Molecular Biology, Vol. 38, No. 1, 2004, pp. 78–88. Translated from Molekulyarnaya Biologiya, Vol. 38, No. 1, 2004, pp. 92–103.

COMPUTER GENOMICS

UDC 577.2.08:681.3

Comparative Analysis of Nucleic Acid and Protein Primary Structures

M. A. Roytberg

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia
E-mail: roytberg@impb.psn.ru

Received August 15, 2003

Abstract—The review considers the original works on the primary structure of biopolymers carried out from 1983 to 2003. Most works were supported by the Russian program Human Genome and earlier similar Russian programs. Little-known publications of 1983–1993 and recent unpublished results are described in detail. In the field of genome comparisons, these concern the OWEN hierarchic algorithm aligning syntenic regions of two genome sequences. The resulting global alignment is obtained as an ordered chain of local similarities. Alignment of megabase sequences takes several minutes. The concept of local similarity conflicts is generalized to multiple comparisons. New algorithms aligning protein sequences are described and compared with the Smith–Waterman algorithm, which is now most accurate. The ANCHOR hierarchic algorithm generates alignments of much the same accuracy and is twice as rapid as the Smith–Waterman one. The STRSWer algorithm takes into account the secondary structures of proteins under study. With the secondary structures predicted using the PSI-PRED software for pairs of proteins having 10–30% similarity, the average accuracy of alignments generated by STRSWer is 15% higher than that achieved with the Smith–Waterman algorithm.

Key words: biopolymer primary structures, biological sequence alignment, 3D structure of proteins, genome comparison

INTRODUCTION

It is a great honor for me to be invited by the leaders of the Russian program Human Genome and to present a review in this special issue dedicated to the memory of Academician A.A. Bayev.

The Russian program Human Genome, which was led by Bayev for many years, created a favorable environment for my work and, I think, for the work of many of my colleagues. I became involved in primary structure analysis of biopolymers in the early 1980s by the offer of A.A. Molchanov, who headed our institute (at that time, the Research Computer Center (CRC) of the USSR Academy of Sciences). As far as I know, Molchanov initiated analysis of biological sequences in the institute by a directive of Bayev, who was Academic Secretary of the Division of Physicochemical Biology of the Academy of Sciences. At the same time, A.S. Kondrashov and several other researchers of the institute joined in working in the field, and collaboration started with other institutions, in particular, the Laboratory of Protein Physics (which was headed by Prof. O.B. Ptitsyn) of the Institute of Protein Research.

I focused on the primary structures (sequences) of biopolymers (nucleic acids and proteins), and the major objectives were to develop methods for their comparative analysis and to employ these in solving biological problems.

By the early 1980s when the first databases of DNA and protein sequences became available, most necessary algorithms had already been prepared. At that time, biological applications were considered similar to technical ones, the same algorithms were used to compare (or "to align") biological sequences and to search for failures in file storage, and no difference was made between comparisons of nucleotide or amino acid sequences. Characteristically, a book containing probably the first review on bioalgorithmics was titled *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (D. Sankoff and J.B. Kruskal, Eds., 1983). Yet it came to be understood quite soon that analysis of biological sequences has its specifics concerning first and foremost the formulation of problems. The progress in mathematical analysis of biological sequences was due to two factors: (1) the problems received a better formal definition and were brought into an agreement with essential biological problems and peculiarities of the subjects under study, and (2) new mathematical ideas were employed. To illustrate the progress in comparative sequence analysis, it is possible to mention the introduction of position-specific scoring matrices (PSSM) and development of methods for their construction and employment in searching for

0026-8933/04/3801-0078

related sequences [2, 3]. Some algorithms, which were only of theoretical importance earlier, came into use as computer speed and memory increased several orders of magnitude in the past 20 years. It should be noted that the International Human Genome Project owes its achievements not only to the progress in sequencing techniques, but also to development of the hardware and software suitable for dealing with sequences sized tens or hundreds of megabases and for maintaining databases with billions of monomers.

My first work, which was influenced by fruitful discussions with A.S. Kondrashov, aimed at developing sequence alignment procedures that are efficient and allow due consideration of the biological specifics of the sequences under study [4]. I proposed a new definition for the problem of aligning biological sequences and a new algorithm to solve it. The main result was isolation of a class of gap-weighting functions, which allowed construction of an efficient algorithm of aligning two symbol sequences. More precisely, deletion weights in an alignment were given by the following set of functions (where L is the deletion length, X is a symbol context of a fixed size at the flank of a deletion, and $i = 1, 2$ is the number of a sequence):

$S_i(L) + T_i(X)$ is the penalty for a gap at the start of sequence i,

$F_i(L) + G_i(X)$ is the penalty for a gap at the end of sequence i, and

$D_i(L) + B_i(X)$ is the penalty for an internal gap in sequence i.

In this set, functions $S_i(L)$, $T_i(X)$, $F_i(L)$, $G_i(X)$, and $B_i(X)$ are arbitrary, whereas function $D_i(L)$ is fragmentarily linear.

Algorithm LINNEUS was proposed for this set of weighting functions. The time it takes for the algorithm to construct an optimal alignment is proportional to $k1 \cdot k2 \cdot L1 \cdot L2$, where $L1$ and $L2$ are the lengths of the sequences under study and $k1$ and $k2$ are the numbers of linearity intervals in functions $D1$ and $D2$, respectively. The memory required for the algorithm is approximately proportional to $m1 \cdot L1$, where $m1$ is the start of the last linearity interval of function $D1$. The fact that the required memory is approximately proportional to the length of the shortest sequence under study, rather than to the product of sequence lengths, was of immense importance in the early 1980s. Further development of bioalgorithmics showed that the algorithm is excessively generalized. In particular, arbitrary weighting functions for terminal gaps are not used in modern alignment programs (i.e., only two variants are possible: the penalty is equal to that for an internal gap or to zero). Although biologically justified in some cases, the context dependence of the penalty is omitted. Functions with a single linearity interval are used in place of fragmentarily linear ones. Mathematical generalization unnecessarily high for real biological objects was characteristic of works of that time. The above result was published only as a preprint (another sign of the time) and was unknown beyond the Soviet Union. In 1988, Myers and Miller [5] independently considered a similar set of weighting functions and a corresponding alignment algorithm.

