Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 31450, 18 pages
doi:10.1155/2007/31450

*Research Article*

# Splitting the BLOSUM Score into Numbers of Biological Significance

**Francesco Fabris,[1,2] Andrea Sgarro,[1,2] and Alessandro Tossi[3]**

[1] *Dipartimento di Matematica e Informatica, Università degli Studi di Trieste, via Valerio 12b, 34127 Trieste, Italy*
[2] *Centro di Biomedicina Molecolare, AREA Science Park, Strada Statale 14, Basovizza, 34012 Trieste, Italy*
[3] *Dipartimento di Biochimica, Biofisica, e Chimica delle Macromolecole, Università degli Studi di Trieste, via Licio Giorgieri 1, 34127 Trieste, Italy*

Received 2 October 2006; Accepted 30 March 2007

Recommended by Juho Rousu

Mathematical tools developed in the context of Shannon information theory were used to analyze the meaning of the BLOSUM score, which was split into three components termed as the BLOSUM spectrum (or BLO*Spectrum*). These relate respectively to the sequence convergence (the stochastic similarity of the two protein sequences), to the background frequency divergence (typicality of the amino acid probability distribution in each sequence), and to the target frequency divergence (compliance of the amino acid variations between the two sequences to the protein model implicit in the BLOCKS database). This treatment sharpens the protein sequence comparison, providing a rationale for the biological significance of the obtained score, and helps to identify weakly related sequences. Moreover, the BLO*Spectrum* can guide the choice of the most appropriate scoring matrix, tailoring it to the evolutionary divergence associated with the two sequences, or indicate if a compositionally adjusted matrix could perform better.



## 1. INTRODUCTION

Substitution matrices have been in use since the introduction of the Needleman and Wunsch algorithm [1], and are referred to, either implicitly or explicitly, in several other papers from the seventies, McLachlan [2], Sankoff [3], Sellers [4], Waterman et al. [5], Dayhoff et al. [6]. These are the conceptual tools at the basis of several methods for attributing a similarity score to two aligned protein sequences. Any amino acid substitution matrix, which is a $20 * 20$ table, has a scoring method that is implicitly associated with a set of *target frequencies* $p(i, j)$ [7, 8], pertaining to the pair $i$, $j$ of amino acids that are paired in the alignment. An important approach to obtaining the score associated with the paired amino acids $i$, $j$, was that suggested by Dayhoff et al. [6], who developed a stochastic model of protein evolution called PAM (points of accepted mutations). In this model, the frequencies $m(i, j)$ indicate the probability of change from one amino acid $i$ to another amino acid $j$, in homologous protein sequences with at least 85% identity, during short-term evolution. The matrix $M$, relating each amino acid to each of the other 19, with an evolutionary distance of 1, would have entries $m(i, j)$ close to 1 on the main diagonal ($i = j$) and close to 0 out of the main diagonal ($i \neq j$). An $M^k$ matrix, which estimates the expected probability of changes at a distance of $k$ evolutionary units, is then obtained by multiplying the $M$ matrix by itself $k$ times. Each $M^k$ matrix is then associated to the scoring matrix PAM$^k$, whose entries are obtained on the basis of the log ratio

$$s(i, j) = \log \frac{m^k(i, j)}{p(i)p(j)}, \tag{1}$$

where $p(i)$ and $p(j)$ are the observed frequencies of the amino acids.

S. Henikoff and J. G. Henikoff introduce the BLOck SUbstitution Matrix (BLOSUM) [9]. While the scoring method is always based on a log odds ratio, as seems natural in any kind of substitution matrices [7], the method for deriving the target frequencies is quite different from PAM; one needs evaluating the *joint* target frequencies $p(i, j)$ of finding the amino acids $i$ and $j$ paired in alignments among homologous proteins with a controlled rate of percent identity. This joint probability is compared with $p(i)p(j)$, the product of the *background frequencies* of amino acids $i$ and $j$, derived from amino acids probability distribution $P = \{p_1, p_2, \ldots, p_{20}\}$.

The target and background frequencies are tied by the equality $p(i) = \sum_{j=1}^{20} p(i, j)$ so that the background probability distribution is the *marginal* of the joint target frequencies [10]. The product $p(i)p(j)$ reflects the likelihood of the independence setting, namely that the amino acids $i$ and $j$ are paired by pure chance. If $p(i, j) > p(i)p(j)$, then the presence of $i$ stochastically induces the presence of $j$, and vice versa ($i$ and $j$ are "attractive"), while if $p(i, j) < p(i)p(j)$, then the presence of $i$ stochastically prevents the presence of $j$, and vice versa ($i$ and $j$ are "repulsive"). The log ratio (taken to the base 2)

$$s(i, j) = \log \frac{p(i, j)}{p(i)p(j)} \tag{2}$$

furnishes the score associated with the pair of amino acids $i$, $j$, when these are found in a certain position $h$ of an assigned protein alignment; it is positive when $p(i, j) > p(i)p(j)$, and negative when the opposite occurs. The $i$, $j$ entry of the BLOSUM matrix is the score of the pair $i$, $j$ (or $j$, $i$, which is the same since the sequences are not ordered; for a different approach see Yu et al. [11]) multiplied by a suitable scale factor (4 for BLOSUM-35 and BLOSUM-40, 3 for BLOSUM-50, and 2 for the remaining). The value so obtained is then rounded to the nearest integer, and the (unscaled) global score of two sequences $X = x_1, x_2, \ldots, x_n$ and $Y = y_1, y_2, \ldots, y_n$ of length $n$ is given by summing up the scores relative to each position

$$S(X, Y) = \sum_{h=1}^{n} s(x_h, y_h) = \sum_{i,j} n(i, j) \log \frac{p(i, j)}{p(i)p(j)}, \tag{3}$$

where $n(i, j)$ is the number of occurrences of the pair $i$, $j$ inside the aligned sequences. This equation weighs the log ratio associated to the $i$, $j$ entry of the BLOSUM matrix with the occurrences of the pair $i$, $j$, and seems intuitive following a heuristic approach, as any reasonable substitution matrix is implicitly of this form [7]. In order to compute the necessary target and background frequencies $p(i, j)$ and $p(i)p(j)$, S. Henikoff and J. G. Henikoff used the database BLOCKS (http://blocks.fhcrc.org/index.html), which contains sets of proteins with a controlled maximum rate of percent identity "$\theta$" that defines the BLOSUM matrix, so that BLOSUM-62 refers $\theta = 62\%$, and so forth.

Scoring substitution matrices, such as PAM or BLOSUM, are used in modern web tools (BLAST, PSI-BLAST, and others) for performing database searches; the search is accomplished by finding all sequences that, when compared to a given query sequence, sum up a score over a certain threshold. The aim is usually that of discovering biological correlation among different sequences, often belonging to different organisms, which may be associated with a similar biological function. In most cases, this correlation is quite evident when proteins are associated with genes that have duplicated, or organisms that have diverged from one another relatively recently, and leads to high values of the BLOSUM (or PAM) score. But in some cases, a relevant biological correlation may be obscured by phenomena that reduce the score, making it difficult to capture. Those that limit the efficiency of the scoring method in finding concealed or weakly correlated sequences are well documented in the literature, the most relevant being:

(1) *Gaps*: insertions or deletions (of one or more residue) in one or both the aligned sequence cause loss of synchronization, significantly decreasing the score;

(2) *Bad* $\theta$: using a BLOSUM-$\theta$ matrix tailored for a particular evolutionary distance on sequences with a different evolutionary distance leads to a misleading score [7, 12, 13];

(3) *divergence in background distribution*: standard substitution matrices, such as BLOSUM-$\theta$, are truly appropriate only for comparison of proteins with standard background frequency distributions of amino acids [11].

We have set out to inspect, in more depth and by use of mathematical tools, what the BLOSUM score really measures from a biological point of view; the aim was to split the score into components, the BLO*Specrum*, that provide insight on the above described phenomena and other biological information regarding the compared sequences, once the alignment has been made using the classical methods (BLAST, FASTA, etc.). We do not propose an alternative alignment algorithm or a method for increasing the performance of the available ones; nor do we suggest new methods for inserting gaps so as to maximize the score (see, e.g., [14, 15]). Ours is simply a diagnostic tool to reveal the following:

(1) if, for an available algorithm, the chosen scoring matrix is correct;

(2) whether the aligned sequences are typical protein sequences or not;

(3) whether the alignment itself is typical with respect to BLOCKS database; and

(4) the possible presence of a weak or concealed correlation also for alignments resulting in a relatively low BLOSUM score, that might otherwise be neglected.

The method is associated with the use of a BLOSUM matrix that has been developed within the context of local (ungapped) alignment statistics [7, 8, 11]. To allow a critical evaluation of our method, we furnish an online software package that provides values for each component of the BLO*Specrum* for two aligned sequences (http://bioinf.dimi.uniud.it/software/software/blosumapplet). Providing a rationale about the biological significance of an obtained score sharpens the comparison of weakly related sequences, and can reveal that comparable scores actually conceal completely different biological relationships. Furthermore, our decomposition helps in selecting the matrix that is correctly tailored for the actual evolutionary divergence associated to the two sequences one is going to compare, or in deciding if a compositionally adjusted matrix might not perform better.

Although we have used the BLOSUM scoring method for our analyses, since it is the most widely used by web tools measuring protein similarities, our decomposition is applicable, in principle, to any scoring matrix in the form of (3),

and confirms that the usefulness of this type of matrix has a solid mathematical justification.

## 2. METHODS

### 2.1. Mathematical analysis of the BLOSUM score

The BLOSUM score (3) can be analyzed from a mathematical perspective using well-known tools developed by Shannon in his seminal paper that laid the foundation for *Information Theory* [16, 17]. The first of these is the *Mutual Information* $I(X, Y)$ (or *relative entropy*) between two random variables $X$ and $Y$,

$$I(X,Y) = \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(i)p(j)}, \tag{4}$$

where $p(i, j)$, $p(i)$, $p(j)$ are, respectively, the joint probability distribution and the marginals associated to the random variables $X$ and $Y$. We can adapt (4) to the comparison of two sequences if we interpret $p(i, j)$ as the relative frequency of finding amino acids $i$ and $j$ paired in the $X$ and $Y$ sequences, and $p(i)$ ($p(j)$) of finding amino acid $i$ ($j$) in sequence $X$ ($Y$). Following this approach, in a biological setting, mutual information ($MI$) becomes a measure of the *stochastic correlation* between two sequences. It can be shown (see the appendix) that $I(X, Y) \leq \log 20 \approx 4.3219$. The second tool is the *informational divergence* $D(P//Q)$ between two probability distributions $P = \{p_1, p_2, \ldots, p_K\}$ and $Q = \{q_1, q_2, \ldots, q_K\}$ [18], where

$$D(P//Q) = \sum_{i=1}^{K} p(i) \log \frac{p(i)}{q(i)}. \tag{5}$$

The informational divergence ($ID$) can be interpreted as a measure of the nonsymmetrical "distance" between two probability distributions. A more detailed mathematical treatment of the properties associated with $MI$ and $ID$ is provided in the appendix. Here, we simply indicate that $ID$ and $MI$ are nonnegative quantities, and that they are tied by the formula

$$I(X,Y) = \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(i)p(j)} = D(P_{XY}//P_X P_Y) \geq 0, \tag{6}$$

so that $MI$ is really a special kind of $ID$, that measures the "distance" between the joint probability distributions $P_{XY}$ and the product $P_X P_Y$ of the two marginals $P_X$ and $P_Y$.

