Construction of compositionally biased amino acid substitution matrices to improve annotation of Plasmodium falciparum proteins

Brick K, Pizzi E

Dipartimento di Malattie Infettive, Parassitarie ed Immunomediate, Istituto Superiore di Sanità, Roma

Motivation

Protein alignment algorithms such as BLAST and FASTA, currently use substitution matrices based on average amino acid distributions of conserved domains to score hits. However, due to the method of construction of these matrices, when proteins of biased amino acid distribution, or with large low complexity domains are aligned, the implicit frequencies in these matrices do not reflect accurately the protein composition. Proteins of the most virulant strain of the human malaria parasite, Plasmodium falciparum, exhibit a strong amino acid bias, as a result of a highly AT-biased genome (~80%). Some 60% of the proteins of this organism remain annotated as hypothetical and efforts to align these proteins with those of other known species have met with little success. Our approach aims to improve the annotation of this group of proteins, through the construction and use of matrices which more closely reflect the implicit amino acid bias.

Methods

Amino acid frequency profiles were generated for each of the 28337 blocks of multiple alignments in the BLOCKS database (Henikoff and Henikoff 1991, Smith 1990). A range of substitution matrices were developed using a perl algorithm based on the same underlying mathematical structure as the commonly used BLOSUM and PAM matrices. This model dictates that a matrix can be built in the log-odds form: sij = 1/λ * (qij/pi*pj), where there is at least one positive score and the expected score is negative. Our algorithm derived the observed amino acid pair frequencies (qij) and expected amino acid background frequencies (pi) for each set of blocks, created a matrix in half bit units, and provided the statistical parameters (expected score and un-gapped entropy) of each matrix. Each set of blocks was developed into a set of matrices clustered at 100% and 62%. Clustering was based on an original hierarchical clustering method, in which each cluster was composed of sequences all sharing a given percentage identity. When a sequence contributed to more than one block, the contribution of that sequence to amino acid counts was scaled by the inverse of the number of blocks in which it was a member. The alignments yielded by each matrix were then compared with those from a BLOSUM matrix with equivalent gapped statistical parameters. A novel procedure was used to quantitatively compare the statistical parameters (bit score, E-value) and length of each hit to evaluate improvements. This allowed us to select a subset of matrices, (SCAMB) which showed the most improved alignments using this technique.

Results

Our matrices, due their biased and symmetrical nature, are suited to the comparison of proteins with others of similarly biased amino acid distributions. We have shown, that both in searches using BLASTP and the Smith-Waterman algorithm (SSEARCH) a substantial improvement is seen in the alignments of a set of P.falciparum antigen proteins using a subset of these matrices. Specifically, it is shown that annotations of all members of the recently categorised family of parasite erythrocyte membrane proteins, the SURFINS (Winter et al. 2005), are improved when aligned using the SCAMB matrices. 1. PFA0725w is a P.falciparum protein which has been suggested to play a dual role in the life cycle of the parasite, both as an antigen presented on the host erythrocyte membrane and as a protein probably involved in invasion of host red blood cells. When aligned against a protein database, using the SCAMB matrix, similarity to several P.falciparum erythrocyte surface proteins is improved, and novel hits to axoneme proteins from other species are found. These results validate the dual role hypothesis and suggest what the role of this protein in the malarial invasion process may be. 2. Alignments between highly similar members of the same family, show slight changes to the alignments, which result in improved statistical parameters (bit score, E-value). 3. All SURFINS had an increased number of novel hits to long low complexity regions in other antigens in a P.falciparum protein database. It should be noted that a large proportion of these hits had an increased Bit score, decreased E value and an increased hit length. We suggest that a refinement of this approach may yield further improvements in the annotation of Plasmodium and other biased genomes.

Contact email: kevbrick@gmail.com

References:

- Henikoff,J.G. and Henikoff,S. (1992) Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA, 89,10915-10919
- Smith,H.O. (1990) Finding sequence motifs in groups of functionally related proteins, Proc. Natl. Acad. Sci. USA, 87,826-830
- Winter,G., et al. (2005) SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes., Journal of Experimental Medicine, 11,1853-1863