DIMACS Technical Report 2008-16
December 2008

Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families

by

Valentina V. Sulimova
Tula State University
Lenin Ave. 92, 300600 Tula, Russia
vsulimova@yandex.ru

Vadim V. Mottl
Computing Center of the Russian Academy of Sciences
Vavilov St. 40, 119333 Moscow, Russia
vmottl@yandex.ru

Casimir A. Kulikowski
Deptartment of Computer Science
Rutgers University
New Brunswick, New Jersey 08903
kulikows@cs.rutgers.edu

Ilya B. Muchnik
DIMACS
Rutgers University
New Brunswick, New Jersey 08903
muchnikilya@yahoo.com

Abstract

Background: Almost all problems of protein analysis must inevitably be based on comparing the types of amino acids from which protein sequences are composed. Similarities between amino acids are most commonly based on two methods derived from very different approaches: the evolutionary based substitution matrixes of the PAM (Point Accepted Mutation) family, derived from phylogenetic trees, and the BLOSUM substitution matrixes which are statistically inferred from multiple alignments of groups of proteins which, according to their authors, S. and J. Henikoff, are essentially different from the PAM family of matrices.

Results: In this paper we prove that the statistical approach for computing substitution matrixes of the BLOSUM family can be explained in terms of the PAM evolutionary model. This means that both of these approaches are actually based on similar types of evolutionary models, and the main difference between them lies in the different initial data for estimating their unknown model parameters. We also show that all PAM substitution matrices can be represented as kernel functions in their mathematical structure, and lose their positive semi-definiteness only because of choice of final representation.

Conclusions: The fact that the PAM and BLOSUM substitution matrices are originally positive semidefinite, allows them to be easily used for constructing kernels over a set of proteins, so, without loss of biological meaning, these similarity measures can be applied without correction. Furthermore, any new substitution matrix will automatically be a kernel if, first, it is estimated by either the Dayhoff or Henikoff techniques and, second, the final representation proposed in the present research is adopted.

Contents

1 Introduction

One of the fundamental issues in bioinformatics is the problem of choosing a quantitative measure of similarity or dissimilarity between amino-acid sequences, which in turn has to be based on measuring the similarities between the amino acids that compose them.

Amino acid similarity is traditionally described by a 20x20 square matrix of pairwise similarity values denoted as a substitution matrix. For protein analysis these substitution matrices should be consistent with the best knowledge available about protein amino acid sequence evolution.

The first most commonly adopted similarity measure involves the family of Point Accepted Mutation (PAM) substitution matrices, introduced by Margaret Dayhoff [1], based on a probabilistic model of evolution. Parameters of this model are estimated from empirical data which generate phylogeny trees for families of closely related proteins.

The second, more popular family of substitution matrices, whose advantages have been confirmed in a number of practical cases, was introduced by Steven and Jorjia Henikoff and called BLOSUM (BLOcks SUbstitution Matrices) [2]. According to the authors these matrices differ significantly from matrices of the PAM family in that they are inferred from local multiple alignments within protein families which produce blocks of conservative fragments of amino acid sequences.

BLOSUM directly calculates frequencies of appearance of different amino acids at the same positions in an extracted block of similar fragments of sequences, requiring no knowledge of phylogeny but only the results of the alignment.

In this paper we show that the family of BLOSUM substitution matrices can be explained in terms of Dayhoff's evolutionary model as was done for PAM. From this analysis one can see that the Henikoffs' statistical approach, applied to their particular BLOCKs data gives exactly the same result which one would have got based on Dayhoff's model using an appropriate hierarchical structure knowledge.

Henikoff's work, together with our explanation of how the same result can be derived through application of Dayhoff's technique, gives a strong hint that it is possible and useful to calculate adaptive substitution matrixes which specify the substitution scale for a particular group of proteins. For instance, the COG-server [3] provides us with a very important classification if one is particularly interested in a specific group of clusters such that each is found in the same group of organisms. These COG-clusters taken together can be considered as a separate group characterized by a particular substitution matrix. The specificity arises from the particular set of organisms arising from having already decided on a group of organisms of interest. It is then very likely that reconstructing COG-clusters from this group (and only within this group) will allow us to build a more detailed clustering. Further, if one is interested in building a classification rule to distinguish pairs of protein functional families, ignoring all other protein families, so one can determine a specific substitution matrix for the two families. Determining such specific substitution matrices has been discussed over the last few years [4, 5] yet all have been so far restricted by heuristic-statistical assumptions. In showing that BLOSUM matrices can be derived in a way that satisfies the rigorous Dayhoff PAM model of evolution, this research can now be recast on a solid theoretical base.

In parallel with discussion about the power of Dayhoff's evolutionary model which is a fundamental base for constructing substitution matrices, we investigate some their algebraic properties. Specifically, we investigate the conditions of their derivation which guarantee that these matrices should be positive semi definite. This aspect we have analyzed in greatest detail, encouraged by the opportunity of building an automatic classifier for protein families by Support Vector Machine (SVM) methods [6, 7, 8]. These methods allow one to construct classifiers at any level of complexity if one is able to build a corresponding similarity function for pairs of objects which satisfies the

property that it is an inner product function in some linear vector space. Such functions are traditionally called kernel functions. From our perspective, it means that a similarity score for pairs of proteins has to satisfy a particular formal property. In the last seven years many bioinformatics researchers have devoted much time and effort to building corresponding kernels [9, 10, 11, 12, 13, 14, 15, 16]. These procedures on the whole are based on similarity functions for amino acids which serve as the principal foundation for constructing such a function for amino acid sequences with the same property of being positive semi-definite. Yet, substitution matrices of the PAM and BLOSUM families in their traditional representation have negative eigenvalues leading numerous publications to attempt correcting traditional substitution matrices, but this has resulted either in the loss of their biological meaning [9, 10] or in revised matrices whose positive semi-definiteness was not guaranteed [11].

The present paper gives a very clear picture of the conditions under which a similarity function between sets of amino acids satisfies the positive semi-definite property. It proves satisfying that those conditions which determine the positive semi-definite property for a matrix of substitutions is completely characterized by a few natural simple properties of Dayhoff's probabilistic evolutionary model.

2 Dayhoff's PAM model and its evolutionary assumptions

This section describes the basic probabilistic Dayhoff evolutionary model in an alternative, non-standard form which is most convenient for further analysis and comparison with the BLOSUM approach.

