Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 328 (2005) 445–453
COMPTES RENDUS BIOLOGIES
http://france.elsevier.com/direct/CRASS3/

Genetics / Génétique

Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions

Olivier Bastien [a,b], Sylvaine Roy [c], Éric Maréchal [a,*]

[a] *Laboratoire de physiologie cellulaire végétale, département « Réponse et Dynamique cellulaire », UMR 5019, CNRS–CEA–INRA–université Joseph-Fourier, CEA Grenoble, 17, rue des Martyrs, 38054 Grenoble cedex 09, France*
[b] *Gene-IT, 147, av. Paul-Doumer, 92500 Rueil-Malmaison, France*
[c] *Laboratoire de biologie, informatique et mathématiques, département « Réponse et Dynamique cellulaire », CEA Grenoble, 17, rue des Martyrs, 38054 Grenoble cedex 09, France*

Received 6 December 2004; accepted after revision 1 February 2005

Available online 25 February 2005

Presented by Roland Douce

Abstract

Automatic comparison of compositionally biased genomes, such as that of the malarial causative agent *Plasmodium falciparum* (82% adenosine + thymidine), with genomes of average composition, is currently limited. Indeed, popular tools such as BLAST require that amino acid distributions be similar in aligned sequences. However, the *P. falciparum* genome is so biased that six amino acids account for more than 50% of the protein composition. One reason for the comparison methods failure lies in the compositional difference between the query and the subject proteomes, which is not taken into account in the amino acid substitution matrices. This paper introduces a method to derive substitution matrices, in particular BLOSUM 62, in the frame of the information theory. It allows the construction of non-symmetrical matrices, taking into account the non-symmetric amino acid distributions. The dirAtPf family of matrices allowing the comparison of *P. falciparum* and *A. thaliana* is given as an example. This paper further provides an analysis of the obtained matrices in the frame of the information theory, supporting the discrimination advantage they bring. ***To cite this article: O. Bastien et al., C. R. Biologies 328 (2005).***

Résumé

La comparaison automatique de génomes biaisés, tel que celui de l'agent du paludisme *Plasmodium falciparum* (82 % adénosine + thymidine), avec des génomes de composition moyenne, est limitée. En effet, les outils populaires, tels que BLAST, imposent que les distributions en amino acides des séquences comparées soient proches. Or le génome de *P. falciparum* est tellement biaisé que six aminoacides constituent plus de 50 % de la composition protéique. Une cause de l'échec des méthodes de comparaison est de ne pas tenir compte de ces différences de distributions entre protéomes « requête » et « sujet », en parti-

* Corresponding author.
E-mail address: emarechal@cea.fr (É. Maréchal).

1631-0691/$ – see front matter
doi:10.1016/j.crvi.2005.02.002

culier au niveau de la matrice de substitution des aminoacides. Cette note présente une méthode pour dériver les matrices de substitution, en particulier BLOSUM 62, dans le cadre de la théorie de l'information. Il est ainsi possible de construire des matrices non symétriques, tenant compte de la non-symétrie des distributions en amino acides. La famille dirAtPf de matrices permettant de comparer *Arabidopsis thaliana* et *Plasmodium falciparum* est proposée comme exemple. Cette note présente, de plus, une analyse de ces matrices dans le cadre de la théorie de l'information, soutenant théoriquement le gain de discrimination qu'elles peuvent apporter. ***Pour citer cet article : O. Bastien et al., C. R. Biologies 328 (2005).***

Keywords: Substitution matrix; BLOSUM; Biased genome; *Plasmodium falciparum*; Information theory; Mutual information

Mots-clés : Matrice de substitution ; BLOSUM ; Génome biaisé ; *Plasmodium falciparum* ; Théorie de l'information ; Information mutuelle