Another algorithm, KARL, was developed almost simultaneously, was substantially faster, but operated with a far narrower set of weighting functions [6]. The algorithm was described in detail in [7]. These works were supplemented with algorithms detecting all local similarities in at least two sequences [8–12]. The relevant results were reported at the First All-Union Conference “Human Genome” (Pereslavl-Zalesskii, October 1990) and described in more detail in [12].

The above programs, along with those for standard (e.g., statistical) analyses of sequences, were included into the SAMSON open software package, which was developed under the guidance of A.S. Kondrashov [13–16]. An important role in its development was played by high programming and mathematical culture, which was characteristic of CRC at that time owing to its founders, A.M. Molchanov and E.E. Shnol’. In particular, it was of great importance for me to collaborate with V.V. Levitin who headed a programming group and, more recently, a laboratory (e.g., see [17]). The programs devised were used in numerous works (e.g., see [18–20]).

In the mid-1980s, interlaboratory seminars were held by O.B. Ptitsyn and E.E. Shnol’, and was valuable for me and, I think, for all other participants. In 1985, Ptitsyn and Shnol proposed that A.V. Finkelstein and I should prepare a comprehensive review of the mathematical methods of biopolymer analysis. Working on the review, we devised a general procedure to describe various methods of biopolymer analysis (a method of dynamic programming for oriented graphs on semirings, see [21] for relevant terms). In particular, the method allows a single viewpoint on the algorithms computing the partition function for a given molecule and those computing its optimal structure. On the basis of the known algorithm predicting the optimal secondary structure of RNA molecules, we extended the procedure to hypergraphs and constructed an algorithm to compute the corresponding statistical sums (the algorithm was not published; more recently, it was developed independently [22]). The review served as a basis for our presentation at the Fourth All-Union Conference “Mathematical Methods for Polymer and Biopolymer Research” [23] and a more recent publication [24].

In the past ten years, I focused on three major subjects: prediction of the exon–intron structure in higher eukaryotic genomes, analysis of long genomic DNA sequences (segmentation into statistically homoge-

nous regions, genome alignment), and comparison of amino acid sequences.

M.S. Gelfand invited me to work on recognition of coding sequences. Most of my works in the field were done in collaboration with him as well as with T.V. Astakhova, A.A. Mironov, and P.A. Pevzner [25–32]. Analysis of statistically homogeneous regions of the genome [33–35] was initiated by V.Yu. Makeev, V.E. Ramensky, and V.G. Tumanyan (Engelhardt Institute of Molecular Biology). The results were published in Russia and abroad and reported at conferences on the Russian program Human Genome [36–45]. Some of the relatively recent results concerning comparative analysis of biological sequences are considered below.

GENOME COMPARISON

The postgenomic era, which was called inevitable at almost every conference on the Russian program Human Genome, has come at last. Most of the problems formulated in general detail early in the program are now defined more accurately, concretized, and, in some cases, separated from one another. In particular, this concerns comparative genomics.

It became clear that comparison of prokaryotic genomes (which mostly consist of coding sequences) and comparison of eukaryotic genomes (in which the gene portion is low) are problems substantially differing from each other. Another important factor is similarity of genomes. Three similarity levels may be recognized: (1) high, when similarity is observed for the whole genome, including both coding and noncoding regions (e.g., human–chimpanzee); (2) intermediate, when genes are highly homologous while noncoding regions have only isolated local similarities (e.g., human–mouse); and (3) low, when gene homology is no more than 30% (e.g., human–fish).

Since 2000, we have collaborated with the group headed by A.S. Kondrashov (who works now in NCBI, United States), and focused on comparisons of intermediately or low-similar eukaryotic genomes. Such genomes contain extended syntenic regions, which have one order of genes [46]. Syntenic regions are several tens of megabases in size in the pair human–mouse. Substantial genome collinearity was observed for vertebrates [47] and for flowering plants [48].

Our approach to solution of the above problem is based on the following observations. First, some local similarities are statistically valid at the level of full-length syntenic genome regions. Most local similarities found in orthologous intergenic regions also take one and the same order and are thereby collinear. Second, “conflicts” of local similarities (i.e., their incompatibility in one alignment) do exist, and it is their resolution that constitutes the major problem of genome alignment. Third, many conflicts are associated with evolutionary events (e.g., those related to transposons or microsatellites), which render the basic statistical model and relevant quality parameters inapplicable. In many cases, such situations may be recognized and filtered out [49].

We proposed and programmed a hierarchic method of genome aligning. Its major principles are the following.

(1) Genome regions to be compared are treated hierarchically. First, a backbone chain of collinear (lacking conflicts with each other) local similarities is constructed. Each similarity included in the backbone chain is statistically valid. The backbone chain is maximal and cannot be augmented with additional statistically valid local similarities. False local similarities (transposons, microsatellites) are recognized and “removed” (masked) before analysis, and are never included in the backbone chain.

(2) A local resolution is provided to conflicts arising between statistically valid local similarities during construction of the backbone chain; i.e., accepted is the similarity having a better significance level (*P*-value). This approach radically differs from the common one (choosing a local alternative by weighting a global alignment), which is often used in the case of protein and relatively short nucleotide sequences. Our approach was grounded in detail elsewhere [50]. Rejected similarities are stored as comments to the major alignment and are accessible for analysis.

(3) When the backbone chain has been constructed, the problem is reduced to aligning regions located between two chosen local similarities. Since these regions are commonly far shorter than the initial sequences, the same significance level may be achieved with lower weights of local similarities. Then the above procedure is hierarchically applied to pairs of relatively short sequences generated at the previous step.

This approach is implemented in the OWEN system [51]. The system operates both in the interactive and in the package mode. With a standard personal computer, it takes several minutes to align two sequences of 10^6 units each. The required memory does not exceed $20L$ bytes, where L is the total length of the sequences to be compared.

The approach may be generalized to multiple genome comparisons (this work is in progress now). Multiple genome comparison is among the most topical problems of modern computational molecular biology. As in pairwise comparisons, we intend to construct a multiple alignment as a chain of nonconflicting local similarities, which may be found in several, though not necessarily all, sequences under study. Again, the hierarchic approach is used to construct a backbone chain of local similarities, and local resolution provided to conflicts of local similarities. It should be noted that the term conflict needs a better

definition. Since a local similarity may concern only some genomes, a situation is possible when three particular local similarities cannot be simultaneously included in the final alignment (triple conflict), while any two of these can (i.e., double conflicts are absent). A definition of the conflict of multiple local similarities is given in Appendix 1.