Given two amino acid sequences, $X$ and $Y$, the corresponding BLOSUM (unscaled) normalized score $S_N(X, Y)$, measured in *bits*, is computed as

$$S_N(X,Y) = \frac{1}{n} \sum_{h=1}^{n} s(x_h, y_h) = \sum_{i,j} f(i,j) \log \frac{p(i,j)}{p(i)p(j)}, \tag{7}$$

where $f(i, j) = n(i, j)/n$ is the relative frequency of the pair $i$, $j$ observed on the aligned sequences $X$ and $Y$. Because one usually deals with sequences that could have remarkably different lengths, we report the normalized perresidue score to permit a coherent comparison. It is important to stress the fact that while $f(i, j)$ is the observed frequency pertaining to the sequences under inspection, the target frequencies $p(i, j)$, together with the background marginals $p(i)$ and $p(j)$, pertain to the database BLOCKS. In a sense, they constitute "the model" of the typical behaviour of a protein, since $p(i)$ or $p(j)$ is in fact the "typical" probability distribution of amino acids as observed in most proteins, while $p(i, j)$ is the "typical" probability of finding the amino acids $i$ and $j$ positionally paired in two protein sequences with a percent identity depending from $\theta$. From an evolutionary point of view, we can say that if $p(i, j)$ is greater than in the case of independence, then it is very likely that $i$ and $j$ are biologically correlated.

Equation (7) is in fact quite similar to (4), which specifies mutual information, the only difference being the use of $f(i, j)$ instead of $p(i, j)$ as the multiplying factor for the logarithmic term, so that the normalized score is a kind of "mixed" mutual information. As a matter of fact, we can define

$$I(A,B) = \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \tag{8}$$

as the mutual information, or relative entropy, of the target and background frequencies associated to the database BLOCKS, or to any other protein model used to find the target frequencies. Here $A$, and $B$ are dummy random variables taken to have generated the data of the database. The quantity $I(A, B)$ was in effect used by Altschul in the case of PAM matrices [7], and by S. Henikoff and J. G. Henikoff [9] for the BLOSUM matrices, and in both cases it can be interpreted as the average exchange of information associated with a pair of aligned amino acids of the data bank, or as the expected average score associated to pairs of amino acids, when they are put into correspondence in alignments that adhere to the protein model over which the matrices are computed. From the perspective of an aligning method, we can state that $I(A, B)$ measures the average information available for each position in order to distinguish the alignment from chance, so that the higher its value, the shorter the fragments whose alignment can be distinguished from chance [7]. Equation (6) (or (A.4) in the appendix) ensures also that this average score is always greater than or equal to zero.

On the other hand, if we compute the expected score when two amino acids $i$ and $j$ are picked at random in an independence setting model, given as

$$E(A,B) = \sum_{i,j} p(i)p(j) \log \frac{p(i,j)}{p(i)p(j)} = -D(P_X P_Y//P_{XY}) \leq 0, \tag{9}$$

the classical assumptions made in constructing a scoring matrix [7] require that this expected score is lower than or equal to zero. Note that all these quantities pertain to the database BLOCKS (in the case of BLOSUM), that is to the particular "protein model" used.

To solely evaluate the stochastic similarity between two sequences $X$ and $Y$, the identity

$$I(X,Y) = \sum_{i,j} f(i,j) \log \frac{f(i,j)}{f_X(i) f_Y(j)}, \tag{10}$$

which measures the degree of stochastic dependence between the protein sequences, would suffice (here $f_X(i) = n(i)/n$ and $f_Y(j) = n(j)/n$ are the relative frequencies of amino acid $i$ observed in sequence $X$ and amino acid $j$ observed in sequence $Y$). But this is not so interesting from the biological point of view, as one has to take into account the possibility that, even if similar from the stochastic point of view, two sequences are far from being an example of a typical protein-to-protein matching (or evolutionary transition). In other words, we need to inspect this stochastic similarity under the "lens" of the protein model used in the BLOCKS database (or by the PAM model, for the matter).

Subjecting the (unscaled) normalized score $S_N(X,Y)$ of (7) to simple mathematical manipulations (see the appendix for details), we can split $S_N(X,Y)$ into the following terms:

$$\begin{aligned} S_N(X,Y) = {} & I(X,Y) - D(F_{XY}//P_{AB}) \\ & + D(F_X//P_A) + D(F_Y//P_B). \end{aligned} \tag{11}$$

Here, $F_{XY}$ is the joint frequency distribution of the amino acids pairs in the sequences, (observed target frequencies), while $F_X$ and $F_Y$ are, respectively, the distribution of the amino acids inside $X$ and $Y$ (observed background frequencies). $P_{AB}$ instead is the joint probability distribution associated to the BLOCKS database, and is the vector of target frequencies. Note also that $P_A = P_B = P$ are the probability distributions of the amino acids inside the same database BLOCKS, that is the database background frequencies; they are equal as a consequence of the symmetry of the BLOSUM matrix entries, since $p(i,j) = p(j,i)$. We define the set $\{I(X,Y), D(F_{XY}//P_{AB}), D(F_X//P), D(F_Y//P)\}$ to be the *BLOSUM spectrum* of the aligned sequences (or BLO*Spectrum*). Notice that (11) holds also when the BLOSUM matrix is decompositionally adjusted following the approach described in Yu et al. [11], that is when the background frequencies are different ($P_A \neq P_B$).

The terms constituting the BLO*Spectrum* have a different order of magnitude, as $D(F_X//P)$ and $D(F_Y//P)$ act with a cardinality of 20, when compared to the joint divergences $I(X,Y)$ and $D(F_{XY}//P_{AB})$, that act on probability distributions whose cardinality is $20 * 20 = 400$. From a practical point of view, this means that the contribution of $I(X,Y)$ and $D(F_{XY}//P_{AB})$ to the score is expected to be roughly double than that of $D(F_X//P)$ and $D(F_Y//P)$. Actually, under the hypothesis of a Bernoullian process (i.e., stationary and memoryless), we have $D(P^2//Q^2) = 2D(P//Q)$ [18] (as in our case $20^2 = 400$), and the sum of the two terms $D(F_X//P) + D(F_Y//P)$ compensates the order of magnitude of the joint divergences.

Finally, it should be recalled that the score actually obtained by using the BLOSUM matrices, whose entries are multiplied by the constant $c$ and rounded to the nearest integer, is an approximation of the exact score $S_N(X,Y)$ of (11), once it has been scaled. The difference is usually quite small (about 2-3% if the score is high), but it becomes more and more significant as the score approaches zero.

### 2.2. Taking gaps into account

An important consideration regarding our mathematical analysis is that it does not formally take gaps into account. From a mathematical perspective, the only way to account correctly for gaps would be to use a $21 * 21$ scoring matrix, in which the gap is treated as equivalent to a 21st amino acid, so that pairs of the form $(i,-)$ or $(-,j)$, where the symbol "$-$" represents the gap, are also contemplated; but from a biological perspective this might not be acceptable, since a gap is not a real component of a sequence. We can nevertheless extend our analysis to a gapped score if we admit the independence between each gap and any residue paired with it. Biologically, independence may be questionable, and would need to be determined case by case, as each gap is due to a chance deletion or insertion event subsequently acted on by natural selection (which may be neutral or positive). Moreover, there is no certainty as to the correct positioning of a gap in any given alignment, as it is introduced a posteriori as the product of an alignment algorithm that takes the two sequences $X$ and $Y$, and tries to minimize (by an exact procedure, or by a heuristic approach) the number of changes, insertions or deletions that allow to transform $X$ into $Y$ (or vice versa). In practice, we consider quite reasonable the idea that gaps in a given position should imply a degree of independence as to which amino acids might occur there in related proteins; this is accepted also in PSI-BLAST [19]. The consequence of assuming independence is that $p(-,j) = p(-)p(j)$ leads to a null contribution of the corresponding score, since $s(-,j) = \log[p(-,j)/p(-)p(j)] = 0$ (see (3)), so that for gapped sequences, we simply assign a score equal to zero whenever an amino acid is paired with a gap. Note that this does not mean that we reduce a gapped alignment to an ungapped one, but that we simply ignore the gap and the corresponding residue, since the pair is not affecting the BLO*Spectrum*, due to its zero contribution to the score. Moreover, it is conceivable that for distant sequence correlations, the use of different algorithms, or of different gap penalties schemes for any given algorithm, could result in a different pattern of gaps and consequently in different sequence alignments, each with a corresponding BLO*Spectrum*. In this case, the likelihood of each alignment might be tested by exploiting the BLO*Spectrum*, that might be quite different even if the numerical scores have approximately the same value; this can help identify the most appropriate one.

## 3. RESULTS AND DISCUSSION

### 3.1. Meaning and biological implications of the BLOSpectrum terms

Let us now analyze the meaning of the terms in (11).

(i) The *mutual information* $I(X,Y)$ is the *sequence convergence*, which measures the degree of *stochastic dependence* (or stochastic correlation) between aligned

sequences $X$ and $Y$; the greater its value, the more statistically correlated are the two. It is highly correlated with, but not identical to, the percent identity of the alignment, as it also includes 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.

(ii) The *target frequency divergence* $D(F_{XY}//P_{AB})$ 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 the $X$ and $Y$ sequences and inside the "protein model" implicit in the BLOCKS database. 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, so that starting from $X$ we obtain an $Y$ (or vice versa) 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 (11), since it is taken with the minus sign.

(iii) The *background frequency divergence* $D(F_X//P_A)$ (or $D(F_Y//P_B)$) of the sequence $X$ (or $Y$) 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 $F_X$ (or $F_Y$) and the vector $P = P_A = P_B$ of background frequencies of the amino acids inside the database BLOCKS. The greater is 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.

