JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 7, 2012

Pp. 879–886
DOI: 10.1089/cmb.2011.0307

Optimization of Profile-to-Profile Alignment Parameters for One-Dimensional Threading

PAWEL GNIEWEK, ANDRZEJ KOLINSKI, and DOMINIK GRONT

ABSTRACT

The development of automatic approaches for the comparison of protein sequences has become increasingly important. Methods that compare profiles allow for the use of information about whole protein families, resulting in more sensitive and accurate detection of distantly related sequences. In this contribution, we describe a thorough optimization and tests of a profile-to-profile alignment method. A number of different scoring schemes has been implemented and compared on the basis of their ability to identify a template protein from the same SCOP family as a query. In addition to sequence profiles, secondary structure profiles were used to increase the rate of successful detection. Our results show that a properly tuned one-dimensional threading method can recognize a correct template from the same SCOP family nearly as well as structural alignment. Our benchmark set, which might be useful in other similar studies, as well as the fold-recognition software we developed may be downloaded (www.bioshell.pl/profile-alignments).

Key words: algorithms.

1. INTRODUCTION

The alignment of protein or genomic sequences is one of the most common research tool for modern molecular biologists. The methods of sequence alignment have obtained much attention since they have been developed in the 1970s and the 1980s (Needleman and Wunsch, 1970; Smith and Waterman, 1981). Global and local alignment algorithms were followed by the development of models describing evolution of one sequence into another. These considerations resulted in the two commonly used approaches that describe point-mutation events: Blossum (Henikoff and Henikoff, 1992) and Pam (Dayhoff et al., 1978) substitution matrices. These matrices—an indispensable part of any sequence alignment tool—have found a wide application in genome annotation, protein classification, studying protein evolution, phylogenetic analysis, and protein design. Sequence alignment is also the foundation of methods for secondary (Jones, 1999) and tertiary (Kolinski, 2004) protein structure prediction.

Due to enormous experimental efforts, the size of available databases has increased drastically. This imposes a considerable challenge for computational methods. Help has come from novel implementations (Oehmen and Nieplocha, 2006) and hardware development (Blazewicz et al., 2011). Considerable progress has also been made on theoretical grounds. In their seminal article, Gribskov et al. (1987) introduced a

Faculty of Chemistry, Warsaw University, Warsaw, Poland.

sequence profile that is a matrix that contains $N \times 20$ amino acid probabilities and describes a family of protein sequences rather than a single sequence of N amino acid residues. Originally, profiles were used to enhance the search in a database of sequences by means of profile-to-sequence alignment methods. These were in turn generalized by Pietrokovski (1996) into profile-to-profile alignments. Since then a search has been performed as alignments of a query profile with a database of precomputed sequence profiles, where each of these profiles describes a single family of proteins. Progress therefore went through three types of methods: sequence-to-sequence, e.g., BLAST (Altschul et al., 1990) and FASTA (Lipman and Pearson, 1985); profile-to-sequence, e.g., PSI-BLAST (Altschul et al., 1997) and SAM-T98 (Karplus et al., 1998); and profile-to-profile alignment. The former two utilize a substitution matrix to assess the effect of mutations. The latter one requires a function that compares two profile columns, where each column is a vector of 20 real values (21 when a profile also describes gap probability). Numerous functions have been proposed for this purpose (Yona and Levitt, 2002; Pietrokovski, 1996; Sadreyev and Grishin, 2003; von Ohsen et al., 2003; Rychlewski et al., 2000; Panchenko, 2003; Panchenko et al., 2000); some of them perform better than the other, as assessed in several independent studies (Ohlson et al., 2004; Wang and Dunbrack, 2004). The specificity of the search can be enhanced by the use of additional information (e.g., secondary structure or surface exposure). Such data may come either from experiments or from machine learning methods, and results in a better distinguishing between pairs of homologous and non- homologous proteins.

