JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 4, 2008
© Mary Ann Liebert, Inc.
Pp. 379–391
DOI: 10.1089/cmb.2007.0145

Reconstruction of Genuine Pair-Wise Sequence Alignment

VALERY POLYANOVSKY,[1] MIKHAIL A. ROYTBERG,[2] and VLADIMIR G. TUMANYAN[1]

ABSTRACT

In many applications, the algorithmically obtained alignment ideally should restore the "golden standard" (GS) alignment, which superimposes positions originating from the same position of the common ancestor of the compared sequences. The average similarity between the algorithmically obtained and GS alignments ("the quality") is an important characteristic of an alignment algorithm. We proposed to determine the quality of an algorithm, using sequences that were artificially generated in accordance with an appropriate evolution model; the approach was applied to the global version of the Smith-Waterman algorithm (SWA). The quality of SWA is between 97% (for a PAM distance of 60) and 70% (for a PAM distance of 300). The percentage of identical aligned residues is the same for algorithmic and GS alignments. The total length of indels in algorithmic alignments is less than in the GS—mainly due to a substantial decrease in the number of indels in algorithmic alignments.

Key words: accuracy of alignment, alignment reliability, confidence of alignment, insertion/deletion, protein sequence alignment.

1. INTRODUCTION

Pair-wise sequence alignment is the basic method of comparative analysis of proteins and nucleic acids. The most popular algorithms (Needleman and Wunch, 1970; Smith and Waterman, 1981; Lipman and Pearson, 1985; Altschul et al., 1990) are based on sequence comparison only. More sophisticated methods (Wallqvist et al., 2000; Litvinov et al., 2006) take into account the secondary structure of proteins. Distance-based methods are now widely used for building evolutionary trees (Saitou and Neil, 1987). The approaches to measuring the distance can be divided into two classes. In the first class, the percentage of non-identical sites is calculated. The second class uses matrices determining the weights of amino acid substitutions (Hollich et al., 2005). In the case of distance estimators of the second type, each pair of sequences requires reliable pair-wise alignments.

For many applications, including evolutionary modeling, it is important to evaluate the "quality" of algorithmically obtained alignments (i.e., how close the algorithmic alignment is to the "evolutionarily true" one). Here, the "evolutionarily true" alignment is an alignment superimposing the positions originating from the same position of the common predecessor (Sunyaev et al., 2004). However, the evolutionarily true alignment of given sequences is usually unknown, and thus an approximation is needed.

[1] Engelhardt Institute of Molecular Biology, Russian Academy of Science (RAS), Moscow, Russia.
[2] Institute of Mathematical Problems in Biology, Russian Academy of Science (RAS), Puschino, Russia.

Papers studying the quality of alignment algorithms typically use the three-dimensional (3D) sequence alignment as such an approximation. The reason for this choice is that the 3D structure of proteins is much more conservative than their amino acid sequences (Doolittle, 1981).

Moreover, Vingron and Argos (1990) have shown a link between the degree of stability of the region of optimal global alignment with respect to the set of suboptimal alignments and their similarity to the structure-based alignment: the regions of optimal alignment frequently repeating in suboptimal alignments display maximal similarity to the structure-based alignment. Based on this observation, Mevissen and Vingron (1996) suggested using the difference between the weights of the optimal alignment and the alignment of maximal weight without matching i-th and j-th residues of proteins as the validity measure of the matching. Another measure for the validity of a matching within the alignment was proposed by Schlosshauer and Olsson (2002). The measure is also based on the comparison of the optimal alignment and an appropriate suboptimal one.

Our approach to the analysis of algorithmic alignments differs from the approaches described above. We are not interested in the validity of individual positions, but rather in the similarity between the algorithmic alignment and the "golden standard" (GS) alignment of the same sequences (Vogt et al. 1995; Domingues et al., 2000; Sunyaev et al., 2004). As mentioned above, the papers use the alignments obtained by the superposition of 3D structures as the golden standard. These alignments may differ from the evolutionarily true alignments, and thus lead to errors of unknown proportions in the quality estimation. To overcome this drawback, we propose using artificial pairs of sequences instead of pairs of homologous proteins. In any pair of artificial sequences, one is a random sequence, and the other is obtained from it by mutations, insertions, and deletions in accordance with an appropriate model of the evolution—namely, the model suggested in Dayhoff et al. (1978), Benner et al. (1993), and Reese and Pearson (2002). Therefore, the true alignment of the sequences is known from the pseudo-evolution process. A similar approach was used in Polyanovskii et al. (1994), but the rules for deletions were too simple, while no insertions were allowed.

We have performed the comparative analysis of algorithmic and GS alignments based on the "artificial evolution process." Within the analysis, we have studied both the measure of similarity between the algorithmic and GS alignments, and the correspondence between the distribution of insertions/deletions (indels) and substitutions in algorithmic and GS alignments. The analysis showed that the differences between algorithmic and GS alignments concern mainly the number and mean lengths of indels.

Similarly to Vogt et al. (1995), Domingues et al. (2000), and Sunyaev et al. (2004), we have studied the Smith-Waterman algorithm (SWA), but we have studied its global version, since the compared sequences as a whole are similar in our case.