Note that the quantities that constitute the decomposition of the BLOSUM score are not independent of one another. For example, $D(F_{XY}//P_{AB}) \approx 0$ implies low values for $D(F//P)$ also. This is because when $F_{XY} \to P_{AB}$ (or $D(F_{XY}//P_{AB}) \to 0$; see the appendix), then also the observed marginals $F_X$ and $F_Y$ are forced to approach the background marginal, that is $F_X \to P$ and $F_Y \to P$, which implies $D(F//P) \to 0$. This is a consequence of the tie between a joint probability distribution and its marginals [10]. For the same reason, if $D(F//P) \gg 0$, then $D(F_{XY}//P_{AB})$ will also be large, although the opposite is not necessarily the case. This leads to (at least partially) a compensation of the effects, due to the minus sign of the target frequency divergence, so that $-D(F_{XY}//P_{AB}) + D(F_X//P_A) + D(F_Y//P_B)$ has a small value. This implies that a significant BLOSUM score can be obtained only when the aligned sequences are statistically correlated, that is, when $I(X, Y)$ has a high value. Since when performing an alignment we are mainly interested in positive or almost positive global scores, it is a straightforward consequence that only alignment characterized by remarkable values of $I(X, Y)$ will emerge.

There are therefore essentially three cases of biological interest, which we can now analyze in terms of the correspondence between mathematical and biological meaning of the terms.

*Case 1.* The joint observed frequencies $F_{XY}$ are *typical*,[1] that is, they are very close to the target frequencies, $F_{XY} \approx P_{AB}$.
In this case, $D(F_{XY}//P_{AB}) \approx 0$ and also $D(F//P) \approx 0$.

*Case 2.* The joint observed frequencies $F_{XY}$ are not typical ($F_{XY} \neq P_{AB}$), but the marginals are typical ($F_X \approx P, F_Y \approx P$).
In this case, $D(F_{XY}//P_{AB}) \gg 0$, but $D(F//P) \approx 0$.

*Case 3.* Both the joint observed $F_{XY}$ and the marginals $F_X$, $F_Y$ are not typical, that is $F_{XY} \neq P_{AB}$, $F_X \neq P$, $F_Y \neq P$.
In this case, $D(F_{XY}//P_{AB}) \gg 0$, but also $D(F//P) \gg 0$.

Case 1 is straightforward; two similar protein sequences with a typical background amino acid distribution; and amino acids paired in a way that complies with the protein model implicit in BLOCKS result in a high score. This is frequently the case for two firmly correlated sequences, belonging to the same family of proteins with standard amino acid content, associated with organisms that diverged only recently.

Case 2 is rather more interesting; the amino acid distribution is close to the background distribution (these are "typical" protein sequences) but the score is highly penalized as the observed joint frequencies are different from the target frequencies implicit in the BLOCKS database. This can have different causes. For example, the chosen BLOSUM matrix may be incorrectly matched to the evolutionary distance of the sequences, or the sequences may have diverged under a nonstandard evolutionary process. For high-scoring alignments involving unrelated sequences, the target frequency divergence $D(F_{XY}//P_{AB})$ will tend to be low, due to the second theorem of Karlin and Altschul [8], when the target frequencies associated to the scoring matrix in use are the correct ones for the aligned sequences being analyzed.[2] This is because any set of target frequencies in any particular amino acid substitution matrix, such as BLOSUM-$\theta$, is tailored to a particular degree of evolutionary divergence between the sequences, generally measured by relative entropy (8) [7], and related with the controlled maximum rate $\theta$ of percent identity. So a low $D(F_{XY}//P_{AB}) \approx 0$ is evidence that the BLOSUM-$\theta$ matrix we are using is the correct one, as a precise consequence of a mathematical theorem, while conversely for positive (or almost positive) scoring alignments with large target frequency divergence, the sequences may be

[1] Recall that the concept of "typicality" always refers to the adherence of the various probability distributions to that of the protein model associated to the database BLOCKS.

[2] Note that in general, choosing the ($\theta$ parameter associated with the) smallest $D(F_{XY}//P_{AB})$ is different from choosing the minimum $E$-value associated with different $\theta$ parameters. Recall that $E = m * n2^{-S}$, where $S$ is the score and $m$ and $n$ are the sequences lengths.

related at a different evolutionary distance than that of the substitution matrix in use. Trying several scoring matrices until "something interesting" is found is a common practice in protein sequence alignment [20]. In our case, scanning the $\theta$ range could thus lead to a significant decrease in $D(F_{XY}//P_{AB})$, as detected in the BLO*Spectrum*, and improve the score [7, 12, 13], taking it back to Case 1. This could in turn result in a better capacity to discriminate weakly correlated sequences from those correlated by chance. If, on the other hand, tuning $\theta$ does not greatly affect $D(F_{XY}//P_{AB})$, and we are comparing typical sequences (low background frequency divergence) with an appropriate $\theta$ parameter, the large target frequency divergence indicates that some nonstandard evolutionary process (regarding the substitution of amino acids) is at work. This cannot adequately be captured by the standard BLOCKS database and BLOSUM substitution matrices. Under these circumstances, Case 2 can never lead to high scores, due to the penalization of the target frequency divergence. We are here likely in the grey area of weakly correlated sequences with a very old common ancestor, or of portions of proteins with strong structural properties that do not require the conservation of the entire sequence. Note that unfortunately we are not able to assess the statistical significance when our method finds a suspected concealed correlation; however, the method still gives us useful information that helps guide our judgment on the possible existence of such correlation, that needs to be further investigated in depth, exploiting other biological information such as 3D structure and biological function.

Case 3 accounts for the situation in which we have two nontypical sequences, with high values of both target and background frequency divergence. This applies, for example, to some families of antimicrobial peptides, that are unusually rich in certain amino acids (such as *Pro* and *Arg*, *Gly*, or *Trp* residues). This means that the high penalty arising from the subtracted $D(F_{XY}//P_{AB})$ is (at least partially) compensated by the positive $D(F_X//P_A)$ and $D(F_Y//P_B)$, and the global score does not collapse to negative values, even if it is usually low. In effect, the background frequency divergence acts as a compensation factor that prevents excessive penalties for those sequences which, even though related by nonstandard amino acid substitutions, also have a nontypical background distribution of the amino acids inside the sequences themselves. In other words, the nontypicality of $F_{XY}$ is (at least in part) forced of by the anomalous background frequencies of the amino acids. This compensation is welcome, since it avoids missing biologically related sequences pertaining to nontypical protein families, and mathematically corroborates the robustness of the BLOSUM scoring method.

The problem of evaluating the best method for scoring nonstandard sequences has been recently tackled by Yu et al. [11, 21], who showed that standard substitution matrices are not truly appropriate in this case, and developed a method for obtaining compositionally adjusted matrices. In general, when background frequencies differ markedly from those implicit in the substitution matrix (i.e., the background frequency divergence is high) is one case when using a standard matrix is nonoptimal. Another is when the background frequencies vary, and the scale factor $\lambda = (\log(p(i,j)/p(i)p(j)))/s(i,j)$ appropriate for normalizing nominal scores varies as well [8]. If the real $\lambda$ is lower than the "standard" one, then the uncorrected nominal score can appear much too high [19, 22]. Our approach offers a different perspective to the problem, that is, the possibility of gaining insight about biological sequence correlation directly from the BLOSUM score. Moreover, the background frequency divergence components of BLO*Spectrum* indicate whether compositionally adjusted matrices could be useful in the case under inspection. Since [21] illustrates three "criteria for invoking compositional adjustment" (length ratio, compositional distance, and compositional angle), we suggest that the occurrence of "Case 3" in the BLOSUM spectrum could be thought of as an additional fourth criterion.

The background divergence of the BLO*Spectrum* decomposition offers a further rationale to confirm the effectiveness of the procedure proposed by Yu et al., since a large background divergence $D(F//P)$ forces the target frequency divergence $D(F_{XY}//P_{AB})$ to be unnaturally large; compositionally adjusted matrices, that minimizes background frequency divergence, tend to remove this effect, leaving it free to assume the value associated to the (correct degree of evolutionary) divergence between the sequences under inspection.

As a consequence of the three cases discussed above, we can suggest the following procedure for analyzing the score obtained from an alignment between two given sequences of the same length, or resulting from a BLAST or FASTA (gapped or ungapped) database search.

*Scoring analysis procedure*

(1) Given the two sequences, evaluate the components of (11) by inserting the sequences in the available software to obtain the BLO*Spectrum* (http://bioinf.dimi.uniud.it/software/software/blosumapplet).
(2) Evaluate the target frequency divergence $D(F_{XY}//P_{AB})$ for each $\theta$.
(3) Choose the $\theta$ value that minimizes $D(F_{XY}//P_{AB})$.
(4) Determine if the alignment falls in Cases 1, 2, or 3 as described.
(5) If the alignment falls in Case 1, we have two strictly correlated proteins.
(6) If, even after tuning $\theta$, the alignment falls in Case 2 ($D(F_{XY}//P_{AB})$ is high, but $D(F//P)$ is low), then we may have a concealed or weak correlation between the sequences.
(7) If the alignment falls in Case 3 (both $D(F_{XY}//P_{AB})$ and $D(F//P)$ are high), we may have correlated sequences belonging to a nontypical family. In this case, the use of compositionally adjusted matrices may provide a sharper score [11, 21].

In analyzing the parameters that compose the BLO*Spectrum*, so as to decide among Cases 1, 2, and 3, we find it useful to use an indicative, if somewhat arbitrary set of guidelines, as summarized in Table 1.

We assign a range of values for each parameter (tag L = Low, tag M = Medium, tag H = High). These values have been

Table 1: Rule of thumb guidelines to decide among low (L), medium (M), and high (H) values of the parameters.

| | L | M | H |
|---|---|---|---|
| $I(X,Y)$ | <0.9 | 0.9–1.1 | >1.1 |
| $D(F_{XY}//P_{AB})$ | <1.1 | 1.1–1.5 | >1.5 |
| $D(F//P)$ | <0.3 | 0.3–0.7 | >0.7 |

derived from a "rule of thumb" approach when analyzing the results of the experiments described in the following sections; but obviously they need to be tuned as soon as new experimental evidence will be available.

The final consideration is that, when comparing biologically related sequences, one has to choose the correct scoring matrix if necessary by means of a compositional adjustment. If, as a result, background and target frequency divergences have low values, the mutual information or sequence convergence $I(X,Y)$ remains as the effective parameter that measures protein similarity. If, after considering the above possibilities, one still observes a residual persistence of the target frequency divergence, then two weakly correlated sequences are presumably identified, that derived from a common remote ancestor after several events of substitution.

