Alignment of Biological Sequences, that Contains no More than One Deleted Fragment in Each Sequence.

I.M. GUKOV, T.V. ASTAHOVA, M.A. ROYTBERG,

The Institute of Mathematical Problems in Biology RAS, 142292,c.Puschino Moscow region,

mroytberg@mail.ru

I.I.TSITOVICH,

The Institute of Information Transferring Problems RAS, 127994, Moscow , B.Karetnij lane,19 , cito@iitp.ru

Arrangement

One of the important methodical problems of biological sequences alignments is theoretical grounds of deletion penalties choice. It was suggested considerations in paper [2] : deletion penalty must be chosen so that [1] incensement of deletions become unprofitable, if brings insignificant (random) incensement of alignment weight; penalty must be minimal, that satisfied condition [1.] The object oa this work was statistical investigation of simple alignments, that are optimal on matches amount in sequences of equal lengths, more than one deletions in each sequence allowed (these fragments are of equal length if it is necessary.) Even such model case permit to obtain new evidence on the character of symbol sequences. All the experiments was carried with sequences in 4-letter alphabet and simulate alignments of nucleic sequences. Identity matrix was considered as a weight matrix.

Methods

In each experiment one sequence was random with uniform symbol distribution in 4-letter alphabet (if not claimed contraries: in several experiments was used sequences in 20-letter alphabet, that simulate proteins.) Second sequence came of first by mean of two under mentioned models of mutation: (1) number of substitutions is fixed; their positions are ; every new symbol gets according substitution matrix [BLOSUM62] for proteins and has equal probabilities for nucleic acids; (2) every symbol transformed independently, probability save initial symbol is fixed: every new symbol gets as in model 1. Our experiments show that results depends on model choosing only with a small part of conservative positions. Values of expectation and dispersion under mentioned variables are little bigger for model 2.

We use follow designations (all them belong to pair of given sequences):

Length – Length of compared sequences (is equal); *Match0* – the weight of comparison of initial sequences; *(p1, p2, Del)* – alignment such, that in every sequence one fragment of length *Del* is deleted(in first sequence – from position *p1*, in 2-nd sequence – from *p2*); *Del* – length of each deleted fragment; *shift(p1, p2, Del)* = |*p1-p2*| - value of shift between sequences; *disrupt(p1,* *p2, Del)* –weight of “disruped” matches when delete fragments, i.e. weight of fragment of initial “trivial” alignment from position *min(p1, p2)* to position *max(p1, p2)+Del.*

Then let *MaxMatch* – maximum Match on all possible pairs of deleted fragments of equal lengths: *Score = MaxMatch – Match0* – “score”, obtained by mean of deletion fragments and corresponding shifting of sequenses; *Shift u Disrupt* –corresponding this optimal alignment values of characteristics *shift* and *disrupt.*

Results.

1. An analisis of empirical distribution functions of values *Score, Disrupt, Shift* show , that Γ-distribution is most profitable for its approximation. It was established , that values *Disrupt u Shift* have distribution close to Γ-distribution when *Length = 500*. Values of coeffitient *alpha* became stable near value 3.
2. Mean values *E Score, E Disrupt, E Shift* grow linear with logarithm of sequence length grows. Such a dependences are significant if level of validity equal 0.99.
3. Logarithms of mean values of variables Shift, Score, Disrupt dependences on conservative level *Save* is linear. Such a dependences are significant if level of validity equal 0.99.

LITERATURE

1. St.Henikoff, J.G.Henikoff (1992) Amino acid substitution matrices from protein blocks, *Biochemitry,* **11:** 10915-10919.
2. М.А.Ройтберг, М.Н. Семионенков, О.Ю. Таболина (1999) Парето-оптимальные выравнивания биологических последовательностей, *Биофизика*, **44.4:** 581-594.
3. М.С.Уотермен (1999) Математические методы для анализа последовательностей ДНК, М.: Мир.