We have investigated different evolution distances, namely, from 60 to 300 point accepted mutations (PAM), which, on average, corresponds to sequence identity between 58% and 16%.

2. METHODS AND ALGORITHMS

2.1. Measures of alignment similarity

To describe the similarity between the algorithmic and the GS alignments of the given pair of protein sequences, we use the following characteristics (Vogt et al., 1995; Domingues et al., 2000; Sunyaev et al., 2004). Alignment *accuracy* is the number I of positions *identically* superimposed in the algorithmic and the GS alignment, divided by the total number G of positions in the GS alignment:

$$Accuracy = I/G \tag{1}$$

Alignment *confidence* is defined analogously to the notion of accuracy (Eq. (1)):

$$Confidence = I/A \tag{2}$$

where I is again the number of residues *identically* aligned in algorithmic and GS alignments, and A is the total number of aligned positions in the *algorithmic* alignment.

Finally, the *overall correctness* is an integral measure of the algorithmic alignment—the average of its accuracy and confidence:

$$Overall\ correctness = 2 * I/(G + A) \tag{3}$$

In the evolution model under study, *accuracy* and *confidence* are similar for global alignments; therefore, *overall correctness* was mainly used.

2.2. General description of the approach

We calculate the above values according to the following procedure. First, the test set consisting of 10,000 pairs of amino acid (aa) sequences is to be generated; then, the set is characterized by (a) sequence length L and (b) evolutionary distance D within the pair. We have used two values of L (200 and 500 aa) and four values of D (60, 100, 200, 300 PAM). Thus, we have considered $2 \times 4 = 8$ test sets.

The first sequence of each pair is a random sequence of length L. All positions are regarded as independent and have the same amino acid composition as described in Dayhoff et al. (1978). The second sequence in the pair was obtained from the first by a mutational process, including both substitutions and insertions/deletions (indels); the frequencies and other parameters of mutations correspond to the given evolutionary distance D (see Section 2.3 for details). The set of mutational events determines the alignment of the sequences; the alignment is regarded as the GS.

For each pair of sequences, we then produce the algorithmic alignment and compute its accuracy, confidence, and overall correctness with respect to the above GS. The average values of accuracy, confidence, and quality for all 10,000 sequence pairs of the given test set will serve as quantitative measures of closeness between the algorithmically obtained alignments, and the true ones for given sequence length and evolutionary distance.

2.3. Modeling evolution

The modification of a random sequence taken as the first element of a sequence pair consists of several steps. The parameters of these steps depend on the chosen evolutionary distance D, measured in PAM units (Dayhoff et al., 1978). Deletions are performed first, followed by insertions, and, finally, by substitutions. The generation of random homologous sequences was done from scratch because the existing tools, e.g., ROSE (Stoye et al., 1998), were designed to generate multiple alignments and do not support the desired model of evolution.

2.3.1. Total length of indels. The total length of all insertions and deletions T equals

$$T = L * P(indel) * M\delta,$$

where L is the length of the sequence. $P(indel)$ is the probability of insertion/deletion appearance, which was calculated using the formula proposed in Reese and Pearson (2002):

$$P(indel) = 0.0224 - 0.0219 * e * *(-0.01168 * D),$$

where D is the distance between the sequences in PAM units. $M\delta$ is the mathematical expectation of the length of insertion/deletion:

$$M\delta = \sum_{d=1,\dots,d\ \max} d * P\{\delta = d\},$$

where $P\{\delta = d\}$ is the probability that the given indel has the length d. According to Benner et al. (1993), this probability follows the Zipf distribution and does not depend on the evolution distance D. For simplicity, we suppose that the total length of insertions and the total length of deletions equal $T/2$.

2.3.2. Deletions. The introduction of deletions was performed as a series of identical tries. At each try, we randomly chose the starting position N of a new deletion and its length d. The value N was obtained according to the uniform distribution, and d was obtained according to the Zipf distribution. If the distance from the position N to the end of sequence or to the nearest previously inserted deletion is less than d, then the try was discarded. Similarly, the try was discarded if, after the introduction of the deletion of length d, the total lengths of deletions became greater than $T/2$. Otherwise, the positions $[N..N + d - 1]$ are marked as deleted.

TABLE 1. CORRESPONDENCE BETWEEN THE EVOLUTIONARY DISTANCE IN PAM UNITS AND THE AVERAGE PERCENTAGE OF IDENTICAL SYMBOLS (%ID) IN THE INITIAL AND MUTATED SEQUENCES

PAM	*10*	*60*	*100*	*200*	*300*
%id	90	58	43	24	16

Note that, due to the discard of certain tries, the procedure leads to a distribution of deletion lengths that differs from the Zipf distribution. To overcome this drawback, the procedure generating deletions uses the modified probabilities obtained by the relaxation procedure (Samarskii and Goolin, 1989), rather than the Zipf probability distribution (see Appendix). Finally, the accuracy of approximation of the obtained indel length distribution with respect to the Zipf distribution is 3–4%.