In this article, we describe the design, optimization, and benchmarking of a protocol that aligns a query protein sequence with a database of protein domains. The main purpose in devising such a computational tool is to provide template structures for comparative modeling protocols. These are based on the assumption that, if similarity between two sequences is high enough, the two proteins share a common fold. Most commonly, comparative modeling methods heavily rely on a query-template alignment. Relevant structural parts of the template are copied to the query according to the alignment. Missing fragments are reconstructed in the subsequent step of the modeling. Reliability of such an approach strongly depends on the quality of the alignment. Therefore, sequence alignment methods used in comparative modeling are usually optimized for alignment accuracy. This however may lower the template detection sensitivity. All the alignment-related problems can be alleviated by a novel comparative modeling method recently proposed by Kolinski and Gront (2007). The method requires only specifying the template structure (or structures, as the method can utilize multiple templates); the a priori alignment is not required. Therefore, in this study, we have optimized an alignment method to maximize the chance for correct selection of a template. To achieve the highest possible sensitivity, we performed alignment of sequence profiles combined with secondary structure profiles. For the sake of speed, we limited ourselves to dynamic programming algorithm with affine gap penalty.

2. METHODS

2.1. Benchmark set

The optimization process described in this contribution has been based on the most recent ASTRAL database (Chandonia et al., 2004). The database consists of protein domain structures extracted from the PDB content according to the SCOP classification (Murzin et al., 1995). Redundancy had already been removed from the set in such a way that amino acid sequences of any two domains from the database are identical in at most 40%. In order to transform the ASTRAL database into a well-balanced benchmark set, we performed the following steps:

(i) We considered only these SCOP families that are represented in ASTRAL by at least 4 domains. Any domain that belongs to a Family which does not satisfy this condition was excluded from the benchmark. In this way, we could easily divide the dataset into a *train* set and a *test* set.
(ii) Then, the best four Family representing structures (according AREOSPACI score) were divided into training and testing sets by putting two randomly selected domains into the first set and the remaining two into the other.

The procedure resulted in a benchmark set comprising two subsets: train and test (1082 domains each). The former set was used to determine the optimal values for the necessary parameters and the latter one to assess the quality of our method. Any of these 1082 domains may be used as a query in a search for homologues domains. The way that the benchmark set was constructed ensured that there was exactly one

correct answer that shared the same SCOP Family with the query. Moreover, the *training* and *testing* sets were of the same size, which helped to avoid any bias during the optimization procedure.

In the course of this study, to provide homology sequence redundancy on the 30% level, several additional domains were excluded from the benchmark. In order to keep the benchmark consistent, we also removed all other domains that belong to the same SCOP Family as the problematic ones. The reduced benchmark set therefore comprised two subsets of 935 domains.

2.2. Profile generation

For each domain, a sequence profile was computed by PsiBlast with the following settings: matrix, BLOSSUM62; gap open, -11; gap extend, -1; number of iterations, 5; and profile e-value threshold, 10^{-5}. Resulting checkpoint files that contained raw observed amino acid frequencies (twenty real values) at each position in a sequence were modified by adding pseudocounts q_i in a manner similar to that used by Tatusov et al. (1994):

$$q_i = \sum_{j=1}^{20} p_j f_j \exp(\lambda \mathrm{B62}_{j,i})$$

where $\lambda = 0.001$, p_j is the observed frequency for j^{th} amino acid in a profile column, f_j is the frequency of j^{th} amino acid in the whole SwissProt database (Boeckmann et al., 2003) and B62 detones BLOSSUM62 matrix.

A sequence profile computed for an amino acid sequence of N_L residues is thus a matrix $N_L \times 20$. Its rows represent twenty-dimensional probability vectors, normalized to 1.0. This study also utilizes secondary structure (SS) profiles, which are $N_L \times 3$ matrices. The three columns provide the probability of finding a helix, strand, or loop at a given position in the protein. Two kinds of secondary structure profiles were derived for each domain: one computed from the structure by the DSSP program (Kabsch and Sander, 1983) and the other predicted by PSIPRED (Jones, 1999). The DSSP program is a definition rather than a predictor: it assigns secondary structure symbols based on the hydrogen bond network observed in a tertiary structure. The resulting probabilities were therefore either 1.0 or 0.0. Profiles of this type were utilized when a domain served as a template in an alignment calculation. Since the secondary structure of a query protein is not known, in this case we used a profile predicted by PSIPRED.

2.3. Scoring methods

During the parameter optimization step, the following profile-to-profile scoring schemes were considered: (1) regular sequence alignment, (2) dot product, (3) L1 score, and (4) Picasso3. The best method with the optimal set of parameters was used in the production version of the protocol.

2.3.1. Sequence alignment. A substitution matrix was used to define the score for aligning the i^{th} position from query sequence and the j^{th} position from the template. A number of substitution matrices (BLOSSUM, PAM) were tested, and BLOSSUM62 (denoted as B62) was selected as the best one.