To measure the similarity or dissimilarity between amino-acid sequences requires that one adopt a representation for measuring the similarity of their constituent amino acids.

Let A be the finite set of 20 natural amino acids (AAs) such that $A=\left\{\alpha^{(1)},...,\alpha^{(20)}\right\}$. A measure of similarity between two amino acids $\alpha^i,\alpha^j \in A$ is usually understood to be their predisposition to having mutually exchanged by mutation from one protein sequence to another over evolutionary time and the simplest and most commonly applied probabilistic model of mutation is Dayhoff's PAM.

The key hypothesis underlying this model is that evolutionary changes in the amino-acid sequence of a protein are the result of random independent mutations at separate points of the AA chain, and that the mutations observed today, were those "accepted" as result of subsequent natural selection processes.

The PAM model represents these predispositions of AAs towards mutual mutative transformations as a square matrix of conditional probabilities

$$\mathbf{\Psi} = (\psi^{ij}, i, j = 1,...,n),\ \psi^{ij} = \psi(\alpha^j \mid \alpha^i),\ \alpha^i,\alpha^j \in A,\ n=20, \tag{1}$$

which are interpreted as the probabilities that at the next step of evolution amino acid i will transform into amino acid j. So, $\sum_{\alpha^j \in A} \psi(\alpha^j \mid \alpha^i) = 1$ for each $\alpha^i \in A$.

The main mathematical notion used in Dayhoff's model is that of a Markov chain h_s, $s = 1,2,3,...$ of evolution of an amino acid, which completely defined by the matrix of probabilities from (1), and applied independently at every separate point of an AA chain.

It is further assumed, that this Markov chain represents an ergodic random process, which is characterized by a final probability distribution $\xi(\alpha^j)$:

$$\sum_{\alpha^i \in A} \xi(\alpha^i)\psi(\alpha^j \mid \alpha^i) = \xi(\alpha^j).$$

This is the formalization of an assumption that evolutionary process started very long ago, with random processes of mutations setting in early, and as result do not depend on unknown (and empirically unknowable) initial probabilities of states.

The second fundamental supposition about PAM model is the assumption that its Markov chain is reversible, i.e. the condition:

$$\xi(\alpha^i)\psi(\alpha^j \mid \alpha^i) = \xi(\alpha^j)\psi(\alpha^i \mid \alpha^j).$$

holds true. This implies that it is impossible to determine in the process of observation which of two amino acids is an ancestor and which is a descendant. While involving simplification from a biological perspective, PAM models have proven remarkably good at predicting observed evolutionary transformations in sequences of rapidly reproducing organisms and enable mathematical tractability of derivations of probabilities of transformations from one AA chain to another.

3 Similarity measures between amino acids on the basis of Dayhoff's model and its interpretation

This section gives a more mathematically rigorous introduction of the similarity functions for pairs of amino acids based on Dayhoff's evolutionary model and their possible probabilistic interpretations.

Let A, as before, be the finite set of amino acids $A = \left\{\alpha^{(1)},...,\alpha^{(20)}\right\}$, which are states of ergodic and reversible Markov chain with a conditional probability density $\psi^{ij} = \psi(\alpha^j \mid \alpha^i)$ and final probability distribution $\xi(\alpha^j)$, $\alpha^j \in A$.

Let us define also a two-step random transformation of amino acids $\alpha^i \to \alpha^k \to \alpha^j$:

$$\psi_{[2]}(\alpha^i \mid \alpha^j) = \sum_{\vartheta \in A} \psi(\alpha^i \mid \vartheta)\psi(\vartheta \mid \alpha^j),$$

which defines a new Markov chain, thinned out in comparison with initial one, defined by the original transformation $\psi(\alpha^j \mid \alpha^i)$, which in contrast can be called a one-step transformation $\psi_{[1]}(\alpha^j \mid \alpha^i)$.

Theorem 1. *The Markov random process produced by the two-step random transformation remains ergodic and reversible with the same final distribution* $\xi(\alpha)$.

Proof.

First we show that the Markov process, defined by the two-step transformation $\psi_{[2]}^{ij}$ is ergodic with the same final distribution $\xi(\alpha)$.

As it can be represented as $\psi_{[2]}^{ij} = \sum_{l=1}^{n} \psi_{[1]}^{il}\psi_{[1]}^{lj}$,

$\sum_{i=1}^{n} \xi^i \psi_{[2]}^{ij} = \sum_{i=1}^{n} \xi^i \sum_{l=1}^{n} \psi_{[1]}^{il}\psi_{[1]}^{lj} = \sum_{l=1}^{n} \left(\sum_{i=1}^{n} \xi^i \psi_{[1]}^{il} \right) \psi_{[1]}^{lj} = \sum_{l=1}^{n} \xi^l \psi_{[1]}^{lj} = \xi^j$. Thus, ergodicity is proved.

The proof of the reversibility of the two-step transformation follows from the reversibility of the original one-step transformation $\xi^i \psi^{ij} = \xi^j \psi^{ji}$:

$$\xi^i \psi_{[2]}^{ij} = \xi^i \sum_{l=1}^{n} \psi^{il}\psi^{lj} = \sum_{l=1}^{n} (\xi^i \psi^{il})\psi^{lj} = \sum_{l=1}^{n} (\xi^l \psi^{li})\psi^{lj} = \sum_{l=1}^{n} (\xi^l \psi^{lj})\psi^{li}$$
$$= \sum_{l=1}^{n} (\xi^j \psi^{jl})\psi^{li} = \xi^j \sum_{l=1}^{n} \psi^{jl}\psi^{li} = \xi^j \psi_{[2]}^{ji}.$$

This completes the proof.