2.3.3. Insertions. The positions and lengths of insertions were determined analogously to those of deletions with two exceptions: (1) the deleted positions were not considered; and (2) the length d could not be discarded, and thus there is no need to improve the length distribution. The inserted symbols were generated analogously to the generation of the first sequence of a pair.

2.3.4. Substitutions. Substitutions were introduced after the introduction of deletions of insertions; substitutions were allowed only in the non-deleted parts of the initial sequence. We have tested two ways of generating substitutions.

In the first case, for each position allowed for the substitution, we generate the new symbol for the position according to the PAM matrix corresponding to the evolutionary distance T (e.g., PAM60, PAM100). In the second case, we use the matrix PAM1 instead, but repeat the substitution procedure the required number of times (e.g., 60, 100). Both methods lead to the ~5% difference between the obtained substitution frequencies and the Dayhoff frequencies. The results below were obtained with the first procedure.

Table 1 shows the correspondence between the evolutionary distance in PAM units and the average percentage of identical symbols (%id) in the initial and mutated sequences. The data are in accordance with the data presented in Dayhoff et al. (1978).

2.4. Alignment algorithm and its parameters

We have tested the SWA (1981) in the global version (Needleman and Wunsch, 1970; Waterman, 1989). In all cases, we show the results obtained with substitution matrix PAM250 and gap penalty parameters 14 (gap opening penalty [GOP]) and 2 (gap elongation penalty [GEP]), which, on average, yield the best results, as demonstrated by our experiments. We have also tested the "native" substitution matrix (e.g., PAM60 for the test set with an evolutionary distance of 60 PAM) and the corresponding gap penalties, but the similarity between the algorithmic alignments and the GS in this case is approximately the same (the difference is about 0.5%; Table 2), like with the "universal" parameter set {matrix PAM250, GOP = 14, GEP = 2}. Contrastingly, the alignment of distant proteins (e.g., with evolutionary distance 250 PAM) with the "close" substitution matrix (e.g., PAM60) leads to significantly worse results.

3. RESULTS AND DISCUSSION

3.1. Quality of global SW-alignments

Figure 1 shows the histograms (empirical densities) of overall correctness of global SW alignments with respect to the GS. The histograms for accuracy and confidence are very similar to those for overall correctness (data not shown). The average values of overall correctness for two random alignments are given for reference. To obtain the data for each initial random sequence, two systems of indels were

Table 2. The Values of Average Accuracy and Confidence for the Global Version of Smith-Waterman Algorithm

		Native matrix			*Matrix PAM250*				
		Optimal penalties			*Optimal penalties*			*GOP = 14, GEP = 2*	
Length	*PAM*	*GOP, GEP*	*Acc*	*Conf*	*GOP, GEP*	*Acc*	*Conf*	*Acc*	*Conf*
200	60	13, 3	98.3	98.0	10, 1–2	98.0	97.7	97.7	97.4
200	60	14–15, 2	98.3	98.0					
200	100	12–13, 2	95.3	94.9	10, 1	95.3	94.9	94.6	94.0
200	100	12, 1	95.2	95					
200	200	15, 2	85.0	84.2	14, 1	85.4	84.7	85.1	84.1
200	200	17, 1	84.9	84.3					
200	300	17, 2	71.2	70.1	15, 2	71.1	69.9	70.5	69.5
500	60	15–16, 1	98.6	98.5	10, 1–2	98.4	98.2	98.2	98.0
500	60				11–12, 1	98.4	98.2		
500	100	14–16, 1	96.6	96.4	10–13, 1	96.5	96.2	96.3	95.9
500	200	18, 2	87.0	86.1	14, 1	87.1	86.5	87.1	86.2
500	200	19, 1	86.9	86.4					
500	300	20, 2	74.3	73.2	18, 2	74.4	73.2	73.2	72.3

For each test set we have checked two substitution matrices: (1) the "native" one, e.g., PAM60 for the set with distance parameter $D = 60$ PAM, and (2) PAM250. For each case, we show the results obtained with the gap penalties optimizing accuracy or confidence. If the maximal value of a characteristic can be obtained with different penalties, we chose the penalty providing the best value for another characteristic. If the best values of accuracy and confidence were achieved with the different sets of parameters, both pairs of the parameters are listed.

independently generated in accordance with the procedure described above. Then, the similarity between the alignments from each pair was calculated. Note that the quality of algorithmic alignment is much better than the quality of random alignment even for an evolutionary distance of PAM300.

The general characteristics of the empirical densities are given in Tables 3 and 4. Table 3 contains the average values of accuracy (Acc), confidence (Conf), and overall correctness (Corr) for all eight test sets. We give the average value of overall correctness of two random alignments from the test set as a background value. In contrast, Table 4 gives an impression of the shape of overall correctness density: all the distributions are unimodal, and Table 4 describes their maxima.

Table 3 allows one to conjecture that the quality of global SW alignments depends only on the evolutionary distance between the sequences and almost does not depend on their lengths (the values for $L = 500$ are slightly greater). In contrast, the similarity between two random alignments significantly depends on their lengths, and the values for $L = 500$ are significantly lower.

Unlike the average values, the shape of maxima does not depend on the evolutionary distance alone; it also correlates with the lengths of the sequences. The maximal values for sequences of length 500 are significantly greater than for sequences of length 200.