1. Introduction

Comparison of biological macromolecules has become an everyday task for biologists, for extremely diverse purposes such as genomic sequencing, structural modelling, functional inference, phylogenetic reconstruction, allelic or mutational analyses, etc. In all cases, comparison methods rely on a fundamental postulate that one can simply state as: "the closer in the evolution, the more alike and reversely, the more alike, *probably* the closer in the evolution" [1]. Numerous computer-based tools are used to estimate the proximity of protein sequences [2]. Alignment of two sequences is typically done by maximizing (or minimizing) a given quantity, named score, which reflects the shared features of the two biological entities. Global alignment algorithms [3] are not accurate to assess homology of domains in modular proteins [4]; local alignments are better suited [5,6]. They use scoring matrices to maximize the summed scores of compared residues and find optimal local alignments, computed with a dynamic programming procedure [2–6]. Scoring matrices have been found to be similarity matrices as well [2,4,6]. Many similarity matrices are available [7–10] and evaluation studies led to the conclusion that those based on a log-odds ratio, like BLOSUM 62 [8], over performed the others [9]. BLOSUM 62 was computed using blocks of aligned sequences with

$$s_{ij} = \frac{1}{\lambda} \log\left(\frac{q_{ij}}{p_i p_j} \right) \tag{1}$$

where i and j are aligned amino acids, q_{ij} the frequency of the observation: "i is aligned with j", i.e. the target frequency, p_i and p_j are respectively the i and j frequency, i.e. the background frequency and λ is a scaling factor.

Karlin and Altschul [11] have shown that substitution matrices depend on a particular set of data in which amino acids are paired with frequencies that correspond to the matrices' target frequencies. If the set of data is made up of aligned homologous sequences, then the matrix is usable to distinguish distant local homologies, from similarities due to chance [12]. Using information theory, Altschul [12] have further reported that substitution matrices can be evaluated using the average information they contained. This average information, known as the *relative entropy* of Shannon (1948) [13], was computed as

$$H = \sum_{i,j} q_{ij} s_{ij} \tag{2}$$

This formula can be trivially applied to similarity matrices, in order to estimate their 'sensitivity'. H computation is therefore a popular parameter when new matrices are proposed. In a recent report [14], it has further been used as a computation constraint to derive substitution matrices, i.e. with constant entropy. Information theory is much more than a practical frame for matrix computation; it allows essentially the translation of biological properties into mathematical models, particularly in probabilistic terms.

Given a probability law P that characterizes a random variable, the Hartley self-information h [15] is defined as the amount of information one gains when an event i occurs:

$$h(i) = -\log\big(P(i)\big) \tag{3}$$

The less likely an event i, the more we learn about the system when i happens. The mutual-information I between two events, is the reduction of the uncertainty of one event i due to the knowledge of the other j:

$$I_{j \to i} = h(i) - h(i/j) \tag{4}$$

Mutual information being symmetrical, $I_{j\to i} = I_{i\to j}$ and is noted $I(i; j)$. Self and mutual information of two events i and j are related:

$$h(i \cap j) = h(i) + h(j) - I(i; j) \quad (5)$$

If the occurrence of one of the two events makes the second impossible, mutual information is equal to $-\infty$. If the two events are fully independent, mutual information is null.

Recently, we rigorously demonstrated that the score $s(i, j)$ between two amino acids was the *mutual information*, in the sense of Hartley, between the two considered amino acid (Bastien et al., submitted), that is to say:

$$s(i, j) = \frac{1}{\lambda} I(i; j) \quad (6)$$

This assertion was implicit in Eq. (2). It first implies that it is impossible to build separable and metric sequence spaces that conserve the mutual information between compared sequences. Second, the fundamental postulate can be reformulated in the information theory framework: "Given two homologous proteins a and b, the closer in the evolution, the greater the mutual information between a and b and reversely, the greater the mutual information between a and b, *probably* the closer in the evolution."

Whereas the BLOSUM model is efficient in most cases, it fails to estimate satisfactorily the alignment between two proteins of very different amino acid composition [16–18]. A major reason lies in Eq. (1) that does not account for the distinct sequences where the amino acids i and j are sampled. This can be of importance when compared proteins are from very different cell environments (like soluble or membrane-bound proteins) or of strongly different amino acids composition [18–21]. In a pioneering study addressing this problem, Müller et al. [22] introduced a non-symmetric substitution matrices model for the comparison of homologous trans-membrane proteins and showed that this kind of matrices had a larger discrimination power, i.e. specificity. We reformulated this model for the comparison of biased and non-biased genomes ([16] and the present work).

Considering the general case of genome comparison with distinct global amino acid compositions, we used mutual information theory to construct non-symmetric substitution matrices dedicated to the detection of homologous sequences. An important application is the computing of non-symmetric matrices for the comparison between complete proteomes with deep differences in their composition in amino acids. Comparison of *Arabidopsis thaliana* and *Plasmodium falciparum* proteomes is given as a case study.