PROTEIN COMPARISON

Methods of protein comparison have been analyzed by our group in the past five years. The objectives are to study how well algorithmic and biologically correct alignments correlate with each other (in particular, to define a biologically correct alignment) and to develop new alignment procedures surpassing the existing ones in "quality" (a correlation between algorithmic and biologically correct alignments) and efficiency. Several approaches were analyzed. One is based on estimating the quality of alignments with multiple criteria. The standard scalar evaluating function with affine penalties for gaps is a linear combination of the summed weights of aligned symbols and of the number and lengths of gaps. In this combination, the coefficients are penalties for gap opening and elongation. We proposed that a vector should be used to characterize alignment *A*. The above linear combination corresponds to characteristic

$$\begin{aligned}&\text{VectorScore}(A)\\ &= (\text{Subst}(A), -\text{NumGap}(A), -\text{LenGap}(A)),\end{aligned} \quad (*)$$

where Subst(*A*) is the total weight of symbols matching each other in alignment *A*, NumGap(*A*) is the number of gaps, and LenGap(*A*) is the total length of gaps.

Other vector characteristics are also possible (relevant results were reported at the Third Conference on the Russian program Human Genome [37] and descried in detail in [52]).

Alignment *B* is considered dominating over alignment *A* when at least one component of the vector characteristic of *B* is better than in *A* and all other components are no worse than in *A*. Alignment *A* is termed Pareto-optimal when no other alignment *B* dominates over *A*. We proposed and realized an algorithm constructing all Pareto-optimal alignments for the given two sequences [52].

In fact, it is Pareto-optimal alignments that agree with the intuitive idea of a "nondegenerate" alignment, which is apparently no worse than another one. Pareto-optimal alignments are analogous to those optimal with respect to the scalar evaluating function. An alignment that is optimal with given substitution matrix *M* and given gap opening and gap elongation penalties is Pareto-optimal with respect to vector function (*). Hence the problem of choosing "correct" penalties for gaps is thereby transformed into the problem of selecting a "biologically correct" alignment among Pareto-optimal ones.

This problem has not been completely solved so far. A way to its solution was proposed in [53, 54], and is based on the following observation. Consider two-component evaluating vector function

$$\text{VectorScore2}(A) = (\text{Subst}(A), -\text{NumGap}(A)), \quad (**)$$

where Subst(*A*) and NumGap(*A*) are the same as in function (*). It is clear that any alignment *A* that is Pareto-optimal with respect to evaluating function (**) has vector characteristic

$$\langle \text{MaxSubst}(n), -n \rangle,$$

where *n* is the number of gaps in alignment *A* and MaxSubst(*n*) is the maximal total substitution weight possible for such alignments of sequences *S*1 and *S*2 with no more than *n* gaps.

Assume that removal of a set of *K* fragments from sequences *S*1 and *S*2 yields equally sized, highly similar sequences (i.e., this is a "correct" set of fragments to be removed). Then removal of any fragment of the set improves the match of extended regions of the initial sequences and, consequently, substantially increases total substitution weight MaxSubst(*i*), where $i = 1, \ldots, K$.

After all *K* fragments of the correct set are removed, deletion of any other fragment does not appreciably increase MaxSubst(*i*). In particular, when sequences resulting from "correct" removal coincide, a further increase in this parameter is impossible, and MaxSubst(*K*) = MaxSubst(*t*) for every $t > K$.

Thus, correct gap number *K* corresponds to a dramatic drop in parameter

$$\text{DelSubst}(i) = \text{MaxSubst}(i + 1) - \text{MaxSubst}(i).$$

The Pareto-optimal alignment corresponding to the drop in DelSubst(*i*) was termed critical. As observed in experiments with artificial test sequences and with several pairs of proteins showing no less than 30% similarity and having a reference alignment, the drop in DelSubst(*i*) does exist, and the critical alignment is a good approximation of a correct one [54].

Another line of our research is comparing amino acid sequence alignments generated via the Smith–Waterman algorithm (which is the most accurate to date) with gold standard alignments [55]. To be used as a gold standard, structurally adequate alignments of protein domains were sampled from BaliBase [56]. In total, the database contains 23 families (multiple alignments) of sequences sized several tens to several hundreds of residues, each family including about 15 domain sequences. Sequence similarity (%ID) varies from several to about 80%.

The agreement between algorithmic and gold standard alignments was characterized by accuracy (the portion of gold standard alignment pairs reproduced

Table 1. Comparison of the accuracy, confidence, and speed for the Smith–Waterman and Anchor algorithms

ID, %	Total protein pairs	Accuracy, %			Confidence, %			Computation time		
		Anchor	SW	An/SW	Anchor	SW	An/SW	Anchor	SW	An/SW
10–30	298	36.1 ± 31.4	35.0 ± 32.1	1.03	49.6 ± 35.5	48.6 ± 37.1	1.02	397.5	731.3	0.54
>30	253	83.2 ± 7.0	84.5 ± 6.6	0.98	89.1 ± 5.7	86.8 ± 6.6	1.03	158.8	393.4	0.4
All	583	39.8 ± 22.6	40.1 ± 24.7	0.99	54.5 ± 26.2	49.7 ± 27.5	1.1	275.7	552.3	0.5

Note: Test pairs of sequences were extracted from BaliBase. Designations: SW, Smith–Waterman algorithm; An, Anchor. Here and in Tables 2 and 3: Comparisons were made for the total sample (at the bottom) and for subsamples differing in the similarity of proteins in a pair.

Table 2. Comparison of the accuracy and confidence for alignments constructed with the Smith–Waterman algorithm (SW) and with our algorithm (STRSWer) and experimental secondary structure data

ID, %	Accuracy		Confidence	
	SW	STRSWer	SW	STRSWer
<10	0.037	0.256	0.076	0.310
10–30	0.306	0.502	0.470	0.521
30–40	0.818	0.852	0.864	0.873
>40	0.893	0.902	0.921	0.919
All	0.521	0.647	0.629	0.667

Note: Experimental secondary structure data were extracted from DSSP. Here and in Table 3: Reference alignments were obtained from BaliBase.

in the algorithmic alignment) and confidence (the portion of algorithmic alignment pairs coinciding with those in the gold standard alignment).