3.2. Detailed comparison of algorithmic and golden standard alignments

To get a deeper understanding of the interrelations between the SW and GS alignments, we have considered the following characteristics: (1) %id, or the fraction of exact matches among the total number of superimposed residues; (2) the total number of insertions and deletions ("indels"); (3) the average length of an indel; and (4) the total length of indels.

Figure 2 represents the scatter-plots of the characteristics for evolutionary distances PAM60, PAM300, and sequences of length 200. The data on the other test sets are similar and are not presented. The cumulative data for all test sets are given in Tables 5 and 6. Table 5 contains the average values of the four considered characteristics, presented separately for SW and GS alignments. Table 6 illustrates the ratios of the indel characteristics of algorithmic and GS alignments.

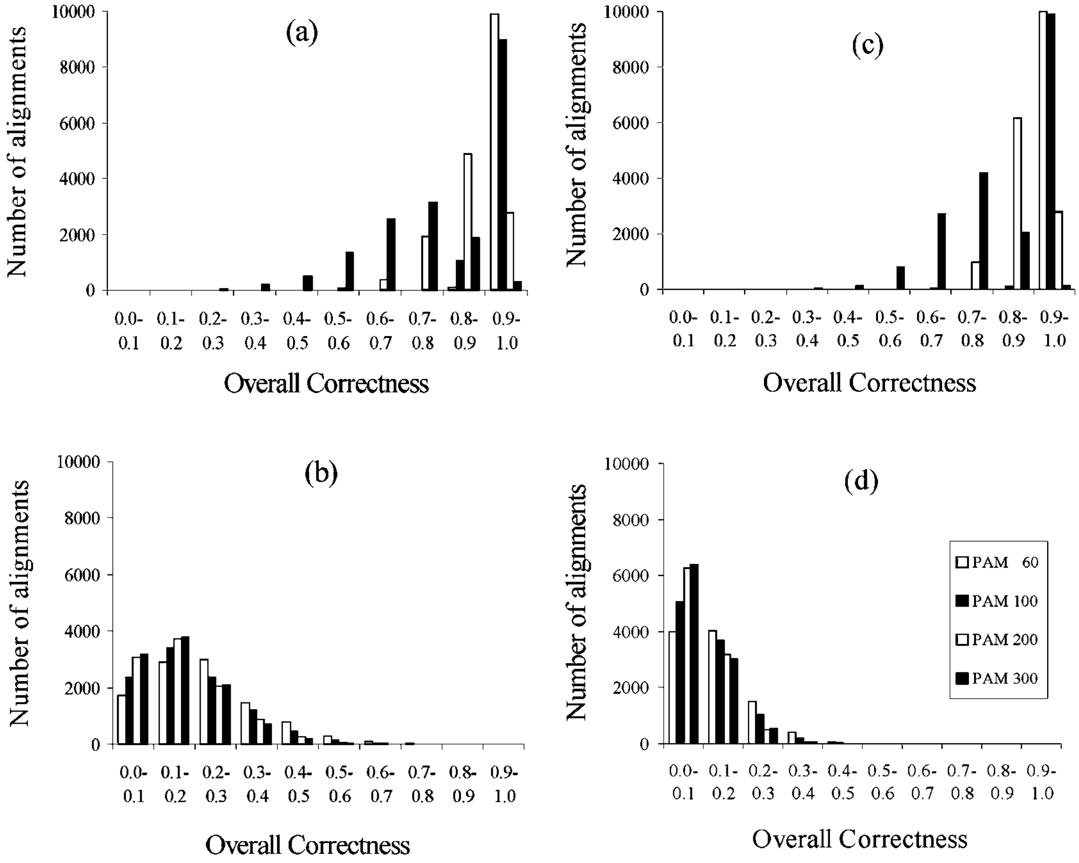

FIG. 1. Numbers of algorithmic alignments having a given overall correctness with respect to the corresponding golden standard alignment **(a,c)** and numbers of pairs of random alignments having given similarity **(b,d)**. The plots a and b correspond to the sequences of length 200; the plots c and d correspond to the sequences of length 500. In all diagrams, the similarity between the alignments is represented with the overall correctness measure.

Table 3. The Average Values of ACCuracy, CONFidence, and Overall CORRectness

			SW vs. GS			*Two random*
Length	*PAM*	*%id*	*Acc*	*Conf*	*Corr*	*Corr*
200	60	58	0.97	0.97	0.97	0.23
200	100	43	0.95	0.94	0.95	0.19
200	200	24	0.85	0.84	0.85	0.17
200	300	16	0.70	0.69	0.70	0.16
500	60	58	0.98	0.98	0.98	0.14
500	100	43	0.96	0.96	0.96	0.11
500	200	24	0.87	0.85	0.87	0.09
500	300	16	0.73	0.72	0.73	0.09

The average values of three measures of alignment quality (ACCuracy, CONFidence, and overall CORRectness) of Smith-Waterman (SW) alignments with respect to the golden standard (GS) alignments. For reference in the last column, we give the average overall correctness of a random alignment with respect to another independent random alignment. The data are given for all eight test sets (sequence lengths 200 and 500; PAM values 60, 100, 200, 300). The "%id" column contains average values of the sequence identity within the generated test sets.

