ISSN 0026-8933, Molecular Biology, 2006, Vol. 40, No. 3, pp. 474–480. © Pleiades Publishing, Inc., 2006.
Original Russian Text © I.I. Litvinov, M.Yu. Lobanov, A.A. Mironov, A.V. Finkelshtein, M.A. Roytberg, 2006, published in Molekulyarnaya Biologiya, 2006, Vol. 40, No. 3, pp. 533–540.

MATHEMATICAL AND SYSTEM BIOLOGY

UDC 577.1

Information on the Secondary Structure Improves the Quality of Protein Sequence Alignment

I. I. Litvinov[a], M. Yu. Lobanov[b], A. A. Mironov[c], A. V. Finkelshtein[b], and M. A. Roytberg[a, d]

[a] *Institute of Mathematical Problems in Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia; e-mail: mroytberg@mail.ru*

[b] *Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*

[c] *Department of Bioengineering and Biotechnology, Moscow State University, Moscow, 119992 Russia*

[d] *Pushchino State University, Pushchino, Moscow Region, 142290 Russia*

Received January 10, 2006

Abstract—The most popular algorithms employed in the pairwise alignment of protein primary structures (Smith–Watermann (SW) algorithm, FASTA, BLAST, etc.) only analyze the amino acid sequence. The SW algorithm is the most accurate, yielding alignments that agree best with superimpositions of the corresponding spatial structures of proteins. However, even the SW algorithm fails to reproduce the spatial structure alignment when the sequence identity is lower than 30%. The objective of this work was to develop a new and more accurate algorithm taking the secondary structure of proteins into account. The alignments generated by this algorithm and having the maximal weight with the secondary structure considered proved to be more accurate than SW alignments. With sequences having less than 30% identity, the accuracy (i.e., the portion of reproduced positions of a reference alignment obtained by superimposing the protein spatial structures) of the new algorithm is 58 vs. 35% of the SW algorithm. The accuracy of the new algorithm is much the same with secondary structures established experimentally or predicted theoretically. Hence, the algorithm is applicable to proteins with unknown spatial structures. The program is available at ftp://194.149.64.196/STRUSWER/.

DOI: 10.1134/S0026893306030149

Key words: protein, amino-acid sequence, alignment, alignment accuracy, secondary structure

INTRODUCTION

Many problems of computer analysis of proteins require pairwise alignments of their amino acid sequences. The ideal ultimate goal of all alignment algorithms is to generate a biologically correct alignment, reflecting the evolutionary history of homologous proteins [1]. An aligned position of two proteins corresponds, in this case, to the same position in the sequence of their common ancestor. Yet such a biologically correct alignment is unknown to us. A possible approximation is an alignment obtained by superimposing protein spatial structures, which are conserved to a far greater extent as compared with primary sequences [2]. Hence, an alignment resulting from superimposition of spatial structures was used as a reference in this work. The quality of an algorithmic alignment of amino acid sequences (i.e., its similarity to the reference alignment) is critical when spatial structures are modeled by homology [3] or identified on the basis of known structures of other proteins [4], protein domains are analyzed [5], or the function is studied for particular regions of a protein [6].

The quality of algorithmic alignments is high only when the amino acid sequences are sufficiently similar. It has been found, for instance, that the accuracy of the alignment by the Smith–Waterman (SW) algorithm is 84% when the protein identity (i.e., the portion of identical positions in two proteins) is not less than 30%; when the identity is lower, the alignment accuracy is about 30% [7]. As already mentioned the reference alignments are obtained by superimposition of the spatial structures; the alignment quality is the portion of the reference alignment positions that are reproduced algorithmically (see Experimental for more detail). Rapid approximate alignment algorithms, such as BLAST and FASTA, are even less accurate. The accuracy of BLAST alignments is 26% with sequences having less than 30% identity and 81% with sequences having more than 30% identity.

Thus, it is of interest to develop methods improving the reliability of analyses of homologous proteins. One way for such an improvement is to consider the physico-chemical properties of a protein, e.g., its secondary structure. First, the secondary structure (as part of the spatial structure) is far more conserved as compared to the primary sequence [8]. Second, efficient methods are available to theoretically predict the secondary structure. The first methods predicting the

secondary structure were developed in the 1970s and the early 1980s and utilized only the primary sequence of the protein under study [9–13]. The accuracy of the predictions with early methods was limited as only few spatial structures had been resolved at that time. With the accumulation of known protein sequences and, more importantly, spatial structures, the accuracy of the predictions was significantly improved. Currently, the best methods (e.g., PSIPRED [14] or Jpred [15]) employ neuronal networks for predicting secondary structures. Neuronal networks are trained with the use of known structures, established experimentally. The general picture of the current methods predicting the secondary structure of proteins is available from the EVA server (http://cubic.bioc.columbia.edu/eva/).