As our and published [57] data demonstrate, algorithmic and reference alignments virtually coincide at %ID > 30% and have almost nothing in common at %ID < 10%. In the intermediate region (10% < %ID < 30%), both the accuracy and the confidence of algorithmic alignment vary greatly. To study the cause of the variation, we analyzed the gap-free regions (islands) of alignments. We observed that up to 30% of islands have negative weight in reference alignments and, consequently, are unrecognizable by the Smith–Waterman algorithm or its analogs. In view of this, we proposed a new algorithm [55], which is as accurate and twice as rapid as the Smith–Waterman one (Table 1). The main idea was to avoid wasting time on scanning the Needleman-Wunsch matrix in low-similarity regions, which account for more than 90% of the matrix cells.

Recent works of our group directed the way to improvements in alignment quality [58]. The gist is taking advantage of the predicted secondary structure. Like the Smith–Waterman algorithm, our algorithm STRSWer recursively constructs similarity matrix W_{ij} (Needleman-Wunsch matrix). However, the recurrent equation differs in having bonus SBON for matching residues in regions similar in secondary structure. In our case, the equation is

$$W_{ij} = \max\begin{cases} W_{i-1,\,j-1} + d(a_i, b_j) + \mathrm{BONUS}[i, j]; \\ W_{i-1,\,j} - \mathrm{GOP} - \mathrm{GEP}; \\ W_{i,\,j-1} - \mathrm{GOP} - \mathrm{GEP}; \\ 0 \end{cases}$$

where $d(a_i, b_j)$ is the weight of a substitution of one amino acid residue for another according to the substitution matrix, BONUS[i, j] is the bonus for matching two residues of one secondary structure type. When structural types are similar, BONUS[i, j] = 0. When these are different, BONUS[i, j] = SBON, where SBON is a predetermined positive value. GOP is the gap opening penalty; when a gap is opened, GOP = 0. GEP is the gap elongation penalty; W_{ij} is the weight of the alignment of initial subsequences 1 and 2 sized i and j, respectively. With the BLOSUM62 substitution matrix, we established that SBON = 11.

Tables 2 and 3 demonstrate the gains in accuracy and confidence achieved with our algorithm as compared with the Smith–Waterman one. We used the secondary structures that were established experimentally (Table 2) or deduced from the amino acid sequence with the PSIPRED software [59]. The gains

Table 3. Comparison of the accuracy and confidence for alignments constructed with the Smith–Waterman algorithm (SW) and with our algorithm (STRSWer) and predicted secondary structure data

ID, %	Accuracy		Confidence	
	SW	STRSWer	SW	STRSWer
<10	0.037	0.197	0.076	0.226
10–30	0.306	0.482	0.470	0.503
30–40	0.818	0.856	0.864	0.878
>40	0.893	0.903	0.921	0.918
All	0.521	0.635	0.629	0.656

Note: Secondary structures were predicted from the amino acid sequences with the PSIPRED software.

are quite high even in the case of predicted secondary structures.

Another original algorithm comparing amino acid sequences was used to analyze the β-structure proteins, in particular, immunoglobulins [60]. A peculiarity of the problem is that certain well-known motifs ("words"), which correspond to β-strands, are contained in sequences to be compared. Description of protein families with words is similar to that with frequency profiles. In either case, a spectrum of possible amino acid residues is specified for every position. However, while a function summarizing symbol weights over all positions is used to evaluate the whole sequence in the case of profiles, the limitations are discrete in the case of word-dependent algorithms. The problem of amino acid sequence alignment with word templates is considered in detail in Appendix 2.

CONCLUSIONS

This review covers about 20 years of research. What has changed over this period in the field of comparative analysis of biological sequences?

The major change is probably a radical increase in data to be analyzed. For instance, nucleotide sequence databases increased two orders of magnitude in the recent decade. Genomic DNA fragments typically subject to analysis were several tens to several hundreds of kilobases in the early 1990s; now their size is several tens or hundreds of megabases. Moreover, the data changed qualitatively. At present, complete sequences are available for about 100 genomes (mostly bacterial ones). Comparative analysis is the main strategy in modern genomics.

Special databases were created to store the results of protein comparisons in the form of global (FSSP, BaliBase) or local (BLOCKS, PRINT) alignments. Whenever possible, sequences are supplemented with experimental data concerning, in particular, the spatial structure of proteins. A database of clusters of orthologous genes (COG), which was made in NCBI under the guidance of E.V. Koonin, came to play a prominent part in current research. The need to accumulate the alignment data was realized rather long ago (for instance, it was discussed at the international conference in Novosibirsk in 1990), yet the relevant databases were created only in the late 1990s, as their role in solving certain biological problems came to be better understood.

Owing to the above achievements, comparative sequence analysis is employed in solving most current problems of computational molecular biology. Three of these—recognition of the exon–intron structure, prediction of the spatial structure of proteins, and prediction of the secondary structure of proteins—were already mentioned. Less known applications of sequence comparisons are exemplified by mass spectrometry [61].

New application fields of sequence comparisons continue to pose new problems; some of these were considered above. Of the others, it is possible to mention spliced sequence alignment, which was proposed by M.S. Gelfand *et al.* [62] for identifying coding regions in eukaryotic genomes. Many sequence comparison problems are complex. For instance, the input data may contain not only the sequences to be compared, but also phylogenetic information, data on the spatial structure, margins of functional regions, etc. I think that "natural selection" of problems in comparative analysis of biological sequences is the second important result of the past years.

Homology search in databases is the most common application of sequence comparison. The major algorithmic ideas of sequence comparison per se—dynamic programming, a search for identical regions and the neighboring strong similarities, step-by-step minimization of a proper target function ("annealing")—were formulated as early as in the 1980s. In various combinations and with various heuristics, these ideas are now employed in all known methods of sequence comparison. Sparse dynamic programming is among the most interesting relatively recent findings [63]. A breakthrough in search quality was achieved in a few past years owing to two factors.

First, it became clear that the search results need careful statistical evaluation [64], and adequate methods were devised [65]. Second, the use was made of multiple comparison methods, in particular, iterative frequency profiling [66, 67]. In addition, a theory of generating substitution matrices was developed, which was also of immense importance.