It is natural to evaluate the similarity of two amino acids by computing the probability that they resulted from two independent branches of evolution followed by descendents of one and the same unknown amino acid

$$K_{[2]}(\alpha^i, \alpha^j) = \sum_{k=1}^{n} \xi(\alpha^k)\psi_{[1]}(\alpha^i \mid \alpha^k)\psi_{[1]}(\alpha^j \mid \alpha^k). \tag{2}$$

However it is more convenient to represent this similarity measure in an alternative way, based on the property of reversibility of the Markov chain:

$$\begin{aligned} K_{[2]}(\alpha^i, \alpha^j) = \sum\nolimits_{k=1}^{n} \underbrace{\xi(\alpha^k)\psi_{[1]}(\alpha^i \mid \alpha^k)}_{\xi(\alpha^i)\psi_{[1]}(\alpha^k \mid \alpha^i)} \psi_{[1]}(\alpha^j \mid \alpha^k) = \\ \xi(\alpha^i) \sum\nolimits_{k=1}^{n} \underbrace{\psi_{[1]}(\alpha^k \mid \alpha^i)\psi_{[1]}(\alpha^j \mid \alpha^k)}_{\psi_{[2]}(\alpha^j \mid \alpha^i)} = \xi(\alpha^i)\psi_{[2]}(\alpha^j \mid \alpha^i), \end{aligned} \tag{3}$$

Such a representation allows us to interpret $K_{[2]}(\alpha^i, \alpha^j)$ differently. It can be considered as the probability that two randomly chosen immediately adjacent positions in this Markov chain with transition probabilities $\psi_{[2]}(\alpha^j \mid \alpha^i)$ are occupied by two preset states (α^i, α^j).

Besides, in a number of cases it is very convenient to use the similarity measure (3), normalized by the final probabilities of the amino acids being compared:

$$\bar{K}_{[2]}(\alpha^i, \alpha^j) = \frac{K_{[2]}(\alpha^i, \alpha^j)}{\xi(\alpha^j)\xi(\alpha^i)} = \frac{\xi(\alpha^i)\psi_{[1]}(\alpha^j \mid \alpha^i)}{\xi(\alpha^j)\xi(\alpha^i)} = \frac{\psi_{[1]}(\alpha^j \mid \alpha^i)}{\xi(\alpha^j)}. \tag{4}$$

4 Dayhoff's approach to estimating unknown parameters of the PAM model

This section presents Dayhoff's method for reconstructing the similarity functions from empirical data (substitution matrices PAM).

For inferring an estimate $\hat{\mathbf{\Psi}} = (\hat{\psi}^{ij}, i, j = 1,...,20)$ of the unknown matrix of random point mutations $\mathbf{\Psi} = (\psi^{ij}, i, j = 1,...,20)$ M. Dayhoff studied 34 protein "superfamilies" of closely related proteins (more than 85% identical to each another), grouped into 71 phylogenetic trees.

From these data 1572 accepted point mutations were observed empirically, and a symmetric matrix $\mathbf{C} = (C^{ij}, i, j = 1,...,20)$ of frequencies with which amino acid α^i is replaced by amino acid α^j was derived, together with a vector $\mathbf{m} = (m^i, i = 1,...,20)$ of so-called relative mutabilities of amino acids $\alpha^i, i = 1,...,20$ – the probabilities that each amino acid will change in a given small evolutionary interval.

An estimate of a vector of occurrence frequencies of amino acids $\hat{\boldsymbol{\xi}} = (\hat{\xi}^{(i)}, i = 1,...,20)$ was also computed by Dayhoff from the accepted point mutation data.

The matrix $\hat{\boldsymbol{\Psi}} = (\hat{\psi}^{ij}, i, j = 1,...,20)$ can be calculated as

$$\hat{\psi}^{ij} = \frac{\lambda m^j C^{ij}}{\sum_i C^{ij}} \text{ for } i \neq j \text{ and } \hat{\psi}^{jj} = 1 - \lambda m^j \text{ for } i = j,$$

where λ is a normalization parameter, which is equal for each column of the matrix $\hat{\boldsymbol{\Psi}}$ and defined in such a way that

$$\sum_{i=1}^{20} \hat{\xi}^{(i)} (1 - \hat{\psi}^{ii}) = 0.01, \tag{5}$$

i.e. that, in average, only one amino acid among 100 randomly chosen amino acids would change. The value $(1 - \hat{\psi}^{ii})$ is proportional to the mutability of the respective amino acid.

The conditional mutation probabilities $\hat{\boldsymbol{\Psi}}$ satisfying (5) are said to define what is denoted as an evolutionary distance of 1 PAM. The most widely used evolutionary distance is that of 250 PAM associated with the 250th degree of the 1 PAM matrix: $\hat{\boldsymbol{\Psi}}^{250} = \underbrace{\hat{\boldsymbol{\Psi}} \times ... \times \hat{\boldsymbol{\Psi}}}_{250}$.

And, if a 1 PAM matrix is multiplied by itself an infinite number of times all columns of the resulting matrix will be equal to one another and to the vector of occurrence frequencies of the amino acids $\hat{\boldsymbol{\xi}} = (\hat{\xi}^{(i)}, i = 1,...,20)$, which in terms of Dayhoff's model of evolution is the vector of final probabilities.

The commonly adopted Dayhoff substitution matrices are computed from $\hat{\boldsymbol{\Psi}}^m$ and $\hat{\boldsymbol{\xi}}$ as given by the rule:

$$d_{[m]}^{ij} = 10 \log_{10} \pi_{[m]}^{ij}, \quad \pi_{[m]}^{ij} = \frac{\hat{\psi}_{[m]}^{ij}}{\hat{\xi}^j} \tag{6}$$

and then rounded to the nearest integer value.

It should be noticed, that the value under logarithm has the same structure as the similarity measure (4).

5 Family of BLOSUM substitution matrices and statistical basis of the Henikoff Derivation

This section describes the statistical method for determining, from empirical data, the BLOSUM (BLOcks SUbstitution Matrices) proposed by Steven and Jorja Henikoffs in 1992 [2] and widely used for protein sequence alignments.

In contrast to the Dayhoff matrices, the BLOSUM ones are inferred not from phylogenetic trees but from local multiple alignments containing much more diverse protein families (starting from a level of 45% sequence similarity). Such multiple alignments produce blocks of evolutionarily conservative amino acid sequence fragments without gaps.

Each element of a BLOSUM matrix is computed as the log-odds ratio between the observed probabilities of occurrence of amino acid pairs among the blocks and those expected by chance.

Let us to consider all blocks from the Henikoff BLOCKS data base as one consolidated block consisting of Q columns, involving of N_k amino acids, $k = 1,...,Q$.

For computing the square matrix of observed probabilities $\mathbf{p}=(p^{ij},i,j=1,...,20)$ of occurrence of amino acid pairs in accordance with the Henikoff technique, one first counts pair frequencies for each non-ordered pair of amino acids α^i and α^j, for each column k of the consolidated block:

$$M_k^{ij}=\begin{cases}N_k^i(N_k^i-1)/2, i=j,\\ N_k^iN_k^j, i\neq j,\end{cases} \tag{7}$$

where N_k^i и N_k^j are numbers of positions in the column k occupied by amino acids α^i or α^j respectively.

