Construction and analysis of compositionally biased substitution matrices for alignment of Plasmodium spp. proteins

ID - 128

Brick Kevin[1], Pizzi Elisabetta[1]

[1]Dipartimento di Malattie Infettive, Parassitarie ed Immunomediate, Istituto Superiore di Sanita, 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. For such proteins, hit distributions often do not adhere to the expected extreme value distribution, resulting in large numbers of hits with artificially low e-values and high bit scores. Proteins of the most virulent strain of the human malaria parasite, Plasmodium falciparum, and a rodent malaria, P.yoelii, exhibit a strong amino acid bias, as a result of a highly AT-biased genome (~75% in coding regions). More than 60% of the proteins of these organisms remain annotated as hypothetical however, among these, orthologs do exist between species. Our approach in this work aims to develop develop substitution matrices from a more representative amino acid background for aligning these proteins, investigate the differences between these and BLOSUM matrices, and finally, examine the effects of using these matrices to align related proteins between two strains of malaria parasite containing such intrinsic bias.

Methods

Amino acid frequency profiles were generated for each of the 28337 blocks of multiple alignments in the BLOCKS database, and blocks with an amino acid distributions matching that of P.falciparum hypothetical proteins were used to construct matrices. A range of substitution matrices were developed in this way 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/lambda * (qij/pi*pj), where there is at least one positive score and the expected score is negative. Using this algorithm, each set of blocks was developed into a set of matrices clustered at 80% and 62%. The ungapped entropy of each matrix was constrained to that of the equivalent BLOSUM matrix 2%. Clustering was based on an original hierarchical clustering method, in which each cluster was composed of sequences all sharing a given percentage identity. Proteins from P.falciparum, and the related rodent malaria parasite P.yoelii were than aligned. Using a novel method based on regressing through the gene ontology tree, we assessed the retrieval accuracy of our matrices compared to BLOSUM62 using (a) regular NCBI BLAST, (b) NCBI BLAST with different adjustments for compositionally biased proteins.

Results

High complexity regions of P.falciparum proteins were found to exhibit a significant amino acid bias from the intrinsic background frequencies of the BLOSUM62 matrix. Using this profile as a template, matrices generated were shown to dramatically improve the specificity of retrieval accuracy for regular BLAST searches (using BLOSUM matrices) on our dataset of 3823 Plasmodium proteins. These alignments are also more specific than those achieved using BLAST with compositionally adjusted matrices, though less sensitive. Upon investigation into the nature of the improvements in the alignments, it is seen that there is a reduction in the number of spurious hits to long low complexity domains, and an increase in shorter, more concise hits to these regions, in agreement with recent results from another study. Application of these matrices to local and multiple alignments tasks will improve confidence in annotation and homology mappings of the proteins of these and related organisms.

Email: kev.brick@gmail.com