A Class of Evolution-Based Kernels for Protein Homology Analysis: A Generalization of the PAM Model

Valentina Sulimova [1], Vadim Mottl [2], Boris Mirkin [3], Ilya Muchnik [4], and Casimir Kulikowski [4]

[1] Tula State University, Lenin Ave. 92, 300600 Tula, Russia
vsulimova@yandex.ru
[2] Computing Center of the Russian Academy of Sciences, Vavilov St. 40, 119333 Moscow, Russia
vmottl@yandex.ru
[3] Birkbeck College, University of London, Malet Street, London, WC1E 7HX, UK
mirkin@dcs.bbk.ac.uk
[4] Rutgers University, New Brunswick, New Jersey 08903, USA
muchnikilya@yahoo.com, kulikows@cs.rutgers.edu

Abstract. There are two desirable properties that a pair-wise similarity measure between amino acid sequences should possess in order to produce good performance in protein homology analysis. First, it is the presence of kernel properties that allow using popular and well-performing computational tools designed for linear spaces, like SVM and k-means. Second, it is very important to take into account common evolutionary descent of homologous proteins. However, none of the existing similarity measures possesses both of these properties at once. In this paper, we propose a simple probabilistic evolution model of amino acid sequences that is built as a straightforward generalization of the PAM evolution model of single amino acids. This model produces a class of kernel functions each of which is computed as the likelihood of the hypothesis that both sequences are results of two independent evolutionary transformations of a hidden common ancestor under some specific assumptions on the evolution mechanism. The proposed class of kernels is rather wide and contains as particular subclasses not only the family of J.-P Vert's local alignment kernels, whose algebraic structure was introduced without any evolutionary motivation, but also some other families of local and global kernels. We demonstrate, via k-means clustering of a set of amino acid sequences from the VIDA database, that the global kernel can be useful in bringing together otherwise very different protein families.

Keywords: Protein homology analysis, evolution modeling, amino acid sequence alignment, evolutionary kernel function, kernel-derived clusters.

1 Introduction

Protein homology is understood as sequence similarity based on recent common ancestry and similar function. The concept of homology is one of the most important in

I. Măndoiu, G. Narasimhan, and Y. Zhang (Eds.): ISBRA 2009, LNBI 5542, pp. 284–296, 2009.

proteomics. However, its operational meaning remains rather obscure and not well represented by computationally sound concepts and tools, and practical definition of protein homologous families remains subject to significant manual curation.

In the absence of structural folding data for the overwhelming majority of proteins, the development of protein homologous families mostly relies on protein amino acid sequence data, under the assumption that homologous proteins should have similar protein sequences. To fully automate the process, a reliable and computationally feasible solution to two principal issues is needed: (a) measuring similarity between protein sequences and (b) clustering proteins into similarity groups.

The most popular approach to assessing the pair-wise similarity of amino acid sequences, which was developed as a means of detecting evolutionary relationships between them, is the approach based on the notions of global or local alignment. The original idea was to determine a similarity measure through finding the best correspondence between successive amino acids in two sequences with possible gaps, which is computed via a version of the dynamic programming procedure, respectively, the global Needleman-Wunsch [1] or the local Smith-Waterman algorithm [2]. In the latter case, fast heuristics BLAST, PSI-BLAST [3 and FASTA [4] are usually applied instead of dynamic programming to accelerate calculations.

However, these approaches do not perform well for remotely homologous proteins [5,6 ,7]. We suppose that this insufficiency arises due to the absence of two very important properties:

(a) classical optimal-alignment-based similarity measures are not based on a biologically-defined evolution mechanism, and
(a) they are not kernel functions, i.e., do not enable usage of such powerful and convenient tools as SVM and k -means clustering developed for linear spaces.

Multiple investigations have explored the idea of endowing the similarity measures with these desirable properties. In particular, a number of kernel functions in the set of amino acid and DNA sequences were introduced in [8, 9, 10, 11, 12, 13]. However, practically all of them remain motivated by purely algebraic considerations, even J.-P. Vert's LA-kernels [8] which average the similarity measures produced by all feasible local alignments. The families of kernels studied in [14, 15], in spite of their probabilistic nature, are also not based on an explicitly formulated model of evolution.

A separate line of investigations has been aimed at forming models of protein evolution. In [16, 17, 18, 19], a number of different models were proposed, but the similarity measures resulting from them, which are based on the notion of statistical alignment and take into account most likely ways of evolution [16] or all possible ways [17,18,19], do not possess, nevertheless, all the kernel properties. Besides, the methods of this kind have very high computational complexity [18] or do not guarantee the mathematical correctness of the similarity measure as the probability of two independent transformations of the protein pair under comparison from the same unknown ancestor [17].

In this paper, we propose a simple probabilistic model of evolution of amino acid sequences that is built as a rather straightforward generalization of Margaret Dayhoff's Point Accepted Mutation (PAM) model developed for single amino acids [20]. The respective pair-wise sequence similarity measure has the strong mathematical meaning of the likelihood of the hypothesis that the two sequences are the results of two independent evolutionary transformations of some hidden sequence considered as

their common ancestor. Each similarity measure of the proposed class possesses all the properties of a kernel function, in particular, the matrix of its values computed for any finite set of amino acid sequences is, at least, positive semidefinite. By its algebraic structure, this class of kernels is an essential generalization of that of local alignment kernels [8] and embraces not only the local but also the global principle of sequence comparison. A kernel function defined on a set of entities of arbitrary kind embeds it into a hypothetical linear space, with the role of the inner product played by the kernel itself. Thus, any linear methods are applicable to the set of amino acid sequences since we have managed to measure some pair-wise relation between them by a kernel function.