The use of the secondary-structure data for aligning amino acid sequences has been considered in the works of bioinformatics since the mid-1990s. Most attention has been focused on detecting distant homologies via sequence alignment, while only fragmentary data had concerned the quality of algorithmic alignments (e.g., see [16]). There are two main approaches of using the secondary structure for comparing amino acid sequences (in particular, for determining the spatial folding of a protein chain). One approach considers the secondary structure by its self. Works have focused on the algorithmic alignment of secondary structures [17, 18], a search for secondary-structure patterns with a subsequent selection of structures with particular physical properties (compact models, similar hydrophobicity, etc.) [19], and the construction of a hidden Markov model (HMM) for protein sequences on the basis of secondary structures [20, 21]. The other approach uses both the secondary structure and the primary sequence [16, 22–28]. Wallqvist et al. [27] and An and Friesner [28] used a linear combination of the amino acid component of the alignment weight and its structural component. To compare the secondary structure, a matrix was constructed with the 3D_Ali spatial alignment databank [29] (see Experimental). Method [28] differs from method [27] in that "improper" structures are excluded during the preliminary selection of the possible homologs and that secondary-structure elements are classed by size (e.g., short helix–long helix).

The objective of this work was to assess the quality (i.e., accuracy and confidence, see Experimental) of alignments constructed with the use of secondary-structure data. We developed and implemented the STRUSWER algorithm, which utilizes an additional bonus for the collation of similar secondary structures. Secondary structures can be determined both experimentally and theoretically. It should be noted that, in contrast to our work, previous studies have been aimed at optimizing the search for homologous proteins in amino acid sequence databases.

Ideologically, our method is most similar to the Wallqvist–Fukunishi–Murphy–Fadel–Levy algorithm (WFMFL) [27]. The major differences are (1) that WFMFL utilizes a special secondary-structure similarity weight matrix, which is obtained by analyzing spatial alignments, while STRUSWER uses only one parameter (bonus for collation of secondary structures) to take the secondary structure into account, and (2) that STRUSWER utilizes relative, rather than absolute, predictions of the secondary structure (i.e., the liability to every particular type of structures is indicated for each residue).

EXPERIMENTAL

Secondary structure: Experimental data. Protein secondary structures determined experimentally were extracted from the DSSP database [30], which utilizes the spatial coordinates of proteins. The eight types of secondary structures, which are used in DSSP, were reduced to three commonly accepted types (H (helix), E (β-strand), and L (loop)) according to the following scheme: (H, G, I) $\longrightarrow$ H; (E, B) $\longrightarrow$ E; and (T, S, blank) $\longrightarrow$ L, where H is an α-helix, G is the DSSP 3/10 helix, I is the DSSP π-helix, E is a β-structure element, B is the DSSP single β-bridge, T is the DSSP sharp turn, S is the DSSP turn that is not stabilized by hydrogen bonding, and blank suggests that the secondary structure has not been determined.

Secondary structure: Predictions. To predict the secondary structure, the PSIPRED program [14] was used in two modes, prediction of the structure for a group of homologous proteins (full version) and prediction of the structure from the amino acid sequence alone (single version). The prediction accuracy with our database was 82 and 65%, respectively, which agrees with the results available from the EVA server (http://cubic.bioc.columbia.edu/eva/). With each version we used two representations of the predicted secondary structure. In one case, structure_type, a particular secondary-structure symbol (H, helix; E, β-structure; or L, loop) was ascribed to every residue of an amino acid sequence. In the other, structure_probability, the probability of belonging to one of the three secondary-structure types was computed for every residue with the PSIPRED program. The methods used were abbreviated as follows: Exp, the structure was established experimentally and extracted from DSSP; PSI_S, the structure was predicted by homology (PSIPRED, full version) and presented in the structure_type form; PSI_%, the structure was predicted by homology (PSIPRED, full version) and presented in the structure_probability form; SIN_S, the structure was predicted from the amino acid sequence (PSIPRED, single version) and presented in the structure_type form; and SIN_%, the structure was predicted from

the amino acid sequence (PSIPRED, single version) and presented in the structure_probability form.

