ISSN 0006-3509, Biophysics, 2010, Vol. 55, No. 6, pp. 891–900. © Pleiades Publishing, Inc., 2010.
Original Russian Text © V.V. Yakovlev, M.A. Roytberg, 2010, published in Biofizika, 2010, Vol. 55, No. 6, pp. 965–975.

MOLECULAR BIOPHYSICS

Increasing the Accuracy of Global Alignment of Amino Acid Sequences by Constructing a Set of Alignment Candidates

V. V. Yakovlev and M. A. Roytberg

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia

Received June 29, 2010

Abstract—The accuracy of global Smith–Waterman alignments and Pareto-optimal alignments depending on the degree of sequence similarity (percent of coincidence, *%id*, and the number of removed fragments *NGap*) has been examined. An algorithm for constructing a set of three to six alignments has been developed of which the best alignment on the average exceeds in accuracy the best alignment that can be constructed using the Smith–Waterman algorithm. For weakly homologous sequences (*%id* 15, *NGap* 20), the increase in accuracy is on the average about 8%, with the average accuracy of the global Smith–Waterman alignments being about 38% (the accuracy was estimated on model test sets).

Keywords: amino acid sequences, dynamic simulation, Pareto-optimal alignment, accuracy and confidence

DOI: 10.1134/S0006350910060011

INTRODUCTION

In bioinformatics the problem of global sequence alignment is traditionally posed as a problem of constructing an optimal alignment relative to some weight function [1]. Therewith the choice of the value of the parameters of the weight function, especially the penalties for deletions and insertions, is insufficiently substantiated from the viewpoint of biological applications. As a standard, de facto use is now made of the earlier proposed [2] affine weights of the type $f + s \cdot l$, where l is the length of deletion (i.e., of the segment removed); f (Gap Opening Penalty, *GOP*) and s (Gap Elongation Penalty, *GEP*) are numerical parameters. The question of adequate choice of the values of numerical parameters has been solved purely empirically (see, for example, monograph [3], pp. 126–127). Let us note that in the case of global alignment, which is considered in our paper, the use of GEP can be avoided at the expense of an appropriate modification of the mutation weight matrix (see p. 1.4).

The algorithm of constructing an optimal alignment relative to an affine weight function (Smith–Waterman algorithm [1]) appears to be the most exact among all algorithms widely used at present [4, 5]. Here the accuracy of algorithmic alignment is understood as the fraction of positions of "reference" or "standard" (evolutionarily consistent) alignment that are restored in the algorithmic alignment (in more detail see section 1). At that even for the Smith–Waterman algorithm the alignments of weakly homologous sequences have low accuracy [4–6]. In works [7, 8] and a number of others it has been shown that the accuracy of amino acid sequence alignment can be raised by using data on the secondary structure of proteins.

In the present work we investigate the question of the possibility of raising the accuracy of alignments without invoking additional information on the nature of the compared symbol sequences. The interest in such a statement of the problem is associated both with the demands of theory development and with the application of alignment algorithms to symbol sequences of other nature than amino acid sequences, for example to nucleotide sequences. The approach proposed below to raising the accuracy of alignments can be also applied to alignment of amino acid sequences enriched in information on secondary structure.

One of the means of raising the accuracy of global alignments consists in altering the problem statement: instead of constructing one alignment a problem is posed of constructing a set of alignments one of which (it is unknown which one particularly) has high accuracy. The accuracy of the set of alignments is therewith taken to be the accuracy of the best alignment of this set. M. Waterman and colleagues have considered the problem of constructing suboptimal alignments (having weight differing from optimal by a specified value) and proposed an algorithm of solving this problem ([9, 10], see also review [11]). Practical application of this approach proved complicated because of that the amount of thus determined suboptimal alignment may be very great even at a very low threshold on the differ-

Editor's Note: I certify that this is the closest possible equivalent of the original publication with all its factual statements and terminology, phrasing and style. *A.G.*

ence in the weight of admissible alignments from the weight of the optimal alignment.

In work [12], devoted to alignments of immunoglobulins, an idea was stated about considering a family of alignments each of which appears as optimal at its own set of weight parameters; however, an algorithm of solving this problem was noå proposed. Later in the works of M. Waterman, D. Gusfield, and other authors [13–15] algorithms were proposed for decomposition of the space of parameters into regions corresponding to one and the same optimal alignment. After constructing such a decomposition all possible optimal alignments can be found with the aid of a standard algorithm with the use of one value of parameters for each region. In work [15] an estimate is given on the quantity of decomposition regions. A drawback of this approach is that for it all parameter values are equal in rights, including those that obviously contradict the biological sense. In addition, for the tasks of bioinformatics of interest are the optimal alignments themselves, rather than the structure of the parameter space, which is paid the main attention in the parametric approach.

In works [16–18] we have proposed a multicriterial approach to the alignment problem, which can be regarded as dualistic to the above-described parametric approach. In the framework of the multicriterial approach every alignment is described with a vector weight; the components of such a vector can be, for example, the sum weight of matchings, the quantity of removed fragments and their sum length. The traditional alignment weight presents as a linear combination of the components of the vector weight.

With the multicriterial approach an analog of the algorithm that finds the maximally possible alignment weight and the optimal alignment having this weight is the algorithm that builds a Pareto-optimal set [19] of weights and the "Pareto-optimal" alignments corresponding to these weights. Informally speaking, an alignment is Pareto-optimal if it appears optimal at a certain monotonic (not necessarily linear!) weight function of elementary weight parameters that are vector weight components. In particular, at any values of penalties the Smith–Waterman alignment is Pareto-optimal relative to a vector weight the components of which are: (1) the sum weight of matchings, (2) the quantity of removed fragments, (3) their sum length.