2. Methods

2.1. Nomenclature for sequences databases (query, subject, Species 1 and 2) and for the non-symmetric $DirSp_1Sp_2$ matrices

In molecular alignments, we used the standard nomenclature for homology searches methods, i.e. *query* for a known probing sequence (or database of sequences) that is compared to another sequence (or database of sequences), termed *subject*. The family of non-symmetric matrices computed here were dedicated to the comparison of a query sequence from a first species (termed Species1) with a subject database, or a single sequence, from a second species (termed Species2). The family of amino acid substitution matrices referred to as $DirSp_1Sp_2$ (Species1 → Species2) were designed to be implemented in the conventional BLASTP alignment algorithm: columns correspond to the query (or Species1) and rows to the subject (or Species2) entries.

2.2. Identification of a non-redundant set of homologous proteins between Arabidopsis thaliana and Plasmodium falciparum

As a source of genomic sequence material, we selected the annotated sequences from *Arabidopsis thaliana* and *Plasmodium falciparum* from Internet databases, respectively the National Center for Biotechnology Information server (ftp://ftp.ncbi.nih.gov) and the Plasmodb genome resource (http://plasmodb.org/). The massive annotation resulting from collaborative efforts, genomic annotations contain errors. We selected therefore annotated sequences that were judged trustworthy at the downloading date in the respective Internet-available databases, i.e. 25 545 sequences from *A. thaliana* as of December 2002, 5334 from *P. falciparum* as of December 2002. According to a method describe previously [18], the two proteomes

were used to identify a non-redundant set of homologous proteins using the BLASTP program and the Smith–Waterman algorithm [5,6], implemented in the Biofacet software package (Gene-IT, France, [23]). To remove the similarity redundancies from *P. falciparum* and *A. thaliana* protein-sequence databases, for each proteome, we built up a random proteome database containing an identical number of protein sequences, of identical size and amino acid distribution, in which each sequence was an obligate shuffling of a corresponding sequence from the original database. Each real protein of a given organism was compared to all the sequences of the random database using BLASTP algorithm; the best alignment P-value was collected. From the distribution of the self x random P-values, a 5-percentile was set to define a cut-off. Then, for each species, the calculated cut-off was used as a criterion to partition the proteome owing to the single-linkage clustering method. Eventually, the longest sequence was drawn from each similarity cluster to build up non-redundant proteomes. All proteins from the *A. thaliana* non-redundant proteome were compared to the *P. falciparum* non-redundant proteome, using the SW algorithm. For each alignment, a Z-value was computed and a Z-value-cut-off of 8 was used to create clusters of aligned *P. falciparum* × *A. thaliana* sequences, owing to the single-clustering method. In this paper, this set was termed 'automatic training set'. Clusters were examined manually to select pairs of sequences whose functional annotation appeared analogous. This set was termed 'manual training set' (see table in [18]).

2.3. Initial construction of a database of protein blocks, from pairs of aligned sequences from Species1 and Species2 (or query and subject)

As described by [24], local alignments can be represented as ungapped blocks with each row a different protein segment and each column an aligned residue position. In the particular case described here, blocks can be simply derived as ungapped segments in pairs of aligned sequences. These 2-line blocks were ordered so that the first sequence always belongs to the same genome. Substitutions of a given amino acid from a first sequence (Species1 or query) with another amino acid from a second (Species2 or subject) were counted in all pairs of matching amino acids in each blocks in the database and then summed. The substitution frequency table is used to calculate matrices representing the odd ratio between these observed frequencies and those expected by chance.

2.4. Families of similarity matrices computed from blocks filtered by segment clustering

Closely related blocks in blocks databases exhibit a high percentage of identity, up to 100% when no amino acid substitution is observed in aligned sequences. Evolutionary divergence is marked by a decrease in the identity percentage. Thus, the distribution of the identity percentages within a block database can lead to a biased calculation of substitution matrices, that over- or underscores alignments of close or distant sequences. Therefore, for each matrix computing, the training set of blocks was filtered using a clustering percentage, so that sequence segments that were identical for at least that percentage of amino acid were kept for the substitution frequency counting. This filtering is an alternative definition of the clustering percentage described by [8], in which the multiple contribution of segments that were identical for at least that percentage, were averaged in calculating pair frequencies. In both cases, the decrease in the clustering percentage implies a decrease in the contribution of the blocks which percentage of identity is higher than the clustering percentage. Like for the BLOSUM family of matrices, varying the clustering percentage leads to a family of matrices.