The proposed class of evolutionary kernels is verified via clustering a given set of amino acid sequences that contains several known groups of homologous proteins from the VIDA database. For this purpose, we correspondingly modified the k-means method of clustering, as one of the most popular linear method, for the kernel-based protein representation. It turned out that the subclass of global kernels demonstrated almost complete coincidence of the clustering result with the true homologous groups in the protein set under processing, in contrast to local alignment kernels and similarity measures based on finding the optimal alignment, which could not bring together different protein families.

2 Evolution-Based Principle of Comparing Amino Acid Sequences

2.1 Similarity of Amino Acids

Measuring similarity of amino-acid sequences must inevitably be based on measuring similarity of amino acids forming them. The most commonly adopted similarity measure involves the family of PAM substitution matrices derived by Margaret Dayhoff [20] from a probabilistic model of evolution. Another popular family of substitution matrices was introduced by Steven and Jorjia Henikoff and called BLOSUM (BLOcks SUbstitution Matrices) [21]. These matrices directly calculate 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. However, it is shown in [22] that the family of BLOSUM substitution matrices can be explained in terms of Dayhoff's evolutionary model as was done for PAM.

The main mathematical notion of Dayhoff's PAM evolution model at a single point of protein sequence is that of a Markov chain over the alphabet of 20 amino acids $A = \{\alpha^1, ..., \alpha^{20}\}$. The model is defined by a matrix of conditional probabilities $\mathbf{\Psi} = \left(\psi(\alpha^j \mid \alpha^i), i, j = 1, ..., 20\right)$ that amino acid α^i will be substituted by amino acid α^j at the next step of evolution (mutation probability matrix). It is assumed that the Markov chain of evolution is ergodic and reversible, i.e., there exists the final probability distribution $\xi(\alpha^j)$

$$\sum_{\alpha^i \in A} \xi(\alpha^i)\psi(\alpha^j \mid \alpha^i) = \xi(\alpha^j), \tag{1}$$

and the equality

$$\xi(\alpha^i)\psi(\alpha^j \mid \alpha^i) = \xi(\alpha^j)\psi(\alpha^i \mid \alpha^j) \tag{2}$$

holds true for all amino acids.

When estimating the matrix $\boldsymbol{\Psi}$ by aligning a set of very similar sequences, it was conventionally accepted that $1-\sum_{i=1}^{20}\xi(\alpha^i)\psi(\alpha^i \mid \alpha^i) = 0.01$, i.e., 1% of amino acids would change at a single step of evolution. This mutation probability matrix is called PAM 1 and associated with the evolutionary distance 1. Margaret Dayhoff considered also the s-foldly thinned-out Markov chain, i.e., derived from the original one by taking away $s-1$ of every s elements, which will be defined by the matrix $\boldsymbol{\Psi}_{[s]} = \left(\psi_{[s]}(\alpha^j \mid \alpha^i), i, j = 1,...,20\right) = \underbrace{\boldsymbol{\Psi}\times...\times\boldsymbol{\Psi}}_{s}$ corresponding to the evolutionary distance s and having the same final probability distribution $\xi(\alpha^j)$. The most popular mutation probability matrix is PAM 250 with $s = 250$.

It is easy to prove [22] that for any s the similarity measure

$$\mu_{[s]}(\alpha^i, \alpha^j) = \psi_{[s]}(\alpha^j \mid \alpha^i)\xi(\alpha^i) = \psi_{[s]}(\alpha^i \mid \alpha^j)\xi(\alpha^j), \tag{3}$$

as well as its normalized version

$$\tilde{\mu}_{[s]}(\alpha^i, \alpha^j) = \mu_{[s]}(\alpha^i, \alpha^j) / \xi(\alpha^i)\xi(\alpha^j) = \psi_{[s]}(\alpha^j \mid \alpha^i) / \xi(\alpha^j) = \psi_{[s]}(\alpha^i \mid \alpha^j) / \xi(\alpha^i) \tag{4}$$

are kernel functions, i.e. form positive definite matrices in the set of amino acids.

The PAM scoring matrices are traditionally represented in the log-odds form as $\pi_{[s]}(\alpha^i, \alpha^j) = 10\log_{10}\tilde{\mu}_{[s]}(\alpha^i, \alpha^j)$. This logarithmic representation is convenient from the viewpoint of scaling similarity values but deprives $\tilde{\mu}_{[s]}(\alpha^i, \alpha^j)$ of the original kernel-function properties.