TABLE 4. MAXIMAL VALUES OF THE HISTOGRAMS OF OVERALL CORRECTNESS

			SW vs. GS		*Two random*	
Length	*PAM*	*%id*	*Range*	*Maximum*	*Range*	*Maximum*
200	60	58	0.9..1.0	99.1	0.1..0.2	29.1
200	100	43	0.9..1.0	89.4	0.1..0.2	34.2
200	200	24	0.8..0.9	48.9	0.1..0.2	37.0
200	300	16	0.7..0.8	31.3	0.1..0.2	37.7
500	60	58	0.9..1.0	100.0	0.1..0.2	40.2
500	100	43	0.9..1.0	99.2	0.0..0.1	50.6
500	200	24	0.8..0.9	61.2	0.0..0.1	62.5
500	300	16	0.7..0.8	42.0	0.0..0.1	63.7

Maximal values ("maximum") of the histograms of overall correctness (Fig. 1) and the bins corresponding to the maxima ("range"). The data are shown both for the comparison of Smith-Waterman (SW) alignments and corresponding golden standard (GS) alignments, and for the comparison of two random alignments ("two random").

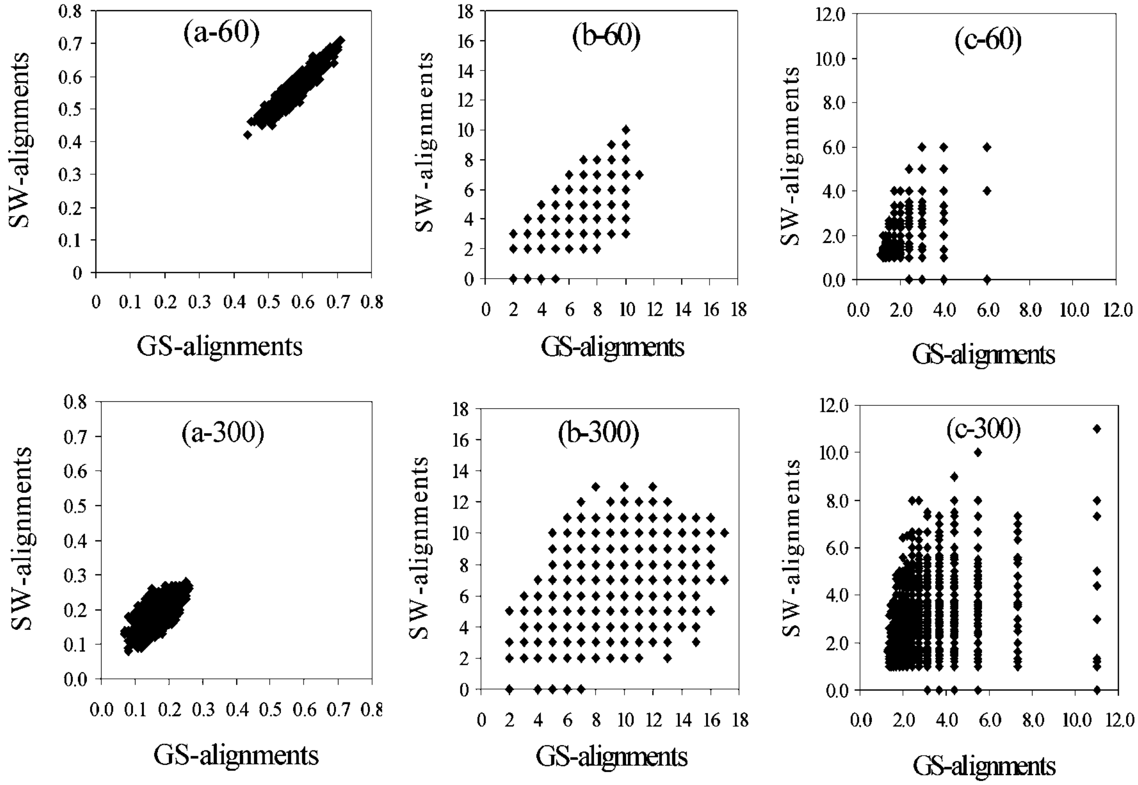

FIG. 2. Scatter-plots representing relations between different characteristics of the Smith-Waterman (SW) and "golden standard" (GS) alignments. Each point corresponds to a pair of proteins; the abscissa represents the value of a characteristic for the GS-alignment; and the ordinate represents the value for the SW-alignment. The following characteristics are represented: %id (**a**); the total number of indels (**b**); and the average indel length (**c**). The upper three plots give the data for the evolutionary distance of 60 point accepted mutations (PAM); the lower three plots give the data for the evolutionary distance of 300 PAM.