For all columns, the pair frequencies for each non-ordered pair (α^i,α^j) is

$$M^{ij}=\sum_{k=1}^{Q}M_k^{ij} \tag{8}$$

and the total number of all ordered pairs in all columns is

$$M=\sum_{k=1}^{Q}M_k=\sum_{k=1}^{Q}N_k(N_k-1)/2\,. \tag{9}$$

Probabilities p^{ij} are counted as

$$p^{ij}=\frac{M^{ij}}{M}\,. \tag{10}$$

The next step is to compute the expected probability of occurrence of amino acid α^i in the (i,j) pair:

$$q^i=p^{ii}+\frac{1}{2}\sum_{i\neq j}p^{ij}\,. \tag{11}$$

The resulting BLOSUM matrix of amino acid similarities $\mathbf{b}=(b^{ij},i,j=1,...,20)$ is defined as

$$b^{ii}=2\log_2\left(\frac{p^{ii}}{q^iq^i}\right)\text{ for }i=j\text{ and }b^{ij}=2\log_2\left(\frac{p^{ij}}{2q^iq^j}\right)\text{ for }i\neq j\,, \tag{12}$$

and rounded for nearest integer value.

There is a series of BLOSUM substitution matrices which differ from each other by the level of identity required of the proteins in the multiply aligned families. So, the matrices BLOSUM 45, BLOSUM 50, BLOSUM 62 and BLOSUM 80 are computed from protein families with levels of similarity of 45%, 50%, 62% and 80%, respectively.

6 Dayhoff's model of evolution for BLOSUM

The above derivation of BLOSUM substitution matrices comes from applying statistical methods to conserved amino acid blocks, and appears completely different in origin from the PAM substitution matrices.

In this section we will show that BLOSUM expresses the same probabilistic model of amino acid mutations as PAM, which is based on the notion of an ergodic and reversible Markov chain defined by the matrix $\mathbf{\Psi}=(\psi^{ij},i,j=1,...,20)$ of conditional probabilities of transformations (1).

Let each column position in the consolidated block $\{\alpha_k,k=1,...,Q\}$ be associated with the hypothesis that all amino acids in it are produced by the same unknown amino acid ϑ (one-step ancestor) as result of independent random transformations with unknown conditional probabilities

$\psi(\alpha \mid \vartheta)$ resulting in an ergodic and reversible Markov chain with final distribution $\xi(\alpha)$. And further we can assume that an ancestor for each column k of the block is chosen independently in accordance with this distribution $\xi(\alpha)$.

The notion of an evolutionary step, which is so important in the PAM framework does not apply here, but let us assume, for definiteness sake, that these conditional probabilities correspond to one step: $\psi(\alpha \mid \vartheta) = \psi_{[1]}(\alpha \mid \vartheta)$.

Let also $\mathbf{\Psi}_{[2]} = (\psi_{[2]}^{ij}, i, j = 1,...,n)$ be a matrix of two-step conditional probabilities, such as

$$\psi_{[2]}^{ij} = \psi_{[2]}(\alpha^j \mid \alpha^i) = \sum_{l=1}^{n} \psi^{il}\psi^{lj},$$

which in accordance with theorem 1 defines an ergodic reversible Markov process with final distribution $\xi(\alpha)$.

Then, in accordance with (3), the observed probability of occurrence of amino acid pair (α^i, α^j) can be expressed as

$$\overline{p}^{ij} = \overline{p}(\alpha^i, \alpha^j) = \xi^i \psi_{[2]}^{ij} = \xi^j \psi_{[2]}^{ji}. \tag{13}$$

Theorem 2. *Statistic (11) proposed by the Henikoffs is an unbiased estimate of* ξ^i.

Proof.

First, it should be noticed, that the statistic (11) can be represented in more convenient form:

$$q^i = p^{ii} + \frac{1}{2}\sum_{\substack{j=1,\\ j\neq i}}^{20} p^{ij} = \frac{M^{ii}}{M} + \frac{1}{2}\cdot\frac{\sum_{i\neq j} M^{ij}}{M} = \frac{\frac{1}{2}\sum_{k=1}^{Q} N_k^i(N_k^i-1)}{\frac{1}{2}\sum_{k=1}^{Q} N_k(N_k-1)} + \frac{1}{2}\cdot\frac{\sum_{j=1,j\neq i}^{20}\sum_{k=1}^{Q} N_k^i N_k^j}{\frac{1}{2}\sum_{k=1}^{Q} N_k(N_k-1)} =$$

$$\frac{\sum_{k=1}^{Q} N_k^i(N_k^i-1)}{\sum_{k=1}^{Q} N_k(N_k-1)} + \frac{\sum_{j=1,j\neq i}^{20}\sum_{k=1}^{Q} N_k^i N_k^j}{\sum_{k=1}^{Q} N_k(N_k-1)} = \frac{\sum_{k=1}^{Q} N_k^i(N_k^i-1) + \sum_{k=1}^{Q}\sum_{j=1,j\neq i}^{20} N_k^i N_k^j}{\sum_{k=1}^{Q} N_k(N_k-1)} = \frac{\sum_{k=1}^{Q} N_k^i(N_k-1)}{\sum_{k=1}^{Q} N_k(N_k-1)} = \frac{\sum_{k=1}^{Q} N_k^i}{\sum_{k=1}^{Q} N_k} = \frac{N^i}{N}.$$

So,

$$q^i = N^i / N. \tag{14}$$

Let's show, that $q^i = N^i/N$ is a maximum likelihood estimate of the final probability ξ^i.