2.5. Iterative process in the computation of non-symmetric matrices

Construction of $\mathrm{DirSp_1Sp_2}$ matrices was stepwise. Frequency tables, matrices, and programs for UNIX machines were primarily designed using the Biofacet multipurpose package (Gene-IT, Rueil-Malmaison, [23]. The initial training set of pairs of sequences derived from alignments using the Smith–Waterman algorithm implemented with BLOSUM 62. For a given clustering percentage, an initial non-symmetric matrix was computed, and indexed *1*: $\mathrm{DirSp_1Sp_{21}}$. The initial training set was then re-aligned using the Smith–Waterman algorithm implemented with this first non-symmetric matrix. From these alignments, ungapped segments were selected and filtered owing

to the defined clustering percentage, and used as a new database of Blocks to compute a new non-symmetric matrix indexed *2*: $\mathrm{DirSp_1Sp_{22}}$. The process was iterated, outputting a convergent family of matrices $\mathrm{DirSp_1Sp_{2,3}}$, $\mathrm{DirSp_1Sp_{2,4}}$, ... $\mathrm{DirSp_1Sp_{2,n}}$. The stable matrices where referred to as un-indexed $\mathrm{DirSp_1Sp_2}$.

3. Results and discussion

3.1. Question of the mutual information between two homologous sequences of distinct amino acid distributions

The physical environment of proteins, (pH, water solubility or association to membranes), the codon use that is required for their synthesis or nucleotidic compositional trends are constraints that can lead to very uncommon amino acid distributions in some families of protein or even in complete proteomes [16–18,22]. Although biased amino acid distributions affect the performance of protein comparison tools built for 'average' amino acid distributions, they can bring useful information to discriminate homologous proteins. To that purpose, we considered two kinds of sequences, or set of sequences, the first named *query*, the second *subject*. For example, *query* sequence can be from a first species, called Species1 (such as *Arabidopsis thaliana*; with an average nucleotidic −55% A+T− and amino acid distribution) and the *subject* sequence from a second species, called Species2 (such as *Plasmodium falciparum*; which nucleotidic bias −82% A+T− leads to a biased compositional proteome).

We can consider two kinds of events: $\{X = i\}$ given the amino acids i in the *subject* sequence and $\{Y = j\}$ given the amino acids j in the *query* sequence (as an application, X can be defined in a given species such as *Arabidopsis thaliana* and Y in another such as *Plasmodium falciparum*). The self-information h, for the occurrence of a given amino acid does not have the same signification in the context of each sequence:

$$h_X(i) = -\log\big(P_X(i)\big) \neq -\log\big(P_Y(i)\big) = h_Y(i) \qquad (7)$$

with P_X and P_Y the probability laws assigned to the random variables X and Y, respectively. From inequality (7), knowing an amino acid in one of the aligned sequences does not bring the same quantity of information concerning an amino acid occurrence at the aligned site of the other sequence. This can be easily verified for Asparagines in the case of *Arabidopsis thaliana* and *Plasmodium falciparum*. In *P. falciparum*, this amino acid is over represented and leads to well-known low-complexity regions, whereas it does not in *Arabidopsis thaliana*. Still, Eqs. (4) and (7) allow the definition of the mutual-information I between two amino acids in two different set of sequences, defined as the reduction of the uncertainty on event j in the query sequence, gained by the knowledge of the occurrence of i in the subject sequence:

$$I_{X=i\to Y=j} = h_Y(j) - h_{Y/X}(j/i) \qquad (8)$$

Using the conditional probability theorem [25], which states that:

$$P_{X/Y}(X = i/Y = j)P_Y(j) = P_{Y/X}(Y = j/X = i)P_X(i) \qquad (9)$$

we can state that the mutual information is symmetric in respect to the amino acid occurrence event *and* to the sequence were this event occurs: $I_{X=i\to Y=j} = I_{Y=j\to X=i}$. This last expression is defined as:

$$I_{XY}(i;j) = I_{YX}(j;i) \qquad (10)$$

It is important to notice that, in general:

$$I_{XY}(i;j) \neq I_{XY}(j;i) \qquad (11)$$

and therefore that the mutual information between two amino acids *is not symmetric* when just permuting the amino acids and not sequences in the two terms of Eq. (11). Using Eqs. (5), (8), (9) and (10), we can now state that:

$$I_{XY}(i;j) = \log\left\{\frac{P_{XY}((X=i)\cap(Y=j))}{P_X(i)P_Y(j)}\right\} \qquad (12)$$