The investigation presented in this work (see Results section) consisted of two parts. First with the aid of computer experiments we have studied the accuracy of alignments obtained in different ways as dependent on the extent of similarity of the compared sequences. We have considered two main classes of alignments: (1) global Smith–Waterman alignments (GSW) obtained at different penalties, (2) subreference Pareto-optimal alignments, i.e., such alignments that have the greatest accuracy among all Pareto-optimal alignments (in other words, the accuracy of a subreference alignment is the accuracy of the set of all Pareto-optimal alignments of the given pair of sequences). Consideration of subreference alignments permits determining the upper limit that can be reached using not only the Smith–Waterman algorithm but also any other algorithm where the penalty for removal is determined by a monotonic function of the quantity of removed fragments and their sum length.

On the basis of these investigations we have for the first time proposed a means of constructing a set of three to six alignments the accuracy of which exceeds the accuracy of Smith–Waterman alignment.

1. DATA AND METHODS

In subsections 1.1 and 1.2 the means is described for generating test sets of sequence pairs. In subsections 1.3 – 1.5 the necessary definitions are given (Pareto-optimal alignment, alignment accuracy and the like). Subsection 1.6 describes the method of conducting computer experiments.

1.1. Generation of model sequences and their alignments. In the capacity of reference alignments we consider the alignments of model (computer generated) amino acid sequences. The algorithm of constructing each sequence pair has the following parameters:

(1) mean sequence length (*LSeq*);

(2) percentage of coincidence in compared sequences, (*%id*);

(3) quantity of insertions (*NGap*);

(4) mutation probability matrix (*Mutation-Prob-Matrix*);

(5) insert length distribution density.

In the case of alignment of two sequences we cannot distinguish an insertion in one sequence from a deletion in the other. Therefore in generating a test pair of sequences we for convenience everywhere performed only insertions. Generation of the test pair consists of the following stages.

1.1.1. Generation of insert lengths. We will call an insert frame a pair (L, s), where L is inserted fragment length, $s \in \{1, 2\}$ is the number of the sequence into which the insertion will be made (the place of insertion in the selected sequence and the inserted symbols are determined separately from the frame and independently of it). At the first stage we independently generate *NGap* insert frames. Insertions into both sequences we consider equiprobable. The insert length L is generated on the basis of an empirical distribution density obtained as a result of analysis of the PREFAB database [20] (see Supplementary materials, Appendix 1 at http://server2.lpm.org.ru/static/papers/pareto/). This distribution is consistent with the results of work [21].

1.1.2. Determination of basic sequence length. We determine the sun lengths *LIns*1, *LIns*2 of inserts pertaining respectively to the first and the second

sequences in accordance with the frames built at the preceding stages. We take the base sequence length *LBase* equal to

$$LBase = LSeq - (LIns1 + LIns2)/2.$$

1.1.3. Generation of basic sequence. We construct a Bernoulli random sequence *SeqBase* of length *LBase*, the probabilities of amino acids correspond to the data from [22].

1.1.4. Generation of mutant sequence—introduction of mutations. A mutant sequence *SeqMute* is constructed in the following way. In accordance with parameters *LBase* and *%id* we calculate the quantity *NMut* of positions of the basic sequence in which mutations must be introduced. With the use of a function from the NumericPython package, we generate an ordered list of *NMut* different random positions from the uniform distribution *R*(0, *LBase-1*). In the chosen positions we perform mutations, the probabilities of the replacing amino acids correspond to *Mutation-ProbMatrix* (see Supplementary materials, Appendix 2 at http://server2.lpm.org.ru/static/papers/pareto/).

1.1.5. Introduction of insertions and construction of reference alignment. Among positions {0, 1, ..., *LBase*} we choose *NGap* independent positions of insertions, the choice of positions is performed as in p. 1.1.4. Let (L_k, s_k) be the frame of k-th insert, p_k its position, $k \in \{1, ..., NGap\}$. At $s_k = 1$ the insertion is made into the basic sequence, at $s_k = 2$, into the mutant one. Then after the k-th symbol into the corresponding sequence we insert a random amino acid sequence of length L_k. Generation of the random sequence is performed as in p. 1.1.3.

The sequences obtained make a test pair of sequences. The reference alignment of this pair is determined by the set of triples (L_k, s_k, p_k), where $k = 1, ..., NGap$, it is obtained by removal of all inserts.

1.2. Test sets. For conducting computer experiments we prepared a series of test sets, each of which consists of 200 sequence pairs. The size of a test set was determined in the course of preliminary computer experiments, proceeding from the requirement of stabilization of the mean values of accuracy and confidence of algorithmic alignments. Each test set is characterized by the following parameters (see above):

(1) mean sequence length (*LSeq*);

(2) percentage of coincidence (*%id*) in compared sequences;

(3) quantity of removed fragments (*NGap*);

(4) mutation probability matrix (*Mutation-ProbMatrix*).

For numerical parameters we have considered the following values:

LSeq: 300, 600, 1000,

%id: 10, 15, 20, 30, 50, 70, 90,

NGap: 5, 10, 15, 20.

The values of different quantities were taken in arbitrary combinations. i.e. we have considered $3 \times 8 \times 3 =$ triples of form (*NGap*, *%id*, *LSeq*). Not to make the presentation cumbersome, data are presented only for sequences of length 300. The results for other lengths are analogous and are presented in Supplementary materials (see \Supplementary\Appendix\).

Let us note that, unlike the authors of work [23], we consider the quantity of inserts an attribute of the test set rather than a random value the distribution of which is determined by the evolutionary distance between the sequences (see [21]). This decision is dictated by the wish to establish the dependence of alignment quality on the insert quantity.