Reference structure alignments. As a reference we used spatial alignments of proteins available from BAliBASE [31]. BAliBASE contains multiple protein domain alignments constructed on the basis of spatial structure alignments and checked by experts. The use of BAliBASE as a test sample was adequate for our objective, to improve the quality of aligning homologous proteins. Experiments were performed with the first set of alignments (Reference 1) from BAliBASE. This set was chosen owing to its universal character; the set includes families of equidistant proteins with a mean sequence identity of 10–50%. The other sets were not used, because BAliBASE is intended for testing multiple sequence alignment algorithms and all sets except Reference 1 include special protein families (only short proteins, transmembrane proteins, highly homologous proteins, etc.; see http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE2/). The set of reference protein pairs included all pairs that met the following two requirements: a pair of proteins belongs to one multiple sequence alignment extracted from BAliBASE Reference 1 and the spatial structure is known for both proteins. In total, 576 pairs of proteins were obtained. Of these, 368 pairs had less than 30% identity. The resulting reference database was divided into training and test sets for correct comparisons of the methods. The training set included all even-numbered (in our list) reference pairs, and the test set, all odd-numbered pairs.

Evaluation of the alignment quality. To compare two alignments (algorithmic and reference ones) and to estimate the agreement between them, we used two parameters, accuracy and confidence.

The alignment accuracy (Acc) was defined as a ratio of the number of positions (I) aligned similarly in reference and algorithmic alignments to the number of aligned positions in the reference alignment (G): Acc = I/G.

The alignment confidence (Conf) was defined as a ratio of the number of positions aligned similarly in reference and algorithmic alignments to the number of aligned positions in the algorithmic alignment (A): Conf = I/A.

An alignment algorithm utilizing the secondary-structure data. Our algorithm is a modification of the SW algorithm. The only difference is that correlation of the i-th amino acid residue of one sequence with the j-th residue of the other involves computation of a bonus as the coefficient SBON, which defines the contribution of the secondary structure, multiplied by the secondary-structure similarity. The complete recursive equations are given below:

$$W(i,j) = \max\begin{cases} W(i-1,j-1) + M(a_i, b_i) + \mathrm{SBON} \times Q(i,j) \\ W(i-1,j) - \mathrm{GOP} - \mathrm{GEP} \\ WA(i-1,j) - \mathrm{GEP} \\ W(i,j-1) - \mathrm{GOP} - \mathrm{GEP} \\ WB(i,j-1) - \mathrm{GEP} \\ 0, \end{cases}$$

$$WA(i,j) = \max\begin{cases} W(i-1,j) - \mathrm{GOP} - \mathrm{GEP} \\ WA(i-1,j) - \mathrm{GEP}, \end{cases}$$

$$WB(i,j) = \max\begin{cases} W(i,j-1) - \mathrm{GOP} - \mathrm{GEP} \\ WB(i,j-1) - \mathrm{GEP}. \end{cases}$$

In these equations a and b are the first and the second protein chains under study; a_i and b_j are the i-th and the j-th residues in chains a and b; $W(i, j)$ is the weight of the best alignment of the initial fragment $a[1\ldots i]$, which includes residues $1 - i$ of sequence a, and the initial fragment $b[1\ldots j]$, which includes residues $1 - j$ of sequence b; $WA(i, j)$ is the weight of the best alignment of the fragments $a[1\ldots i]$ and $b[1\ldots j]$, in which the last residue i in $a[1\ldots i]$ is not correlated with any residue of $b[1\ldots j]$; $WB(i, j)$ is the weight of the best alignment of the fragments $a[1\ldots i]$ and $b[1\ldots j]$, in which the last residue j in $b[1\ldots i]$ is not correlated with any residue of $a[1\ldots j]$; $M(a_i, b_j)$ is the weight of the correlation of amino acid residues according to the substitution matrix used (in this work, we used blosum62 [32]); SBON is the coefficient defining the contribution of the secondary structure to the alignment; and $Q(i, j)$ is the function characterizing the similarity of the secondary structures of residues i and j in chains a and b. When structure types $T_a(i)$ and $T_b(j)$ are ascribed to amino acid residues i and j, then

$$Q(i,j) = \begin{cases} 1, \text{ if } \{[T_a(i) = T_b(j) = 'H'] \\ \text{or } [T_a(i) = T_b(j) = 'E']\} \\ 0 \text{ in other cases.} \end{cases}$$