Among all open problems of comparative sequence analysis, the most pressing are comparison (especially multiple one) of whole genomes and functional annotation of genomes with the use of sequence comparisons. In the coming years, the information on intraspecific variation will substantially increase; as a result, the existing problems will be more accurately formulated and new ones posed. As for protein sequence comparison, it seems most promising to utilize the external (relative to the amino acid sequence) information, such as the structural or evolutionary data.

ACKNOWLEDGMENTS

I am grateful to T.V. Astahova for help in preparation of this article. The relevant works were supported by the Russian Foundation for Basic Research (project nos. 03-04-49469, 02-07-90412) and by grants from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003) and NWO.

APPENDIX 1

Conflicts of Local Similarities in Multiple Genome Comparisons

Let there be t syntenic regions, each belonging to one of the t genomes to be compared.

Multiple local similarity (MLS) is fragment set $L = \{[a_1, b_1], \ldots, [a_t, b_t]\}$ such that

(1) some fragments may be empty (in the last case $a_k > b_k$); i.e., similarity L does not concern the k-th sequence; and

(2) any two nonempty fragments are similar to each other.

To define what is MLS conflict, it is necessary to introduce the term MLS graph.

Definition. Let S be a set of MLSs. Its graph $G(S)$ is defined as follows. Vertices of the graph are in a one-to-one correspondence with MLSs of set S. Graph $G(S)$ has an edge leading from $p1$ to $p2$ if and only if an initial sequence W_k meets the following requirements:

(1) both similarities $p1$ and $p2$ concern sequence W_k and

(2) the projection of $p1$ onto W_k is not located after the projection of $p2$ onto W_k.

Note. When the projections of $p1$ and $p2$ onto W_k intersect, both the rib leading from $p1$ to $p2$ and that leading from $p2$ to $p1$ are included into $G(S)$.

Definition. Set S of MLSs is conflicting if all MLSs cannot be included in one multiple alignment.

Statement. Set S of MLSs is conflicting if and only if graph $G(S)$ contains an oriented cycle.

Demonstration of the statement is simple, and is omitted here.

APPENDIX 2

Alignment of Amino Acid Sequences and Word Templates

Word templates provide a means to describe a family of similar proteins. Aligning a protein sequence vs. the template of a particular family, it is possible to check whether the given protein belongs to the family. Below word template is formally defined, the problem of sequence vs. template alignment formulated, and an algorithm to solve it discussed in brief.

1. Amino acid groups. Let there be amino acid classification G, that is, set $G = \{G_1, \ldots, G_n\}$ of n amino acid groups. Informally speaking, amino acids of one group are interchangeable with each other in one position of a protein. An amino acid may belong to several groups. This agrees with the fact that similarity of amino acids is not absolute, but rather depends on the amino acid position in a protein. In analysis of the β-structural proteins, amino acids were classed into seven groups as (1) hydrophobic (V, L, I, M, A, and C), (2) aromatic (Y, F, and W), (3) hydrophobic or aromatic (pooled groups (1) and (2)), (4) neither hydrophobic nor aromatic, (5) polar (R, K, E, D, Q, and N), (6) neutral (P, G, S, and T), and (7) nonpolar (other than R, K, E, D, Q, or N).

2. Words, multiwords, and their images. Let there be fixed alphabet $R = \{A, \ldots, W\}$ of amino acid residues and classification alphabet $A = \{a_1, \ldots, a_n, x\}$, where a_1 is an amino acid of group G_1 and 'x' is an arbitrary amino acid. Hereafter word is understood as word in the classification alphabet.

Arbitrary word w of length L corresponds to set $S(w)$ of amino acid sequences of length L. In other words, $S(w)$ harbors all amino acid sequences that may be derived from w by substituting a letter other than 'x' with an amino acid of the corresponding group and letter 'x', with any amino acid.

Thus, a set of sequences corresponding to a particular structural fragment of a multiple alignment may be described with a set of words of length L. This set is termed multiword of length L. If a multiword is $M = \{w_1, \ldots w_t\}$, then the corresponding set of sequences is a pool of sets $S(w_1), \ldots, S(w_t)$.

The correspondence between words and sequences may be defined more formally.

Let w be a word of length L and u be an amino acid sequence of the same length. We will consider that u and w have a mismatch in the i-th position if $w[i] \neq$ 'x' and $u[i]$ does not belong to the same group as $w[i]$ ($w[i]$ is the i-th symbol of word w).

We will consider u as a d-image of word w if u and w have no more than d mismatches.

Let $M = \{w^1 \ldots w^k\}$ is a multiword. We will consider u as a d-image of M if u is a d-image of word $w^j \in M$.

It is clear that every d-image of a word or a multiword is also its $(d + 1)$-image, $(d + 2)$-image, etc.

3. Word templates.

Definition. Word template of length N is a chain (an ordered set) of four-component elements $t^i = \langle m^i, d^i, r^i, q^i \rangle$ ($i = 1, \ldots, N$), where m^i is a multiword; d^i is an integer equal to the maximal allowable number of mismatches; r^i is an integer equal to the minimal distance to the image of the next, $(i + 1)$-th multiword in the amino acid sequence (see below); and q^i is an integer equal to the maximal distance to the next, $(i + 1)$-th multiword in the amino acid sequence (see below).

Let $P = \{\langle m^i, d^i, r^i, q^i \rangle \mid i = 1, \ldots, N\}$ be a word template of length N and S be an amino acid sequence.

Let $U = \{u_1 = S[x_1, y_1], \ldots, u_N = S[x_N, y_N]\}$ be a set of fragments of sequence S (x_i and y_i are respectively the first and last positions of the i-th fragment u_i).

Definition. Fragment chain U is an image of word template P with correspondence K if

(1) $r^I \leq x_{i+1} - y_i + 1 \leq q^i$ ($i \in \{1, \ldots, N - 1\}$) (informally speaking, the interval between u_i to u_{i+1} is within the interval between r^i and q^i), and

(2) there are exactly K indices i such that u_i is a d^i-image of multiword m^i ($i \in \{1, \ldots, N\}$).

4. Alignment of word templates. We focus on the problem of word template vs. amino acid sequence (WT–AS) alignment.

Problem 1. Given: Word template P of length N and amino acid sequence S of length L.