The mutation probability matrix was chosen taking into account the value of parameter *%id*. At *%id* = 90, 70, 50, 30 in the capacity of a basis matrix we used the PAM120 matrix [22]; for *%id* = 20 the basis matrix was PAM240 matrix; for *%id* = 15, PAM360 matrix; *%id* = 10, PAM480 matrix. Such a choice is determined by that, as shown in [23] with the aid of computer experiments, mutation with the aid of the PAM120 matrix leads to about 30% coincidences, 20% coincidences correspond to the PAM240 matrix and so on. We did not use matrices PAM100, PAM60 etc. for analysis of highly homologous sequences (*%id* = 50 and higher), because it is known [23], that in these cases the PAM120 matrix gives food enough results, which are only insignificantly worse than the results obtained in matrices more exactly corresponding to the evolutionary distance.

1.3. Measures of alignment quality. In the capacity of a quantitative measure of the quality of algorithmically obtained alignments we used two mutually complementary measures (see [24, 25]):

Alignment Accuracy (designation: *Acc*) is equal to the ratio of the quantity of matchings that are present in both alignments (*I*) to the total quantity of matchings in the reference alignment (*G*):

$$Acc = I/G \cdot 100\ \%.$$

Alignment Confidence (designation: *Conf*) is equal to the ratio of the quantity of matchings that are present in both alignments (*I*) to the total quantity of matchings in the algorithmically constructed alignment (*A*):

$$Conf = I/A \cdot 100\ \%.$$

Informally speaking, accuracy *Acc* shows what fraction of the reference alignment could be restored, while confidence *Conf*, to what extent one can trust the constructed alignment.

In accordance with the abovesaid, accuracy (confidence) of a set of alignments of the given pair of test sequences we will call the maximal accuracy (confidence) for alignments of the given set.

As our computer experiments show, in the case of global alignments (and as distinct from local align-

Table 1. Values of parameters *GOP* and *GEP* with which GSW alignments were constructed for different test sets with mean sequence length 300

%id	5 indels		10 indels		15 indels		20 indels	
	GOP	*GEP*	*GOP*	*GEP*	*GOP*	*GEP*	*GOP*	*GEP*
10	37	1	29	1	21	1	16	1
15	27	1	20	1	18	1	16	1
20	16	1	16	1	14	1	11	1
30	17	1	13	1	11	1	10	1
50	12	1	11	1	9	1	10	0.5
70	12	0.5	9	0.5	9	0.5	10	0.5
90	7	0	6	0	5	0	6	0

Note: For sequences with *%id* = 30 and higher the indicated parameter values were used with mutation weight matrix PAM120, at *%id* = 20 with matrix PAM240, at *%id* = 15 with matrix PAM360, at *%id* = 10 with matrix PAM480.

ments) the accuracy and confidence of Pareto-optimal alignments (in particular global Smith–Waterman alignments, see p. 1.5) are approximately equal. Therefore in the Results section we present data only for the accuracy of the investigated alignments.

1.4. Global Smith–Waterman alignments. *1.4.1. General information.* In the capacity of a standard algorithm of constructing a global alignment we used the Smith–Waterman algorithm [1, 24], the most exact one among the similar algorithms broadly used at present. We should note that in the original work [24] they described a version of the algorithm for constructing an optimal local alignment, while in our work we consider only global alignments. To underscore this, we will write «GSW alignments» and «GSW algorithm.»

The GSW algorithm builds an optimal alignment for the given sequence pair at preset mutation weight matrix, Gap Opening Penalty (*GOP*) and Gap Elongation Penalty (*GEP*). The alignment weight is determined by formula:

$$Score = W - GOP \cdot g - GEP \cdot d.$$

Here and below, W is the sum weight of matchings, g is the quantity of removed fragments, d is th quantity of removed symbols. In the case of global alignment, which is considered in our paper, the use of GEP can be avoided at the expense of an appropriate modification of the mutation weight matrix.

Indeed, let the alignment of sequences of lengths $L1$, $L2$ respectively be performed on the basis of mutation weight matrix M and penalties *GOP* (for fragment removal) and *GEP* (for symbol removal). Let some alignment A contain t matchings of symbols, the sum weight of which is W, and g removed fragments. Then the quantity of removed symbols for alignment A equals $L1 + L2 - 2t$, i.e. the weight of alignment A equals:

$$\begin{aligned} Score(A) &= W - GOP \cdot g - GEP \cdot (L1 + L2 - 2t) \\ &= W + t \cdot 2 \cdot GEP - GOP \cdot g - GEP \cdot (L1 + L2). \end{aligned} \quad (1)$$

Let M' be a weight matrix that is obtained from matrix M by addition of value 2 *GEP* to each element; $Score'(A)$ the weight of alignment A with the use of weight matrix M', penalty *GOP* for fragment removal and zero penalty for symbol removal. Obviously, for any alignment A fulfilled is:

$$Score'(A) = Score(A) + GEP \cdot (L1 + L2).$$

Since the value $GEP \cdot (L1 + L2)$ does not depend on alignment A, then the choice of optimal alignment in the initial system of weights and penalties is equivalent to a choice of optimal alignment in a new system where the penalty for symbol removal is zero.

1.4.2. Choice of mutation weight matrix and removal penalties. In the capacity of a mutation weight matrix we used a matrix of the PAM family corresponding to the mutation probability matrix used in constructing the test set (see p. 1.2). Matrices of the PAM family were taken from the supply of the EMBOSS package [25] (see also Supplementary materials, subdirectory \Supplementary\Matrices).