When structure probabilities $Hp_a(i)$, $Ep_a(i)$, and $Lp_a(i)$ are ascribed to residue i, and $Hp_b(j)$, $Ep_b(j)$, and $Lp_b(j)$ to residue j, then

$$Q(i,j) = Hp_a(i) \times Hp_b(j) + Ep_a(i) \times Ep_b(j).$$

With the given gap penalties, the structural weight of the alignment differs from the SW weight only in using weight $M(a_i, b_j)$ + SBON*$Q(i, j)$ instead of the weight of correlation according to substitution matrix $M(a_i, b_j)$, which is used in the SW algorithm. In place of SBON*$Q(i, j)$, the WFMFL algorithm utilizes a cor-

relation matrix of secondary-structure elements (Table 1).

Parameter optimization of the program. Nine alignments were constructed for each pair of proteins from the training set and for each set of parameters:

(1) SW alignment (secondary structure disregarded);

(2) STRUSWER_SIN_S, which was a STRUSWER alignment with the secondary structure predicted using PSIPRRED (single version) with selection of a major secondary structure;

(3) STRUSWER_SIN_%, which was a STRUSWER alignment with the secondary structure predicted using PSIPRRED (single version) and the secondary-structure probabilities;

(4) WFMFL_SIN, which was a WFMFL alignment with the secondary structure predicted using PSIPRRED (single version) with selection of a major secondary structure;

(5) STRUSWER_PSI_S, which was a STRUSWER alignment with the secondary structure predicted using PSIPRRED (full version, see Experimental) with the selection of a major secondary structure;

(6) STRUSWER_PSI_%, which was a STRUSWER alignment with the secondary structure predicted using PSIPRRED (full version) and secondary-structure probabilities;

(7) WFMFL_PSI, which was a WFMFL alignment with the secondary structure predicted using PSIPRRED (full version) with a major structure presentation;

(8) STRUSWER_Exp, which was a STRUSWER alignment using the secondary structure determined experimentally; and

(9) WFMFL_Exp, which was a WFMFL alignment using the secondary structure determined experimentally.

Each of the algorithmic alignments constructed with these methods was compared with the reference alignment, using the above definitions of accuracy and confidence. The results obtained for all protein pairs of the training set (which included only even-numbered protein pairs), the accuracy, and confidence were averaged over all pairs to yield $\langle \text{Acc} \rangle = \langle I/G \rangle$ and $\langle \text{Conf} \rangle = \langle I/A \rangle$, respectively. The parameters were optimized by trial and error, changing SBON from 1 to 30, GOP from 4 to 20, and GEP from 1 to 7 with an increment equal to unity. Thus, the training set was analyzed $30 \cdot 17 \cdot 7 = 3570$ times. Depending on the purpose of the optimization, the parameters ensuring the highest accuracy or confidence were used to examine each method with the test set. It should be noted that all three parameters (SBON, GOP, and GEP) can be optimized only in the STRUSWER program (methods (2)–(3), (4)–(5), and (8)). Only GOP and GEP can be optimized in the other cases, because the WFMFL algorithm utilizes a fixed matrix for the correlation of the secondary-structure elements (Table 1) and the SW algorithm only compares amino acid sequences.

Table 1. Weight matrix used for correlation of the secondary structure elements in the WFMFL algorithm

	H	E	L
H	2	–15	–4
E	–15	4	–4
L	–4	–4	2

Algorithm testing. Each algorithm was tested using the test set (only including the odd-numbered pairs of proteins), with the parameters selected as a result of optimization. In addition to accuracy and confidence averaged over the total set, these parameters were computed for low-homologous pairs (identity < 30%).

RESULTS AND DISCUSSION

We estimated the accuracy and the confidence (Tables 1, 2), and selected the SBON, GOP, and GEP values to ensure the highest accuracy (Table 2) or highest confidence (Table 3). In addition, Table 4 characterizes the accuracy and confidence obtained for the WFMFL and SW algorithms with their standard parameters. Data are given for the total test set and for "twilight area" protein pairs, in which protein identity is less than 30%. Although optimization was not performed with a subset of low-homologous proteins, it is advantageous to consider such pairs separately in order to evaluate the algorithms as applied to such cases, which are common in practice. Optimization was performed differentially for three variants of the secondary-structure data: (i) the secondary structure predicted from the amino acid sequence, (ii) the secondary structure predicted using data on homologous proteins, and (iii) the secondary structure established experimentally (for more detail, see Experimental).