Required: Chain U consisting of N fragments of sequence S, which is an image of P with maximal possible correspondence K_{max}.

Note that problem 1 is analogous to the problem of global sequence alignment. Analogs of other sequence alignment problems (e.g., searching for all suboptimal local alignments) may be formulated and solved with a proper modification of the algorithm used to solve problem 1.

We introduced word templates to describe a set of aligned sequences and formalized the problem of WT–AS alignment for the purposes that are usually served by frequency profiles [2, 3]. The major differences between word templates and frequency profiles are the following. First, with word templates, we explicitly isolate functionally or structurally important fragments (multiword images) on the sequence, and count mismatches in each individual fragment. Second, we forbid gaps within multiword images, and limit the distance between words, rather than using standard gap penalties.

5. Algorithm to solve the problem of WT–AS alignment. This problem may be solved with a proper modification of dynamic programming. Such an algorithm is outlined below.

Let $t \in \{1, \ldots, N\}$, $i \in \{1, \ldots, L\}$; P_t be a word template consisting of the first t multiwords of template P; and S_i be the start of sequence S of length i.

Let $U = \{u_1 = S[x_1, y_1], \ldots, u_t = S[x_t, y_t]\}$ be a set of nonoverlapping fragments of S_i, with $x_1 < y_1 < \ldots < x_t < y_t$.

As the end of U, we will consider position y_t, which is the last position of the last fragment of U.

Definition. Let $F[t, i]$ be the maximal possible k for which chain $U = \{u_1, \ldots, u_t\}$ of fragments of sequence S exists and meets the following requirements:

(1) i is the end of U, and

(2) U is an image of P_t with correspondence k.

It is clear that the following recurrent equation is true for $F[t, j]$ (compare with the corresponding recurrent equations for sequence alignment [4, 5]):

$$F[t, j] = \max\{F[t-1, x] \mid j - 1^t - q^{t-1} \leq x \leq j - l^t - r^{t-1}\} + \mathrm{Curr}[t, j], \quad (1)$$

where l^t is the length of multiword m^t.

If fragment $S[j - l^t + 1, j]$ is a d^t-image of multiword m^t, $\mathrm{Curr}[t, j] = 1$; otherwise, $\mathrm{Curr}[t, j] = 0$ (end effects are disregarded here for simplicity).

An algorithm based on Eq. (1) solves the problem of WT–AS alignment in time

$$T_1 + T_2, \quad (2)$$

where T_1 is the total time of computing $\max\{F[t-1, x] \mid j - 1^t - q^{t-1} \leq x \leq j - l^t - r^{t-1}\}$, and T_2 is the total time of finding all d^i-images of multiword m^i in sequence S ($i = 1, \ldots, N$).

Note that independent methods may be used to compute $\mathrm{Curr}[t, j]$, which is responsible for T_2 in Eq. (2), and $\max\{F[t-1, x] \mid j - 1^t - q^{t-1} \leq x \leq j - l^t - r^{t-1}\}$, which is responsible for T_1 in Eq. (2). The method to compute $\mathrm{Curr}[t, j]$ depends on the properties of the given set of multiwords. For instance, multiwords of the cadherin family contain many 'x'. Hence it is optimal to compute $\mathrm{Curr}[t, j]$ by direct comparison, which yields inequality

$$T_2 < L \cdot H,$$

where H is the total number of non-'x' positions in all multiwords.

As for T_1, it is possible to demonstrate that

$$T_1 < c \cdot L \cdot N,$$

where c is a constant (demonstration is omitted here).