The choice of values of penalties *GOP* and *GEP* was conducted in the following way. For each test set of length 300 we calculated the mean accuracy of alignments with the PAM family weight matrix indicated in p. 1.4.1 and all pairs of parameters *GOP* and *GEP* such that $GEP = 0; 0.5; 1.0; 1.5; 2.0; 2.5; 3.0$, $GEP = 1, 2, \ldots, 30$. After this we selected the pair of values corresponding to the greatest mean accuracy. The values of parameters *GOP* and *GEP* used in alignment of test sets of mean sequence length 300 are given in Table 1. Let us note that the mean accuracy of GSW alignments weakly depends on the choice of parameter *GEP*. In the case of approximate equality of mean accuracies for different *GEP* values we chose the value that corresponded to the optimum for several sets. Completely

the materials on the choice of penalty values are presented in Supplementary materials, see Appendix 3.

1.4.3. Software realizations. In conducting computer experiments we constructed Smith–Waterman alignments with the aid of the ParetoSW program, a modification of the program for constructing Pareto-optimal alignments Pareto (see p. 1.5). Preliminarily we conducted experiments on comparison of the performance of the ParetoSW program with the known realization of the Smith–Waterman algorithm—program stretcher from the EMBOSS package [25] (note that the authors of the EMBOSS program not quite correctly call the used algorithm a Needleman & Wunsch algorithm). Experiments have shown full identity of the results of the programs. For more detail see Supplementary materials, Appendix 4.

1.5. Pareto-optimal alignments. The notions of vector weight function and Pareto-optimal alignments are presented in [17]. Here we briefly and in simplified form present the information that we used below. In more detail about Pareto-optimal alignments see [17].

Let $k \geq 2$ be an integer. *Vector alignment function* is a function matching to each alignment A a k-dimensional vector $V(A)$ called the vector alignment of alignment A.

An example of vector weight is weight

$$V(A) = (M(A), -g(A)), \qquad (1.1)$$

where $k = 2$; $M(A)$ and $g(A)$, sum weight of matchings and quantity of removed fragments in alignment A respectively. Vector weight components (in the given case $M(A)$ and $-g(A)$) are called elementary weight functions. Below we will consider only the vector weight function (1.1).

Let S_1 and S_2 be sequences; V, vector weight function. Alignment $A1$ of sequences S_1 and S_2 *dominates* alignment $A2$ if

$$M(A1) \geq M(A2); \quad g(A1) \leq g(A2) \qquad (1.2)$$

and at least one of inequalities (1.2) is strict. Alignment A is called Pareto-optimal if no alignment of sequences S_1 and S_2 dominates alignment A. The notions of a Pareto-optimal set of alignments are an analog of the notion of optimal alignment for the case of vector weights.

The Pareto-optimal alignment relative to weight (1.1) that has the greatest accuracy we will call a *subreference* alignment of the given pair of sequences. In accordance with what was said in p.1.3, the accuracy of a set of Pareto-optimal alignments coincides with the accuracy of the subreference alignment.

The importance of the Pareto-optimal alignment set is determined by the following observation. Let $V(A) = \langle V_1(A), ..., V_k(A)\rangle$ be a vector weight function; $g(x_1, ..., x_k)$, a function of k variables monotonically nondecreasing in each of the arguments; $\varphi(A) = g(V_1(A), ..., V_k(A))$, a scalar weight function; B, an optimal alignment of some sequences relative to function φ. Then B is the optimal alignment of these sequences relative to vector function V. In particular, if B is the optimal alignment of some sequences relative to a linear combination of functions $V_1(A)$, ..., $V_k(A)$ with positive coefficients, then B is the optimal alignment of these sequences relative to vector function $V(A)$.

In [17] we presented an algorithm of constructing all Pareto-optimal alignments of the given sequences S_1 and S_2 relative to the given weight function $V(A)$. This algorithm is a dynamic programming algorithm and is bases on the relationship of distributivity between the operation of vector addition and the operation of taking a Pareto subset of the sum of two Pareto-optimal sets. The algorithm for the case of vector weight function (1.1) is implemented in the Pareto program.

1.6. Method of conducting computer experiments. *1.6.1. General information.* For each test set we have conducted several series of experiments. A series of experiments consisted in that for each pair of test sequences we constructed a set of alignments, and then conducted analysis of these alignments. In this way, each series is characterized by the

(1) test set;

(2) method of constructing the set of alignments;

(3) parameters of the chosen method.

The method of alignment analysis was completely determined by the chosen method of constructing alignments (see p. 1.6.2 and p. 2).

For all test sets we used two methods of constructing the set of alignments:

(1) constructing all Pareto-optimal alignments relative to vector weight function $(M, -g)$, see p.1.4;

(2) constructing GSW accuracies with prior-chosen "optimal" penalties for indels (insertions–deletions) and mutation weight matrix.

The choice of mutation weight matrix and penalties for indels in constructing GSW alignments is described in p. 1.4.2.

In constructing Pareto-optimal alignments the single parameter is the mutation weight matrix. In such capacity we chose a modification of the same matrix of the PAM family that was used in constructing GSW alignments. The modification of the matrix (see p. 1.4.1) consisted in that to each its elements we added the number 2 *GEP*, where *GEP* is the weight of gap elongation penalty used in constructing GSW alignments. As it follows from pp. 1.4, 1.5, at the preset weights of mutations and penalty *GEP,* a GSW alignment obtained at any value of penalty *GOP* is Pareto-optimal.