### 3.2. Practical implementation of the method

As stated in the Introduction, we recall that the analysis based on the BLO*Spectrum* evaluation is not aimed at increasing the performance of available alignment algorithms, nor at suggesting new methods for inserting gaps so as to maximize the score. The BLO*Spectrum* only gives added information of biological and operative interest, but only once two sequences have already been aligned using current algorithms, such as BLAST, BLAST2, FASTA, or others. The ultimate biological goal of the method is that of revealing the possible presence of a weak or concealed correlation for alignments resulting in a relatively low BLOSUM score, that might otherwise be neglected. Another operative merit is that the knowledge of the target frequency divergence helps identify the best scoring matrix, that is the one tailored for the correct evolutionary distance.

In order to perform automatic computation of the four terms of (11), we have developed the software BLO*Spectrum*, freely available at http://bioinf.dimi.uniud.it/software/software/blosumapplet. Given two sequences with the same length, with or without gaps, the software derives the vectors $F_X$, $F_Y$, and $F_{XY}$ by computing the relative frequencies $f(i) = n(i)/n$, $f(j) = n(j)/n$, and $f(i,j) = n(i,j)/n$, that is the relative frequency of amino acid $i$ observed in sequence $X$, of amino acid $j$ observed in sequence $Y$, and the relative frequency of the pair $i$, $j$. The vectors $P_{AB} = \{p(i,j)\}_{i,j}$ and $P = \{p(i)\}_i$, needed to decompose the score, are those derived from BLOCKS database and used by S. Henikoff and J. G. Henikoff [9] to extract the score entries of the $20 * 20$ BLOSUM matrices (35, 40, 50, 62, 80, 100); they have been kindly provided by these authors on request. The software computes also the exact BLOSUM normalized score, that is the algebraic sum of the four terms, together with the rough BLOSUM score, directly obtained by summing up the integer values of the BLOSUM-$\theta$ matrix. As already observed in Section 2.2 the pairs containing a gap, such as $(-, j)$ or $(i, -)$, are not considered in the computation, since their contribution to the score is zero when one assumes the independence between a gap and the paired amino acid.

There are essentially two ways for employing the BLO*Spectrum*. The first one is that of performing a BLAST or FASTA search inside a database, given a query sequence. The result is a set of $h$ possible matches, ordered by score, in which the query sequence and the corresponding match are paired for a length that is respectively $n_1, n_2, \ldots, n_h$. The user can extract all matches of interest within the output set and compares them with the query sequence by using BLO*Spectrum* software. The second one is that of comparing two assigned sequences with a program such as BLAST2, so as to find the best gapped alignment. Also in this case we can use BLO*Spectrum* on the two portions of the query sequences that are paired by BLAST2 and that have the same length $n$. It is obvious that the next step would be that of integrating the BLO*Spectrum* tool inside a widely used database search engine.

Even if the correct way for using the BLO*Spectrum* software is that of supplying it with two sequences of the same length, derived from preceding queries of BLAST, BLAST2, FASTA or others, the BLO*Spectrum* applet accepts also two sequences of different length $n$ and $m > n$; in this case the program merely computes the scores associated to all possible alignments of $n$ over $m$, showing the highest one, but it does not insert gaps.

### 3.3. Biological examples

To illustrate the behavior of the BLO*Spectrum* under the perspective of the above three cases, we have chosen groups of proteins from several established protein families present in the SWISSPROT data bank http://www.expasy.uniprot.org (see Table 2), together with some specific examples of sequences, taken from the literature, that are known to be biologically related, even if aligning with rather modest scores.

The first set contains sequences from the related *Hepatocyte nuclear factor* 4$\alpha$ (HNF4-$\alpha$), *Hepatocyte nuclear factor* 6 (HNF6), and GAT binding protein 1 (globin transcription factor 1 families). These represent typical protein families coupled by standard target frequencies. Furthermore, sequences within each family are quite similar to one another, with a percent identity greater than 85%. All these proteins are expected to fall in Case 1.

The second set of sequences is expected to fall in Case 2. A first example is taken from the *serine protease family*, containing paralogous proteins such as trypsin, elastase, and chymotrypsin, whose phylogenetic tree constructed according to the multiple alignment for all members of this family [23] is consistent with a continuous evolutionary divergence from a common ancestor of both prokaryotes and eukaryotes. Another example pertaining to weakly correlated sequences that show distant relationships is the one originally used by

Table 2: The three sets of protein families used in testing the BLO*Spectrum*. The UniProt ID is furnished (with the sequence length). For the defensins and Pro-rich peptides, only the mature peptide sequences were used in alignments. In the following tables, sequences are indicated by the corresponding numbers 1–4.

| | Sequence | | | |
|---|---|---|---|---|
| Family | 1 | 2 | 3 | 4 |
| | First set | | | |
| HNF4-$\alpha$ | P41235 (465) *H. sapiens* | P49698 (465) *Mus musculus* | P22449 (465) *Rattus norv.* | |
| HNF6 | Q9UBC0 (465) *H. sapiens* | O08755 (465) *Mus musculus* | P70512 (465) *Rattus norv.* | |
| GAT1 | P15976 (413) *H. sapiens* | P17679 (413) *Mus musculus* | P43429 (413) *Rattus norv.* | |
| | Second set | | | |
| Serine proteases | P07477 (247) *H. sapiens* trypsin | P17538 (263) *H. sapiens* chymotrypsin | Q9UNI1 (258) *H. sapiens* elastase1 | |
| | P00775 (259) *Streptomyces griseus* trypsin | P35049 (248) *Fusarium oxysporum* trypsin | | |
| Hemoglobins | P02232 (92) *Vicia faba* leghemoglobin I | S06134 (92) *P. chilensis* hemoglobin I | | |
| Transposons | A26491 (41) *D. mauritiana* mariner transposon | NP493808 (41) *C. elegans* transposon TC1 | | |
| Beta defensins | BD01 (36) *H. sapiens* | BD02 (41) *H. sapiens* | BD03 (39) *H. sapiens* | BD04 (50) *H. sapiens* |
| | Third set | | | |
| Pro/Arg-rich peptides | BCT5 (43) bovin | BCT7 (59) bovin | PR39PRC (42) pig | PF (82) pig |

Altschul [7] to compare PAM-250 with PAM-120 matrices, that is, the 92 length residue *Vicia faba* leghemoglobin I and *Paracaudina chilensis* hemoglobin I, characterized by a very poor percent identity (about 15%), with pairs of identical amino acids residues that are spread fairly evenly along the alignment. A further example considers the sequences associated to *Drosophila mauritiana* mariner transposon and *Caenorhabditis elegans* transposon TC1, with a length of 41 residues, used by S. Henikoff and J. G. Henikoff [9] to test the performance of their BLOSUM scoring matrices. The last example derives from *human beta defensins*. This family of host defense peptides have arisen by gene duplication followed by rapid divergence driven by positive selection, a common occurrence in proteins involved in immunity [24]. They are characterized by the presence of six highly conserved cysteine residues, which determines folding to a conserved tertiary structure, while the rest of the sequence seems to have been relatively free of structural constraints during evolution [25, 26]. Even if clearly related, these peptides have a percentage sequence identity less than 40%.

All these families represent the case of nonstandard target frequencies, while the amino acid frequency distribution does not appear, at first sight, to be too abnormal. The sequence comparisons score are modest at best, even though members are known to be biologically correlated.

The third set contains sequences that are expected to fall in Case 3. These are members of the *Bactenecins* family of linear antimicrobial peptides, with an unusually high content of *Pro* and *Arg* residues, and an identity of about 35% [27], representing sequences with a highly atypical amino acid frequency distribution.

If we analyze the alignments inside all these sets of protein families, we effectively find examples for each of the three cases illustrated in the preceding section. The alignments of human and mouse HNF4-$\alpha$ sequences (as illustrated in Table 3), and the BLO*Spectrum* of HNF4-$\alpha$, HNF6, and GAT1 sequence comparisons (see Figure 1), are clear examples of Case 1, with high correlation between all respective couples of sequences and a target frequency divergence that is strongly sensitive to the BLOSUM-$\theta$ parameter, so we stop the *scoring procedure* at step 5.

For example, the HNF4-$\alpha$ alignment has a target frequency divergence that varies from 2.41 to 0.93 when passing from BLOSUM-35 (a matrix tailored for a wrong

Table 3: BLOSUM decomposition for intrafamily alignments for proteins of the first set.

| HNF4-$\alpha$ human versus HNF4-$\alpha$ mouse | | | | | | | |
|---|---|---|---|---|---|---|---|
| BLOSUM | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| *100* | *3.939* | *0.929* | *0.050* | *0.057* | *3.118* | *2833* | *95.9* |
| 80 | 3.939 | 1.297 | 0.046 | 0.053 | 2.741 | 2537 | 95.9 |
| 62 | 3.939 | 1.582 | 0.046 | 0.052 | 2.456 | 2330 | 95.9 |
| 50 | 3.939 | 1.861 | 0.043 | 0.050 | 2.171 | 3003 | 95.9 |
| 40 | 3.939 | 2.226 | 0.039 | 0.047 | 1.800 | 3381 | 95.9 |
| 35 | 3.939 | 2.414 | 0.036 | 0.044 | 1.605 | 2982 | 95.9 |
| HNF4-$\alpha$ (BLOSUM-100) | | | | | | | |
| Sequences | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 1–3 | 3.955 | 0.930 | 0.050 | 0.056 | 3.132 | 2846 | 96.3 |
| 2-3 | 4.141 | 1.008 | 0.057 | 0.056 | 3.246 | 2952 | 99.5 |

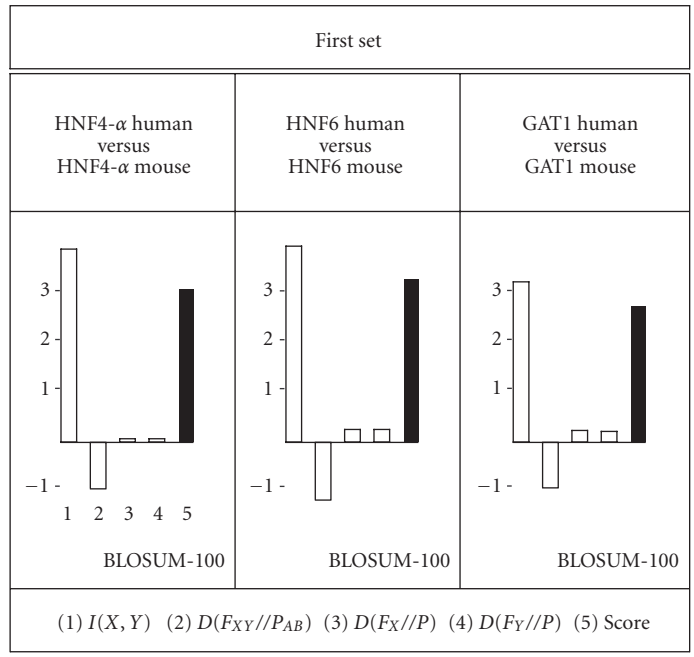


Figure 1: BLO*Spectrum* for sequences of the first set.

evolutionary distance), to BLOSUM-100 (the matrix tailored for a correct evolutionary distance) so that minimizing the frequency divergence (rows in italic) helps identify the best $\theta$ parameter for comparing the analyzed sequences; it corresponds to $\theta = 100$, coherent with the high percent identity (86–96%). In this case, the compensation factor $D(F_X//P) + D(F_Y//P)$ corresponding to background frequency divergence is almost zero, since observed background and target frequencies are very near to those implicit in the BLOCKS database, leading to the conclusion that these are typical sequences that correspond closely to the protein model associated with BLOCKS. The global (normalized) score is high (3.12 in the HNF4-$\alpha$ example), due to a high degree of stochastic similarity ($I(X,Y) \approx 3.94$), which is not greatly penalized. Other members of the HNF4-$\alpha$, HNF6, or GAT1 families behave similarly (see Figure 1).