Eq. (12) can be estimated from observed homologous aligned sequences and allows computation of a substitution amino acid sequence with Eq. (6). This matrix is non-symmetric (Eq. (12)) and implementation in optimization alignment algorithm should therefore be carried out paying attention to the *query* and the *subject* sequence order. An important property of this matrix is that the application inverse (transformation of the *query* into a *subject*, and vice-versa) is done by the transposition of the scoring matrix. The matrix $S_{XY}(i,j)$ in this order (i is taken from the Species2 and is reading on the row) is called dirSp1Sp2.

3.2. Training sets to compute non-symmetrical substitution matrices

Determining sets of homologous sequences is a difficult question. Here, we examined the possibilities of an automatic or manual selection of a training set of pairs of similar sequences, in the given example of *Arabidopsis thaliana* and *Plasmodium falciparum* proteomes, obtained after an all-against-all comparison of non-redundant protein sequence databases. We selected genomic sequences from *A. thaliana* (25 545 annotated sequences as of December 2002) and *P. falciparum* (5334 annotated sequences as of December 2002). The strong nucleotidic bias of the *P. falciparum* genome (82% AT) strikingly affects the amino acid distribution within encoded proteins (Fig. 1). Six amino acids (N, K, I, L, E, and S) account for 51% of the total amino acid content in proteins. Fig. 1 shows that the amino acid distribution in *A. thaliana* is strongly divergent, with a more balanced contribution of individual amino acids to the overall composition of proteins. On top of the very strong amino acid bias found in *P. falciparum*, the protein sequences exhibit a very low complexity, marked by long stretches of repeated amino acids. It is still not known whether the very low complexity is solely due to the amino acid bias or if a generic mechanism dedicated to the insertions of amino acid repeats would contribute to this striking occurrence of repeated amino acid portions in proteins. For an in-depth discussion of the biological rational of *P. falciparum* bias as compared to *A. thaliana*, see [18]. We selected a training set of

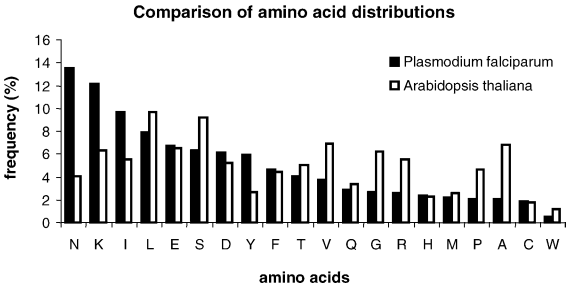

Fig. 1. Comparison of the amino acid distribution in the *Plasmodium falciparum* and *Arabidopsis thaliana* proteomes. Frequencies were calculated using the set of 71 pairs of homologous sequences selected with the method described in [17] from the two complete proteomes (see material and methods). Amino acids were ranked owing to their frequencies in *P. falciparum*.

similar sequences, which was either directly used for matrices' calculations (automatic sampling) or with the restriction that the analogy of the protein function could be assessed by inspection of an expert curator (manual sampling). The advantage of the manual training set, described in [18], lies in its quality, but it is costly. Although one might question the quality of the automatic sample, the case study in the present paper proved that no major difference between the converging matrices calculated from the manual and the automatic training sets could be noticed. Because a model generalization and a pragmatic computation of matrices would benefit mostly from a fully automated method, we therefore detailed results obtained in that context.

