D.G.]. Funding for open access charge: Foundation for Polish Science TEAM project (TEAM/2011-7/6) co-financed by the European Regional Development Fund operated within the Innovative Economy Operational Programme.

Conflict of interest statement. None declared.

REFERENCES

1. Wolf,Y.I., Grishin,N.V. and Koonin,E.V. (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.*, **299**, 897–905.
2. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Nat. Acad. Sci.*, **84**, 4355–4358.
3. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
4. Godzik,A., Kolinski,A. and Skolnick,J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, **227**, 227–238.
5. Lathrop,R. (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.*, **7**, 1059–1068.
6. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
7. Bryant,S.H. (1996) Evaluation of threading specificity and accuracy. *Proteins*, **26**, 172–185.
8. Panchenko,A.R., Marchler-Bauer,A. and Bryant,S.H. (2000) Combination of threading potentials and sequence profiles improves fold recognition. *J.Mol. Biol.*, **296**, 1319–1331.
9. Söding,J., Biegert,A. and Lupas,A. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
10. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
11. Karplus,K. (2009) SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res.*, **37**, W492–W497.
12. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W. and Godzik,A. (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res.*, **33**, W284–W288.
13. Ginalski,K., Pas,J., Wyrwicz,L., von Grotthuss,M., Bujnicki,J. and Rychlewski,L. (2003) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res.*, **31**, 3804–3807.
14. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
15. Wu,S. and Zhang,Y. (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, **72**, 547–556.
16. Mariani,V., Kiefer,F., Schmidt,T., Haas,J. and Schwede,T. (2011) Assessment of template based protein structure predictions in CASP9. *Proteins*, **79(Suppl. 10)**, 37–58.
17. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.
18. Rychlewski,L. and Fischer,D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci.*, **14**, 240–245.
19. Kolinski,A. and Gront,D. (2007) Comparative modeling without implicit sequence alignments. *Bioinformatics*, **23**, 2522–2527.
20. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
22. Pollastri,G. and McLysaght,A. (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**, 1719–1720.
23. Katzman,S., Barrett,C., Thiltgen,G., Karchin,R. and Karplus,K. (2008) PREDICT-2ND: a tool for generalized protein local structure prediction. *Bioinformatics*, **24**, 2453–2459.
24. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

25. Mittelman,D., Sadreyev,R. and Grishin,N. (2003) Probabilistic scoring measures for profile–profile comparison yield more accurate short seed alignments. *Bioinformatics*, **19**, 1531–1539.
26. Gront,D., Kulp,D., Vernon,R., Strauss,C. and Baker,D. (2011) Generalized fragment picking in rosetta: design, protocols and applications. *PLoS ONE*, **6**, e23294.
27. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
28. Gront,D. and Kolinski,A. (2006) BioShell—a package of tools for structural biology computations. *Bioinformatics*, **22**, 621–622.
29. Gront,D. and Kolinski,A. (2008) Utility library for structural bioinformatics. *Bioinformatics*, **24**, 584–585.
30. Gniewek,P., Kolinski,A. and Gront,D. (2012) Optimization of profile-to-profile alignment parameters for one-dimensional threading. *J. Comput. Biol.*, doi: 10.1089/cmb.2011.0307.
31. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, **33**, 2302–2309.