The situation changes considerably when we compute the BLOSUM decomposition for the different examples listed for the second set, for example, comparing human trypsin, elastase and chymotrypsin to one another, or comparing these enzymes in distantly related species, such as *human*, *streptomyces griseus* (a bacterium), and *Fusarium oxysporum* (a fungus). Following the *Scoring Procedure*, and starting with ungapped alignments, we have a case of high target frequency divergence, with a low level of background frequency divergence, corresponding to the situation outlined in step 6. However, as soon as we use gapped alignments, we observe a remarkable increment in the score, due to a reduced

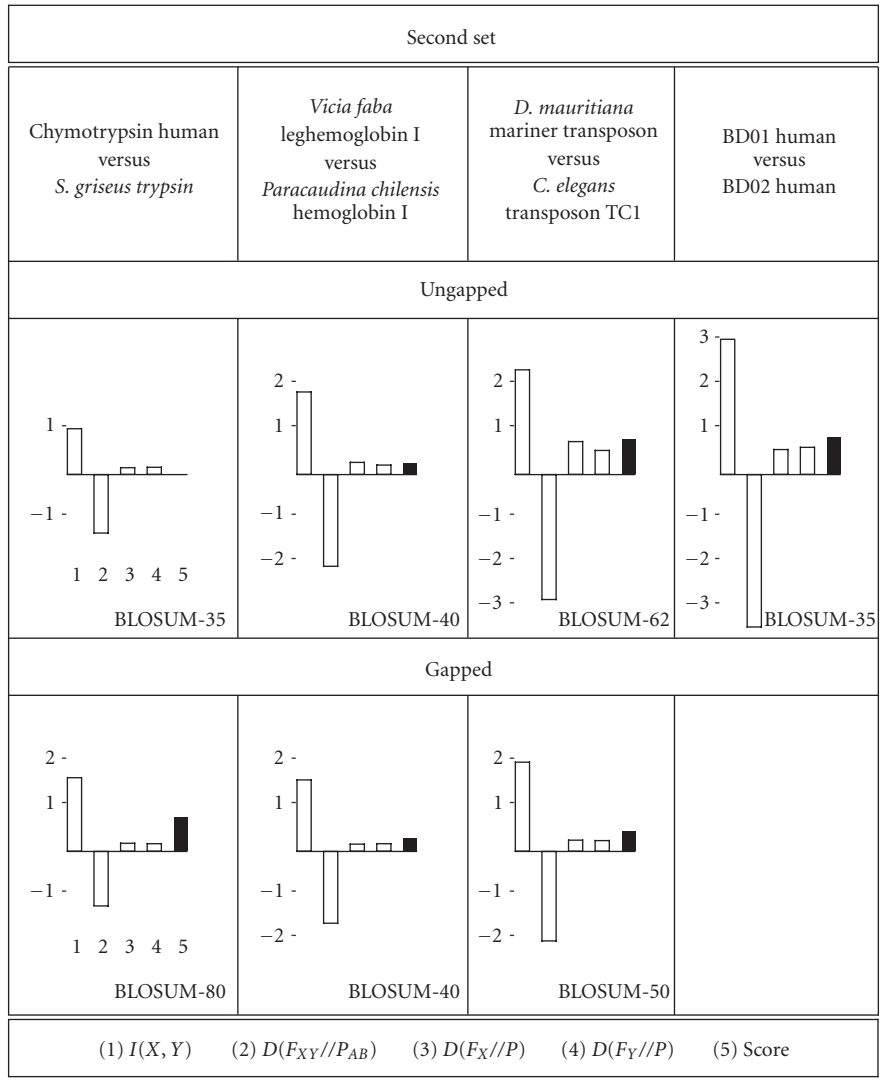


Figure 2: BLO*Spectrum* for (ungapped and gapped) sequences of the second set.

penalization factor associated to target frequency divergence (see Figure 2, first column, and Table 4). This is the obvious case when the bad matching is a consequence of deletions and/or insertions that occurred during evolution, which is resolved once gaps are introduced, so that the sequence comparison falls into Case 1

A different situation occurs aligning *Vicia faba* leghemoglobin I and *Paracaudina chilensis* hemoglobin I. $D(F_{XY}//P_{AB})$ minimization (step 3) leads to a narrower spread of values (2.48–2.07) when passing from BLOSUM-100 to BLOSUM-35, with minimum (2.05) at $\theta = 40$, which is consequently the best parameter to compare the sequences. The global score (0.24) is rather low, despite these sequences being clearly evolutionarily related. In fact, the BLO*Spectrum* shows that the stochastic correlation $I(X,Y)$ is quite high (1.84), but is killed by the heavy penalty derived from the negative contribution of $D(F_{XY}//P_{AB})$, while the compensation factors due to background frequency divergence are less significant (0.25 and 0.19, resp.), as the sequences are typical proteins under the BLOCKS model. Furthermore, extending the size of the alignment or including gaps does not significantly alter the spectrum (see Table 5 and Figure 2, second column), so we leave the *Scoring Procedure* at step 6; we simply have weakly related sequences.

The *Drosophila mauritiana* and *Caenorhabditis elegans* transposons provide a similar example, with only a weak minimization for $\theta = 62$ ($D(F_{XY}//P_{AB}) = 2.80$). The other BLO*Spectrum* components are respectively $I(X,Y) = 2.34$, $D(F_X//P) = 0.53$, and $D(F_Y//P) = 0.72$. The sequences thus have a high stochastic correlation, but the target frequencies are rather atypical, so that the divergence entirely kills the contribution derived from mutual information, and if the score is weakly positive (0.79) it is only due to the terms associated to background frequency divergence. In fact, the biological relationship of these atypical sequence fragments is effectively captured only due to the presence of this compensation factor. In this case, a gapped alignment including a wider portion of the sequences, actually reduces the

Table 4: BLOSUM decomposition for ungapped and gapped serine proteases.

| Serine proteases | | | | | | | |
|---|---|---|---|---|---|---|---|
| BLOSUM | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| human chymotrypsin versus *Streptomyces griseus* trypsin (ungapped) | | | | | | | |
| 100 | 1.014 | 2.023 | 0.134 | 0.132 | −0.742 | −398 | 11.5 |
| 80 | 1.014 | 1.739 | 0.141 | 0.137 | −0.446 | −230 | 11.5 |
| 62 | 1.014 | 1.570 | 0.146 | 0.145 | −0.264 | −121 | 11.5 |
| 50 | 1.014 | 1.437 | 0.134 | 0.141 | −0.147 | −120 | 11.5 |
| 40 | 1.014 | 1.321 | 0.132 | 0.138 | −0.035 | −42 | 11.5 |
| *35* | *1.014* | *1.305* | *0.136* | *0.145* | *−0.008* | *−7* | *11.5* |
| human chymotrypsin versus *Streptomyces griseus* trypsin (gapped) | | | | | | | |
| 100 | 1.645 | 1.213 | 0.164 | 0.156 | 0.753 | 326 | 35.9 |
| *80* | *1.645* | *1.138* | *0.170* | *0.164* | *0.842* | *382* | *35.9* |
| 62 | 1.645 | 1.149 | 0.178 | 0.171 | 0.845 | 416 | 35.9 |
| 50 | 1.645 | 1.176 | 0.171 | 0.159 | 0.800 | 557 | 35.9 |
| 40 | 1.645 | 1.270 | 0.170 | 0.158 | 0.703 | 640 | 35.9 |
| 35 | 1.645 | 1.346 | 0.177 | 0.163 | 0.640 | 584 | 35.9 |

background frequency divergences to remarkably lower values (0.237 and 0.226), neutralizing the compensation (see Table 6 and Figure 2, third column).

In both the preceding examples, we are in the situation where the parameter $\theta$ of the substitution matrix is appropriate for the sequence divergence of the sequences in question, the background frequency divergence is small, but the target frequency divergence is still large: this is a signal that we are dealing with weakly related sequences, characterized by several events of substitution that occurred during evolution. It is usually difficult to capture these weakly related sequences using standard scoring matrices, such as BLOSUM or PAM, since the common ancestor could be very old. As a matter of fact, this difficulty was used to respectively test the PAM-250 versus PAM-120 matrices (Altschul [7], hemoglobin) and BLOSUM-62 versus PAM-160 matrices (S. Henikoff and J. G. Henikoff [9], transposons). Here, we cannot remove the cause of mismatching and we leave the *Scoring Procedure* at step 6.

The last example from this group derives from *human beta defensins*, and even if these sequences are known to be evolutionarily related, some couples actually show a negative normalized score (1–4, 2-3, 2–4, see Table 7 and Figure 2, last column), suggesting that they are not. In fact, a normal BLOSUM-62 BLAST search using the human beta defensin 1 sequence, picks up several homologues from other mammalian species, whereas those with the three paralogous human sequences are below the cutoff score. BLO*Spectrum* analysis reveals a high stochastic correlation $I(X,Y)$ (2.00–3.03), neutralized by an even higher-penalty factor due to the target frequency divergence (3.28–3.56), partly compensated by the substantial background frequency divergences (0.54–0.79), and with little effect of the BLOSUM-$\theta$ parameter, or of introducing gaps. These are fairly typical proteins, whose score is heavily penalized by a remarkable target frequency divergence. Only the compensation factor induced by background frequency divergence can, in some cases, sustain the score over positive values, allowing the identification of a biological correlation that would otherwise have been lost.