2.3.2. Dot product. Dot product (DP), one of the simplest formulas used to assess the similarity between two sequence profile columns i^{th} and j^{th}, is defined as:

$$S_{\mathrm{DP}}(i,j) = \sum_{a=1}^{20} Q_{i,a} T_{j,a}$$

where $Q_{k,a}$ or $T_{k,a}$ is the frequency of the appearance of amino acid a in k^{th} column from the query and the template profile, respectively.

2.3.3. L1-score. L1-score similarity measure is simply the L_1 distance between profiles' corresponding columns:

$$S_{\mathrm{L1}}(i,j) = 1 - \sum_{a=1}^{20} |Q_{i,a} - T_{j,a}|$$

Despite its very simple formulation and great computational efficiency, we are not aware of any study that compares the L1 score with other functions that are commonly used in the field. The function however has been used to guide fragment selection for modeling (Gront et al., 2011) with the Rosetta program.

2.3.4. PICASSO3. This is one of the log-odds based method, introduced by Heger and Holm (2001). The original function is not symmetric, which means that results depend on which sequence is treated as a query and which one as a template. This ambiguity was removed by Mittelman et al. (2003) by introducing a symmetrical score equation for aligning the i^{th} and j^{th} amino acids from query and template sequences:

$$S_{\mathrm{Pic}}(i,j) = \sum_{a=1}^{20} T_{j,a} \ln \frac{Q_{i,a}}{p_a} + \sum_{a=1}^{20} Q_{j,a} \ln \frac{T_{i,a}}{p_a}$$

where p_a is the expected frequency of appearance of the a^{th} amino acid, computed in this work by averaging the whole SwissProt database.

2.3.5. Structure alignment. Besides the methods for profile-to-profile alignments, we used a structure alignment algorithm to match the query-template pairs. In a real-life application, the three-dimensional structure for a query protein remains unknown; the structure-based comparison was used in this study as a reference. Since three-dimensional structure provides much more information than just a sequence profile, we had expected it would define the upper bound for profile-based methods and serve as a reference point for the comparisons. Structure alignments were computed with TM-align—a method that employs a heuristic iterative algorithm (Zhang and Skolnick, 2005) which maximizes TM-score, a parameter defined as:

$$S_{\mathrm{Str}} = \frac{1}{L_T} \sum_{i=1}^{L_A} \frac{1}{1 + (\frac{d_i}{d_0})^2}$$

where L_T is the length of the shorter protein, L_A is the length of the alignment, d_i is the distance of two corresponding amino acids in the alignment (after structures superposition), and d_0 is estimated from the following formula:

$$d_0 = 1.24 \sqrt[3]{L_T - 15} - 1.8$$

2.4. Accuracy measures

Two numerical measures of success were used. The first one, denoted further as a hits ratio (HR), is the fraction of correctly matched Family or Superfamily members for the best scoring query-template pair. In other words, if one wants to assign a given query sequence to a SCOP Family or Superfamily just by copying the annotation of the top scoring template, HR is the chance for this annotation to be correct. The second parameter, Area Under Curve (AUC), measures the area under Receiver-Operator Curve (ROC). AUC quantifies the probability of scoring a randomly chosen positive case higher than a negative one (Fawcett, 2006). A random predictor would feature AUC = 0.5. The first criterion fit the overall purpose of our work better; hence, it was used as the criteria for all the parameter optimization procedures. The second parameter was used for additional characterization of the results.

2.5. Implementation and optimization

In the optimization process, any domain from the train subset was aligned with any other domain in this set, which required 935 × 934 alignment calculations. Subsequently, Family HR was reported based on these 935 cases. This procedure was applied to assess any alignment scheme attempted in the experiment. An alignment scheme is defined by two affine gap penalty parameters (open, extend) and a scoring system: one of the four profile-to-profile similarity measures or one of several substitution matrices. Moreover, each of these combinations was used both in a global and in a local alignment mode. An example of such an optimization for a single scoring method is shown in Figure 1, which presents the Family HR value as a function of affine gap penalty parameters. Each square in this plot denotes the HR found after 935 × 934 local alignment computations with Picasso3 scoring scheme. In order to test all the combinations of the parameters, the optimization required the all-versus-all profile alignment procedure to be performed more