Table 5A. The Mean Values of %id of Smith-Waterman (SW) and Golden Standard (GS) Alignments for Different Test Sets

		%id		
Length	*PAM*	*SW*	*GS*	*Ratio*
200	60	57.5	58.1	0.99
200	100	42.6	43.1	0.99
200	200	24.5	24.3	1.01
200	300	17.9	16.2	1.10
500	60	57.7	58.1	0.99
500	100	42.9	43.1	1.00
500	200	24.6	24.3	1.01
500	300	18.0	16.3	1.10

The scatter-plots (Fig. 2A) are symmetrical; i.e., the difference between %id of algorithmic and GS alignments is insignificant. However, the plots in Figures 2B and 2C are asymmetric, which reflects the difference in the number of indels and the total indel length in the considered groups of alignments.

The %id value is approximately the same for algorithmic and GS alignments; there is a slight increase in the %id ratio for distant homologs (the case of PAM300). The standard deviation of the ratios is very low (Tables 5A and 6).

The main difference between the SW and GS alignments is the number of indels (Tables 5B and 6). The algorithmic alignments contain significantly fewer indels; the difference tends to grow with the growth of evolutionary distance.

The average indel lengths of SW and GS alignments are approximately same; the lengths of SW alignments are slightly greater. The standard deviations of ratios for all indel characteristics grow with the growth of evolutionary distance; the deviations for sequences of length 200 are larger than those for sequences of length 500 (Table 6).

In order to clarify the dependencies between the characteristics of indels (e.g., total number, mean length) and the degree of alignment reconstruction, we divided the overall set of aligned pairs of sequences into six subsets: the average length of indels in algorithmic alignments is less than in GS alignments (AL); the average length of indels in algorithmic alignments is greater than in GS alignments (AG); the average

Table 5B. The Mean Values of the Indel Characteristics (Number, Average Length, Total Length) of Smith-Waterman (SW) and Golden Standard (GS) Alignments for Different Test Sets

		Number of indels			*Average length of indels*			*Total length of indels*		
L	*PAM*	*SW*	*GS*	*Ratio*	*SW*	*GS*	*Ratio*	*SW*	*GS*	*Ratio*
200	60	4.85	5.69	0.85	2.17	2.11	1.03	10	12	0.83
200	100	5.49	6.85	0.80	2.43	2.34	1.04	14	16	0.88
200	200	5.78	7.98	0.72	2.61	2.51	1.04	16	20	0.8
200	300	6.10	8.54	0.71	2.67	2.58	1.03	16	22	0.73
500	60	9.37	10.58	0.89	2.92	2.83	1.03	25	30	0.83
500	100	10.81	12.72	0.85	3.26	3.14	1.04	35	40	0.88
500	200	12.67	16.06	0.79	3.41	3.36	1.01	45	55	0.82
500	300	13.46	16.12	0.83	3.29	3.47	0.95	45	55	0.82

Each "ratio" column contains the ratio of values of the respective Smith-Waterman (SW) over golden standard (GS) columns.

Table 6. Correspondence Between %id, Number of Indels, Indel Average Length, and Indel Total Length for SW and GS Alignments

		Algorithmic/true							
		%id (A)		*Number of indels (B)*		*Average length of indels (C)*		*Total length of indels (D)*	
L	*PAM*	*Mean*	*σ*	*Mean*	*σ*	*Mean*	*σ*	*Mean*	*σ*
200	60	0.99	0.02	0.86	0.17	1.02	0.18	0.88	0.19
200	100	0.99	0.03	0.82	0.17	1.03	0.21	0.83	0.2
200	200	1.01	0.06	0.75	0.2	1.03	0.28	0.76	0.23
200	300	1.11	0.14	0.75	0.25	1.02	0.35	0.74	0.27
500	60	0.99	0.01	0.89	0.11	1.03	0.13	0.92	0.13
500	100	0.99	0.02	0.86	0.12	1.03	0.15	0.89	0.14
500	200	1.01	0.04	0.81	0.15	1.01	0.2	0.79	0.16
500	300	1.11	0.08	0.87	0.22	0.94	0.23	0.80	0.19

The table describes the behavior of four alignment characteristics for Smith-Waterman (SW) and golden standard (GS) alignments. The characteristics are as follows: %id (A), number of indels (B), average length of indels (C), or total length of indels (D). For each characteristic f, we give the mean value ("mean") and standard deviations ("σ") of the ratio $f(\mathrm{SW}(S_1, S_2))/f(\mathrm{GS}(S_1, S_2))$ for eight test sets. Here, $\mathrm{SW}(S_1, S_2)$ and $\mathrm{GS}(S_1, S_2)$ are SW and GS alignments of the test sequences S_1 and S_2. Note, that, unlike in Table 5, we give the average values of ratios, not the ratio of average values. Thus, the numbers are slightly different.

lengths of indels in algorithmic alignments and in GS alignments are equal (AE); the same for the number of indels (NL, NG, NE).

Table 7A shows that the maximal number of sequence pairs falls into subset NG, characterized by the number of indels in GS alignments greater than in algorithmic ones (70% for sequences of length 200 and 74% for sequences of length 500), and in subset AL, characterized by an inverse relation of mean lengths of indels in GS and algorithmic alignments (42% for sequences of length 200 and 47% for sequences of length 500).