REFERENCES

1. Sankoff D., Kruskal J.B. 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Chicago: The Univ. of Chicago Press.
2. Eddy S.R. 1998. Profile hidden Markov models. *Bioinformatics*. **14**, 755–763.
3. Sunyaev S.R., Eisenhaber F, Rodchenkov I.V., Eisenhaber B., Tumanyan V.G., Kuznetsov E.N. 1999. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Engineering*. **12**, 387–394.
4. Roytberg M.A. 1984. *Algoritm opredeleniya gomologii pervichnykh struktur* (An algorithm to estimate the primary structure homology). Pushchino: NTsBI.
5. Myers E., Miller W. 1988. Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50**, 97–120.
6. Roytberg M.A. 1988. Fast algorithm for optimal aligning of symbol sequences. *Teoreticheskie issledovaniya i banki dannykh v molekulyarnoi biologii i genetike* (Theoretical studies and databases in molecular biology and genetics). Novosibirsk: Nauka, 69–70.
7. Roytberg M.A. 1992. Fast algorithm for optimal aligning of symbol sequences. In *Mathematical methods of the analysis of biopolymer sequences*. Providence: AMS, pp. 103–117.
8. Roytberg M.A. 1990. A Search for Common Patterns in Many Sequences. In *Proc. of the Workshop "Computer Applications in Biosciences"*. Novosibirsk: Nauka, p. 132.
9. Roytberg M.A. 1990. Similarity search in two biological sequences. *Int. Congr. Modelling and Computer Methods in Molecular Biology.* 1990. Novosibirsk: Nauka, pp. 7–8.
10. Roytberg M.A. 1990. A search for similar fragments in several sequences. *Trudy pervoi vsesoyuznoi konferentsii "Genom cheloveka" (Pereslavl'-Zalesskii, oktyabr' 1990 g.)* (Proc. 1st All-Union Conf. "Human Genome" (Pereslavl-Zalesskii, October 1990)). Moscow: Nauka. 209.
11. Roytberg M.A. 1992. Similarity Search in Biological Sequences. In *Modelling and Computer Methods in Molecular Biology and Genetics*. Eds. Ratner V.A., Kolchanov N.A. N.Y.: Nova Science Publishers, Inc., pp. 81–86.
12. Roytberg M.A. 1992. A Search for Common Patterns in many Sequences. *Comput. Appl. Biosci*. **8**, 57–64.
13. Vernoslov S.E., Kondrashov A.S., Roytberg M.A., Shabalina S.A., Yur'eva O.V., Nazipova N.N. 1989. Software package Samson for primary structure analysis of biopolymers. *Materialy po matematicheskomu obespecheniyu EVM. Seriya ES EVM* (Materials on computer software: Ser. ES computers). Pushchino: ONTI NTsBI. Part 1.
14. Vernoslov S.E., Kondrashov A.S., Roytberg M.A., Shabalina S.A., Yur'eva O.V., Nazipova N.N. 1989. Software package Samson for primary structure analysis of biopolymers. *Materialy po matematicheskomu obespecheniyu EVM. Seriya ES EVM* (Materials on computer software: Ser. ES computers). Pushchino: ONTI NTsBI. Part 2.
15. Vernoslov S.E., Kondrashov A.S., Roytberg M.A., Shabalina S.A., Yur'eva O.V., Nazipova N.N. 1990. Software package Samson for primary structure analysis of biopolymers. *Mol. Biol.* **24**, 524–529.
16. Nazipova N.N., Shabalina S.A., Ogurtsov A.Yu., Kondrashov A.S., Roytberg M.A., Buryakov G.V., Vernoslov S.E. 1995. "SAMSON: a software package for the biopolymer primary structure analyses". *Comput. Appl. Biosci*. **11**, 423–426.
17. Levitin V.V., Roytberg M.A. 1991. A program to compare biological sequences. *Trudy vtoroi vsesoyuznoi konferentsii "Genom cheloveka" (Pereslavl'-Zalesskii, oktyabr' 1991 g.)* (Proc. 2nd All-Union Conf. "Human Genome" (Pereslavl-Zalesskii, October 1991)). Moscow: Nauka, 159–160.
18. Shabalina S.A, Roytberg M.A., Kondrashov A.S.,Vernoslov S.E. 1990. Some characteristic features of 5'-regulatory regions of heat shock protein genes. In *Modelling and Computer Methods in Molecular Biology. Abstracts of the Int. Congr.* Novosibirsk: Nauka, pp. 23–24.
19. Matveeva O.B., Roytberg M.A., Shabalina S.A. Textual and statistical similarities in RNA nucleotide sequences. 1991. In *Abstracts of the Int. conf. "Protein biosynthesis"*. Pushchino, USSR, August 26–September 3. Pushchino: ONTI, p. 81.
20. Beridze T., Tsirekidze N., Roytberg M.A. 1992. On the tertiary structure of satellite DNA. *Biochimie*. **74**, 187–194.
21. Aho A.V., Hopcroft J.E., Ullman J.D. 1974. *The Design and Analysis of Computer Algorithms*. Reading, MA: Addison-Wesley.
22. McCaskill J.S. 1990. The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*. **26**, 1105–1119.
23. Finkelstein A.V., Roytberg M.A. 1985. Mathematical methods to analyze the primary structures of biopolymers: The spatial structures of biopolymers and their sequences. *Chetvertoe Vsesoyuznoe soveshchanie "Matematicheskie metody dlya issledovaniya polimerov i biopolimerov", 17–19 iyulya 1985 g.* (4th All-Union Conf. "Mathematical methods to study polymers and biopolymers." July 17–19, 1985). Pushchino: NTsBI.
24. Finkelstein A.V., Roytberg M.A. 1993. Computation of biopolymers: a general approach to different problems. *BioSystems*. **30**, 1–19.
25. Gelfand M.S., Roytberg M.A. 1993. A dynamic programming algorithm for prediction of the exon–intron structure. *BioSystems*. **30**, 173–182.
26. Gelfand M.S., Podolsky L.I., Astakhova T.V., Roytberg M.A. 1995. Prediction of the exon–intron structure and multicriterial optimization. In *Bioinformatics and Genome Research.* Eds. Lim H.A., Cantor C.R. Singapore: World Scientific Publ., Co., pp. 173–183.

27. Gelfand M.S., Podolsky L.I., Astakhova T.V., Roytberg M.A. 1996. Recognition of genes in human DNA sequences *J. Comput. Biol*. **3**, 223–234.

28. Roytberg M.A., Astahova T.V., Gelfand M.S. 1997. An algorithm for highly specific recognition of protein-coding regions in sequences of higher eukaryotes. *Mol. Biol.* **31**, 26–32.

29. Roytberg M.A., Astahova T.V., Gelfand M.S. 1997. Combinatorial approaches to gene recognition. *Comput. and Chem*. **21** (4), 229–236.

30. Sze S.-H., Roytberg M.A., Gelfand M.S., Astahova T.V., Mironov A.A., Pevzner P.A. 1998. Algorithms and software for support of gene identification experiments. *Bioinformatics*. **14**, 14–19.

31. Mironov A.A., Roytberg M.A., Pevzner P.A., Gelfand M.S. 1998. Performance guarantee gene predictions via spliced alignment. *Genomics*. **51**, 332–339.

32. Mironov A.A., Koonin E.V., Roytberg M.A., Gelfand M.S. 1999. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res*. **27**, 2981–2989.

33. Ramensky V., Makeev V., Gelfand M., Roytberg M., Tumanyan V. 2000. Bayesian approach to DNA segmentation into regions with different average nucleotide composition. In I*nformatique Mathematique JOBIM 2000*. Montpellier, France. 241–248.

34. Ramensky V.E., Makeev V.Ju., Roytberg M.A., Tumanyan V.G. 2000. DNA segmentation through the Bayesian approach. *J. Comput. Biol*. **7** (1–2), 215–231.

35. Ramensky V.E., Makeev V.Ju., Roytberg M.A., Tumanyan V.G. 2001. Segmentation of long genomic sequences into domains with homogeneous composition with BASIO software. *Bioinformatics*. **17**, 1065–1066.

36. Roytberg M.A. 1993. One more approach to sequence alignment: More similarities, less gaps, and no weighting coefficients. *Trudy konferentsii "Genom cheloveka-93"* (Proc. Conf. "Human Genome-93"). Moscow: Nauka, 136.

37. Podol'skii L.I., Roytberg M.A., Gelfand M.S. 1993. Prediction of the exon–intron structure and dynamic programming on distributive semirings. *Trudy konferentsii "Genom cheloveka-93"* (Proc. Conf. "Human Genome-93"). Moscow: Nauka, 135.

38. Roytberg M.A., Astahova T.V., Gelfand M.S. 1996. Reliable recognition of protein-coding regions and construction of oligonucleotide probes. *Trudy konferentsii "Genom cheloveka-96"* (Proc. Conf. "Human Genome-96"). Moscow: Nauka, 103–104.