2.2 The Main Idea of Comparing Two Amino Acid Sequences

Let Ω be the set of all finite amino acid sequences $\boldsymbol{\omega} = (\omega_t, t = 1,...,N)$, $\omega_t \in A$. We shall use also the notation $\Omega_n = \{\boldsymbol{\omega} = (\omega_t,\ t = 1,...,N),\ \omega_t \in A,\ N = n\}$ for the set of all sequences having the fixed length n, and $\Omega_{\geq n} = \{\boldsymbol{\omega} = (\omega_t, t = 1,...,N),\ \omega_t \in A,\ N \geq n\}$ for the set of sequences that are not shorter than n. Let us, further, consider a random sequence $\boldsymbol{\vartheta} = (\vartheta_i \in A, i = 1,...,n) \in \Omega_n \subseteq \Omega$ of random length n, such that the pair $(n, \boldsymbol{\vartheta})$ is jointly defined by a pair of probability distributions $\left(r(n), n = 0,1,2,...\right)$ and $\left(p_n(\boldsymbol{\vartheta}), \boldsymbol{\vartheta} \in \Omega_n\right)$.

It appears natural to evaluate the similarity of two amino acid sequences $\boldsymbol{\omega}', \boldsymbol{\omega}'' \in \Omega_{\geq n}$ by computing the probability of the hypothesis that they originate from the same random ancestor $\boldsymbol{\vartheta} \in \Omega_n$ as results of two independent branches of evolution defined by a known random transformation $\left(\varphi_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}), \boldsymbol{\omega} \in \Omega_{\geq n}, \boldsymbol{\vartheta} \in \Omega_n\right)$:

$$\mathcal{K}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} r(n) \sum_{\boldsymbol{\vartheta}\in\Omega_n} p_n(\boldsymbol{\vartheta})\varphi_n(\boldsymbol{\omega}' \mid \boldsymbol{\vartheta})\varphi_n(\boldsymbol{\omega}'' \mid \boldsymbol{\vartheta}). \tag{5}$$

Theorem 1. *Any choice of distribution* $\left(r(n), n = 0,1,2,...\right)$ *and families of conditional distributions* $\left(\varphi_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}), \boldsymbol{\omega} \in \Omega_{\geq n}, \boldsymbol{\vartheta} \in \Omega_n\right)$ *and* $\left(p_n(\boldsymbol{\vartheta}), \boldsymbol{\vartheta} \in \Omega_n\right)$ *leads to the fact that (5) is a kernel function in the set of all amino acid sequences.*

The proof of Theorem 1 is based on checking whether Mercer's conditions [23] are met.

2.3 The Model of Random Evolutionary Transformation of an Amino Acid Sequences into Another One

We consider here only two-step random transformations $\varphi_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta})$ of the ancestor sequence $\boldsymbol{\vartheta} \in \Omega_n$ into a resulting sequence $\boldsymbol{\omega} \in \Omega_{\geq n}$.

Step 1. Random choice of an n-length structure $\mathbf{v} = (v_1, ..., v_n)$ of transformation $\boldsymbol{\vartheta} \to \boldsymbol{\omega}$ with distribution $q(\mathbf{v})$ defined in the set of all n-length structures $\mathbf{v} = (v_1, ..., v_n) \in V_n$, $\sum_{\mathbf{v} \in V_n} q_n(\mathbf{v}) = 1$, where increasing sequence of natural numbers $1 \leq v_1 < v_2 < ... < v_n$, indicates the positions of the resulting sequence into which the respective elements of the ancestor $\boldsymbol{\vartheta} = (\vartheta_1, ..., \vartheta_n)$ will be transformed. So, for any given sequence $\boldsymbol{\omega} = (\omega_1, ..., \omega_N)$ of length $N \geq n$, a specific structure $\mathbf{v} = (v_1, ..., v_n)$ explicitly defines, first, a subsequence $\overline{\boldsymbol{\omega}}_{\mathbf{v}} = (\omega_{v_i}, i = 1, ..., n)$ of elements, which we shall name the key subsequence, and, second, the additional subsequence $\overline{\overline{\boldsymbol{\omega}}}_{\mathbf{v}} = (\omega_t,$ $t \neq v_i, i = 1, ..., n)$, such that $\boldsymbol{\omega} = \overline{\boldsymbol{\omega}}_{\mathbf{v}} \bigcup \overline{\overline{\boldsymbol{\omega}}}_{\mathbf{v}}$.

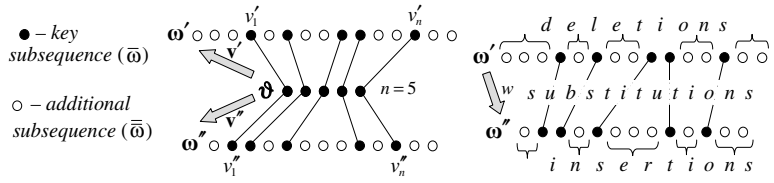

Fig. 1. The structure of random transformation of symbolic sequences

Step 2. For each structure $\mathbf{v} \in V_n$, a structure-dependent random transformation is assumed to be defined $\eta_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}, \mathbf{v}) \geq 0$ such that $\eta_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}, \mathbf{v}) = 0$ if $\boldsymbol{\omega} \notin \Omega_{\geq v_n}$, i.e. $\sum_{\boldsymbol{\omega} \in \Omega_{\geq v_n}} \eta_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}, \mathbf{v}) = 1$. As a result, distribution $\varphi_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta})$ is the mixture

$$\varphi_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}) = \sum_{\mathbf{v} \in V_n} q_n(\mathbf{v}) \eta_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}, \mathbf{v}). \quad (6)$$

In this work, we make some additional assumptions on the distributions forming the class of kernel functions (5).

