Statistics of simulated homologous protein sequences pairwise alignment.

(Anothter variant of title:
Statistical approach to estimation of pairwise alignment reliability.)

V.O. Polyanovsky[1]1 , M.A. Roytberg[2]2 and V.G. Tumanyan[3]1
[4]1 Engelhardt Institute of Molecular Biology, RAS, Moscow,117984, Russia
[5]2 Institute of Mathematical Problems in Biology, RAS, Puschino, 142290, Russia

Statement of problem. The object of investigation was the pair-wise alignments of artificial homologous amino acid sequences. One of the sequences was a random Bernoulli sequence in 20-letter alphabet; the other was obtained from it by random sequence of mutation and indels; the set of indels determined the "true" alignment of the sequences. We tried to estimate the ability of the standard alignment procedure to reconstruct the "true" alignment. The similar problem for the case of alignments of real homologous proteins have been studied in Sunyaev et al, 2004 [6], where structural alignments were used as a "true" ones. However, structural alignments only approximate the alignments based on real evolutionary events. The above model gives an opportunity to study a situation when a "true" alignment is known exactly [7].
Methods. As the alignment procedure the widely used for global alignment Needleman-Wunsch algorithm [4] was employed.

The model pairs of sequences were generated in the following way. The first sequence of each pair was obtained using random number generator. Alphabet included 20 Latin letters, standing for amino acids in single-letter code; all positions were considered as independent and identically distributed. [??? Как выбирались вероятности символов???] At the next step the obtained Bernoulli sequence was modified by introducing substitutions according to Dayhoff model [1,3] with different pre-defined evolutionary distances (PAM= 60, 100, 200, 300) [??? Разве модель Dayhoff описывает deletions, insertions? Я убрал про них. См. следующую фразу.] Then we have chosen the random number of total deletion lengths D and introduced randomly the deletions of total length D into each of sequences, see details in the APPENDIX. [!!! Напишите это подробно –этот текст уже был. Если трудно перевести – пришлите по-русски]. Therefore, the final sequences to be aligned are of equal length.

All numerical experiments were performed with two sets of sequences; one set consisting of sequences of length 200; the other consisting of sequences of length 500 [??? Это длины после делеций или до ???] The numbers of sequences in different subsets were 100, 1000 and 10000 for initial and modified sequences.

The similarity between two alignments A[6]1 and A[7]2 was defined as (see [5]) :

---

[1]1
[2]2
[3]1
[4]1
[5]2
[6]1
[7]2

$$S(A^{8}1, A^{9}2) = 2\, C^{10}M / (M^{11}1 + M^{12}2),$$

where $M^{13}1$, $M^{14}2$ are the numbers of matches in the compared alignments;

$C^{15}M$ is the number of identically aligned residues in the alignments.

If $A^{16}1$ is "true" alignment and $A^{17}2$ is an algorithmic alignment of the same sequences, we refer to $S(A^{18}1, A^{19}2)$ as *reconstruction ability*.

Results. Table 1 shows the "most popular" value of reconstruction ability and its frequency for different evolutionary distances and sequence lengths (A). The column B gives analogous data for two random alignments

Table 1.

		A		B	
Length of Sequence	PAM	Reconstruction reliability Range of reconstr.	The number of alignments, %	Random alignments similarity Range of reconstr.	The number of alignments, %
200	60	0.9 - 1.0	92	0.1 - 0.2	39
	100	0.9 - 1.0	52	0.0 - 0.1	44
	200	0.7 - 0.8	36	0.0 - 0.1	53
	300	0.5 - 0.6	26	0.0 - 0.1	54
500	60	0.9 - 1.0	99	0.0 - 0.1	66
	100	0.9 - 1.0	70	0.0 - 0.1	77
	200	0.7 - 0.8	48	0.0 - 0.1	85
	300	0.5 - 0.6	36	0.0 - 0.1	87

We have also compared another characteristics of "true" and algorithmic alignment (see Table 2): (1) %id - "sequence identity" value; (2) average length of indels; (3) the number of indels. Correspondent data are given in the table 2. Evidently, the %id values of "true" and algorithmic alignments are quite similar; in contrast the number and total [??? average как раз отличается мало!!! – см. таблицу. Тут ошибки

[8] 1
[9] 2
[10] M
[11] 1
[12] 2
[13] 1
[14] 2
[15] M
[16] 1
[17] 2
[18] 1
[19] 2

нет?] length of indels are lower essentially in algorithmic in comparison to "true" alignments. ??? Хорошо бы дать и средние абсолютные цифры по “true” alignments

Table 2.

Length of sequence	PAM	ID: Alg / True		Number of indels: Alg / True		Indel average length: Alg / True	
		Mean value	σ	Mean value	σ	Mean value	σ
200	60	0.9773	0.0281	0.7984	0.1510	1.0424	0.2110
	100	0.9654	0.0429	0.7271	0.1581	1.0568	0.2497
	200	0.9866	0.0886	0.6113	0.1711	1.0315	0.3249
	300	1.1145	0.1695	0.5927	0.2016	0.9526	0.3679
500	60	0.9849	0.0176	0.8418	0.1058	1.0412	0.1501
	100	0.9760	0.0267	0.7783	0.1153	1.0503	0.1855
	200	0.9967	0.0561	0.6927	0.1467	0.9941	0.2439
	300	1.1168	0.1053	0.6795	0.1849	0.8926	0.2573

The work was supported by Molecular and Cell Biology program of RAS.

References:

1. Dayhoff,M., Schwartz,R. and Orcutt,B. (1978) A model of evolutionary change in proteins. In Dayhoff,M. (ed.), Atlas of protein sequence and structere, National Biomedical Research Foundation, Washington, pp. 345-352.
2. 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.
3. Reese,J.T., Pearson,W.R. (2002) Empirical determination of effective gap penalties for sequence comparison. Bioinformatics, vol.16, no.11, 1500-1507.
4. Needleman,S.B., 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.
5. Vogt,G., Etzold,T., Argos,P. (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twinlight zone revisited. J. Mol.Biol., 249, 816-831.
6. Sunyaev,S.R., Bogopolsky,G.A., Oleynikova,N.V., Vlasov,P.K., Finkelstein, A.V. and Roytberg,M.A. (2004) From analysis of protein structural alignments toward a novel approach to align protein sequences. Proteins: Structure, Function and Bioinforrmatics, 54(3), 569-582.
7. Polyanovskii,V.O., Demchuk,E.Ya., Tumanyan, V.G. (1995) Efficiency of alignment procedure in respect of reconstruction reliability. Molecular Biology, vol.28, no.6, part 2, 833-835.