The maximal similarity of algorithmic and GS alignments is attained at the equivalence between mean lengths and the number of indels in algorithmic and GS alignments (Table 7B). The number of such

Table 7A. Subsets of Test Sets by Indel Characteristics in SW and GS Alignments

			Size of the subset, % of the full test set	
Indel characteristics	*Description of the subset*	*Subset code*	*L = 200*	*L = 500*
Average length of indels	SW < GS	AL	32.2	38.3
	SW > GS	AG	42.1	47.0
	SW = GS	AE	25.7	14.7
Number of indels	SW < GS	NL	70.6	74.0
	SW > GS	NG	3.8	7.4
	SW = GS	NE	25.6	18.5

Subsets of test sets meeting the following conditions: average length of indels in algorithmic alignments is less than in golden standard alignments (AL); average length of indels in algorithmic alignments is greater than in golden standard alignments (AG); average lengths of indels in algorithmic alignments and in golden standard alignments are equal (AE); same for the numbers of indels (NL, NG, NE). For each subset, we give its size as percentage of the full test set. Here, we consider joint test sets obtained as unions of test sets with PAM60, 100, 200, 300, and the given sequence length.

TABLE 7B. INTERSECTIONS OF THE SUBSETS

Description of the intersection	*Size of the subset, % of the full test set*		*Overall correctness, %*	
	L = 200	*L = 500*	*L = 200*	*L = 500*
AL and NL	25.7	26.7	83.2	84.9
AL and NG	3.4	7.1	77.5	79.9
AG and NL	40.9	46.1	84.9	89.5
AG and NG	0.4	0.4	64.8	70.7
AE and NE	21.6	13.4	96.7	97.9

Intersections of the following pairs of subsets: AL and NL, . . . AE and GE. For each intersection, we give (1) its size as percentage of the full test set (i.e., joint test sets obtained as unions of test sets with PAM60, 100, 200, 300, and the given sequence length) and (2) the average overall correctness over the intersection.

sequences decreases with the evolution distance increasing (data not shown). Most significant among subset intersections (Table 7B) are the subsets of alignments with the number of indels less than that in the GS alignment (AL and NL, AG and NL). The lowest correctness was observed for alignments with the number of indels greater than that in GS alignments (AL and NG, AG and NG); independently from the mean length of indels, this case is quite rare.

4. CONCLUSION

We have investigated the quality of global SW alignments (i.e., their similarity to the GS alignments); the latter alignments were obtained through modeling in accordance with the evolution model described in Dayhoff et al. (1978) and Benner et al. (1993). In other words, we have investigated if, for a protein P1 and its descendant P2, the pair-wise sequence alignment of P1 and P2 can reconstruct the evolutionary events leading from P1 to P2. As the next step, we plan to study the same question for alignments of descendants P1′ and P2′ of the same ancestor. The problems are equivalent for the symmetric model of evolution described in the above papers, but the asymmetric models (Jordan et al., 2005) can lead to a significant difference in the results.

We have performed the computer experiments for sequences of length 200 and 500, and for evolutionary distances from 60 to 300 PAM. Previous works (Sunyaev et al., 2004; Domingues et al., 2000) have investigated local alignments and have used the 3D-structure-based alignments as a golden standard. The 3D structures of proteins are much more conservative than their sequences; however, they also can lead to unpredictable errors in the estimation of quality of algorithmic alignments.

In line with the previous works, we have used two measures of alignment quality: (1) accuracy, reflecting what part of the true alignment was correctly restored, and (2) confidence, which shows what part of the algorithmic alignment coincides with the GS. Unlike local alignments, global ones have almost identical characteristics; therefore, we mainly used the average characteristic called overall correctness. As intuition suggests, the growth of PAM distance leads to a decrease in the alignment correctness and makes the maxima of the corresponding distributions less sharp. The length of compared sequences almost does not affect the average alignment correctness. The dispersion of alignment correctness for the test sets of length 500 is significantly less than for the sets of length 200 (Table 4).

The main difference between the GS and algorithmic alignments is the number of gaps; the average lengths of gaps are approximately the same. The total length of indels for all PAM distances is less in algorithmic alignments than in GS ones. This is caused by a substantial decrease in the number of indels in algorithmic alignments along with an insignificant increase in the mean length of indels. The more distinctly this trend is seen for a given pair of proteins, the greater the difference between their algorithmic and GS alignments.

TABLE 8A. THE DISTORTION OF ZIPF DISTRIBUTION AS A CONSEQUENCE OF SIEVE EFFECT

Length	200	200	200	500	500	500
PAM	50	100	200	50	100	200
$\|f - Z\|/\|Z\|$	0.31	0.29	0.26	0.25	0.21	0.19

f, obtained distribution of indel lengths; Z, Zipf distribution; $\|X\| = \sum |xi|$, norm.

TABLE 8B. ERRORS OF APPROXIMATION OF EMPIRICAL INDEL LENGTH DISTRIBUTIONS WITH ZIPF DISTRIBUTION

PAM	6	10	19	35	48	87
$\|E - Z\|/\|Z\|$	0.31	0.22	0.14	0.15	0.12	0.10

E, empirical distribution of indel lengths; Z, Zipf distribution; $\|X\| = \sum |xi|$, norm.