Conditional independence of the key-subsequence elements. The symbols of the key subsequence are randomly generated from those of the original sequence in

accordance with Dayhoff's mutation probabilities $\psi_{[s]}(\alpha^j \mid \alpha^i)$ in the set of amino acids with a conventional evolution step s (Section 2.1)

$$\bar{\eta}_n(\bar{\boldsymbol{\omega}}_{\mathbf{v}} \mid \boldsymbol{\vartheta}, \mathbf{v}) = \prod_{i=1}^{n} \psi_{[s]}(\omega_{v_i} \mid \vartheta_i). \tag{7}$$

Independence of the ancestor-sequence elements. The original sequence is formed by independent symbols chosen in accordance with Dayhoff's final probabilities $\xi(\alpha^i)$ in the set of amino acids

$$p_n(\boldsymbol{\vartheta}) = p_n(\vartheta_1, ..., \vartheta_n) = \prod_{i=1}^{n} \xi(\vartheta_i). \tag{8}$$

Completely random runaway of the resulting sequence. The runaway τ of the length $N_{\boldsymbol{\omega}} = v_n + \tau$ over v_n in the transformation structure $\mathbf{v} = (v_1, ..., v_n)$ is determined through the "completely random" length of the final part of the additional subsequence $(\omega_{v_n}, ..., \omega_{N_{\boldsymbol{\omega}}})$. So, the distribution of τ is considered as an improper "almost uniform" distribution

$$z(\tau) \to \approx 0\,, \ \sum_{\tau=0}^{\infty} z(\tau) = 1\,. \tag{9}$$

Independence of the additional subsequence from the key one. It is assumed that

$$\eta_n(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}, \mathbf{v}) = \bar{\eta}_n(\bar{\boldsymbol{\omega}}_{\mathbf{v}} \mid \boldsymbol{\vartheta}, \mathbf{v})\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}}). \tag{10}$$

The notation $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}})$ instead of $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}} \mid \boldsymbol{\vartheta}, \mathbf{v})$ means that, first, there is no dependence on the original sequence $\boldsymbol{\vartheta}$ and, second, if the symbolic compositions of different subsequences coincide $\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}'} = \bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}''}$ then $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}'}) = \bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}''})$.

2.4 The General Kernel Structure

After these assumptions, it remains only to choose the family of distribution $q_n(\mathbf{v})$ over the set of n-length structures $\mathbf{v} = (v_1, ..., v_n) \in V_n$ and the family $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}})$ that determines the additional subsequence. Below, in Section 3, we shall see that these two choices not only accomplish the definition of the class of kernels but also essentially affect its properties. However, the already made assumptions allow for representing the class of kernels (5) in a more structurally explicit form.

Any pair of n-length structures $\mathbf{v}' = (v_1', ..., v_n') \in V_n$ and $\mathbf{v}'' = (v_1'', ..., v_n'') \in V_n$ of transformations $\boldsymbol{\vartheta} \to \boldsymbol{\omega}'$ and $\boldsymbol{\vartheta} \to \boldsymbol{\omega}''$, $\boldsymbol{\vartheta} \in \Omega_n$, $\boldsymbol{\omega}', \boldsymbol{\omega}'' \in \Omega_{\geq n}$ defines a pair-wise alignment of the two sequences (Figure 1):

$$\mathbf{w} = (\mathbf{v}'_{\mathbf{w}}, \mathbf{v}''_{\mathbf{w}}) = \left[\begin{pmatrix} v'_{1,\mathbf{w}} \\ v''_{1,\mathbf{w}} \end{pmatrix}, ..., \begin{pmatrix} v'_{n,\mathbf{w}} \\ v''_{n,\mathbf{w}} \end{pmatrix} \right].$$

We shall call it the pair-wise alignment of order n (n-order alignment) because exactly n pairs of amino acids will be immediately compared. The set of all n-order pair-wise alignments is $W_n = V_n \times V_n$, and distribution $q_n(\mathbf{v})$ in V_n defines distribution $q_n(\mathbf{w}) = q_n(\mathbf{v}'_{\mathbf{w}})q_n(\mathbf{v}''_{\mathbf{w}})$. Vice-versa, any n-order pair-wise alignment $\mathbf{w}$ defines

a pair of n-length transformation structures $(\mathbf{v}'_{\mathbf{w}}, \mathbf{v}''_{\mathbf{w}})$. It should be noticed that not all pair-wise alignments can define the pair of sequences $\boldsymbol{\omega}'$ and $\boldsymbol{\omega}''$ of lengths N' and N'', but only those of them which satisfy the conditions $v'_n(\mathbf{w}) \le N'$ and $v''_n(\mathbf{w}) \le N''$. The set of all such pair-wise alignments of sequences $\boldsymbol{\omega}'$ and $\boldsymbol{\omega}''$ will be denoted as $\mathbf{w} \in W_{nN'N''} \subset W_n$.