The best accuracy (Table 2) and confidence (Table 3) was observed for methods utilizing the secondary structures established experimentally (iii) and methods utilizing the secondary structures predicted using data on homologous proteins (ii). However, this result is mostly of a technical interest. Since the experimental secondary structure usually suggests a known spatial structure, it is better to use a program aligning proteins by their spatial structures in such cases. On the other hand, the tests performed with the secondary structures determined experimentally showed that the approximate threshold that can be achieved with the

Table 2. Accuracy (Acc) and confidence (Conf) of several alignment algorithms tested with the test set (after their parameters were optimized with respect to confidence with the use of the training set)

Algorithm	Bonus	GOP	GEP	Acc	Conf	Acc, ID < 30%	Conf, ID < 30%
SW	–	7	1	0.525	0.585	0.353	0.429
(i) Secondary structure predicted from the primary sequence							
STRUSWER_SIN_S	2	10	1	0.578	**0.622**	0.428	**0.482**
STRUSWER_SIN_%	7	8	2	**0.602**	0.618	**0.461**	0.477
WFMFL_SIN	Matrix	13	1	0.399	0.488	0.263	0.346
(ii) Secondary structure predicted with data on homologous proteins							
STRUSWER_PSI_S	8	9	1	0.659	0.683	0.546	0.573
STRUSWER_PSI_%	17	6	2	**0.683**	**0.695**	**0.579**	**0.589**
WFMFL_PSI	Matrix	16	1	0.631	0.672	0.503	0.560
(iii) Secondary structure established experimentally							
STRUSWER_EXP	8	10	1	**0.677**	**0.700**	**0.577**	0.601
WFMFL_EXP	Matrix	15	1	0.638	0.698	0.527	**0.602**

Note: The parameters bonus, GOP, and GEP were selected for each algorithm with a training set to achieve maximal accuracy (Acc). Here and in Tables 3 and 4, the results are shown for the total test set (288 pairs of proteins) and for the twilight area (182 pairs of proteins with an identity less than 30%). The algorithms are designated as in Experimental (Secondary structure: Predictions). The matrix utilized in the WFMFL algorithm is described in Experimental (Alignment algorithm utilizing the secondary structure data). The accuracy and confidence are given as fractions of unity. For each variant (i–iii) the highest value in a column is in bold.

Table 3. Accuracy (Acc) and confidence (Conf) of several alignment algorithms tested with the test set (after their parameters were optimized with respect to accuracy with the use of the training set)

Algorithm	Bonus	GOP	GEP	Acc	Conf	Acc, ID < 30%	Conf, ID < 30%
SW	–	20	6	0.380	0.706	0.189	0.607
(i) Secondary structure predicted from the primary sequence							
STRUSWER_SIN_S	1	15	7	0.433	**0.707**	0.246	**0.620**
STRUSWER_SIN_%	1	10	6	**0.458**	0.700	**0.262**	0.596
WFMFL_SIN	Matrix	19	7	0.314	0.646	0.158	0.535
(ii) Secondary structure predicted with data on homologous proteins							
STRUSWER_PSI_S	1	17	6	0.468	0.715	0.286	0.630
STRUSWER_PSI_%	1	14	6	0.465	**0.717**	0.282	**0.631**
WFMFL_PSI	Matrix	12	4	**0.606**	0.694	**0.467**	0.596
(iii) Secondary structure established experimentally							
STRUSWER_EXP	1	13	6	0.483	0.71	0.303	0.615
WFMFL_EXP	Matrix	15	7	0.553	**0.748**	**0.400**	**0.676**

Note: The parameters bonus, GOP, and GEP were selected for each algorithm with the training set to achieve maximal confidence (Conf).