1.6.2. Output data of a series of experiments: alignments. As a result of a series of experiments, files are built with the set of alignments (own file for each test pair) and also a summary table. In the table each row

Table 2. Accuracy of subreference alignments (I) and GSW alignments (II) for different test sets of length 300

%id	5 indels		10 indels		15 indels		20 indels	
	Mean accuracy	σ	Mean accuracy	σ	Mean accuracy	σ	Mean accuracy	σ
				I				
10	0.80	0.12	0.61	0.15	0.44	0.15	0.34	0.13
15	0.89	0.07	0.74	0.11	0.62	0.13	0.49	0.14
20	0.95	0.04	0.87	0.05	0.79	0.07	0.71	0.09
30	0.97	0.02	0.94	0.03	0.90	0.04	0.86	0.05
50	0.99	0.01	0.97	0.02	0.95	0.02	0.92	0.03
70	0.99	0.01	0.98	0.01	0.97	0.01	0.96	0.02
90	1.00	0.00	0.99	0.01	0.99	0.01	0.98	0.01
				II				
10	0.73	0.15	0.53	0.16	0.35	0.16	0.27	0.14
15	0.87	0.09	0.69	0.13	0.55	0.15	0.41	0.15
20	0.93	0.05	0.84	0.07	0.75	0.09	0.67	0.10
30	0.97	0.02	0.92	0.03	0.88	0.05	0.83	0.05
50	0.98	0.01	0.96	0.02	0.94	0.03	0.91	0.03
70	0.99	0.01	0.98	0.01	0.97	0.02	0.95	0.02
90	1.00	0.00	0.99	0.01	0.99	0.01	0.98	0.01

Note: For each set indicated are the mean value of accuracy (Mean accuracy) and its rms deviation σ.

corresponds to one alignment of a particular test pair. The row indicates:

(1) characteristics of the test pair (identifier, sequence lengths and the like)

(2) number of the given alignment among alignments of the given test pair; for GSW alignments, additionally the penalty values (for Pareto alignments the «substantial» identifier is presented by the quantity of removed fragments, see below);

(3) characteristics of the reference alignment (quantity of removed fragments, quantity of matched symbols, fraction of coincidences among them and the like);

(4) analogous characteristics of the given algorithmic alignment;

(5) characteristics of similarity of the algorithmic and the reference alignment—accuracy and confidence (see p. 1.5);

(6) flag—whether the given alignment is subreference (see p. 1.5) for the given series;

(7) for series of experiments with Pareto-optimal alignments, information highlighting alignments that appear preferable from the viewpoint of various algorithms of choosing the preferable alignment (see section 2). In more detail on summary tables see (Supplementary materials, subdirectory \Experiments\\Docs\).

2. RESULTS

2.1. Subreference Pareto-optimal alignments. Let a set of alignments be given for a test sequence pair. A *sequence* alignments we will call such an alignment of the set that has the most high accuracy. Table 2.I presents the mean values and rms deviations for the accuracy of subreference alignments for a set of Pareto-optimal alignments. Below for brevity an alignment of two sequences that is subreference on the set of Pareto-optimal alignments of these sequences we will call a subreference one.

As noted in section 1, the accuracy of subreference alignments is the best of all possible with the use of any criterion that is specified by a monotonic (not necessarily linear) function of the sum weight of matchings and quantity of removed fragments.

As it should be expected, the accuracy of subreference alignments declines with declining level of similarity and increasing quantity of indels. At that the dependence of accuracy on indel number is practically linear (see Fig. 1). An analogous phenomenon takes place also for Smith–Waterman alignments (see Table 2.II). Data were obtained on test sets described in p. 1.2.

The drawback of the set of Pareto-optimal alignments is that it may contain several tens of alignments. Below we will consider two means of narrowing this set (if possible) without a decrease in accuracy. The first means is Smith–Waterman alignments, the second

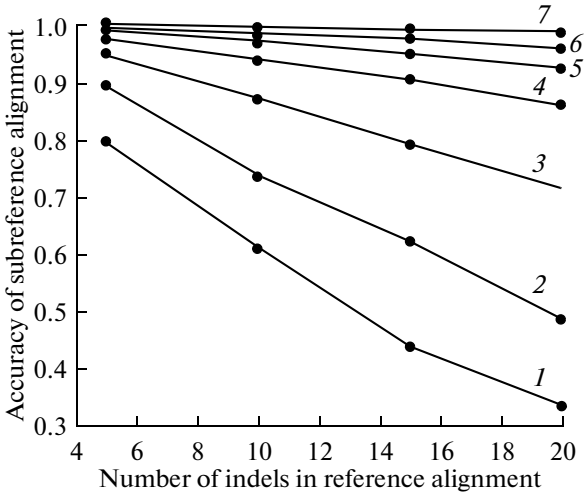

Fig. 1. Decline of accuracy of subreference alignments with growing number of indels in the reference alignment, i.e. the number of fragments inserted into sequences during generation. The curves correspond to different values of the percentage of coincidences, *%id*: *1* – 10; *2* – 15; *3* – 20; *4* – 30; *5* – 50; *6* – 70; *7* – 90%. Data are presented for 5, 10, 15 and 20 indels (corresponding points on abscissa axis) and correspond to data presented in Table 2.I. The dependence for Smith–Waterman alignments (see Table 2.II) bears an analogous character.