For each pair-wise alignment $\mathbf{w} \in W_{nN'N''}$, we define a real-valued symmetric function over all pairs of sequences $\boldsymbol{\omega}' \in \Omega_{\ge N'}$ and $\boldsymbol{\omega}'' \in \Omega_{\ge N''}$

$$\bar{K}_n(\boldsymbol{\omega}', \boldsymbol{\omega}'' | \mathbf{w}) = \bar{K}_n(\bar{\boldsymbol{\omega}}'_{\mathbf{v}'_{\mathbf{w}}}, \bar{\boldsymbol{\omega}}''_{\mathbf{v}''_{\mathbf{w}}}) = \prod_{i=1}^{n} \mu_{[2s]}\left(\omega'_{v'_{\mathbf{w},i}}, \omega''_{v''_{\mathbf{w},i}}\right), \tag{11}$$

where $\mu_{[2s]}(\omega'_{v'_{i,\mathbf{w}}}, \omega''_{v''_{i,\mathbf{w}}})$ is the kernel on the set of amino acids (3) for the double evolution step $2s$ with respect to the step s taken in the one-side model (7). Since any product of kernels remains to be a kernel, the function $\bar{K}_n(\boldsymbol{\omega}', \boldsymbol{\omega}'' | \mathbf{w})$ is also a kernel. We shall call it alignment-dependent key kernel of order n.

Further, we define the alignment-dependent additional kernel as

$$\bar{\bar{K}}_n(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) = \bar{\bar{K}}_n(\bar{\bar{\boldsymbol{\omega}}}'_{\mathbf{v}'_{\mathbf{w}}}, \bar{\bar{\boldsymbol{\omega}}}''_{\mathbf{v}''_{\mathbf{w}}}) = \bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}'_{\mathbf{v}'})\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}''_{\mathbf{v}''}) . \tag{12}$$

Theorem 2. *Under assumptions (7)-(10) and notations (11) and (12), the kernel (5) is representable as*

$$K(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} r(n) \sum_{\mathbf{w} \in W_{nN'N''}} q_n(\mathbf{w}) \bar{K}(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) \bar{\bar{K}}(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) . \tag{13}$$

Proof. Elementary substitution of assumptions (7)-(10) in (5) yields (13) for the function

$$\bar{K}_n(\boldsymbol{\omega}', \boldsymbol{\omega}'' | \mathbf{w}) = \sum_{\boldsymbol{\vartheta} \in \Omega_n} p_n(\boldsymbol{\vartheta}) \prod_{i=1}^{n} \xi(\vartheta_i) \psi_{[s]}(\omega'_{v'_{\mathbf{w},i}} | \vartheta_i) \psi_{[s]}(\omega''_{v''_{\mathbf{w},i}} | \vartheta_i) .$$

Its equivalence to (11) immediately follows from Dayhoff's main assumption on the ergodicity (1) and reversibility (2) of the PAM Markov chain.

Thus, we have come to the class of kernels on the set of amino acid sequences

$$K(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} r(n) \sum_{\mathbf{w} \in W_{nN'N''}} q_n(\mathbf{w}) \bar{\bar{K}}(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) \prod_{i=1}^{n} \mu_{[2s]}\left(\omega'_{v'_{\mathbf{w},i}}, \omega''_{v''_{\mathbf{w},i}}\right), \tag{14}$$

where the distributions $\left(r(n), n = 0,1,2,...\right)$, $\left(q_n(\mathbf{w}), \mathbf{w} \in W_n\right)$ and $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}})$ over all symbolic sequences of any length $\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}} \in \Omega$ are not defined as yet.

To work up some policy of choosing these remaining elements of the kernel construction, it is required to clarify their influence on the result of sequence comparison. The distribution $\left(r(n), n = 0,1,2,...\right)$ is meant to express some assumption on the length of the hidden ancestor, i.e., the number of pairs of amino acid positions in $\boldsymbol{\omega}'$ and $\boldsymbol{\omega}''$, which will be taken into account when comparing these sequences. The similarity of the given sequences from the viewpoint of the presence of some "almost common" subsequence of length n in them is measured just by the key kernel (11), whereas the choice of the additional kernel (12) defined by $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}})$ gives the possibility to dilute

this assessment, if desirable, through involving other unpaired elements into comparison. The role of distribution $\left(q(\mathbf{w}), \mathbf{w} \in W_n\right)$ is, actually, regulation of whether, how and to what extent the gaps between paired positions will affect the comparison.

3 Some Particular Kinds of Kernels

3.1 Kernels of Fixed and Unfixed Order

In particular, if there is no reason to constrain the tentative length of the hypothetical "almost common" subsequence, the distribution $\left(r(n), n = 0,1,2,...\right)$ should be taken as an improper "almost uniform" distribution $r(n) \to const \cong 0$, $\sum_{n=0}^{\infty} r(n) = 1$. In this extreme case, the respective item will fall out of (14) completely, and we obtain key-length indifferent kernels, which we call **kernels of absolutely unfixed order**:

$$K(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} \sum_{\mathbf{w} \in W_{nN'N''}} q_n(\mathbf{w}) \bar{\bar{K}}(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) \prod_{i=1}^{n} \mu_{[2s]}\left(\omega'_{v'_{\mathbf{w},i}}, \omega''_{v''_{\mathbf{w},i}}\right). \tag{15}$$