Table 4. Accuracy and confidence of the WFMFL and SW algorithms with their standard parameters

Algorithm	Bonus	GOP	GEP	Acc	Conf	Acc, ID < 30%	Conf, ID < 30%
SW	–	7	1	0.525	0.585	0.353	0.429
WFMFL_SIN	Matrix	12	2	0.386	0.517	0.234	0.381
WFMFL_PSI	Matrix	12	2	0.620	0.664	0.490	0.551
WFMFL_EXP	Matrix	12	2	**0.632**	**0.697**	**0.520**	**0.603**

given method by using the secondary structure in addition to the primary sequence. Second, the use of homologs and their secondary structures, even predicted ones, to align a pair of proteins contradicts the idea of pairwise alignment. It seems more proper that the alignment obtained for two proteins with the use of homology-based predictions is compared with a multiple sequence alignment of the corresponding group of proteins. We intend to study this problem in the future. Thus, we consider the secondary-structure prediction from the amino acid sequence to be the main method. We used the single version of PSIPRED in this work. Although it is less accurate than the full version of PSIPRED, this method has several advantages. One is that the single version of PSIPRED does not involve a homology search. Consequently, the corresponding modes of the STRUSWER algorithm (STRUSWER_SIN_S and STRUSWER_SIN_%) can be used even when a homology search is unfeasible for some other reasons. Another advantage is a consequence of the first, as it takes less time to predict the secondary structure. This circumstance may be crucial for large-scale computational projects. Alignments constructed by the STRUSWER_SIN algorithm with secondary structures predicted from the primary sequences surpass similar alignments obtained with the SW or WFMFL_SIN algorithm both in accuracy (Table 2) and confidence (Table 3). The WFMFL_SIN algorithm showed the lowest accuracy, even when compared with the SW algorithm. A possible cause is that to compare secondary structures the WFMFL algorithm utilizes a matrix that is based on experimental data and, consequently, is more sensitive to the quality of predictions.

The WFMFL_EXP and WFMFL_PSI algorithms surpassed the SW algorithm in accuracy (the difference was 0.113 and 0.106, respectively) and were exceeded by the STRUSWER_EXP and STRUSWER_PSI algorithms. It is of interest that, with the secondary structure predicted by homology, the accuracy of the alignments was comparable or even higher than with the secondary structure established experimentally. Thus, the highest accuracy was observed for the STRUSWER_PSI_% algorithm, which utilizes the secondary structure presented in terms of probability. A comparison of the algorithms utilizing the secondary-structure states with those utilizing secondary-structure probabilities showed that the latter had a 2% higher accuracy. After optimization with respect to confidence, the WFMFL_EXP algorithm surpassed the STRUSWER_EXP algorithm, the WFMFL_PSI algorithm was exceeded by the STRUSWER_PSI_S and STRUSWER_PSI_% algorithms, and the WFMFL_SIN_S algorithm was exceeded by the STRUSWER_SIN_S and STRUSWER_SIN_% algorithms. All relationships remained the same when the set was restricted to low-homologous proteins. The relative gain in accuracy and confidence increased substantially when the secondary-structure data were used (especially in the case of experimental secondary structures, although this case is hardly of any applied interest). The quality of WFMFL and SW alignments was only slightly lower when the programs were run with standard parameters (Table 4). However, we optimized their parameters to make the comparisons correct.

CONCLUSIONS

The use of the secondary structure considerably improves the quality of amino-acid sequence alignments. In the case of related sequences, alignments having a maximal weight when the secondary structure is considered are more accurate than alignments constructed using the SW algorithm. It is equally possible in this case to use the secondary structure established experimentally or predicted theoretically with the PSIPRED server. Thus, the method is applicable to proteins with unknown spatial structures. The STRUSWER algorithm surpasses its close analog WFMFL in alignment quality, and especially in accuracy. Further analysis is necessary to determine how these advantages can be used to improve the quality of searching databases.

ACKNOWLEDGMENTS

We are grateful to D. Frishman for fruitful discussion.

This work was supported by the Ludwig Institute of Cancer (CRDF RB0-1268); the programs Molecular and Cell Biology and Leading Scientific Schools; the Russian Foundation for Basic Research (project nos. 04-04-49682, 04-04-49438, 03-04-49369, and 02-07-90412); the RF Ministry of Industry, Science, and Technology (project nos. 20/2002, 5/2003); the program of the RF Ministry of Science and Education (project no. 02.434.11.1008); and grants from NWO (the Scientific Foundation of the Netherlands), ECONET (France), and the Howard Hughes Medical School (to M.S. Gelfand and A.V. Finkelshtein).