The third set of sequences are *Pro/Arg* rich antimicrobial peptides of the *Bactenecins* family, with about 35% identity [27, 28]. The obtained scores are clearly positive, despite the poor stochastic correlation (0.40–0.60, see Table 8 and Figure 3).

The penalty factor due to target frequency divergence is remarkably high in this case (4.15–4.49) and should drag the score to quite negative values, but the compensation factor due to background frequency divergence is even greater and fully compensates it. We thus leave the *scoring procedure* at step 7. This is the typical case of poorly conserved sequences with singular key structural aspects that are however highly preserved (c.f. the pattern of proline and arginine residues). As the background frequencies $F_X$ and $F_Y$ are far from the standard background $P$ associated with the BLOCKS database, the evaluation of a more realistic score for these sequences pass through the use of a decompositionally adjusted BLOSUM matrix [11]. Such matrices are built in such a way as to reduce background frequency divergence, so as to eliminate the portion of target divergence that is induced by it. In this way, the residual target divergence accounts only for effective evolutional divergence between sequences.

As a final example, we obtained BLOSUM spectra also for sequences from obviously uncorrelated families. The results are reported in Table 9 and Figure 4. In these cases we generally obtain a poor stochastic correlation $I(X,Y)$, and a high value for the penalty factor $D(F_{XY}//P_{AB})$, leading to a globally negative score, which is not compensated by background

Table 5: BLOSUM decomposition for ungapped and gapped hemoglobins.

```
P02232: 49 SAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVVLDGKDGSIHIQKGVLDPHFVVVKEALLKTIKE 115
           ++ +  S ++ AHA +V  ++ +   +L +     L       H    V   H+ +  + L++ ++
S06134: 61 ASQLRSSRQMQAHAIRVSSIMSEYVEELDSDILPELLATLARTHDLNKVGADHYNLFAKVLMEALQA 127

P02232: 116 ASGDKWSEELSAAWEVAYDGLATAI 140
               G  ++E+   AW  A+
S06134: 128 ELGSDFNEKTRDAWAKAFSIVQAVL 152
```

| *Vicia faba* leghemoglobin I versus *Paracaudina chilensis* hemoglobin I (ungapped) | | | | | | | |
|---|---|---|---|---|---|---|---|
| BLOSUM | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 100 | 1.839 | 2.478 | 0.264 | 0.207 | −0.166 | −31 | 15.2 |
| 80 | 1.839 | 2.240 | 0.264 | 0.199 | 0.063 | 12 | 15.2 |
| 62 | 1.839 | 2.128 | 0.260 | 0.192 | 0.163 | 35 | 15.2 |
| 50 | 1.839 | 2.077 | 0.255 | 0.185 | 0.203 | 54 | 15.2 |
| *40* | *1.839* | *2.051* | *0.255* | *0.194* | *0.237* | *83* | *15.2* |
| 35 | 1.839 | 2.070 | 0.263 | 0.202 | 0.235 | 82 | 15.2 |
| *Vicia faba* leghemoglobin I versus *Paracaudina chilensis* hemoglobin I (gapped) | | | | | | | |
| 100 | 1.597 | 1.962 | 0.166 | 0.172 | −0.026 | −10 | 18.1 |
| 80 | 1.597 | 1.759 | 0.161 | 0.163 | 0.162 | 40 | 18.1 |
| 62 | 1.597 | 1.661 | 0.154 | 0.153 | 0.243 | 65 | 18.1 |
| 50 | 1.597 | 1.618 | 0.145 | 0.145 | 0.268 | 104 | 18.1 |
| *40* | *1.597* | *1.606* | *0.145* | *0.155* | *0.291* | *152* | *18.1* |
| 35 | 1.597 | 1.623 | 0.154 | 0.163 | 0.283 | 148 | 18.1 |

```
P02232: 2 FTEKQEALVNSSSQLFKQNPSNYSVLFYTIILQKAPTAKAMFSFLK--DSAGVVDSPKLGAHAEKVF 68
           T  Q+ +V  +     +N +++    +  I    P+A+  F  +    ++ +  S ++ AHA +V
S06134: 12 LTLAQKKIVRKTWHQLMRNKTSFVTDVFIRIFAYDPSAQNKFPQMAGMSASQLRSSRQMQAHAIRVS 78

P02232: 69 GMVRDSAVQLRATGEVVLDGKDGSIHIQKGVLDPHFVVVKEALLKTIKEASGDKWSEELSAAWEVAY 135
            ++ +   +L +     L       H    V   H+ +  + L++ ++   G  ++E+   AW  A+
S06134: 79 SIMSEYVEELDSDILPELLATLARTHDLNKVGADHYNLFAKVLMEALQAELGSDFNEKTRDAWAKAF 145
```

frequency divergences. Note that in two cases, a mildly positive score could suggest a distant relationship. Analysis of the BLO*Spectrum* helps in evaluating this possibility. The PF12 versus GAT1 alignment is simply a case of overcompensation for a nontypical sequence (the background frequency divergence for one of the sequences is very high). In the second case, however, the $I(X,Y)$ value for the BD04 versus GAT1 human alignment is surprisingly quite high, suggesting that a closer look might be appropriate.

## 4. CONCLUSIONS

A standard use of scoring substitution matrices, such as BLOSUM-$\theta$, is often insufficient for discovering concealed correlations between weakly related sequences. Among other causes, this can derive from (i) the introduction of gaps during evolution (ii) use of a BLOSUM-$\theta$ matrix tailored for a different evolutionary distance than that pertaining to the aligned sequences, and/or (iii) the use of standard matrices

Table 6: BLOSUM decomposition for ungapped and gapped transposons.

```
NP_493808: 243 VFQQDNDPKHTSLHVRSWFQRRHVHLLDWPSQSPDLNPIEH 283
                +F  DN P HT+  VR   +  +  +L   + SPDL P +
A26491: 245 IFLHDNAPSHTARAVRDTLETLNWEVLPHAAYSPDLAPSDY 285
```

| *Drosophila mauritiana* mariner transposon versus *C. elegans* transposon TC1 (ungapped) | | | | | | | |
|---|---|---|---|---|---|---|---|
| BLOSUM | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 100 | 2.339 | 2.926 | 0.740 | 0.531 | 0.685 | 55 | 34.1 |
| 80 | 2.339 | 2.849 | 0.733 | 0.531 | 0.754 | 60 | 34.1 |
| *62* | *2.339* | *2.800* | *0.724* | *0.526* | *0.789* | *67* | *34.1* |
| 50 | 2.339 | 2.831 | 0.721 | 0.516 | 0.746 | 90 | 34.1 |
| 40 | 2.339 | 2.935 | 0.716 | 0.509 | 0.630 | 104 | 34.1 |
| 35 | 2.339 | 2.969 | 0.714 | 0.505 | 0.590 | 92 | 34.1 |
| *Drosophila mauritiana* mariner transposon versus *C. elegans* transposon TC1 (gapped) | | | | | | | |
| 100 | 1.991 | 2.244 | 0.244 | 0.243 | 0.235 | 40 | 25.0 |
| 80 | 1.991 | 2.110 | 0.246 | 0.234 | 0.362 | 67 | 25.0 |
| 62 | 1.991 | 2.021 | 0.245 | 0.227 | 0.443 | 91 | 25.0 |
| *50* | *1.991* | *2.009* | *0.237* | *0.226* | *0.445* | *123* | *25.0* |
| 40 | 1.991 | 2.043 | 0.227 | 0.228 | 0.404 | 152 | 25.0 |
| 35 | 1.991 | 2.066 | 0.226 | 0.229 | 0.381 | 144 | 25.0 |

```
NP_493808: 243 VFQQDNDPKHTSLHVRSWFQRRHVHLLDWPSQSPDLNPIE-HLWEELERRLGGIRASNAD 301
                +F  DN P HT+  VR   +  +  +L   + SPDL P + HL+  +   L   R  + +
A26491: 245 IFLHDNAPSHTARAVRDTLETLNWEVLPHAAYSPDLAPSDYHLFASMGHALAEQRFDSYE 304


NP_493808: 302 AKFNQLENAWKAIPMSVIHKLIDSMPRRCQAVIDANG 338
                +    L+  + A       + I  +P R +  + ++G
A2649: 305 SVKKWLDEWFAAKDDEFYWRGIHKLPERWEKCVASDG 341
```

for comparison of proteins with nonstandard background frequency distributions of amino acids. All these well-known effects can be better evidenced and quantified by decomposition of BLOSUM score (BLO*Spectrum*) according to (11). This equation highlights the core of the biological correlation measured by the BLOSUM score, that is mutual information $I(X,Y)$, or sequence convergence. If gaps are taken into account (such as in BLAST), and the correct $\theta$ parameter is chosen with the help of BLO*Spectrum*, and if the background frequencies of sequences are near to the standard ones, then the global score is given by sequence convergence plus a residual penalization factor due to target frequency divergence. This residual value implicitly takes into account that numerous substitution events may have occurred during sequence evolution, and so is a coherent measure of the biological relationship and distance between the sequences. If the background frequencies of sequences are not standard, then we have shown the BLOSUM scoring method has an in-built capacity to correct for anomalies in amino acid distributions using background frequency divergence as a compensation factor. One can also choose to compositionally adjust the matrix, so as to reduce the compensation factor together with the component of target frequency divergence that is induced by a bad background frequency distribution. This systematic method is illustrated in the *scoring analysis procedure* of Section 2.