On the extreme contrary, if it desirable to strictly preset the length of the key subsequence, this distribution should turn into an absolutely concentrated one $r(n) = 1$ and $r(k) = 0$ with any $k \neq n$. The kernels of this kind are called here **kernels of a fixed order**:

$$K_n(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{\mathbf{w} \in W_{nN'N''}} q_n(\mathbf{w}) \bar{\bar{K}}(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) \prod_{i=1}^{n} \mu_{[2s]}\left(\omega'_{v'_{\mathbf{w},i}}, \omega''_{v''_{\mathbf{w},i}}\right). \tag{16}$$

3.2 Local and Global Kernels

As a rule, the desirability of that of other kinds of alignments is expressed by the mathematical assumption that the likelihood $q_n(\mathbf{w}) = q_n(\mathbf{v}')q_n(\mathbf{v}'')$ depends only on the lengths of the gaps at the left $(v_1 - 1)$, in the middle $\left((v_2 - v_1), ..., (v_n - v_{n-1})\right)$ and at the right $(N_\omega - v_n)$ of each of the two sequences. It is the usual practice to assume that the random lengths of the gaps are a priori independent, and each of them has a probability distribution monotonically diminishing as the length grows:

$$g_i(d_i \mid a, b) \propto \begin{cases} 1, & d_i = 1, \\ \exp\left[-\beta\left(a + bd_i\right)\right], & d_i > 1, \end{cases} \quad d_0 = v_1,\ d_i = v_i - v_{i-1}, \text{ or } d_{n+1} = N_\omega - v_n. \tag{17}$$

If $a = 0$, the "cost" of two gaps d_i and d_j is the same as that of one gap of the summed up length $d_i + d_j$. Otherwise, if $a > 0$, one long gap is considered as more preferable than two short gaps making the same length. However, the distributions may be taken, if required, as position-dependent, i.e., different for different i.

The kernel function is said to be **local** if only middle parts of the two sequences participate in comparison. In this case, we define $\left(q(\mathbf{v}), \mathbf{v} \in V_n\right)$, for instance, by putting the improper joint distribution as the product of the identical single distributions (17) within the range of the key part and "absolutely random" ones beyond it

$q_n(\mathbf{v}) \propto \prod_{i=2}^{n} g(v_i - v_{i-1} \mid a,b)$, and the distribution of the additional subsequences is taken as completely improper one $\bar{\bar{\eta}}(\bar{\bar{\omega}}_v) = const = 1$, which defines neither the lengths nor the compositions of the additional subsequences.

From the inverse point of view, the additional parts of the two sequences should be compared with the same attention as the key ones when judging of sequence similarity. In this case, the a priori models of both gaps and additional symbols should be extended onto the entire lengths of the sequences, for instance, as

$$\begin{cases} q_n(\mathbf{v}) = g(v_1 \mid a,b)\left(\prod_{i=2}^{n} g(v_i - v_{i-1} \mid a,b)\right) g\left(N_\omega - v_n\right), \\ \bar{\bar{\eta}}(\bar{\bar{\omega}}_{\mathbf{v}}) = \prod_{\substack{1 \le t < v_n \\ t \ne v_i}} \xi(\omega_t). \end{cases} \tag{18}$$

The a priori distributions (18) define the family of **global kernels**.

4 Kernel Computation: A Slight Modification of the Algorithm for Local Alignment Kernels

It should be noticed that the properties of the proposed model of evolution allow to express the initial kernel (19) in absolutely equivalent but essentially more simple form (14), which does not contain the sum over all possible hidden ancestor sequences (see Theorem 1). This circumstance provides the possibility for sufficiently simple and quick computation of this kernel with complexity $O(|\boldsymbol{\omega}'||\boldsymbol{\omega}''|)$.

Despite the fact that the local alignment kernels proposed by J.-P. Vert and his colleagues in [8] for classification of biological sequences are not motivated by any explicitly formulated evolution model, they fall, by their algebraic structure, into the class considered here. More exactly, the local alignment kernels belong to the family of **local kernels of absolutely unfixed order**. So, the dynamic-programming algorithm described in [8] computes a kernel of this kind. The other particular cases of the proposed class of evolution-based kernels, namely, **global kernels of absolutely unfixed order** and **local and global kernels of fixed order**, require a slight modification of the algorithm [8]. In particular, the global kernel of absolutely unfixed order can be computed using (17) by recurrent expressions:

$M_{i,j} = \mu_s(\omega'_i, \omega''_j)(M_{i-1,j-1} + \mathrm{e}^{-b} X_{i-1,j-1} + \mathrm{e}^{-b} Y_{i-1,j-1} + e^{-2b} Z_{i-1,j-1} + e^{-g(i-1|a,b)-g(j-1|a,b)})$,

$X_{i,j} = e^{-a} M_{i-1,j} + \mathrm{e}^{-b} X_{i-1,j}$; $Y_{i,j} = \mathrm{e}^{-a} M_{i,j-1} + \mathrm{e}^{-b} Y_{i,j-1}$, $Z_{i,j} = e^{-a}(e^{-a} M_{i-1,j-1} + X_{i,j-1} + Y_{i-1,j}) + e^{-2} Z_{i-1,j-1}$,

starting with $M_{i,0} = M_{0,j} = 0$, $X_{i,0} = X_{0,j} = 0$, $Y_{i,0} = Y_{0,j} = 0$, $Z_{i,0} = Z_{0,j} = 0$. The resulted value of the global kernel of absolutely unfixed order is given by the formula

$$K(\boldsymbol{\omega}', \boldsymbol{\omega}'') = M_{|\boldsymbol{\omega}'|,|\boldsymbol{\omega}''|} + e^{-b}(X_{|\boldsymbol{\omega}'|,|\boldsymbol{\omega}''|} + Y_{|\boldsymbol{\omega}'|,|\boldsymbol{\omega}''|} + e^{-b} Z_{|\boldsymbol{\omega}'|,|\boldsymbol{\omega}''|}).$$

