Jordi Cohen

Essay #3
PHYCS 498BIN

Non-symmetric score matrices and membrane proteins

Homologous proteins search engines such as BLAST use *score matrices* to assign a score to potential protein alignments (which are then used to find homology matches). The scoring matrices that are currently trendy, such as PAM and BLOSUM, are very well suited for general-purpose searches, but they perform sub-optimally when they are used to compare hydrophobic protein domains such as those found in membrane proteins. Hydrophobic domains have different distributions of amino acids than those found in water-soluble domains. It has been suggested that new scoring matrices should be developped, that would be especially suited for the comparison of the hydrophobic parts of proteins. Recently, two scoring matrix systems, SLIM [1] and PHAT [2] have been proposed, that perform noticeably better in queries that involve such protein segments. These matrices have unusual properties such as asymmetric off-diagonal components as well as negative diagonal elements. In this essay, I will present these matrices and describe their properties.

Score Matrix Theory

Scoring matrices are a fundamental component of most of the currently available protein comparison and alignment tools. A score matrix S is a 20x20 matrix in which each row and column corresponds to a given amino acid residue. Each element S_{ij} therefore corresponds to an amino acid pair (i,j) and represents the likelihood of the specified pair to occur in related sequences as opposed to random ones. This can be written as,

$$S_{ij} = \lambda \cdot \ln\left(\frac{P_{target}(i,j)}{P_{background}(i,j)} \right)$$

where $P_{background}$ is the natural probability of occurrence of the pair (i,j) in two unrelated sequences, P_{target} is the probability of occurrence of the pair in the alignment of two related sequences, and λ is a normalization factor. The latter probability is a function of the evolutionary distance between the proteins and can be either modeled, or found by analysing correlations in the protein databases. The local scores S_{ij} can be used to assign a global score to protein alignments and comparisons in order to find the best match.

Hydrophobicity and the background frequencies

When looking for homologous proteins, we wish to repeatedly compare a *subject* protein, whose function we ignore, with *query* proteins taken from a database. The query proteins can be parsed and subdivided into hydrophobic and –phylic regions, and different specialized score matrices should be used for each (e.g.: BLOSUM for the water-soluble parts and SLIM for the transmembrane parts). The main difference between the hydrophobic and –phylic analyses is that the background probabilities of amino acid pairs $P_{background}(i,j)$ are drastically different for both cases and the new specialized matrices should take that into account. The target probabilities (for related sequences) $P_{target}(i,j)$ are not significantly affected since they relate to the functional similarities between the amino acids in the pair, rather than their natural occurance.

The background frequencies for amino acid pairs can be computed from the distribution of the natural occurrence of amino acids in proteins. Hydrophobic protein regions will have a much higher concentration of polar residues as opposed to water soluble ones. We therefore need to use a completely different amino acid distribution when computing the background frequencies for "hydrophobic" score matrices.

The PHAT matrices use this skewed background distribution for both the query and subject proteins instead of the more general one and outperforms BLOSUM in protein searches. The SLIM matrices, on the other hand, use a more complicated correlation for amino acid pairs. For SLIM, the authors note that, while the query proteins are hydrophobic and their constituent amino acids should be taken from the new hydrophobic distribution, the subject protein's function, on the other hand is unknown, and its constituent acids should conform to the general distribution of all proteins. We thus have an asymetry in our matrix. The subject and query amino acids have different background distributions and in general we get that $P_{bkgnd}(i,j) \neq P_{bkgnd}(j,i)$.

A strange bi-product of "hydrophobic" matrices are possible negative scores along the diagonal. The diagonal elements represent the score of having an amino acid align with itself, and we expect these quantities to be strongly positive. However, the opposite is sometimes observed in SLIM matrices for comparing sequences that are evolutionarily distant. We expect extremely distant proteins to be almost uncorrelated (beyond what can be expected from the background). However, the backgrounds are different for hydrophobic and –phylic sequences, so we expect that unrelated hydrophobic protein pairs would score better than a hydrophylic-hydrophobic pair. It is then possible, at large evolutionary distances, that a pair of identical polar residues would be *less probable* to be found in related aligned sequences than in random sequences (with the appropriate amino acid biases).

Performance

The SLIM and PHAT proteins are derived from the same amino acid frequencies, except that the SLIM matrices have asymmetric distributions for the subject and query proteins, as opposed to the PHAT matrices. Testing has been done on these proteins (which I will not cover, but the details can be found in the references). It is shown that

SLIM consistently outperforms PHAT, which in turn outperforms BLOSUM for finding homologous proteins to subject proteins with inter-membrane domains. Performance is measured in different highly technical ways, but what it translates to in common language is that for a given fixed percentage of false positives that one is ready to accept, the SLIM matrix will find between 0-15% more matches than a comparable BLOSUM matrix, when the subject protein is an intermembrane proteins.

Conclusion

We have seen that asymetric score matrices based on specially selected natural amino acid distributions can enhance sequence comparisons for membrane proteins. These methods could technically be extended to other situations, such as the hydrophylic part of proteins, or any other specialized domains that would statistically contain a skewed distribution (maybe proteins could be subdivided into α-helices, β-sheets, etc. ?). An extensive set of specialized matrices, combined with a sorted database, would improve current comparison searches. The unfortunate consequence of this method, though, is that we would be adding a tremendous amount of complexity into our algorithms, as opposed to finding a simpler general one, which is ideally what we ultimately would like to find.

References

[1] Müller T., Rahmann S., Rehmsmeier M., *Bioinformatics* **17** (S1), S182-S189 (2001)
[2] Ng P., Henikoff J., Henikoff S., *Bioinformatics* **16** (9), 760-766 (2001)