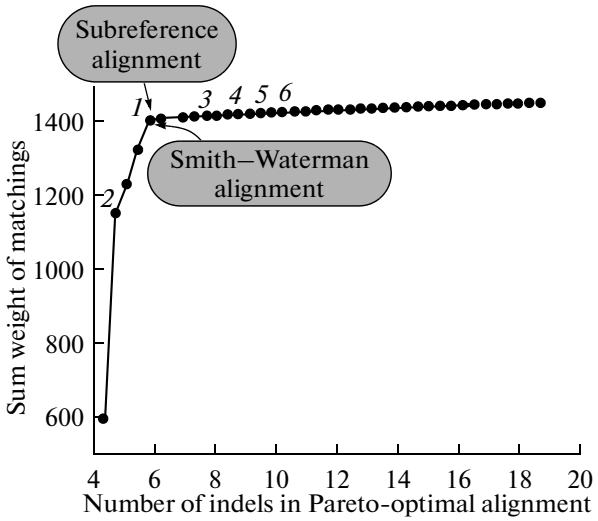

Fig. 2. Plot of $M(g)$ dependence for highly homologous sequences (*%id* = 70, 5 fragments inserted in sequences during generation). Here $M(g)$ is sum weight of matchings in Pareto-optimal alignment obtained upon removing g-fragments ("indels") from compared sequences. In alignment use was made of PAM120 matrix modified in accordance with the value of parameter $GEP = 0.5$ (see p. 1.4.2). Subreference alignment corresponds to the break at $g = 5$.

means (see p. 2.3) is proposed by us on the basis of the conducted analysis of Pareto-optimal alignments.

2.2. Choice of preferable Pareto-optimal alignment with the aid of penalties for deletions. Smith–Waterman alignments. The Smith–Waterman alignment obtained at a given value of parameter *GOP* is a Pareto-optimal alignment with the greatest value of quantity

$$W = M - GOP \cdot g, \qquad (2.1)$$

where M is sum weight of matchings, g is quantity of removed fragments. Recall that at the expense of modification of standard mutation matrices we reduced the value of parameter *GEP* to zero, this is possible in the case of global alignment (see p. 1.2).

Table 2.II presents the mean values and rms deviations for the accuracy of Smith–Waterman alignments (values of parameter *GOP* chosen so as to maximize the mean accuracy, see p. 1.4.2). In Table 3, shown is the fraction of cases in which the Smith–Waterman alignment has the same accuracy as the subreference alignment.

As we can see, at *%id* = 30 and higher the mean accuracy of Smith–Waterman alignments differs from the mean accuracy of subreference alignments approximately not more than by 2%. At the same time at *%id* = 10 the difference in the mean accuracy of subreference alignments and Smith–Waterman alignments can reach almost 10%. Below we will consider the way of overcoming this shortcoming.

2.3. Choice of preferences on the basis of «breaks» of the plot of value *M*(*g*). Our means of narrowing the set of Pareto-optimal alignments is based on distinguishing "breaks" on the plot of value $M(g)$. This plot presents as a broken line (see Fig. 2). Consider some pair of test sequences. Let $A(g)$ be a Pareto-optimal alignment of this pair, containing g removed fragments, $M(g)$, assessment of alignment $A(g)$; $T(g) = (g, M(g))$, the point on the plot corresponding to it. Take (for > 1) $dM(g) = M(g) - M(g-1)$.

The slopes of segments adjacent to point $T(g)$, would be respectively tg_left $= dM(g)/a$ and tg_right $= dM(g+1)/a$, here a is a scale coefficient taken by us to be 20 (this number was selected empirically, results do not depend on changing the coefficient in a very broad range).

The tangent of angle between segments adjacent to point $T(g)$ equals:

Table 3. Fraction of test pairs for which the accuracy of Smith–Waterman alignment coincides with the accuracy of subreference alignment

%id	5 indels	10 indels	15 indels	20 indels
10	0.44	0.26	0.16	0.13
15	0.59	0.26	0.19	0.12
20	0.61	0.32	0.23	0.17
30	0.72	0.44	0.33	0.21
50	0.78	0.55	0.41	0.24
70	0.91	0.65	0.43	0.37
90	0.96	0.80	0.61	0.47

Table 4. Comparison of mean accuracy of sets *Extr*(3), *Extr*(6) with mean accuracies of sequence alignments and Smith–Waterman alignments

%id	10 indels				15 indels				20 indels			
	SW	*Extr*(3)	*Extr*(6)	Subref. algnt.	SW	*Extr*(3)	*Extr*(6)	Subref. algnt.	SW	*Extr*(3)	*Ex tr*(6)	Subref. algnt.
10	0.53	0.56	0.58	0.61	0.34	0.39	0.41	0.44	0.27	0.29	0.31	0.34
15	0.69	0.70	0.71	0.74	0.55	0.57	0.59	0.62	0.38	0.44	0.46	0.49
20	0.84	0.85	0.85	0.87	0.75	0.76	0.77	0.79	0.65	0.68	0.69	0.71
30	0.92	0.93	0.93	0.94	0.88	0.88	0.89	0.90	0.83	0.83	0.84	0.86
50	0.96	0.96	0.96	0.97	0.94	0.94	0.94	0.95	0.90	0.90	0.91	0.92
70	0.98	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.95	0.95	0.95	0.96

$$Tg(g) = (tg\_left - tg\_right)/(1 + tg\_left \cdot tg\_right)$$
$$= (dM(g) - dM(g+1))/(a^2 + dM(g) \cdot dM(g+1)).$$

The Pareto-optimal alignment $A(g)$ we will call *extremal* if g is the point of local maximum in sequence $\{Tg(g)\}$. In the capacity of distinguished sets of Pareto-optimal alignments we will consider sets of type *Extr*(n), where *Extr*(n) consists of n extremal alignments with the greatest values of quantity $Tg(g)$. Computer experiments (see Table 4) show that for weakly homologous alignments at $n = 2 - 3$ the mean accuracy of set *Extr*(n) (definition of alignment set accuracy see in p. 1.3) exceeds the mean accuracy of GSW alignments, while at $n = 5$–6 of set *Extr*(n) is close to the accuracy of the set of all Pareto-optimal alignments. The data of Table 5 confirm the data of Table 4 and show that the superiority of the accuracy of sets *Extr*(n) at $n > 1$ and *%id* = 10, 15 and 20 over the accuracy of Smith–Waterman alignments bears a nonrandom character.