Our decomposition becomes important when we consider sequences for which the BLOSUM score indicates a weak or no correlation. A critical evaluation of the BLO*Spectrum* components can help corroborate or identify an underlying biological correlation and whether the matrices being used are the most appropriate ones for measuring it. In other words, when considering the grey area of BLOSUM scores with a marginal significance, it could help to

Table 7: The BLOSUM terms for beta defensins.

| BD01 human versus BD02 human | | | | | | | |
|---|---|---|---|---|---|---|---|
| BLOSUM | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 100 | 3.030 | 3.566 | 0.564 | 0.618 | 0.646 | 45 | 41.6 |
| 80 | 3.030 | 3.453 | 0.568 | 0.623 | 0.768 | 58 | 41.6 |
| 62 | 3.030 | 3.438 | 0.604 | 0.652 | 0.849 | 65 | 41.6 |
| 50 | 3.030 | 3.418 | 0.615 | 0.663 | 0.891 | 99 | 41.6 |
| 40 | 3.030 | 3.378 | 0.577 | 0.626 | 0.855 | 129 | 41.6 |
| *35* | *3.030* | *3.320* | *0.539* | *0.588* | *0.837* | *120* | *41.6* |
| human beta defensins (BLOSUM-35) | | | | | | | |
| Sequences | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 1–3 | 2.731 | 3.325 | 0.539 | 0.751 | 0.697 | 101 | 30.5 |
| 1–4 | 2.532 | 3.658 | 0.539 | 0.728 | 0.141 | 22 | 16.6 |
| 2-3 | 2.009 | 3.466 | 0.794 | 0.616 | −0.045 | −10 | 10.2 |
| 2–4 | 2.334 | 3.522 | 0.609 | 0.568 | −0.009 | 0 | 12.1 |
| 3-4 | 2.122 | 3.286 | 0.794 | 0.655 | 0.286 | 44 | 20.5 |

Table 8: The BLOSUM terms for Pro/Arg-rich peptides.

| BCT5 bovin versus BCT7 bovin | | | | | | | |
|---|---|---|---|---|---|---|---|
| BLOSUM | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 100 | 0.424 | 4.935 | 2.329 | 2.460 | 0.279 | 28 | 34.8 |
| 80 | 0.424 | 4.724 | 2.317 | 2.449 | 0.467 | 42 | 34.8 |
| 62 | 0.424 | 4.637 | 2.301 | 2.430 | 0.518 | 37 | 34.8 |
| 50 | 0.424 | 4.533 | 2.264 | 2.389 | 0.544 | 68 | 34.8 |
| 40 | 0.424 | 4.407 | 2.221 | 2.338 | 0.576 | 97 | 34.8 |
| *35* | *0.424* | *4.368* | *2.199* | *2.301* | *0.556* | *98* | *34.8* |
| Pro/Arg-rich peptides (BLOSUM-35) | | | | | | | |
| Sequences | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 1–3 | 0.516 | 4.434 | 2.095 | 2.205 | 0.382 | 63 | 30.9 |
| 1–4 | 0.446 | 4.491 | 2.199 | 2.488 | 0.643 | 110 | 39.5 |
| 2-3 | 0.584 | 4.156 | 2.095 | 2.257 | 0.780 | 133 | 47.6 |
| 2–4 | 0.406 | 4.350 | 2.256 | 2.251 | 0.563 | 134 | 37.2 |
| 3-4 | 0.609 | 4.260 | 2.095 | 2.347 | 0.792 | 132 | 45.2 |

decide if an evolutionary relationship actually exists. We provide online software at http://bioinf.dimi.uniud.it/software/software/blosumapplet which integrates a BLO*Spectrum* histogram with the score obtained by a classical BLAST engine working on two input sequences, which allows an immediate visual analysis of the score components. The systematic use of BLO*Spectrum* parameters to permit a more sensitive filtering of scores inside a BLAST or similar engine could be the logical next operative step. We have provided several biological examples indicating the potential of our method, but it is clear that it needs a massive biological experimentation to completely test its effective usefulness.

## APPENDIX

*Proof of* (11). By multiplying inside the log function of (7) by $f(i,j)/f(i,j)$ and by $f(i)f(j)/f(i)f(j)$ and rearranging the terms, we obtain

$$\begin{aligned} S_N(X,Y) &= \sum_{i,j} f(i,j)\log\frac{p(i,j)}{p(i)p(j)}\frac{f(i,j)}{f(i,j)}\frac{f(i)f(j)}{f(i)f(j)} \\ &= \sum_{i,j} f(i,j)\log\frac{f(i,j)}{f(i)f(j)} - \sum_{i,j} f(i,j)\log\frac{f(i,j)}{p(i,j)} \\ &\quad + \sum_{i,j} f(i,j)\log\frac{f(i)f(j)}{p(i)p(j)} \end{aligned}$$

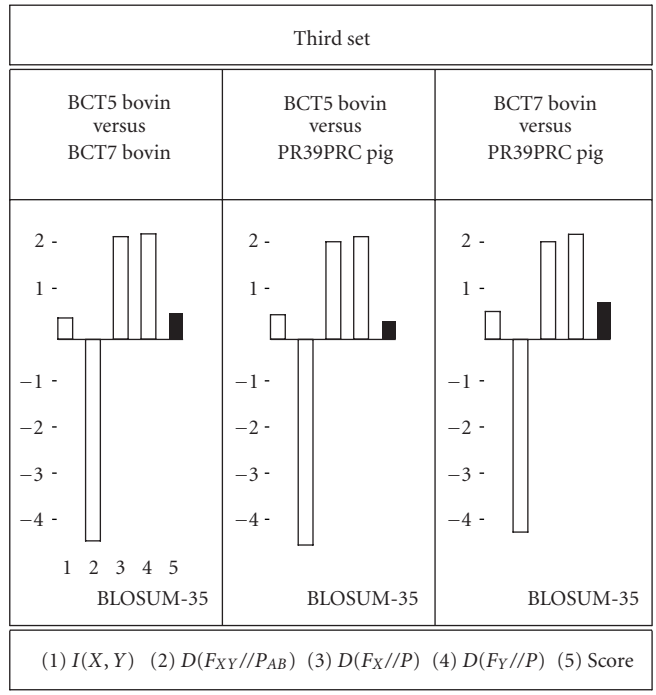


Figure 3: BLO*Spectrum* for sequences of the third set.

Table 9: Some examples of BLOSUM-35 terms for sequences belonging to noncorrelated families.

| BLOSUM-35 | | | | | | | |
|---|---|---|---|---|---|---|---|
| HNF4-$\alpha$ human versus HNF6 human | | | | | | | |
| Sequences | $I(X,Y)$ | $D(F_{XY}//P_{AB})$ | $D(F_X//P)$ | $D(F_Y//P)$ | $S_N(X,Y)$ | Score | % Identity |
| 1-1 | 0.578 | 0.986 | 0.036 | 0.205 | −0.165 | −312 | 5.37 |
| HNF4-$\alpha$ human versus GAT1 human | | | | | | | |
| 1-1 | 0.712 | 1.033 | 0.038 | 0.193 | −0.088 | −144 | 8.71 |
| HNF6 human versus GAT1 human | | | | | | | |
| 1-1 | 0.622 | 1.122 | 0.230 | 0.193 | −0.076 | −143 | 8.47 |
| BD04 human versus BCT7 bovin | | | | | | | |
| 4–2 | 1.010 | 3.887 | 0.460 | 2.220 | −0.195 | −36 | 10.0 |
| PF12 pig versus GAT1 human | | | | | | | |
| 4–1 | 0.686 | 3.486 | 2.182 | 0.709 | 0.091 | 24 | 18.2 |
| BD04 human versus GAT1 human | | | | | | | |
| 4–1 | 2.243 | 3.033 | 0.460 | 0.465 | 0.136 | 25 | 12.0 |

$$
\begin{aligned}
&= I(X,Y) - D(F_{XY}//P_{AB}) \\
&\quad + \sum_{i,j} f(i,j)\log\frac{f(i)}{p(i)} + \sum_{i,j} f(i,j)\log\frac{f(j)}{p(j)} \\
&= I(X,Y) - D(F_{XY}//P_{AB}) + D(F_X//P_A) \\
&\quad + D(F_Y//P_B). \qquad \text{(A.1)}
\end{aligned}
$$

□

A fuller understanding of the mathematical tools used in Section 2 requires some definitions and mathematical properties pertaining to *ID* and *MI*; they are summarized as follows.

Let us start by considering some probability distributions [10] over an alphabet $\mathcal{A}$ with $K$ symbols, for example $P = \{p_1, p_2, \ldots, p_K\}$, $Q = \{q_1, q_2, \ldots, q_K\}$, and so on. In our context, $K = 20$, as there are 20 amino acids, and the alphabet letters correspond to the 1-letter amino acid standard coding (D = Asp, E = Glu, W = Trp, etc.). If we imagine the space of all possible $K$ dimensional probability distributions, it is right to ask what is the "distance" from $P$ to $Q$ (or vice

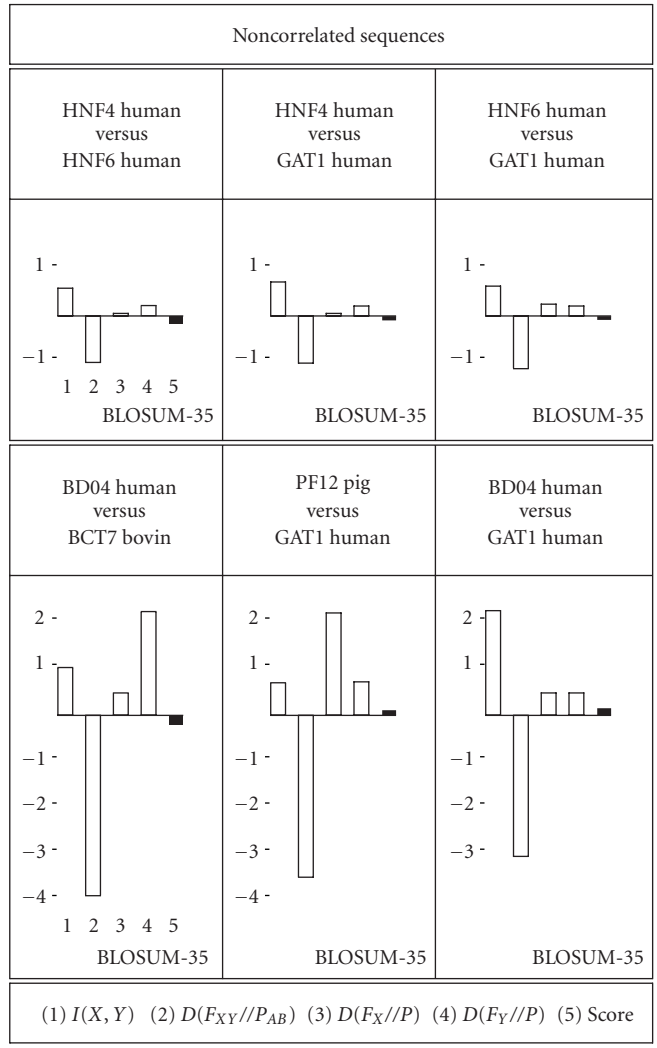


Figure 4: BLO*Spectrum* for noncorrelated sequences.

versa). The most popular (pseudo-)distance is the *informational divergence* $D(P//Q)$,

$$D(P//Q) \triangleq \sum_{i=1}^{K} p(i) \log \frac{p(i)}{q(i)}, \tag{A.2}$$

introduced by Kullback in 1954 in the context of statistics [29]; here $p(i) \geq 0$ and $q(i) > 0$. It is easy to verify [18] that the informational divergence (*ID*) is nonnegative, and it is equal to 0 if and only if $P$ is coincident with $Q$ ($P \equiv Q$). Furthermore, *ID* is not boundable, since $D(P//Q) \to +\infty$ if an $i$ exists such that $q(i) \to 0$. All this can be summarized in the following way:

$$\begin{aligned} &0 \leq D(P//Q) \leq +\infty \qquad (= 0 \text{ when } P \equiv Q) \\ &(= +\infty \text{ when there exists } i \text{ such that } 2(i) = 0). \end{aligned} \tag{A.3}$$

Note that *ID* is the sum of positive and negative terms, and the fact that the average is always greater than zero is not obvious (it is a consequence of the convexity property of the logarithm). Since $D(P//Q) = 0$ if and only if $P \equiv Q$, this allows us to interpret the *ID* as a measure of *(pseudo)distance* between probability distributions. It is only "pseudo" (from the mathematical point of view) since the concept of "distance" is well defined in mathematics, and requires also symmetry between the variables and the validity of the so-called triangular inequality. But *ID* lacks both these last two properties, since, in general, $D(P//Q) \neq D(Q//P)$ (it is asymmetric) and, if $R$ is a third probability distribution, we are not sure that $D(P//R) + D(R//Q)$ is greater than $D(P//Q)$ (the triangular inequality does not hold). We underline that such a distance is not symmetric (and so the order in which $P$ and $Q$ are specified does matter), that is, it is a distance "from" rather than a distance "between."