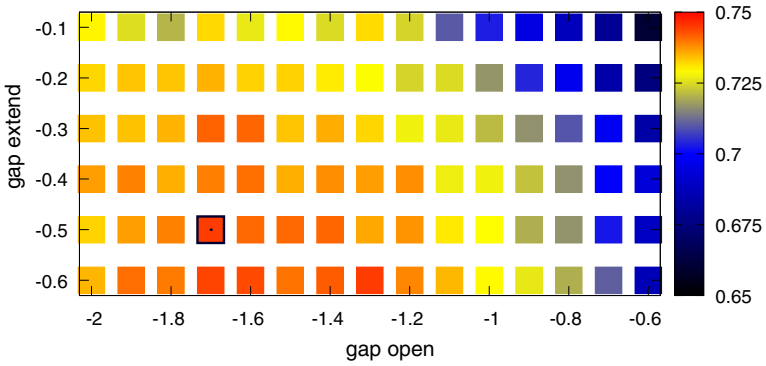

FIG. 1. Optimization of gap penalty (open, extend) parameters (*x* and *y* axis, respectively). Color scale denotes Hit Rate at Family level. A square corresponding to the best performing parameters' combination is marked with a black border.

than 1500 times. Each of these runs resulted in a single Family HR value. The best scoring sequence-to-sequence alignment as well as the best scoring profile-to-profile alignment were also assessed on the test subset. All these alignment calculations as well as data analysis were performed with BioShell package (Gront and Kolinski, 2006, 2008). In fact, after initial tests, the relevant BioShell routines were optimized, which resulted in significant reduction in the CPU time required for this project. In addition, all-versus-all structural alignments were computed with TM-align program. Based on obtained TM-score values, HR and AUC parameters were evaluated in the same manner as for the sequence-based alignment.

3. RESULTS AND DISCUSSION

The performance of profile-to-profile alignment methods has been discussed previously (Ohlson et al., 2004; Wang and Dunbrack, 2004). Although the benchmark set as well as the optimization procedure utilized in this experiment differs considerably from previous studies, in general our findings summarized in Table 1 agree with the previous results. The first section of the table presents the best results obtained after exhaustive optimization of gap parameters, the second part (i.e., the last three rows) provide validation on the *test* set. Each method was applied both in global and local alignment (**g** or **l**, respectively).

In the course of optimization, we found BLOSSUM62 to be the best performing substitution matrix. Using the matrix in the global alignment of domain, sequences yields HR of 35.4%. Not surprisingly, any profile-to-profile

TABLE 1. PERFORMANCE RESULTS FOR THE ALIGNMENT METHODS

		Gap		*Hit Rate*		*ROC area*	
Method		*Open*	*Extend*	*Family*	*Superfamily*	*Family*	*Superfamily*
B62	g	−10.0	−1.0	0.354	0.361	0.709	0.587
B62	l	−6.5	−2.5	0.320	0.324	0.698	0.604
DP	g	−0.3	−0.2	0.583	0.602	0.909	0.784
DP	l	−0.2	−0.1	0.582	0.546	0.868	0.720
L1	g	−0.3	−0.2	0.664	0.693	0.944	0.814
L1	l	−0.4	−0.2	0.660	0.688	0.940	0.781
Pic	g	−2.0	−0.6	0.729	0.700	0.948	0.817
Pic	l	−1.7	−0.5	0.729	0.779	0.964	0.865
Pic	ls	−1.5	−0.6	0.790	0.836	0.974	0.907
Str	–	–	N/A	0.738	0.817	0.990	0.939
Pic	l	−1.7	−0.5	0.748	0.780	0.955	0.870
Pic	ls	−1.5	−0.6	0.779	0.820	0.957	0.908
Str	–	–	N/A	0.752	0.835	0.984	0.936

Assessment of the scoring schemes: **B62**, sequence alignment (with the best performing matrix BLOSUM62); **DP**, Dot Product; **L1**, L_1 measure, and **Pic**, Picasso3, used in global and local (**g** and **l**) alignment, respectively. The best performing method (Pic l) was combined with secondary structure information (**ls** variant). The first section of the table refers to parameter optimization runs done on a *train* set, the second (last three rows) to a *test* set. The sequence-based methods were also compared to structure alignment (**Str**). All the quality assessment parameters (Hit Rate and area under ROC curve) are in the range [0.0, 1.0].