39. Roytberg M.A., Semionenkov M.N., Fomin D.O. 1996. SimSim: An open system for biopolymer analyses. *Trudy konferentsii "Genom cheloveka-96"* (Proc. Conf. "Human Genome-96"). Moscow: Nauka, 102.

40. Gelfand M.A., Mironov A.A., Roytberg M.A., Pevzner P.A., Astahova T.V. 1998. A complex of programs to recognize the protein-coding regions in human DNA. *Vos'maya itogovaya konferentsiya "Genom cheloveka"* (8th Conf. "Human Genome"). Moscow: Nauka, 43.

41. Gelfand M.S., Roytberg M.A., Tverskaya S.M., Evgrafov O.V. 1998. Methods to identify and to analyze new genes on the basis of computer DNA analysis. *Vos'maya itogovaya konferentsiya "Genom cheloveka"* (8th Conf. "Human Genome"). Moscow: Nauka, 44.

42. Roytberg M.A., Petrova S.V., Astahova T.V., Kondrashov A.S. 2001. Recognition of the exon–intron structure by aligning DNAs of related organisms. *Sbornik otchetov po GNTP "GENOM CHELOVEKA-2000"* (Collection of reports on the Russian program Human Genome of 2000). Moscow: Nauka, 163.

43. Bogopolsky G.V., Vlasov P.K., Oleynikova N.V., Roytberg M.A., Sunyaev Sh.R. 2001. Detection of spatial structural similarity of proteins by comparing their amino acid sequences. *Sbornik otchetov po GNTP "GENOM CHELOVEKA-2000"* (Collection of reports on the Russian program Human Genome of 2000). Moscow: Nauka, 145.

44. Makeev V.Yu., Sunyaev Sh.R., Ramenskii V.E., Vlasov P.K., Bogopolsky G.A., Rogulenkova V.N., Esipova N.G., Roytberg M.A., Tumanyan V.G. 2002. Functional homology of genes lacking distinct homology. *Sbornik otchetov po GNTP "GENOM CHELOVEKA-2001"* (Collection of reports on the Russian program Human Genome of 2001). Moscow: Nauka, 45.

45. Kondrashov A.S., Ogurtsov A.Yu., Roytberg M.A., Tsitovich I.I. 2001. Searching for local similarity in genomic DNA with distinctly formulated statistical hypotheses. *Sbornik otchetov po GNTP "GENOM CHELOVEKA-2000"* (Collection of reports on the Russian program Human Genome of 2001). Moscow: Nauka, 43.

46. Zafar N., Mazumder R., Seto D. 2001. Comparisons of gene colinearity in genomes using GeneOrder 2.0. *Trends. Biochem. Sci*. **26**, 514–516.

47. Venkatesh B., Gilligan P., Brenner S. 2000. Fugu: a compact vertebrate reference genome. *FEBS Lett*. **476**, 3–7.

48. Eckardt N.A. 2001. Everything in its place: Conservation of gene order among distantly related plant species. *Plant Cell*. **13**, 723–725.

49. Miller W. 2001. Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics*. **17**, 391–397.

50. Roytberg M.A., Ogurtsov A.Yu., Shabalina S.A., Kondrashov A.S. 2002. A hierarchical approach to aligning collinear regions of genomes. *Bioinformatics*. **18**, 1673–1680.

51. Ogurtsov A.Yu., Roytberg M.A., Shabalina S.A., Kondrashov A.S. 2002. OWEN: aligning long collinear regions of genomes. *Bioinformatics*. **18**, 1703–1704.

52. Roytberg M.A. 1994. *Pareto-optimal alignments of symbol sequences*. Pushchino: ONTI PSC.

53. Roytberg M.A., Semionenkov M.N., Tabolina O.U. 1998. How to find gaps without gap penalty? In *Proceedings of the Int. Conf. RECOMB-98*. N.-Y.

54. Roytberg M.A., Simeonenkov M.N., Tabolina O.Yu. 1999. Pareto-optimal alignments of symbol sequences. *Biofizika*. **44**, 581–594.

55. Sunyaev S., Bogopolsky G., Oleynikova N.V., Vlasov P.K., Finkelstein A.V., Roytberg M. A. 2003. From analysis of protein structural alignments towards a novel approach to align protein sequences. Proteins. In press.

56. Thompson J.D., Plewniak F., Poch O. 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*. **15**, 87–88. http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2/

57. Vogt G., Etzold T., Argos P. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol*. **249**, 816–831.

58. Litvinov I.I., Finkelstein A.V., Roytberg M.A. 2002. The method of the amino acid sequences alignment, taking into account the information about proteins' secondary structures. In *Proc. of 5th Int. Congr. of Mathematical Modeling. Sept. 30–Octob. 6, 2002, Dubna*, Moscow: Nauka, p. 211.

59. McGuffin L.J., Bryson K., Jones D.T. 2000 The PSIPRED protein structure prediction server. *Bioinformatics*. **16**, 404–405.

60. Kister A.E, Roytberg M.A., Chothia C., Gelfand I.M. 2001. The sequence determinants of cadherin molecules. *Protein Science*. **10**, 1801–1810.

61. Shevchenko A., Sunyaev S., Loboda A., Shevchenko A., Bork P., Ens W., Standing K. 2001. Charting the proteomes of organisms with unsequenced genomes by MALDI-Quadrupole Time-of-Flight mass spectrometry and BLAST homology searching. *Anal. Chem*. **73**, 1917–1926.

62. Gelfand M.S., Mironov A.A., Pevzner P.A. 1996. Gene recognition via spliced sequence alignment. *Proc. Natl. Sci. USA*. **93**, 9061–9066.

63. Eppstein D., Galil Z., Giancarlo R, Italiano G.F. 1992. Sparse dynamic programming I: Linear cost functions. *J. ACM*. **39**(3), 519–545.

64. Karlin S., Altschul S.F. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA*. **90**, 5873–5877.

65. Mott R. 2000. Accurate Formula for P-values of gapped local sequence and profile alignments. *J. Mol. Biol*. **300**, 649–659.

66. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**, 3389–3402.

67. Schaffer A.A., Wolf Y.I., Ponting C.P., Koonin E.V., Aravind L., Altschul S.F. 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*. **15**, 1000–1011.