3.3. Convergence of the non-symmetric matrices obtained after iterative computation

Blocks used to calculate matrices were filtered as described in the Methods section, according to a clustering percentage. Matrices were termed dirSp1Sp2, with Sp1 being the query species, and Sp2 the subject species. The matrices devoted to the comparison of *A. thaliana* and *P. falciparum* are therefore called dirAtPf matrices. We generated matrices corresponding to eleven block clustering percentages (dirAtPf100$_n$ for a clustering percentage = 100% and n iterations, dirAtPf90$_n$, ..., dirAtPf50$_n$). We analysed the evolution of the matrices obtained after each iterative round (summarized in Fig. 2). Fig. 3 shows the number of amino acids that are initially aligned by the generalist matrix BLOSUM 62 in the training set and after alignments using matrices computed after 1, 4, 7 and 9 iterations. From the very first matrix computation, one notice a decrease in the number of aligned amino acid, with a very rapid convergence, as early as $\sim$ 10 iterations. Interestingly, in the present result, convergence appears as a decrease in the number of amino acids 'detected' by the non-symmetrical matrices. That decrease would suggest that, in the case of well-assessed alignments, the non-symmetrical matrices are more specific. By contrast, although they converge to a close result, matrices computed with a manual training set exhibit an increase in the number of aligned amino acid along the training process. That increase would conversely suggest that in the case of hypothetical alignments,

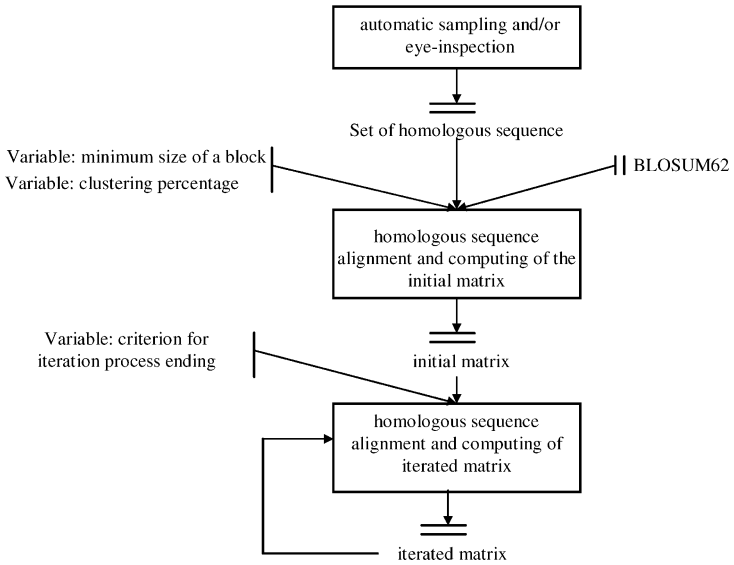

Fig. 2. dirSp1Sp2 matrices iterative computing Workflow. After sampling a set of pairs of homologous sequences between Species1 and Species2, these sequences were primarily aligned using BLOSUM 62 in order to determine conserved block. BLOSUM 62 is therefore used as a way to initially 'anchor' homologous regions. These blocks are considered only if they have both the required minimum size and a maximum given percentage of identity, named clustering percentage. The first deduced matrix, called *initial matrix*, is then used to iterate the process and lead to a sequence of dirSp1Sp2 matrices. A convergence criterion is applied on the sequence dirSp1Sp2_n so as to end the process.

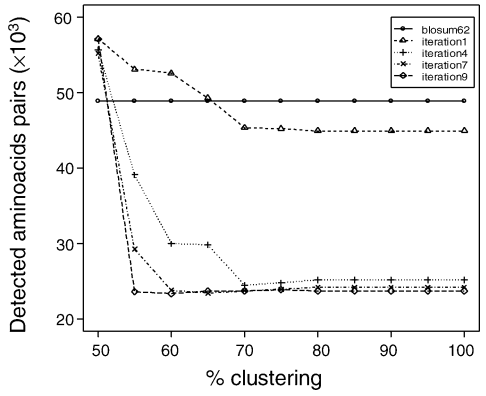

Fig. 3. Convergence of the matrices computing iterating process. Convergence of the iterating process was studied according to the number of detected aligned amino acids pairs. For all clustering percentage, convergence was also observed for the number of detected blocks and the value of the matrices (data not shown). Except for the 50% clustering percentage, convergence leads to a number of aligned amino acids pairs lower than that obtained with the BLOSUM 62 matrix. As described by Müller et al. [22], non-symmetric matrices lead to an increase of the discrimination power of the matrices, an expression of a more accurate mutual information values.

the non-symmetrical matrices would lead to a gain in sensitivity. Thus it appears that an apparent gain in selectivity and specificity would be obtained. To that extent, and as mentioned by Müller et al. [22], definition of a specificity/sensitivity gain when deriving substitution matrices is difficult to rigorously assess. The matrices obtained from automatic or manual samples exhibited identical trends for each s_{ij} terms: no opposite deviations were observed. For all clustering percentages, convergence was also observed for the number of detected blocks and the value of the matrices (data not shown).