In accordance with the proposed model, the chance variable N_k^i is the number of occurrence of the event, which has probability ξ^i, in N_k independent tests. This chance variable is distributed according to a Binomial distribution:

$$\eta_k(N_k^i \mid \xi^i) = C_{N_k}^{N_k^i} (\xi^i)^{N_k^i} (1-\xi^i)^{N_k - N_k^i}.$$

Random variables $(N_k^i, k = 1,...,Q)$ are independent in accordance with the accepted model:

$$\eta(N_k^i, k = 1,...,Q \mid \xi^i) = \prod_{k=1}^{Q} \eta_k(N_k^i \mid \xi^i) = \prod_{k=1}^{Q} C_{N_k}^{N_k^i} (\xi^i)^{N_k^i} (1-\xi^i)^{N_k - N_k^i}. \tag{15}$$

For observed values $(N_k^i, k=1,...,Q)$ the distribution (15) is the likelihood function relative to an unknown value of the probability ξ^i. The maximum likelihood estimate of this probability is defined by the expression:

$$\hat{\xi}^i = \arg\max_{\xi^i} \log \eta(N_k^i, k=1,...,Q \mid \xi^i) = \arg\max_{\xi^i} \left[\log \prod_{k=1}^{Q} C_{N_k}^{N_k^i} + \sum_{k=1}^{Q} \left(N_k^i \log \xi^i + (N^k - N_k^i) \log(1-\xi^i) \right) \right]. \quad (16)$$

The maximum of this likelihood function is:

$$\frac{d}{d\xi^i} \log \eta_k(N_k^i, k=1,...,Q \mid \xi^i) = \sum_{k=1}^{Q} \left(\frac{N_k^i}{\xi^i} - \frac{N^k - N_k^i}{1-\xi^i} \right) = \sum_{k=1}^{Q} \left(\frac{(1-\xi^i)N_k^i - \xi^i(N^k - N_k^i)}{\xi^i(1-\xi^i)} \right) =$$

$$\frac{1}{\xi^i(1-\xi^i)} \left[\left(\sum_{k=1}^{Q} N_k^i \right)(1-\xi^i) - \left(\sum_{k=1}^{Q} N_k \right)\xi^i + \left(\sum_{k=1}^{Q} N_k^i \right)\xi^i \right] =$$

$$\frac{1}{\xi^i(1-\xi^i)} \left[\sum_{k=1}^{Q} N_k^i - \cancel{\left(\sum_{k=1}^{Q} N_k^i \right)\xi^i} - \left(\sum_{k=1}^{Q} N_k \right)\xi^i + \cancel{\left(\sum_{k=1}^{Q} N_k^i \right)\xi^i} \right] =$$

$$\frac{1}{\xi^i(1-\xi^i)} \left[\sum_{k=1}^{Q} N_k^i - \left(\sum_{k=1}^{Q} N_k \right)\xi^i \right] = 0,$$

i.e.

$$\sum_{k=1}^{Q} N_k^i - \left(\sum_{k=1}^{Q} N_k \right)\xi^i = 0 .$$

So, the maximum likelihood estimate for ξ^i is:

$$\hat{\xi}^i = \frac{\sum_{k=1}^{Q} N_k^i}{\sum_{k=1}^{Q} N_k} = \frac{N^i}{N} . \quad (17)$$

Comparing to (14), we see that the estimate (17) is identical to the statistic (11), proposed by the Henikoffs. Furthermore, this estimate is unbiased:

$$E\left(\frac{N^i}{N} \right) = \frac{1}{N} E\left(\sum_{k=1}^{Q} N_k^i \mid \xi^i \right) = \frac{1}{N} \sum_{k=1}^{Q} E(N_k^i \mid \xi^i) .$$

Here $E(N_k^i \mid \xi^i)$ is the average of the distribution of the chance variable under a Bernoulli distribution, i.e. $E(N_k^i \mid \xi^i) = N_k \xi^i$. So,

$$E(\hat{\xi}^i) = E\left(\frac{N^i}{N} \right) = \frac{\sum_{k=1}^{Q} N_k}{\sum_{k=1}^{Q} N_k} \xi^i = \xi^i$$

So, the Henikoffs' statistic (11) is an unbiased maximum likelihood estimate of ξ^i:

$$q^i = \hat{\xi}^i = N^i / N .$$

The proof is complete.

Theorem 3. *Statistic (10)* M^{ij}/M *for* $i=j$ *and* $M^{ij}/2M$ *for* $i\neq j$ *is an unbiased estimate of* $\overline{p}^{ij}=\overline{p}(\alpha^i,\alpha^j)$ *:*

$$\hat{\overline{p}}^{ij}=\begin{cases}M^{ij}/M\,,\ i=j,\\ M^{ij}/2M\,,\ i\neq j\end{cases},\ E(\hat{\overline{p}}^{ij}\mid\boldsymbol{\xi},\boldsymbol{\Psi})=\overline{p}^{ij}\,, \tag{18}$$

Proof.

Let us consider the two cases $i=j$ and $i\neq j$ separately.

Let $i=j$.

$$\frac{M^{ii}}{M}=\frac{\frac{1}{2}\sum_{k=1}^{Q}N_k^i(N_k^i-1)}{\frac{1}{2}\sum_{k=1}^{Q}N_k(N_k-1)}=\frac{\sum_{k=1}^{Q}N_k^i(N_k^i-1)}{\sum_{k=1}^{Q}N_k(N_k-1)}.$$

The value of the random product $N_k^i(N_k^i-1)$ in the k-th column of the consolidated block is associated with n hypothesis about the random choice of amino acid $a_k\in A=\{\alpha^i,i=1,...,n\}$ in accordance with the probability distribution $\boldsymbol{\xi}=(\xi^i,i=1,...,n)$, therefore

$$\begin{aligned}&E\left(N_k^i(N_k^i-1)\mid\boldsymbol{\xi},\boldsymbol{\Psi}\right))=\sum_{l=1}^{n}\xi^l E\left(N_k^i(N_k^i-1)\mid a_k=\alpha^l,\boldsymbol{\psi}^l\right)=\\ &\sum_{l=1}^{n}\xi^l\left[E\left((N_k^i)^2\mid a_k=\alpha^l,\boldsymbol{\psi}^l\right)-E(N_k^i\mid a_k=\alpha^l,\boldsymbol{\psi}^l)\right].\end{aligned} \tag{19}$$

The chance variable N_k^i is distributed under the Bernoulli Law with the probability ψ^{li} of occurrence of the event under consideration (amino acid α^i in each of N_k independent tests).