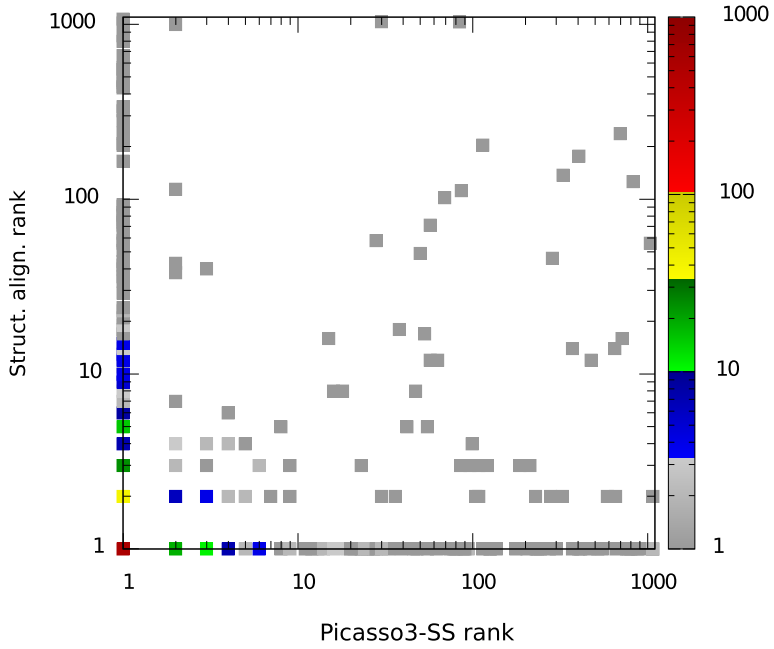

FIG. 2. Comparison between the best method, Picasso3 + secondary structure (Picasso3-SS), with a structural alignment. This two-dimensional histogram shows at $x = i$ and $y = j$ how often a correct query-template pair of two proteins from the same family has been ranked as i^{th} (j^{th}) by Picasso3-SS (structure alignment), respectively.

alignment method is superior to the sequence alignment. The best of the profile similarity scoring methods considered here, Picasso3 yields HR of 0.73 (0.78) on the Family (Superfamily) level, respectively. The combination of Picasso3 (local alignment with optimal gap parameters) score with SS profiles similarity boosts these rates further to 0.79 (0.84), respectively. This simply means that when a query protein is aligned as described above with representative protein domains from SCOP database, there is 84% chance that the top scoring template comes from the same SCOP Superfamily as the query. The scoring function in this case was defined as $S_{\mathrm{PicSS}}(i,j) = S_{\mathrm{Pic}}(i,j) + c_{\mathrm{L1}}^{\mathrm{SSS}}(i,j)$. The value of the constant $c = 0.5$ was optimized on the *train* set.

In general, these are the results one could expect before conducting the experiment. The most unexpected result however comes with the performance of structural alignment, which is very well comparable to the Picasso3 + SS combination, according to the HR parameter. Note, however, that HR is based only on the best scoring query-template pair. On the contrary to HR, according to AUC parameter structural alignment is still better than any other method. In fact, structural alignment tends to rank the correct query-template pairs higher than the best profile alignment method. The latter one however has been extensively trained to maximize the first-rank hits rate. The ability for ranking the correct query-template pairs of the two methods: Picasso3 + SS and structure alignment have been compared in the Figure 2. The x and y axes provide the rank of a correct query-template pair as assigned by Picasso3 + SS and structure alignment, respectively. Each point in this plot represents a bar of a two-dimensional histogram aggregating the ranked pairs. The highest peak in the histogram is located at (1,1) and represents all those correct query-template pairs which both methods (Picasso3 + SS and structure alignment) managed to rank at the first place. There are also other bars on the x (y) axis. They correspond to the pairs which only structural alignment (only Picasso3 + SS, respectively) ranked best. The points scattered over the plot tend to locate in its lower triangle which confirms that structure alignment assigns better ranks to the correct pairs than Picasso3 + SS. Besides the differences in ranking abilities, the two methods offer similar performances.

Another interesting finding is the good performance of the L1 score. Despite its very simple formulation, it achieves HR on the Family level just 7% lower than the best method, Picasso3, and it might be considered for applications where high computational efficiency is required.

4. CONCLUSION

In this contribution, we described an optimization and benchmarking of a profile-to-profile alignment method. The carefully derived benchmark set we used in this study can be freely downloaded (www.bioshell.pl). For each protein domain, the archive provides a structure in PDB format, a sequence profile, and two secondary structure profiles (predicted by PsiPred and defined by DSSP).