5 Kernel-Based Clustering of Proteins

The task of clustering is to partition the given set of amino acid sequences $\Omega^* = \{\boldsymbol{\omega}_j, j = 1,...,M\}$ into k disjoint subsets $\Omega^* = \Omega_1^* \cup \Omega_2^* \cup ... \cup \Omega_k^*$, $\Omega_i^* \cap \Omega_l^* = \varnothing$,

$i,l=1,...,k,\ i\neq l$, each of which consists of similar sequences with respect to the accepted similarity measure. We use the well known k-means method [24, 25], adopted by us for the case when the similarity is measured by a kernel function.

Any kernel $K(\boldsymbol{\omega}',\boldsymbol{\omega}'')$ defined in the given set of amino acid sequences Ω^* embeds it into a hypothetical linear space $\tilde{\Omega}^* \supset \Omega^*$ with Euclidean metric

$$\rho^2(\boldsymbol{\omega}',\boldsymbol{\omega}'') = K(\boldsymbol{\omega}',\boldsymbol{\omega}') + K(\boldsymbol{\omega}'',\boldsymbol{\omega}'') - 2K(\boldsymbol{\omega}',\boldsymbol{\omega}''),\ \boldsymbol{\omega}',\boldsymbol{\omega}'' \in \tilde{\Omega}^*. \tag{20}$$

The k-means iteration procedure consists in implementation of the following two steps at each $(s+1)$th iteration:

1) finding k fixed abstract class centers $\boldsymbol{\vartheta}_k^{s+1},...,\boldsymbol{\vartheta}_k^{s+1} \in \tilde{\Omega}^*$ on the basis of the known partition $\{\Omega_i^{*(s)}, i=1,...,k\}$ at the previous iteration by the rule $\boldsymbol{\vartheta}_i^{s+1} = \arg\min_{\boldsymbol{\vartheta}\in\tilde{\Omega}^*} \sum_{\boldsymbol{\omega}_l\in\Omega_i^{*(s)}} \rho^2(\boldsymbol{\omega}_l,\boldsymbol{\vartheta})$, which, with respect to (20) and the properties of the Frechet differential in the linear space $\tilde{\Omega}^*$, leads to the explicit expression:

$$\boldsymbol{\vartheta}_i^{s+1} = \left(1/|\,\Omega_i^{*(s)}|\right)\sum_{\boldsymbol{\omega}_l\in\Omega_i^{*(s)}} \boldsymbol{\omega}_l\ . \tag{21}$$

2) finding the new partition defined by these centers:

$$\Omega_i^{*(s+1)} = \left\{\boldsymbol{\omega}_j \in \Omega^* : \rho^2(\boldsymbol{\omega}_j,\boldsymbol{\vartheta}_i^{s+1}) = \min_{l=1,...,k} \rho^2(\boldsymbol{\omega}_j,\boldsymbol{\vartheta}_l^{s+1})\right\},\ i=1,...,k\ . \tag{22}$$

In accordance with (20) and on the force of the linearity property of the inner product $K(\boldsymbol{\omega}',\boldsymbol{\omega}'')$ in the linear space $\tilde{\Omega}^*$, we have:

$$\rho^2(\boldsymbol{\omega}_j,\boldsymbol{\vartheta}_l^{s+1}) = K(\boldsymbol{\omega}_j,\boldsymbol{\omega}_j) + \frac{1}{|\,\Omega_i^{*(s)}|^2}\sum_{\boldsymbol{\omega}_j\in\Omega_i^{*(s)}}\sum_{\boldsymbol{\omega}_l\in\Omega_i^{*(s)}} K(\boldsymbol{\omega}_j,\boldsymbol{\omega}_l) - 2\frac{1}{|\,\Omega_i^{*(s)}|}\sum_{\boldsymbol{\omega}_j\in\Omega_i^{*(s)}} K(\boldsymbol{\omega}_j,\boldsymbol{\omega}_l)$$

Substitution of the obtained formula into (22) allows to avoid explicit computation of the abstract centers and, so, to avoid step 1.

To start the k-means algorithm, we apply the procedure of finding "anomalous patterns" proposed in [25], which automatically identifies the number and compositions of the initial clusters.

6 Protein Homology Analysis. Data, Experiments and Results

One of the particular subclasses of the class of evolution-model-based kernels proposed in this paper is that of so-called local kernels of absolutely unfixed order (Section 3) which coincides with the family of local alignment kernels by J.-P. Vert and his colleagues [8]. It is shown in [8] that the kernels of this kind essentially outperform the protein similarity measures based on finding the optimal alignment and other non-evolutionary kernels in detecting remote homology of proteins. The aim of the experiment presented in this Section is to demonstrate that there exists, at least, one subclass in our class of kernels, namely that of global kernels of absolutely unfixed order (Section 3), that can be useful in bringing together proteins from indubita-

bly the same homologous group which appear, nevertheless, as very different from the viewpoints of other similarity measures.