REFERENCES

1. Li W.H. 1997. *Molecular Evolution*. Sunderland: Sinauer Associates.
2. Lesk A.M. 2001. *Introduction to Protein Architecture*. Oxford, New York: Oxford Univ. Press.
3. Sanchez R., Sali A. 2000. Comparative protein structure modeling. Introduction and practical examples with modeller. *Methods Mol. Biol*. **143**, 97–129.
4. Jones D.T., Taylor W.R., Thornton J.M. 1992. A new approach to protein fold recognition. *Nature*. **358**, 86–89.
5. Bateman A., Birney E. 2000. Searching databases to find protein domain organization. *Adv. Protein Chem*. **54**, 137–157.

6. Bork P., Koonin E.V. 1998. Predicting functions from protein sequences: Where are the bottlenecks? *Nat. Genet*. **18**, 313–318.
7. Sunyaev S.R., Bogopolsky G.A., Oleynikova N.V., Vlasov P.K., Finkelstein A.V., Roytberg M.A. 2004. From analysis of protein structural alignments toward a novel approach to align protein sequences. *Proteins*. **54**, 569–582.
8. Russell R.B., Barton G.J. 1994. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol*. **244**, 332–350.
9. Ptitsyn O.B., Finkelstein A.V. 1970. Dependence of the secondary structure of globular proteins on their primary structure. *Biofizika*. **15**, 757–767.
10. Chou P.Y., Fasman G.D. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry.* **13**, 211–222.
11. Lim V.I. 1974. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J. Mol. Biol*. **88**, 857–872.
12. Garnier J., Osguthorpe D.-J., Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol*. **120**, 97–120.
13. Ptitsyn O.B., Finkelstein A.V. 1983 Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*. **22**, 15–25.
14. Jones D. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*. **292**, 195–202.
15. Cuff J.A., Barton G.J. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*. **34**, 508–519.
16. Fischer D., Eisenberg D. 1996. Protein fold recognition using sequence-derived predictions. *Protein Sci*. **5**, 947–955.
17. Sheridan R.P., Dixon J.S., Venkataraghavan R., Kuntz I.D., Scott K.P. 1985. Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. *Biopolymers*. **24**, 1995–2023.
18. Aurora R., Rose G.D. 1998. Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons. *Proc. Natl. Acad. Sci. USA*. **95**, 2818–2823.
19. Russell R.B., Copley R.R., Barton G.J. 1996. Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol*. **259**, 349–365.
20. Di Francesco V., Garnier J., Munson P.J. 1997. Protein topology recognition from secondary structure sequences: Application of the Hidden Markov Models to the alpha class proteins. *J. Mol. Biol*. **267**, 446–463.
21. Di Francesco V., Geetha V., Garnier J., Munson P.J. 1997. Fold recognition using predicted secondary structure sequences and Hidden Markov Models of protein folds. *Proteins*. **1**, 123–128.
22. Fischel-Ghodsian F., Mathiowitz G., Smith T.F. 1990. Alignment of protein sequences using secondary structure: A modified dynamic programming method. *Protein Eng*. **3**, 577–581.
23. Luthy R., McLachlan A.D., Eisenberg D. 1991. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*. **10**, 229–239.
24. Rice D., Eisenberg D. 1997. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol*. **267**, 1026–1038.
25. Rost B., Schneider R., Sander C. 1997. Protein fold recognition by prediction-based threading. *J. Mol. Biol*. **270**, 471–480.
26. Jaroszewski L., Rychlewski L., Zhang B., Godzik A. 1998. Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci*. **7**, 1431–1440.
27. Wallqvist A., Fukunishi Y., Murphy L.R., Fadel A., Levy R.M. 2000. Iterative sequence/secondary structure search for protein homologs: Comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics*. **16**, 988–1002.
28. An Y., Friesner R.A. 2002. A novel fold recognition method using composite predicted secondary structures. *Proteins*. **48**, 352–366.
29. Pascarella S., Milpetz F., Argos P. 1996. A databank (3D-ali) collecting related protein sequences and structures. *Protein Eng*. **9**, 249–251.
30. Kabsch W., Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. **22**, 2577–2637.
31. Bahr A., Thompson J.D., Thierry J.-C., Poch O. 2001. BAliBASE (Benchmark Alignment dataBASE): Enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res*. **29**, 323–326.
32. Henikoff S., Henikoff J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*. **89**, 10915–10919.