3.4. *Asymmetry of the $\text{DirSp}_1\text{Sp}_2$ matrices: case study of DirAtPf*

We generated all the convergent dirAtPf matrices after a 10-iteration computation process dirAtPf100_{10}, dirAtPf90_{10}, …, dirAtPf50_{10}, indistinctly called dirAtPf100, dirAtPf90, …, dirAtPf50. All the matrices we obtained were non-symmetric, as shown in Fig. 4 for dirAtPf100. In detail, the sub-matrices $(\text{N}, \text{R}) \times (\text{N}, \text{R})$ giving the mutual information between all substitution available between Asparagines (N) and Arginines (R) stresses the different roles played by these two amino acids in the two proteomes,

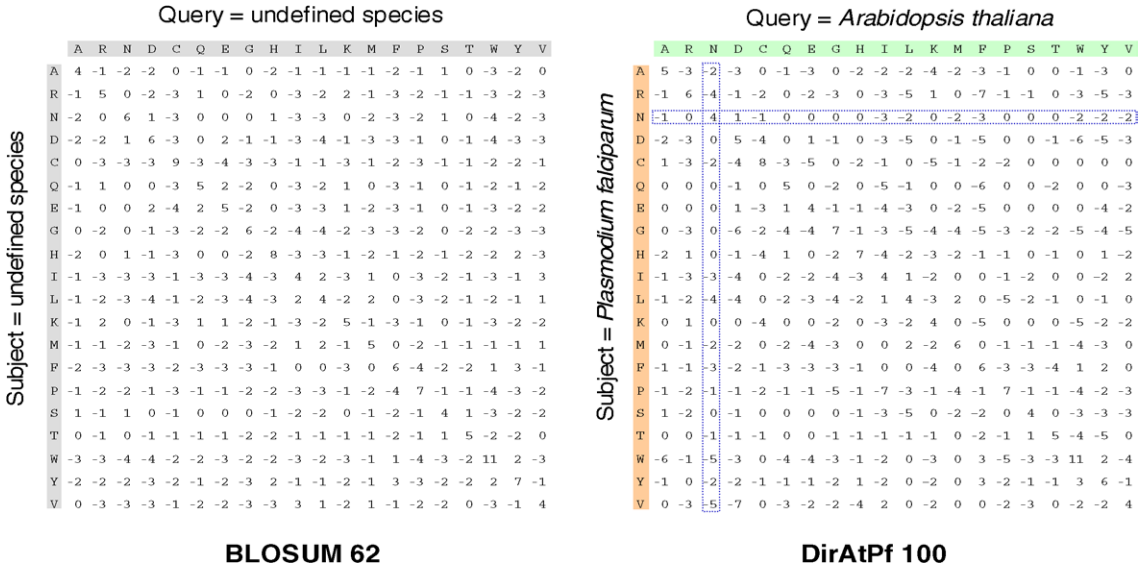

Fig. 4. BLOSUM 62 and dirAtPf100 substitution matrices.

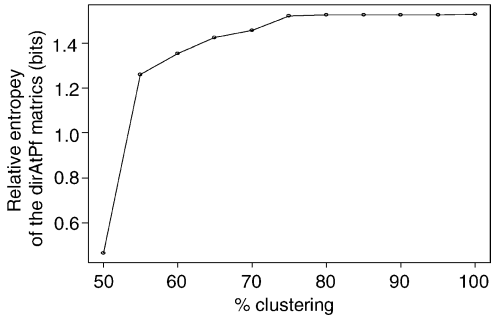

Fig. 5. Evolution of the relative Entropy as a function of the clustering percentage. As intuitively predicted by Müller et al. [22], the better definition of the mutual information between aligned amino acids leads to a higher relative entropy than this of BLOSUM 62 ($H \cong 0.69$).

as recorded by Singer and Hickey [20] and Bastien et al. [17].

3.5. *Analysis of the relative entropy H of the family of DirAtPf matrices*

As described earlier for the family of BLOSUM matrices [8], the relative entropy H derived from Eq. (2) decreases with the blocks clustering percentage (Fig. 5). Interestingly, relative entropy values in dirAtPf matrices (0.5–1.5 bits) are slightly higher than those of the BLOSUM or PAM matrices (0.2–1.2 bits) [7,8]. Following the theory of information, this higher relative entropy would suggest a higher sensitivity of dirAtPf matrices as compared to symmetrical matrices.