The fold recognition method that we developed in this study allows rapid assignment of SCOP Family or Superfamily to a given query sequence with a success rate close to the one achieved by structural alignment. This can be understood from the point of view of how evolutionary relationship is defined in the SCOP database. Homology between proteins is evaluated based on sequence analysis and then manually curated. It expresses the fact that, at least in some cases, the structure similarity is neither the strongest nor the only signal for common origin of two proteins. This can also suggest that fold recognition algorithms have already achieved a fair level of maturity, and there is small room for further improvement (at least on the benchmark posed by this study). There are however two other very important issues that have not been addressed here: (i) assessment of statistical significance for a given result and (ii) alignment quality. As for the first problem, we assumed that the set of templates always contains the correct answer, that is a member of the same SCOP family as the query. Our benchmark set, due to the way it has been constructed, obviously satisfies that assumption. The assumption will also hold true in the majority of real-life application since the protein fold space is already well covered by structures determined experimentally. Finally, the problem of the actual alignment between query and template proteins will be handled by an algorithm used for construction of a structural model for a query protein.

ACKNOWLEDGMENTS

This project has been funded from the TEAM/2011-7/6 Iuventus Plus research grant.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., et al. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Blazewicz, J., Frohmberg, W., Kierzynka, M., et al. 2011. Protein alignment algorithms with an efficient backtracking routine on multiple GPUs. *BMC Bioinform.* 12, 181.

Boeckmann, B., Bairoch, A., Apweiler, R., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370.

Chandonia, J.M.M., Hon, G., Walker, N.S., et al. 2004. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* 32, D189–D192.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. 1978. A model of evolutionary change in proteins. *Atlas Prot. Seq. Struc.* 5, 345–351.

Fawcett, T. 2006. An introduction to ROC analysis. *Patt. Recog. Lett.* 27, 861–874.

Gribskov, M., McLachlan, A.D., and Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proc. Nat. Acad. Sci. USA* 84, 4355–4358.

Gront, D., and Kolinski, A. 2006. BioShell—a package of tools for structural biology computations. *Bioinformatics* 22, 621–622.

Gront, D., and Kolinski, A. 2008. Utility library for structural bioinformatics. *Bioinformatics* 24, 584–585.

Gront, D., Kulp, D.W., Vernon, R.M., et al. 2011. Generalized fragment picking in rosetta: design, protocols and applications. *PloS ONE* 6, e23294+.

Heger, A., and Holm, L. 2001. Picasso: generating a covering set of protein family profiles. *Bioinformatics* 17, 272–279.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89, 10915–10919.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195–202.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14, 846–856.

Kolinski, A. 2004. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.* 51, 349–371.

Kolinski, A., and Gront, D. 2007. Comparative modeling without implicit sequence alignments. *Bioinformatics* 23, 2522–2527.

Lipman, D.J., and Pearson, W.R. 1985. Rapid and sensitive protein similarity searches. *Science* 227, 1435–1441.

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.

Murzin, A.G., Brenner, S.E., Hubbard, T., et al. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Oehmen, C., and Nieplocha, J. 2006. ScalaBLAST: A Scalable Implementation of BLAST for high-performance data-intensive bioinformatics analysis. *IEEE Trans. Parallel Distrib. Syst.* 17, 740–749.

Ohlson, T., Wallner, B., and Elofsson, A. 2004. Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* 57, 188–197.

von Ohsen, N., Sommer, I., and Zimmer, R. 2003. Profile-profile alignment: a powerful tool for protein structure prediction. *Pac. Symp. Biocomput.* 252–263.

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.

Panchenko, A.R. 2003. Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.* 31, 683–689.

Pietrokovski, S. 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* 24, 3836–3845.

Rychlewski, L., Jaroszewski, L., Li, W., et al. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* 9, 232–241.

Sadreyev, R., and Grishin, N. 2003. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* 326, 317–336.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.

Tatusov, R.L., Altschul, S.F., and Koonin, E.V. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. USA* 91, 12091–12095.

Wang, G., and Dunbrack, R.L. 2004. Scoring profile-to-profile sequence alignments. *Prot. Sci.* 13, 1612–1626.

Yona, G., and Levitt, M. 2002. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* 315, 1257–1275.

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

Address correspondence to:
Dr. Dominik Gront
Faculty of Chemistry
Warsaw University
Pasteura 1
02-093 Warsaw, Poland

E-mail: dgront@gmail.com