The relatively low number of indels in the algorithmic alignments follows from the form of the target function of the SW algorithm. One possible way to overcome the drawback is to penalize the deviation of the number of indels from a given desired number, rather than the appearance of an indel.

In our further work, we plan to implement the proposed approach to investigate the local alignments and their relations to the global ones. To do this, we will place the similar sequences into a much longer random environment. It seems to be interesting to compare the effectiveness of global and local approaches for different sizes of environment and different lengths of initial similar sequences.

5. APPENDIX

As a result of elimination of extended deletions during introducing (so called *sieve effect*), the short deletions will be overrepresented in final distribution; i.e., the obtained distribution differs from Zipf distribution (Table 8A). Following Benner et al. (1993), errors of approximation of empirical distributions of indel lengths with Zipf distribution are shown in Table 8B.

Thus, for PAM of >50, the distortion of distribution becomes greater than the approximation distribution error. With the aim of compensating for this distortion, correction of the actual indel length distribution vector can be undertaken.

Notations: $X = (x^1, \ldots, x^n)$, or vector of indel lengths distribution
$F(X) = (f^1(X), \ldots, f^n(X))$, or function of deletion introduction

Problem: Find $X^* : F(X^*) \approx Z$

For finding X^*, the iterative method of relaxation (Samarskii and Goolin, 1989) was applied:

Let us consider the sequence of vector X approximations:

$X^{i+1} = X^i$, or $\tau F(X^i)$,

X^i, or current value of lengths distribution,

$0 < \tau < 1$, or constant multiplier,

$X^0 = Z$.

Procedure of vector X approximation:

1. Introducing of indels in original sequence, $X^0 = Z$.
2. Computation of generated distribution of indel lengths X^1.
3. Check the condition: $\|F(X^1) - Z\| < \varepsilon$, where $\|X\| = \sum |x^j|$, or vector norm, and $\varepsilon > 0$, or small value.

If condition is not satisfied, then

4. Correction of vector X: $X^{i+1} := X^i - \tau F(X^i)$,

Continue step 1.

ACKNOWLEDGMENTS

This work was supported by RFBR (grant 06-04-49249), INTAS (grant 05-1000008-8028), and the Molecular and Cell Biology program of the Russian Academy of Sciences. The authors thank Leonard Okhotchinski for his great help in preparing the manuscript.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

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

Benner, S.A., Cohen, M.A., and Gonnet, G.H. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* 229, 1065–1082.

Dayhoff, M., Schwartz, R., and Orcutt, B. 1978. A model of evolutionary change in proteins, 345–352. *In* Dayhoff, M., ed., *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC.

Domingues, F.S., Lackner, P., Andreeva, A., et al. 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.* 297, 1003–1013.

Doolittle, R.F. 1981. Similar amino acid sequences: chance or common ancestry? *Science* 214, 149–159.

Hollich, V., Milchert, L., Arvestad, L., et al. 2005. Assessment of protein measures and tree building methods for phylogenetic tree reconstruction. *Mol. Biol. Evol.* 22, 2257–2264.

Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., et al. 2005. A universal trend of amino acid gain and loss in protein evolution. *Nature* 433, 633–638.

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

Litvinov, I.I., Lobanov, M.Yu., Mironov, A.A., et al. 2006. Information on the secondary structure improves the quality of protein sequence alignment. *Mol. Biol.* 40, 474–480.

Mevissen, H.T., and Vingron, M. 1996. Quantifying the local reliability of a sequence alignment. *Prot. Eng.* 9, 127–132.

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

Polyanovskii, V.O., Demchuk, E.Ya., and Tumanyan, V.G. 1995. Efficiency of alignment procedure in respect of reconstruction reliability. *Mol. Biol.* 28, 833–835.

Reese, J.T., and Pearson, W.R. 2002. Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics* 18, 1500–1507.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing of phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Samarskii, A.A., and Goolin, A.V. 1989. *Numerical Methods*. Nauka, Moscow [in Russian].

Schlosshauer, M., and Ohlsson, M. 2002. A novel approach to local reliability of sequence alignments. *Bioinformatics* 18, 847–854.

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

Stoye, J., Evers, S., and Meyer, F. 1998. ROSE: Generating sequence families. *Bioinformatics* 14, 157–163.

Sunyaev, S.R., Bogopolsky, G.A., Oleynikova, N.V., et al. 2004. From analysis of protein structural alignments toward a novel approach to align protein sequences. *Proteins* 54, 569–582.

Vingron, M., and Argos, P. 1990. Determination of reliable regions in protein sequence alignments. *Prot. Eng.* 3, 565–569.

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

Wallqvist, A., Fukunishi, Y., Murphy, L.R., et al. 2000. Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics* 16, 988–1002.

Waterman, M.S., ed. 1989. *Mathematical Methods for DNA Sequences*. CRC Press, Boca Raton, FL.

Address reprint requests to:
Dr. Vladimir G. Tumanyan
Engelhardt Institute of Molecular Biology
Russian Academy of Science (RAS)
Vavilova str., 32
119991, Moscow, Russia

E-mail: tuman@imb.ac.ru