Suppose now that $P_X = \{p_X(1), p_X(2), \ldots, p_X(K)\}$ and $P_Y = \{p_Y(1), p_Y(2), \ldots, p_Y(K)\}$ are the probability distributions associated to the (random) variables $X$ and $Y$, which take their values in the same alphabet $\mathcal{A}$. Here, $p_X(i) = \Pr\{X = i\}$ means the probability that the variable $X$ assumes

the value $i$. In our framework, $X$ and $Y$ are two protein sequences of the same length $n$, and $p_X(2) = \Pr\{X = 2\} = 0.09$ (e.g.) is interpreted as the relative frequency of the second amino acid of the alphabet $\mathcal{A}$; so, the overall occurrence of the 2nd amino acid in sequence $X$ is equal to $0.09n$. In this context, we can introduce also a *joint probability distribution* associated to the sequences, $P_{XY} = \{p_{XY}(i,j),\ i,j \in \mathcal{A}\} = \Pr\{X = i,\ Y = j,\ i,j \in \mathcal{A}\}$, where $p_{XY}(i,j)$ corresponds to the relative frequency of finding the amino acids $i$, $j$ paired in a certain position of the alignment between $X$ and $Y$. It is well known that $\sum_{i,j} p_{XY}(i,j) = 1$ ($P_{XY}$ is a probability distribution) and that the sum of the joint probabilities over one variable gives the *marginal* of the other variable $\sum_j p_{XY}(i,j) = p_X(i)$. For example, given that the ninth and the fifth amino acid in the alphabet are Arginine and Leucine, respectively, $p_{XY}(9,5) = p_{XY}(\text{Arg}, \text{Leu}) = 0.01$ means that the relative frequency of finding Arg in $X$ paired with Leu in $Y$ is equal to 0.01. In practice, we avoid the use of the subscripts, and use the simpler notation $p(i)$ and $p(i,j)$ instead of $p_X(i)$ and $p_{XY}(i,j)$.

Since the condition of independence between two variables (protein sequences) $X$ and $Y$ is fixed by the formula $p_{XY}(i,j) = p_X(i)p_Y(j)$ (for each pair $i,j \in \mathcal{A}$), then, once assigned a certain $P_{XY}$, it could be interesting to attempt to evaluate the distance of $P_{XY}$ *from the condition of independence* between the variables. Making use of the *ID* (A.2), we need to evaluate the quantity $D(P_{XY}//P_XP_Y)$, that is the stochastic distance between the joint $P_{XY}$ and the product of the marginals $P_XP_Y$. If we have independence, then $P_{XY} \equiv P_XP_Y$, and the divergence equals zero. On the contrary, if it appears that $X$ and $Y$ are tied by a certain degree of dependence, this can be measured by

$$D(P_{XY}//P_XP_Y) = \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \triangleq I(X,Y) \geq 0. \tag{A.4}$$

This quantity is called also the *mutual information* (or *relative entropy*) $I(X,Y)$ between the random variables (the protein sequences, in our setting) $X$ and $Y$. It is symmetric in its variables ($I(X,Y) = I(Y,X)$) and is always nonnegative, since it is an informational divergence. Note also that $MI$ is upper bounded by the logarithm of the alphabet cardinality, that is $I(X,Y) \leq \log 20$ [18]. Moreover, since it equals zero if and only if the joint probability distribution coincides with the product of the marginals, that is, when we have *independence* between the two variables, we can interpret the mutual information ($MI$) as a measure of *stochastic dependence between X and Y*. From another point of view, we can say that independence is equivalent to the situation in which the variables $X$ and $Y$ do not exchange information. So, the meaning of $I(X,Y)$ can be read also as the *degree of dependence* between the variables, or as the *average information exchanged* between the same variables. Mutual information is one of the pillars of Shannon information theory, and was introduced in the seminal paper by Shannon [16, 17].

## ACKNOWLEDGMENTS

The authors thank Jorja Henikoff, who provided the matrices of joint probability distributions associated to the database BLOCKS, and an anonymous referee of a previous version of this paper, who made several key remarks. This work has been supported by the Italian Ministry of Research, PRIN 2003, FIRB 2003 Grants, by the Istituto Nazionale di Alta Matematica (INdAM), 2003 Grant, and by the Regione Friuli Venezia Giulia (2005 Grants).

## REFERENCES

[1] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[2] A. D. McLachlan, "Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome $c_{551}$," *Journal of Molecular Biology*, vol. 61, no. 2, pp. 409–424, 1971.

[3] D. Sankoff, "Matching sequences under deletion-insertion constraints," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 69, no. 1, pp. 4–6, 1972.

[4] P. H. Sellers, "On the theory and computation of evolutionary distances," *SIAM Journal on Applied Mathematics*, vol. 26, no. 4, pp. 787–793, 1974.

[5] M. S. Waterman, T. F. Smith, and W. A. Beyer, "Some biological sequence metrics," *Advances in Mathematics*, vol. 20, no. 3, pp. 367–387, 1976.

[6] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed., vol. 5, supplement 3, pp. 345–352, National Biomedical Research Foundation, Washington, DC, USA, 1978.

[7] S. F. Altschul, "Amino acid substitution matrices from an information theoretic perspective," *Journal of Molecular Biology*, vol. 219, no. 3, pp. 555–565, 1991.

[8] S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 6, pp. 2264–2268, 1990.

[9] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915–10919, 1992.

[10] W. Feller, *An Introduction to Probability and Its Applications*, John Wiley & Sons, New York, NY, USA, 1968.

[11] Y.-K. Yu, J. C. Wootton, and S. F. Altschul, "The compositional adjustment of amino acid substitution matrices," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 26, pp. 15688–15693, 2003.

[12] S. F. Altschul, "A protein alignment scoring system sensitive at all evolutionary distances," *Journal of Molecular Evolution*, vol. 36, no. 3, pp. 290–300, 1993.

[13] D. J. States, W. Gish, and S. F. Altschul, "Improved sensitivity of nucleic acid database searches using application-specific scoring matrices," *Methods*, vol. 3, no. 1, pp. 66–70, 1991.

[14] S. R. Sunyaev, G. A. Bogopolsky, N. V. Oleynikova, P. K. Vlasov, A. V. Finkelstein, and M. A. Roytberg, "From analysis of protein structural alignments toward a novel approach to align protein sequences," *Proteins: Structure, Function, and Bioinformatics*, vol. 54, no. 3, pp. 569–582, 2004.

[15] M. A. Zachariah, G. E. Crooks, S. R. Holbrook, and S. E. Brenner, "A generalized affine gap model significantly improves protein sequence alignment accuracy," *Proteins: Structure, Function, and Bioinformatics*, vol. 58, no. 2, pp. 329–338, 2005.

[16] C. E. Shannon, "A mathematical theory of communication—part I," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[17] C. E. Shannon, "A mathematical theory of communication—part II," *Bell System Technical Journal*, vol. 27, pp. 623–656, 1948.

[18] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, NY, USA, 1981.

[19] A. A. Schäffer, L. Aravind, T. L. Madden, et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[20] F. Frommlet, A. Futschik, and M. Bogdan, "On the significance of sequence alignments when using multiple scoring matrices," *Bioinformatics*, vol. 20, no. 6, pp. 881–887, 2004.

[21] S. F. Altschul, J. C. Wootton, E. M. Gertz, et al., "Protein database searches using compositionally adjusted substitution matrices," *FEBS Journal*, vol. 272, no. 20, pp. 5101–5109, 2005.

[22] A. A. Schäffer, Y. I. Wolf, C. P. Ponting, E. V. Koonin, L. Aravind, and S. F. Altschul, "IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices," *Bioinformatics*, vol. 15, no. 12, pp. 1000–1011, 1999.

[23] W. R. Rypniewski, A. Perrakis, C. E. Vorgias, and K. S. Wilson, "Evolutionary divergence and conservation of trypsin," *Protein Engineering*, vol. 7, no. 1, pp. 57–64, 1994.

[24] A. L. Hughes, "Evolutionary diversification of the mammalian defensins," *Cellular and Molecular Life Sciences*, vol. 56, no. 1-2, pp. 94–103, 1999.

[25] F. Bauer, K. Schweimer, E. Klüver, et al., "Structure determination of human and murine $\beta$-defensins reveals structural conservation in the absence of significant sequence similarity," *Protein Science*, vol. 10, no. 12, pp. 2470–2479, 2001.

[26] A. Tossi and L. Sandri, "Molecular diversity in gene-encoded, cationic antimicrobial polypeptides," *Current Pharmaceutical Design*, vol. 8, no. 9, pp. 743–761, 2002.

[27] R. Gennaro, M. Zanetti, M. Benincasa, E. Podda, and M. Miani, "Pro-rich antimicrobial peptides from animals: structure, biological functions and mechanism of action," *Current Pharmaceutical Design*, vol. 8, no. 9, pp. 763–778, 2002.

[28] M. E. Selsted, M. J. Novotny, W. L. Morris, Y.-Q. Tang, W. Smith, and J. S. Cullor, "Indolicidin, a novel bactericidal tridecapeptide amide from neutrophils," *Journal of Biological Chemistry*, vol. 267, no. 7, pp. 4292–4295, 1992.

[29] S. Kullback, *Information Theory and Statistics*, Dover, Mineola, NY, USA, 1997.