Application of evolutionary model in the evaluation of the quality of pair wise alignment of amino acid sequences

Valery Polyanovsky, V. G. Tumanyan

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia,

e-mail: polyanovskyvo@mailfrom.ru, tuman@eimb.ru

The purpose of the work is to develop a universal method for estimating the pairwise alignment quality depending on the evolutionary distance (degree of homology) between the sequences being compared, and the type of alignment procedure. 3D alignments or any data on 3D-protein structure are not used in the study. Based on the accepted protein sequences evolution model, it is possible to estimate the capability of the concrete alignment algorithm to recover the genuine alignment. In this study a classical Needleman and Wunsch global alignment algorithm has been tested on a set of sequences from the Prefab database. Accuracy and confidence of a global alignment procedure were calculated as a function of the shares of insertions/deletions and mutations.

Keywords: pairwise alignment, quality of alignment, sequence homology.

INTRODUCTION

The «quality» of alignment algorithms, i.e. the correspondence between algorithmic and «reference» alignments, has been considered from different viewpoints, at that in the role of reference use was usually made of alignments based on collation of spatial structures. The basis for that is that protein 3D structures are essentially more conservative than the sequences corresponding thereto [1]. In work [2] a connection is shown between stability of the region of optimal global alignment in a set of suboptimal alignments and its similarity with structural alignment. The aim of the given work was application of the developed evolution model to estimation of the accuracy and confidence of a concrete alignment of two sequences obtained by one or another of the described algorithms.

METHODS

The procedure of determining the quality of a pairwise alignment consists of the following steps:

1) Determination of the values of characteristics of the estimated algorithmic alignment (mutations, insertions, deletions). 2) Generation of multiple test sets of pairs of sequences with close characteristics and their reference alignments. 3) Alignment of sequence pairs of each test set by a global algorithm. 4) Choice of a test set in which the mean characteristics of the algorithmic alignment minimally differ from the mean characteristics of algorithmic alignment of the tested pair of sequences. 5) Determination of the characteristics of quality of

algorithmic alignments: accuracy and confidence of the chosen test set on the basis of comparison of algorithmic alignments with the reference. The obtained values of accuracy and confidence present as an estimate of the specified alignment.

RESULTS

To test the proposed method we have chosen 37 pairs of protein sequences from the PREFAB 4.0 database intended for testing multiple alignment algorithms [3]. All chosen pairs had accuracy and confidence of global alignment in relation to reference alignment in the interval from 0.7 to 1.0. At that the share of point changes (mutations) in the estimated alignments turned out to be in the interval from 0.59 to 0.81, while the share of indels in the interval 0.014–0.289. As a result of testing we have obtained the dependencies of accuracy and confidence estimates on the share of indels. Upon a rise in the share of indels from 0.014 to 0.289, the values of accuracy estimates change in the range from 0.98 to 0.72, the values of confidence estimates change from 0.98 to 0.67. Obtained by the least square method, the angular coefficients of straight lines approximating the values of accuracy and confidence constitute –0.77 and –0.92, at a mean relative error of 0.020 and 0.027 respectively. In conclusion a comparison was conducted for the values of accuracy and confidence obtained on test alignments with the same values obtained from comparison of the estimated algorithmic alignments with their reference analogs from PREFAB 4.0 base. The results of comparison show a tendency to increase in the difference between values obtained from tests and values obtained from comparison with reference alignments with the growth of the share of indels. At that the mean value of relative deviation of the values of accuracy constitutes 7.6%, while of confidence – 8.7%.

REFERENCES

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

2. M.Vingron, P. Argos (1990) Determination of reliable regions in protein sequence alignments, *Prot Eng*, 3:565-569.

3. M. S. Edgar, C. Robert (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucl. Acids Res.,* 32:1792-97