Splitting the BLOSUM score into numbers of biological significance

FRANCESCO FABRIS * [(1)], ANDREA SGARRO[(1)] and ALESSANDRO TOSSI[(2)]

June 15, 2005

[(1)] Dipartimento di Matematica e Informatica, Università di Trieste, via Valerio 12b, 34127, Trieste, Italy

[(2)] Dipartimento di Biochimica, Biofisica e Chimica delle Macromolecole, Università di Trieste via Giorgieri 1, 34127 Trieste, Italy.

We have analyzed the meaning of the BLOSUM score by using mathematical tools developed in the context of Shannon Information Theory. In particular, we split the BLOSUM score into three terms, that we call the *BLOSUM spectrum* (or BLO*Spectrum*), related respectively to the stochastic similarity of the two protein sequences (sequence convergence), to the typicality of the amino acid probability distribution in each sequence (background frequency divergence), and to the compliance of the amino acid variations between the two sequences to the protein model implicit in the BLOCKS database (target frequency divergence). This sharpens the protein sequence comparison, giving a rationale to the biological significance of the obtained score, and helping to find weakly related sequences. Moreover, the BLO*Spectrum* can guide in choosing the most appropriate scoring matrix tailored to the evolutionary divergence associated with the two sequences, or in deciding if a compositionally adjusted matrix could perform better.

The meaning of the BLO*Spectrum* terms is the following:

- *Mutual Information* $I(X,Y)$ is the *sequence divergence* between the aligned sequences. It measures the degree of *stochastic dependence* (or stochastic correlation) between X and Y; the greater its value, the more statistically correlated are the two sequences. This measure is highly correlated with, though of course not identical to, the percent identity of the alignment under consideration; this in

* frnzfbrs at dsm.univ.trieste.it, tel. +39 040 558 2625, fax +39 040 558 2636

the sense that the greater the percent identity, the greater the stochastic correlation between the aligned sequences, but the opposite does not hold, since high correlation means the propensity of finding certain amino acids paired, even if different.

This term enhances the overall BLOSUM score, since it is taken with the plus sign.

- $D(F_{XY}//P_{AB})$ is the *target frequency divergence*. It measures the difference between the ”observed” target frequencies, and the target frequencies implicit in the substitution matrix. In mathematical terms it measures the *stochastic distance* between F_{XY} and P_{AB}, that is the distance between the mode in which amino acids are paired in sequences X and Y, *and* the mode in which amino acids are paired inside the “protein model” implicit in the database BLOCKS. When the vector of observed frequencies F_{XY} is far from the vector of target frequencies P_{AB} exhibited by the protein model, then the divergence is high. On the contrary, if F_{XY} is almost equal to P_{AB}, then the divergence is very low. Another interpretation is the following: if the divergence is high, then starting from X we obtain a Y (or *viceversa*) that is not that we would expect on the basis of the target frequencies of the database; in other words, the amino acids are paired following relative frequencies that are not the standard ones.

  The term $D(F_{XY}//P_{AB})$ is a penalty factor in equation (**??**), since it is taken with the minus sign.

- $D(F_X//P_A)$ ($D(F_Y//P_B)$) is the *background frequency divergence* of sequence X (Y). It measures the difference between the ”observed” background frequencies, and the background frequencies implicit in the substitution matrix. In mathematical terms it measures the *stochastic distance* between the observed frequencies F_X (F_Y) and the vector $P = P_A = P_B$ of background frequencies of the amino acids inside the database BLOCKS. The greater its value, the more different are the observed frequencies from the background frequencies exhibited by a typical protein sequence.

  This term enhances the score, since it is taken with the plus sign.