Table 5. Comparison of the accuracies of sets of alignments *Extr*(n) at $n = 2, 3, 6$ with the accuracy of Smith–Waterman alignments for test sets with *%id* = 10, 15, 20

%id	*Ngaps*	*Extr*(2) vs. SW			*Extr*(3) vs. SW			*Extr*(6) vs. SW		
		+	=	−	+	=	−	+	=	−
10	10	0.41	0.43	0.16	0.49	0.39	0.13	0.57	0.37	0.07
10	15	0.37	0.42	0.22	0.52	0.37	0.12	0.64	0.29	0.08
10	20	0.35	0.47	0.19	0.49	0.41	0.10	0.61	0.35	0.04
15	10	0.34	0.43	0.24	0.43	0.43	0.15	0.51	0.39	0.11
15	15	0.36	0.46	0.19	0.48	0.43	0.10	0.63	0.33	0.05
15	20	0.48	0.40	0.12	0.65	0.27	0.08	0.76	0.21	0.03
20	10	0.25	0.51	0.24	0.34	0.51	0.16	0.39	0.49	0.13
20	15	0.24	0.44	0.33	0.35	0.47	0.19	0.52	0.40	0.09
20	20	0.29	0.44	0.27	0.47	0.40	0.13	0.64	0.32	0.05

Note: For each test set and each set of alignments *Extr*(n) shown is the fraction of cases in which the accuracy of set *Extr*(n) exceeds the accuracy of Smith–Waterman alignment ("+"), accuracies are equal ("–") and the accuracy of set *Extr*(n) is smaller than the accuracy of Smith–Waterman alignment.

3. DISCUSSION

We have considered a problem of constructing a set of variants of alignments of a given pair of amino acid sequences, which represents a generalization of the problem of constructing an optimal alignment. It is motivated by that an optimal alignment essentially depends on the choice of numerical parameters, first of all the fragment removal penalty. Though the problem of constructing a set of variants of alignment (instead of one alignment) is long known, it has been given significantly less attention than the problem of constructing an optimal alignment. Possible, this is associated with that, proceeding from the content of the problem, the criterion of quality of the set must depend not only on the constructed set of alignments, but also on the reference alignment.

In work [17] we formulated the notion of a set of Pareto-optimal alignments and proposed an algorithm for constructing it. In addition, we formulated the problem of distinguishing among Pareto-optimal alignments the most "biologically adequate" alignment and traced the approaches to solving this problem.

In the present work we have for the first time proposed an algorithm of constructing on a given pair of amino acid sequences an ordered list of candidate alignments. Cutting short this list after the n-th element, one can obtain a set of alignments *Extr*(n) (see p. 2.3) of size n (n = 1,2,...). Computer experiments have shown that the mean accuracy of set *Extr*(*3*) (and for many sets, set *Extr*(*2*) as well) exceeds somewhat the best mean accuracy of GSW alignments, while the accuracy of set *Extr*(*6*) differs from the accuracy of the set of all Pareto-optimal alignments not more than by 2%. The latter circumstance is essential from the practical viewpoint, because the quantity of Pareto-optimal alignments may reach several tens. The algorithm is implemented in the form of a server, which is situated at the address: http://server2.lpm.org.ru/bio/.

At the same time in our investigation there were a number of limitations, which set the boundaries to the possible increase of alignment accuracy. The first limitations was concerned with that no use is made of biological sequences, in particular, the possibility of secondary structure prediction. In works [7, 8] etc. It is shown that in this way one can radically improve the alignment quality. It appears to be of interest to apply the proposed technique of constructing a set of candidate alignments to alignment of sequences enriched in data on secondary structure. From the mathematical viewpoint taking into account the secondary structure means taking into account the dependence between positions in the given sequences. This can be modeled with the aid of Markovian models in constructing and mutating sequences and also can be a subject of additional investigation.

Further, if we digress from the biological specifics of sequences, a limitation of the method consists in that in the capacity of alignment weight we consider only monotonic functions of the sum weight of matchings and quantity of removed fragments. In the framework of such an approach the Smith–Waterman method shows mean accuracy not greatly differing from the maximally possible accuracy—the mean accuracy of reference alignments. At $\%id = 30$ and higher the difference does not exceed 2%, at $\%id = 10$ it constitutes from 6 to 9%.

The procedure of choosing candidate alignments is based on the following informal consideration. Let P be a pair of test sequences, g the quantity of indels, $M(g)$ the sum weight of matchings in the Pareto-optimal alignment of sequence pair P that contains g indels. In work [17] it was noted that addition of an indel leading to improvement of alignment quality corresponds to matching of homologous stretches and, consequently, the value $dM(g) = M(g) - M(g-1)$, as a rule, will be great. Whereas if addition of an indel does not lead to improvement of alignment quality, then the newly matched stretches are not homologous and, consequently, the value $dM(g)$, as a rule, will be small. In this way, choosing one of the Pareto-optimal alignments of the given sequence pair P, we must decide which $dM(g)$ values should be considered great, and which small, which can just be a theoretical basis of the choice of penalty in the Smith–Waterman method.

Figures 2 and 3 present the dependences of $M(P, g)$ on g for two pairs of sequences. In Fig. 2 there are data on two highly homologous ($\%id = 70$) sequences with a small (five) number of indels in the reference alignment. The sequences in Fig. 3 are low-homologous ($\%id = 10$), their reference alignment containing 20 indels. In the former case (Fig. 2) in the plot one can see a break separating the "large" $dM(P, g)$ values from «small» ones, and this break corresponds to the subreference alignment. In the latter case (Fig. 3) there is no explicit break and distinguishing "by eye" a subreference alignment is impossible. Therefore we rely on the value of the tangent of angle between neighboring segments of the plot of function $M = M(g)$, a local characteristic of the break value.