The mean of the Bernoulli distribution is defined as the probability of occurrence of an event in a separate test (in this case ψ^{li}) multiplied by the number of tests (in this case N_k), i.e.

$$E(N_k^i\mid a_k=\alpha^l,\boldsymbol{\psi}^l)=N_k\psi^{li}. \tag{20}$$

The deviation of this distribution is the same

$$E\left[\left(N_k^i-E(N_k^i\mid a_k=\alpha^l,\boldsymbol{\psi}^l)\right)^2\mid a_k=\alpha^l,\boldsymbol{\psi}^l\right]=N_k\psi^{li}(1-\psi^{li}). \tag{21}$$

The average of any chance variable squired is the sum of its deviation and its squared average, i.e.

$$\begin{aligned}&E\left((N_k^i)^2\mid a_k=\alpha^l,\boldsymbol{\psi}^l\right)=N_k\psi^{li}(1-\psi^{li})+(N_k\psi^{li})^2=N_k\psi^{li}\{1-\psi^{li}+N_k\psi^{li}\}=\\ &N_k\psi^{li}\{1+(N_k-1)\psi^{li}\}=N_k(N_k-1)\psi^{li}\psi^{li}+N_k\psi^{li}.\end{aligned} \tag{22}$$

So,

$$E\left(N_k^i(N_k^i-1)\mid\boldsymbol{\xi},\boldsymbol{\Psi}\right)=N_k(N_k-1)\sum_{l=1}^{n}\xi^l\psi^{li}\psi^{li}=N_k(N_k-1)\overline{p}^{ii}.$$

Therefore,

$$E(\hat{\overline{p}}^{ii}\mid\boldsymbol{\xi},\boldsymbol{\Psi})=\frac{\sum_{k=1}^{Q}N_k(N_k-1)}{\sum_{k=1}^{Q}N_k(N_k-1)}\overline{p}^{ii}=\overline{p}^{ii}.$$

Let us consider now the case $i \neq j$.

Here there are n hypotheses about the random choice of an ancestor amino acid $a_k \in A$ for generating all amino acids of the k-th column of the consolidated block, yielding the equality

$$E(N_k^i N_k^j \mid \boldsymbol{\xi}, \boldsymbol{\Psi}) = \sum_{l=1}^{n} \xi^l E(N_k^i N_k^j \mid a_k = \alpha^l, \boldsymbol{\psi}^l).$$

The chance variable N_k^i can take $N_k + 1$ values $N_k^i = s$, $s = 0,1,2,...,N_k$ with probabilities $P(N_k^i = s)$:

$$E(N_k^i N_k^j \mid a_k = \alpha^l, \boldsymbol{\psi}^l) = \sum_{s=0}^{N_k} P(N_k^i = s) s E(N_k^j \mid a_k = \alpha^l, N_k^i = s, \boldsymbol{\psi}^l).$$

Here the conditional distribution of the chance variable N_k^j is Bernoulli distributed with a probability of occurrence of the event (amino acid α^j) in each of $(N_k - s)$ independent tests, which is equal to $\psi^{lj} / (1 - \psi^{li})$. The average of Bernoulli distribution is:

$$E(N_k^j \mid a_k = \alpha^l, N_k^i = s, \boldsymbol{\psi}^l) = \frac{\psi^{lj}}{1 - \psi^{li}} (N_k - s),$$

consequently,

$$E(N_k^i N_k^j \mid a_k = \alpha^l, \boldsymbol{\psi}^l) = \frac{\psi^{lj}}{1 - \psi^{li}} \sum_{s=0}^{N_k} P(N_k^i = s \mid a_k = \alpha^l, \boldsymbol{\psi}^l) s (N_k - s) =$$
$$\frac{\psi^{lj}}{1 - \psi^{li}} \left\{ N_k \sum_{s=0}^{N_k} P(N_k^i = s \mid a_k = \alpha^l, \boldsymbol{\psi}^l) s - \sum_{s=0}^{N_k} P(N_k^i = s \mid a_k = \alpha^l, \boldsymbol{\psi}^l) s^2 \right\}.$$

Here the first sum in braces is the average of the chance variable N_k^i and is equal to $\psi^{li} N_k$. The second sum is the average of squared variable N_k^i:

$$E(N_k^i N_k^j \mid a_k = \alpha^l, \boldsymbol{\psi}^l) = \frac{\psi^{lj}}{1 - \psi^{li}} \left\{ N_k (\psi^{li} N_k) - E\left((N_k^i)^2 \mid a_k = \alpha^l, \boldsymbol{\psi}^l \right) \right\},$$

i.e.

$$E(N_k^i N_k^j \mid a_k = \alpha^l, \boldsymbol{\psi}^l) = \frac{\psi^{lj}}{1 - \psi^{li}} \left\{ (N_k)^2 \psi^{li} - N_k \psi^{li} (1 - \psi^{li}) - (N_k \psi^{li})^2 \right\} =$$
$$\frac{\psi^{lj}}{1 - \psi^{li}} N_k \psi^{li} \left\{ N_k - (1 - \psi^{li}) - N_k \psi^{li} \right\} = \frac{\psi^{li} \psi^{lj}}{1 - \psi^{li}} N_k \left\{ N_k (1 - \psi^{li}) - (1 - \psi^{li}) \right\} =$$
$$\frac{\psi^{li} \psi^{lj}}{1 - \psi^{li}} N_k (N_k - 1)(1 - \psi^{li}) = N_k (N_k - 1) \psi^{li} \psi^{lj}.$$

The full distribution over the set of values of the random variable $N_i^k N_j^k$ is the linear combination

$$E(N_k^i N_k^j) = \sum_{l=1}^{n} \xi^l E(N_k^i N_k^j \mid a_k = \alpha^l, \boldsymbol{\psi}^l) = N_k (N_k - 1) \sum_{l=1}^{n} \xi^l \psi^{li} \psi^{lj}.$$

It follows that

$$E(\hat{\overline{p}}^{ij}) = \frac{1}{2}E\left(\frac{M^{ij}}{M}\right) = \frac{1}{2}\cdot\frac{\sum_{k=1}^{Q}E(N_i^k N_j^k)}{\frac{1}{2}\sum_{k=1}^{Q}N^k(N^k-1)} = \frac{\sum_{k=1}^{Q}N^k(N^k-1)}{\sum_{k=1}^{Q}N^k(N^k-1)}\sum_{l=1}^{n}\xi_l\psi_{li}\psi_{lj} =$$

$$\frac{\sum_{k=1}^{Q}N^k(N^k-1)}{\sum_{k=1}^{Q}N^k(N^k-1)}\overline{p}^{ij} = \overline{p}^{ij}.$$

So, for any i and j

$$E(\hat{\overline{p}}^{ij} \mid \boldsymbol{\xi}, \boldsymbol{\Psi}) = \overline{p}^{ij}.$$

The proof is complete.

