TRANSITIVE SUBSET SEEDS FOR PROTEIN ALIGNMENT

Furletova E.[1], Kucherov G.[2], Noe L.[2], Roytberg M.*[1], Tsitovich I.[3]
[1]Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia
[2] LIFL,Villeneuve d'Ascq, Cedex, France
[3] Institute for information transmission problems, Moscow, Russia
*Corresponding author: e-mail: mroytberg@impb.psn.ru

Motivation and Aim: In [1] we have introduced the class of subset seeds and have demonstrated their advantages for the DNA comparison. In case of protein seeds the number of possible seed letters (i.e. subsets of the set of amino acid pairs *AP*) is $\sim 2^{210}$ thus one has to start with the choice of seed alphabet. Here we consider hierarchical and non-hierarchical transitive alphabets. A letter *D* is *transitive* if *D* contains all pairs *(a,a)* and for all $p_1, p_2, p_3 \in AP$ if $(p_1, p_2), (p_2, p_3) \in D$ then $(p_1, p_3) \in D$. Transitive letters are in one-to-one correspondence with partitions of the amino acid alphabet A. The *transitive* alphabet is a set of transitive letters including *Match* that corresponds to the partition where all amino acids are separated. The *hierarchical* transitive alphabet (HTA) is a set of embedded letters. In a *non-hierarchical* transitive alphabet (NHTA) letters are not necessarily embedded. Transitive letters allow for a direct hashing scheme within the database search. Our aim is to find out if it is possible to design subset seeds with better or equal selectivity (for a given sensitivity) than more complicated BLASTP-like seeds or vector seeds.

Methods and Algorithms: To design the desired seeds we use the three stage procedure: (A) design the transitive alphabet (hierarchical or not); (B) reveal seeds with maximal selectivity for the given set of sensitivity values; (C) same as (B) for the multi seeds. To perform the step (A) every amino acid pair is associated with both background and foreground probabilities. The *background* distribution is the distribution of letters in the aligned independent sequences, and the *foreground* distribution corresponds to the really interesting alignments. Let $D = \{p_1, \ldots, p_k\} \subseteq AP$ be a subset seed letter; $b(D) = \sum_{i=1,k} b(p_i)$ and $f(S) = \sum_{i=1,k} f(p_i)$ be its background and foreground probabilities. To find a good HTA we start with $R_1 = Match$ and then recursively produce R_{k+1} from R_k according to the following algorithm. For all pairs of classes C_1, C_2 from the partition R_k we find $W(C_1, C_2) = f(Br(C_1, C_2))/ f(Br(C_1, C_2))$, where $Br(C_1, C_2) = \{(a,b) \in AP \mid a \in C_1, b \in C_2\}$. Then we unite C_1, C_2 having maximal value of $W(C_1, C_2)$ into one class of R_{k+1}. To find a good NHTA we use similar algorithm. To produce successors of a member of *k*-th generation we maintain not one but *P=50* partitions having best value of *W*. The *(k+1)*-th generation is formed as Q=200 best successors of k-th generation. Thus we have ~10000 and to pick out of NHTA we used different heuristics based on the two following ideas: (1) we prefer the letters with high likelihood ratio; (2) the alphabet should contain letters of different weights. Given a transitive alphabet, steps (B) and (C) can be performed with proper genetic algorithms.

Experiments and Results: We have shown that the transitive seeds of size 4 (both hierarchical and non-hierarchical) can outperform the efficiency of BLASTP-like seeds.

The seed selectivity was computed by the algorithm from [1] according to the Bernoully model with probabilities corresponding to the BLOSUM62 substitution matrix. The same result was also shown for the experiments with real testing sets (BAliBase, HOMESTRAD, etc.)

References:

1. G. Kucherov, L. Noe, M. Roytberg. A unifying framework for seed sensitivity and its application to subset seeds. Journal of Bioinformatics and Computational Biology 4 (2006) 553–570