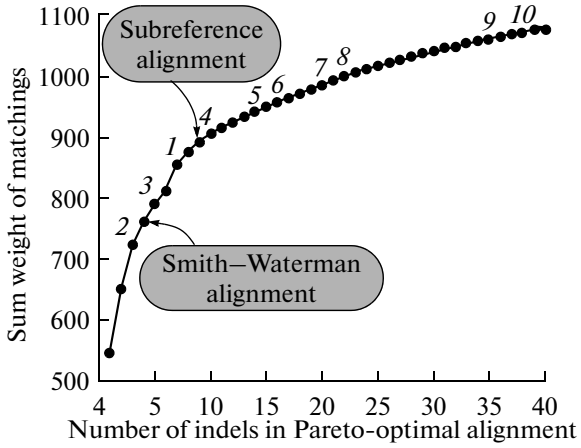

Fig. 3. Plot of $M(g)$ dependence for low- homologous sequences ($\%id = 10$, 20 fragments inserted in sequences during generation). Here $M(g)$ is sum weight of matchings in Pareto-optimal alignment obtained upon removing g-fragments ("indels") from compared sequences. In alignment use was made of PAM480matrix modified in accordance with the value of parameter $GEP = 1$ (see p. 1.4.2). Subreference alignment corresponds to $g = 9$. At this point there is no obvious break.

Let us note that in choosing the subset of Pareto-optimal alignments we did not analyze the inner structure of alignments, in particular the weights of deletionless fragments (see [4]). This probably is a reserve for improving the method of distinguishing candidate alignments. Another direction of improving the method is distinguishing in the alignments the supposedly most reliable fragments. In such a capacity we can consider the stretches identically aligned in all candidate alignments.

ACKNOWLEDGMENTS

The authors thank V.O. Polyanovsky and V.G. Tumanyan for discussing the method of generating sequences.

The work was supported by the Russian Foundation for Basic Research (08-01-92496, 09-0401053) and the Program of scientific exchange between scientists of Russia and France.

REFERENCES

1. *Mathematical Methods for DNA Sequences*, Ed. by M. S. Waterman (CRC Press, Boca Raton, FL, 1989).
2. G. H. Gonnet, M. A. Cohen, and S. A. Benner, Science **256** (5062), 1443 (1992).
3. M. Zvelebil and J. O. Baum, *Understanding Bioinformatics* (Garland Science, London, 2007).

4. S. R. Sunyaev, G. A. Bogopolsky, N. V. Oleynikova, et al., Proteins **54**, 569 (2004).
5. G. Vogt, T. Etzold, and P. Argos, J. Mol. Biol. **249**, 816 (1995).
6. V. Polyanovsky, M. A. Roytberg, and V. G. Tumanyan, Comput. Biol. **15** (4), 379 (2008).
7. I. I. Litvinov, M. Yu. Lobanov, A. A. Mironov, et al., Mol. Biol. **40** (3), 533 (2006).
8. A. Wallqvist, Y. Fukunishi, L. R. Murphy, et al., Bioinformatics **16**, 988 (2000).
9. M. S. Waterman, Proc. Natl. Acad. Sci. USA **80**, 10 3123 (1983).
10. T. M. Byers and M. S. Waterman, Oper Res. **32**, 1381 (1984).
11. M. Vingron, Curr. Opinion Struct. Biol. **6** (3), 346 (1996).
12. W. M. Fitch and T. F. Smith, Proc. Natl. Acad. Sci. USA **80**, 1382 (1983).
13. M. S. Waterman, M. Eggert, and E. Lander, Proc. Natl. Acad. Sci. USA **89**, 6090 (1992).
14. D. Fernandez-Baca and S. Srinivasam, Operat. Res. Letters **10**, 87 (1991).
15. D. Gusfield, K. Balasubramian, and K. Naor, in *Proc. 3rd Ann. ACM-SIAM Discrete Algorithms* (1992), pp. 432–439.
16. M. A. Roytberg, Preprint (ONTI NTsBI, Pushchino, 1994) [in Russian].
17. M. A. Roytberg, M. N. Simeonenkov, and O. Yu. Tabolina, Biofizika **44** (4). 581 (1998).
18. I. M. Gukov, T. V. Astahova, M. A. Roytberg, et al., *Proc. Moscow Conf. Computat. Molecular Biol.* (MCCMB'05, Moscow, 2005), p. 136.
19. V. Pareto, *Manual of Political Economy* (A. M. Kelley, New York, 1972).
20. R. C. Edgar, Nucl. Acids Res. **32** (5), 1792 (2004).
21. S. A. Benner, M. A. Cohen, and G. H. Gonnet, J. Mol. Biol. **229**, 1065 (1993).
22. M. Dayhoff, R. Schwartz, and B. Orcutt, Ed. M. Dayhoff, in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, 1978), pp. 345–352.
23. V. O. Polyanovsky, M. A. Roytberg, and V. G. Tumanyan, Biofizika **53** (4), 533 (2008).
24. G. Vogt, T. Etzold, and P. Argos, J. Mol. Biol. **249**, 816 (1995).
25. F. S. Domingues, P. Lackner, A. Andreeva, and M. J. Sippl, J. Mol. Biol. **297**, 1003 (2000).
26. T. F. Smith and M. S. Waterman, J. Mol. Biol. **147**, 195 (1981).
27. http://www.be.embnet.org/embosshelp/stretcher.html