According to theorems 2 and 3, BLOSUM represents the same probabilistic model of amino acid mutations as PAM, based on the notion of an ergodic and reversible Markov chain so that each element of the BLOSUM matrix (12) can be expressed in corresponding terms as:

$$b^{ii} = 2\log_2\left(\frac{p^{ij}}{q^i q^j}\right) = 2\log_2\left(\frac{\hat{\overline{p}}^{ii}}{\hat{\xi}^i\hat{\xi}^i}\right) \text{ for } i=j, \text{ and } b^{ij} = 2\log_2\left(\frac{p^{ij}}{2q^i q^j}\right) = 2\log_2\left(\frac{\hat{\overline{p}}^{ij}}{\hat{\xi}^i\hat{\xi}^j}\right) \text{ for } i\neq j,$$

or following (13):

$$b^{ij} = 2\log_2\left(\frac{p^{ij}}{2q^i q^j}\right) = 2\log_2\left(\frac{\xi^i\psi_{[2]}^{ij}}{\xi^i\xi^j}\right) = 2\log_2\left(\frac{\psi_{[2]}^{ij}}{\xi^j}\right). \tag{23}$$

7 Mathematical properties of substitution matrices of the PAM and BLOSUM families: kernels in the set of amino acids on the basis of Dayhoff's model of evolution

This section is completely focused on the analysis of the semi-positive definite properties for substitution matrices and their relation with the original Dayhoff model of point evolution processes.

In a number of problems of protein analysis it is useful that the similarity measure over the set of amino acids should possess the properties of an inner product. Functions of this kind are called kernel functions. For the finite set of amino acids, a kernel function is any real-valued two-argument function forming a positive semi-definite matrix. Each kernel function, defined over the set of amino acids embeds the set of amino acids in some hypothetical linear space and lets us to consider each amino acid as a point in it.

However, it should be noted that the substitution matrices of the PAM and BLOSUM families in their traditional log-odds representation (6) and (23) have negative eigenvalues, i.e. they are not valid kernels.

In this section we prove an interesting fact: Dayhoff's evolutionary model based on the notion of an ergodic and reversible Markov chain, is sufficient for constructing a valid kernel over the set of amino acids. From this it follows that the PAM and BLOSUM families are kernel functions by mathematical structure, but have lost their natural positive definiteness simply because of an unsuitable choice of final representation.

It should be noted that the similarity measure (2) can be easily represented as an inner product

$$K_{[2]}(\alpha^i,\alpha^j) = \sum_{k=1}^{n}\left(\sqrt{\xi(\alpha^k)}\psi_{[1]}(\alpha^i \mid \alpha^k)\right)\left(\sqrt{\xi(\alpha^k)}\psi_{[1]}(\alpha^j \mid \alpha^k)\right) = \sum_{k=1}^{n}x_{ik}x_{jk} = \mathbf{x}_i^T\mathbf{x}_j,$$

where $\mathbf{x}_i = \left(\sqrt{\xi(\alpha^k)}\psi_{[1]}(\alpha^i \mid \alpha^k), k = 1,...,n\right) \in R^n$ can be considered as a feature vector of the i-th amino acid.

So, the similarity measure (2) and its representation through two-step random transformation are kernels.

It is easy to show that the normalized similarity measure (4) is also a kernel.\ as follows:

Theorem 4. *The two-argument function* $\bar{K}_{[2]}(\alpha^i, \alpha^j) = \dfrac{K_{[2]}(\alpha^i, \alpha^j)}{\xi(\alpha^j)\xi(\alpha^i)} = \dfrac{\psi_{[2]}(\alpha^j \mid \alpha^i)}{\xi(\alpha^j)}$ *is a kernel.*

Proof.

$$\bar{K}_{[2]}(\alpha^i, \alpha^j) = \frac{K_{[2]}(\alpha^i, \alpha^j)}{\xi(\alpha^j)\xi(\alpha^i)} = \frac{1}{\xi(\alpha^j)\xi(\alpha^i)} \sum_{k=1}^{n} \xi(\alpha^k)\psi_{[1]}(\alpha^i \mid \alpha^k)\psi_{[1]}(\alpha^j \mid \alpha^k) =$$

$$\sum_{k=1}^{n} \left(\frac{\sqrt{\xi(\alpha^k)}}{\xi(\alpha^i)} \psi_{[1]}(\alpha^i \mid \alpha^k) \right) \left(\frac{\sqrt{\xi(\alpha^k)}}{\xi(\alpha^j)} \psi_{[1]}(\alpha^j \mid \alpha^k) \right) = \sum_{k=1}^{n} \bar{x}_{ik}\bar{x}_{jk} = \bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j$$

So, the function $\bar{K}_{[2]}(\alpha^i, \alpha^j)$ is a kernel.

The proof is complete.

It should be noted that the expression within the logarithm term in the definition of the BLOSUM substitution matrix (23) is absolutely equal to the normalized similarity measure $\bar{K}_{[2]}(\alpha^i, \alpha^j)$ and so, in accordance with theorem **4** is a kernel.

As to the PAM substitution matrix, the expression within the logarithm in its definition (6) has the same structure as $\bar{K}_{[2]}(\alpha^i, \alpha^j)$.

It is evident, that if $m = 2$ these functions $\pi_{[m]}^{ij} = \dfrac{\psi_{[m]}^{ij}}{\xi^j}$ and $\bar{K}_{[2]}(\alpha^i, \alpha^j) = \dfrac{\psi_{[2]}^{ij}}{\xi^j}$ are equal and, so $\pi_{[2]}^{ij}$ is a kernel.

The question arises: will the functions $\pi_{[m]}^{ij}$ be kernels for any m?

The answer is that they will, if there exists a random transformation $\psi_{[m/2]}(\alpha^j \mid \alpha^i)$ such that

$$\psi_{[m]}(\alpha^i \mid \alpha^j) = \sum_{\vartheta \in A} \psi_{[m/2]}(\alpha^i \mid \vartheta)\psi_{[m/2]}(\vartheta \mid \alpha^j).$$

This transformation $\psi_{[m/2]}(\alpha^j \mid \alpha^i)$ also defines an ergodic and reversible Markov random process. If the transformation $\psi_{[m/2]}(\alpha^j \mid \alpha^i)$ exists, we say that an ergodic and reversible Markov random process $\psi_{[m]}(\alpha^j \mid \alpha^i)$ is divisible.