4. Conclusion

This article describes a method to compute a novel family of substitution matrices that are dedicated to the comparison of proteins, which amino acid composition deviates from the average distribution. They exhibit remarkable features such as (*i*) the possibility of computing reliable matrices from automatically selected pairs of similar sequences (automatic training sets), (*ii*) a rapidly convergent iterative process, and (*iii*) an increase in relative entropy. Still the selectivity/sensitivity gain is difficult to assess besides pragmatic use. Families of matrices for pairwise proteome comparisons including biased genomes such as that of *Plasmodium flaciparum* (AT rich) or *Chlamydomonas reinhardtii* (GC rich) are expected to be improved.

References

[1] F. Dardel, F. Képès, Bioinformatique: Génomique et post-génomique, Les Éditions de l'École polytechnique, Paris, 2002.
[2] M.S. Waterman, Introduction to Computational Biology, CRC Press, Boca Raton, FL, USA, 1995.

[3] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[4] J. Setubal, J. Meidanis, Introduction to Computational Molecular Biology, PWS Publishing Compagny, New York, 1997.
[5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[6] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[7] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, A model of evolutionary change in proteins, Atlas of protein sequence and structure 5 (1978) 345–352.
[8] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl Acad. Sci. USA 89 (1992) 10915–10919.
[9] S. Henikoff, J.G. Henikoff, Performance evaluation of amino acid substitution matrices, Proteins 17 (1993) 49–61.
[10] J.L. Risler, M.O. Delorme, H. Delacroix, A. Henaut, Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix, J. Mol. Biol. 204 (1988) 1019–1029.
[11] S. Karlin, S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc. Natl Acad. Sci. USA 87 (1990) 2264–2268.
[12] S.F. Altschul, Amino acid substitution matrices from an information theoretic perspective, J. Mol. Biol. 219 (1991) 555–565.
[13] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.
[14] Y.K. Yu, J.C. Wootton, S.F. Altschul, The compositional adjustment of amino acid substitution matrices, Proc. Natl Acad. Sci. USA (2003) 15688–15693.
[15] R.V.L. Hartley, Transmission of Information, Bell Syst. Tech. J. 3 (1928) 535–564.
[16] O. Bastien, J.-C. Aude, K. Métayer, S. Roy, J.-J. Codani, É. Maréchal, Method for automatic pairwise alignment of protein sequences from biased and non-biased genomes: generalized model for substitution matrices and theoretical significance of Z-value statistics, in: Proc. Eur. Conf. on Computational Biology, Paris, France, 2003, pp. 525–526.
[17] O. Bastien, J.-C. Aude, S. Roy, É. Maréchal, Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics, Bioinformatics 20 (2004) 534–537.
[18] O. Bastien, S. Lespinats, S. Roy, K. Metayer, B. Fertil, J.-J. Codani, É. Maréchal, Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using *Arabidopsis thaliana* as a reference, Gene 336 (2004) 163–173.
[19] A.K. Chamberlain, J.U. Bowie, Asymmetric amino acid compositions of transmembrane beta-strands, Protein Sci. 13 (2004) 2270–2274.
[20] G.A. Singer, D.A. Hickey, Nucleotide bias causes a genomewide bias in the amino acid composition of proteins, Mol. Biol. Evol. 17 (2000) 1581–1588.
[21] G.E. Tusnady, I. Simon, Principles governing amino acid composition of integral membrane proteins: application to topology prediction, J. Mol. Biol. 283 (1998) 489–506.
[22] T. Müller, S. Rahmann, M. Rehmsmeier, Non-symmetric score matrices and the detection of homologous transmembrane proteins, Bioinformatics 17 (Suppl. 1) (2001) S182–S189.
[23] J.-J. Codani, J.-P. Comet, J.-C. Aude, E. Glémet, A. Wozniak, J.-L. Risler, A. Hénaut, P.P. Slonimski, Automatic analysis of large-scale pairwize alignments of protein sequences, Methods Microbiol. 28 (1999) 229–244.
[24] S. Henikoff, J.G. Henikoff, Automated assembly of protein blocks for database searching, Nucleic Acids Res. 19 (1991) 6565–6572.
[25] A.J. Valleron, Introduction à la biostatistique, Masson, Paris, 1998.