# QUALITY OF LOCAL AND GLOBAL PAIR-WISE ALIGNMENTS OF AMINO ACID SEQUENCES

***Polyanovsky V.*[1], *Roytberg M.**[2], *Tumanyan V.*[1]
[1] Institute of Molecular Biology, Moscow, Russia
[2] Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia
* Corresponding author: e-mail: mroytberg@impb.psn.ru

*Motivation and Aim*: In many applications, the algorithmic alignment of two protein sequences ideally should restore the genuine one, i.e. the alignment that superimposes positions originating from the same position of the common ancestor of the proteins. Thus, it is important to know the average accuracy and confidence of the algorithmic alignments depending on the sequence similarity and to understand when the global alignment leads to better results than the local one and *vice versa.*

*Methods and Algorithms*: The local and global versions of Smith-Waterman alignment algorithm with standard values of parameters were considered. We have performed computer experiments; each experimental set consists of 1000 pairs of artificial amino acid sequences. Given the set, we align locally and globally all its pairs, and then compute average values of alignment accuracy and confidence. Each sequence pair $S_1$, $S_2$ in the experiment was generated with the following procedure, depending on three parameters ***(P, a, b);*** the parameters are the same for all sequences of the set**.** First, a random amino acid sequence $C_0$ of length $\boldsymbol{L}$ = 200 was generated. Second, we have produced two its independent "ancestors" $C_1$ and $C_2$ using the evolutionary model [1]; the model allows both replacements and indels, the frequencies of mutation events depend on the PAM value ***P.*** Third, we independently generate random "wings" $B_1$, $B_2$, $E_1$, $E_2$, the sequences to be compared are $S_1 = B_1 \cdot C_1 \cdot E_1$ and $S_2 = B_2 \cdot C_2 \cdot E_2$. The genuine alignment of $S_1$ and $S_2$ superimposes only positions of the "cores" $C_1$, $C_2$ corresponding to the same position of $C_0$. The wings reflect that $S_1$ and $S_2$ may have only local similarity. The lengths of the wings meet two conditions: (1) the total length $|B_1|+|E_1| = |B_2|+|E_2| = \boldsymbol{a \cdot L}$; (2) the cores Ci within Si (i=1, 2) are shifted into different directions: $|B_1| = |E_2| = (1- \boldsymbol{b}) \cdot (|B_1|+|E_1|)/2$ ( $\boldsymbol{b}$ = 0 corresponds to the absence of the shift). We have checked all combinations (***P, a, b)*** where ***P*** = 30, 60, 120, 240 (that is ~ 60, 40, 20 and 10 %id)***; a***= 0.1, 0.2, 0.5, 1.0, 2.0; ***b***= 0, 0.1, **…,** 0.9, 1**.**

*Experiments and results*: (1) Average confidence and accuracy are almost equal both for local and global alignments in all experiments. (2) Given ***P*** and ***a,*** the accuracy of the global alignment is almost equal to that with ***b*** =0 or is almost zero (as for ***b*** =1). The cut-off value of ***b*** depends on ***P*** and ***a*** **(**e.g. for ***P*** = 120 the cut-off is 0.1 if ***a*** =2.0; 0.2 for if ***a*** =1.0; 0.4 if ***a*** =0.5). (3) If ***b*** =0, then the global alignment has better accuracy than the local one. This suggests the 2-step alignment procedure: (1) perform *local* alignment to find similar fragments $F_1$, $F_2$ of the compared sequences $S_1$, $S_2$; let $F_i = S_i[\boldsymbol{c_i}, \boldsymbol{g_i}]$; $\boldsymbol{d_i = g_i - c_i}$. (2) perform *global* alignment of enlarged fragments $H_i = S_i[\boldsymbol{c_i - k d_i}, \boldsymbol{g_i + k d_i}]$ where ***k*** depends on sequence similarity, e.g. ***k***=0.5 if ***P=120***. The computer experiments show that the procedure improves the accuracy of local alignment on 5-10% , e.g. for ***P=120; a = 1; b=0.6.*** it gives 69 % instead of 64%

*Conclusion*: The global alignment gives better accuracy than the local one if the similar fragments are placed in the same positions of their sequences even if the non-similar wings are as long as the similar cores. This observation leads to the improvement of the accuracy of local alignment.

*Reference:*

1. 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.