It should be noted that divisibility does not automatically follow from the fact that a Markov random process is ergodic and reversible. It is evident that $\pi_{[m]}^{ij}$ is a kernel, at least for any even number m. As to other, non-even values of m, the divisibility of the respective Markov random processes $\psi_{[m]}(\alpha^j \mid \alpha^i)$ can be checked experimentally. Particularly, $\psi_{[1]}(\alpha^j \mid \alpha^i)$, which participates in forming PAM1 substitution matrix, is divisible and so, PAM1, or rather the expression under the logarithm $\pi_{[1]}^{ij}$ is also a kernel.

8 Results and discussions

One of the main results of this paper consists in proving an interesting and useful fact: that the statistical approach for computing substitution matrices of the BLOSUM family, introduced by the Henikoffs, can be explained in terms of the PAM evolutionary model, proposed by M. Dayhoff. So, both of these commonly adopted ways of comparing amino acid sequences are based on the same model of evolution and the main difference between them lies only in the different initial data used for estimating their unknown parameters.

Another interesting result from proving the above is that the model of PAM evolution with its main assumption of an ergodic and reversible Markov chain of point mutations of amino acids, is sufficient for constructing kernel functions over the set of amino acids. Moreover, we have shown that the natural similarity measure of amino acids, based on Dayhoff's model and commonly adopted in bioinformatics, which is the probability that two amino acids being compared might have resulted from two independent random transformations from one and the same unknown source amino acid, possesses all the properties of a kernel function. So, the substitution matrices of the PAM and BLOSUM families naturally are kernels and they have lost this property only as the result of a specific and unsuitable choice of final representation. The review of the traditional representation of substitution matrixes raises the possibility of applying more general methods for constructing kernels over the set of proteins, which is often needed to obtain a positive semi-definite substitution matrix, while at the same time using a source similarity matrix justified from the point of view of evolution.

From a mathematical perspective there are several interesting questions which remain to be answered, but which should be solvable in the near future. One set of questions is related to the problem of how to approximate an original substitution matrix by one in simple integer form (as is done now) so that the new representation will conserve the positive semi-definite property.

Other open questions arise from the above analysis which are likely to be harder to answer, but are important in practice. The best example is how to construct a pair-wise alignment for a couple of protein sequences based on the positive semi-definite substitution matrices whose score function also satisfies the same positive semi-definite property. There exist a few publications [10,11,14,15,16] which have described such alignment procedures, but these procedures don't have a rigorous evolutionary foundation.

9 Conclusions

One of the foundations for solving many problems in bioinformatics, such as protein homology detection, prediction of protein-protein interactions, prediction of biological functions, secondary and 3D structure of proteins etc., involves using a similarity measure over the set of amino acids.

The commonly used families of substitution matrices PAM and BLOSUM define similarity measures which are adequate from the biological point of view but have been here shown to fail to satisfy kernel properties, which has obscured the fact that they are so closely related, and only differ in the choice of statistical parameter choices. In numerous publications, attempts at correcting traditional substitution matrices have resulted either in the loss of their biological meaning or in revised matrices whose positive semi-definiteness was not guaranteed.

In this paper we prove that all PAM substitution matrices are kernel functions by their mathematical structure and lose their positive definiteness only because of an unsuitable final representation.

Moreover, we have proved that BLOSUM substitution matrices can also be justified in terms of the PAM evolutionary model. As result, the BLOSUM kernel family only needs a modified final representation that does not destroy the original biological rationale.

The results obtained are especially significant in view of the growing popularity of constructing data-specific amino acid substitution matrices. Any new substitution matrix will automatically be a

kernel if, first, it is estimated by Dayhoff's or the Henikoffs' techniques both based on the PAM evolutionary model, and, secondly, if the final representation guaranteeing positive definiteness proposed in the present research is applied.

References

1 Dayhoff M.O., Schwartz R.M., Orcutt B.C. A model for evolutionary change in proteins. *Atlas for Protein Sequence and Structure*, 1978, Vol. 5, 345-352.

2 Henikoff S., Henikoff J. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, 1992, 10915-10919.

3 Tatusov R.L., Galperin M.Y., Natale D.A. and Koonin E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 2000, Vol. 28, No. 1 33-36.

4 Xin Liu1, Wei-Mou Zheng. Amino acid substitution matrices for protein conformation identification. *arXiv:q-bio.BM/0406032*, v. 1, 15 June 2004.

5 Ng, P.C., Henikoff, J.G., Henikoff, S. PHAT: a transmembrane-specific substitution matrix. *Bioinformatics*, 2000, 16, 760–766.

6 Vapnik V. *Statistical Learning Theory. New York: John-Wiley&Sons, Inc.*,1998, 732 p.

7 Burges CJC: A Tutorial on Support Vector Machines for Pattern Recognition. Data *Mining and Knowledge Discovery* 1998, 2:121–167.

8 M.A. Aizerman, E.M. Braverman, L.I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 1964, Vol. 25, pp. 821-837.

9 Vanschoenwinkel B., Manderic B. Substitution matrix based kernel functions for protein secondary structure prediction. *Proceedings of the 2004 International Conference on Machine Learning and Applications.* December16-18, 2004, pp. 388 – 396.

10 Wu F., Oslon B., Dobbs D., Honavar V. Comparing kernels for predicting protein binding sites from amino acid sequence. *Newral Netwarks*, 2006, IJCNN'OC, pp. 1612-1616.

11 Vert JP, Saigo H, Akutsu T. Local alignment kernels for biological sequences. *In B. Schölkopf, K. Tsuda, and J. P. Vert, editors, Kernel Methods in Computational Biology. MIT Press,* 2004.

12 Mottl V.V., Seredin O.S., Sulimova V.V. Mathematically correct methods of similarity measurement on sets of signals and symbol sequences of different length. *Pattern Recognition and Image Analysis,* Vol.15, No. 1, 2005, pp. 87-89.

13 Caragea C., Sinapov J., Silvescu A., Dobbs D., Honavar V. Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics,* November 2007, **8**:438 doi:10.1186/1471-2105-8-438.

14 Vert J.-P., Qiu J., Noble W.S. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics,* December 2007, **8** (Suppl 10): S8 doi:10.1186/1471-2105-8-S10-S8.

15 Vincent M., Passerini A., Labbe M., Frasconi P. A simplified approach to disulfide connectivity prediction from protein sequences. *BMC Bioinformatics,* January 2008, 9:20 doi:10.1186/1471-2105-9-20.

16 *Kernel methods in computational biology.* Edited by B. Scholkoph, K.Tsuda and J.-p.Vert. A Bradford book, The MIT press, Cambridge, Massachusetts, London, England, 2004, 397p.