The data set consists of 233 membrane glycoproteins comprising 8 herpesvirus Homology Protein Families (HPF) divided in the VIDA database in three subsets according to their function. The structure of the data set and the results of its clustering are shown in Figure 2.

In the experiments, we tested four different similarity measures:

(a) PSI-BLAST tool [3], (b) Needleman-Wunsch algorithm [1], (c) local alignment kernel of absolutely unfixed order [8], and (d) global kernel of absolutely unfixed order. Traditional measures (a) and (b) are based on the optimal local and global alignment respectively. The kernels (c) and (d) considered in this paper are based on multiple alignments. For methods (a)-(c), we used the values of parameters recommended by authors.

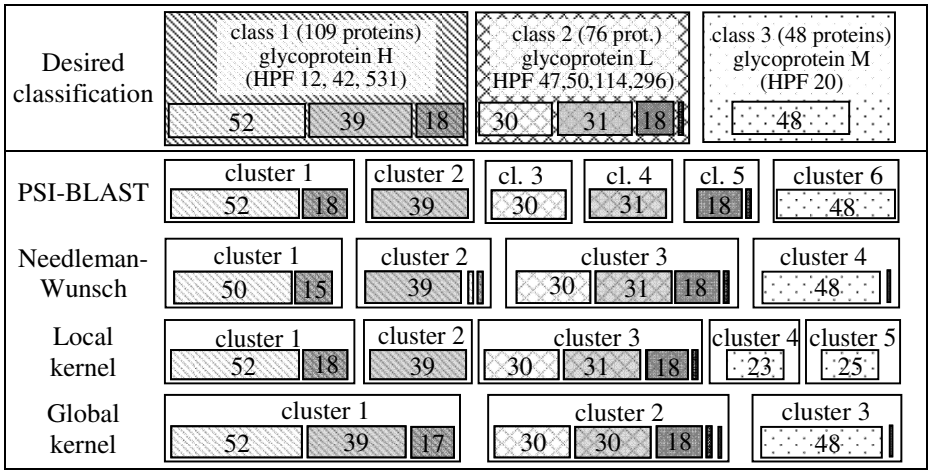

Fig. 2. Results of clustering the set of 233 membrane glycoproteins

For each of these similarity measures, we solved the problem of clustering the 233 proteins into an unfixed number of clusters. In all the cases, the number of initial classes k was identified on the basis of the procedure of finding "anomalous patterns" [25]. For cases (a) and (b), we used the standard dissimilarity-based k -centers algorithm of clustering in which some real object plays the role of the approximate center of each class, whereas for cases (c) and (d) the kernel-based k -means procedure was applied in which the center of the respective class is represented by the arithmetic mean of objects forming it (21) in accordance with the linear operations induced by the kernel in $\tilde{\Omega}^*$.

The results of clustering presented in Figure 2 show that only the global kernel of completely unspecified order yields the clustering which practically coincides with the actual structure of the three homology groups of proteins. In fact, the global kernel correctly identifies the similarity between otherwise dissimilar homologous protein families that bear the same function in the organisms under consideration.

This is the obvious example of superiority of the global kernel. However, it should be noticed that this superiority is not absolute. For a number of classes, the other particular cases of the proposed evolutionary kernel demonstrate higher performance compared to the global kernel. In particular, we detected that the global kernel, as any global similarity measure, is very sensitive to the length of proteins and cannot detect the similarity of sequences of very different lengths. At the same time, it allows comparing whole sequences instead of their parts and, as a result, detecting similarity in some cases when local kernels fail.

7 Conclusions

In this paper, we have proposed a simple probabilistic model of amino acid sequence evolution, which is built as a straightforward generalization of the PAM evolution model developed by Margaret Dayhoff for single amino acids. The respective pair-wise sequence similarity measure possesses the properties of a kernel function computed as the likelihood of the hypothesis that both sequences are results of two independent evolutionary transformations of some hidden common ancestor.

Under some particular assumptions on the model of protein evolution, the proposed kernel has the same structure as the well-known local alignment kernel introduced by J.-P. Vert [8]. So, on one hand, we have found a probabilistic justification of Vert's local alignment kernels, and, on the other, an essential generalization of them is proposed, which embraces not only the local but also the global principle of sequence comparison and a number of other particular cases, which are specified by the choice of the parameters in the evolution model. We also show that the proposed evolution-based pair-wise similarity measure can be useful in the analysis of some difficult distant homology sets of proteins and help in computationally resolving situations in which other measures may fail.

The particular subclass of fixed-order kernels, which are based on alignments with only a fixed number of substitutions, attract a special interest. This kind of kernels may be very useful when the a priori information is available on the length of the unknown ancestor sequence.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research, Grants 08-01-12023 and 08-01-00695-a, and INTAS grant YSF 06-1000014-6563 to V. Sulimova for her visiting Birkbeck College in 2007-2008. The authors are grateful to Dr. P. Kellam of the Department of Virology UCL London for making the VIDA database contents available for the analysis and advising us on substantive matters. The authors are indebted to anonymous referees whose multiple comments helped us to improve the presentation.

References

1. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino-acid sequence of two proteins. J.of Molecular Biology 48, 443–453 (1970)
2. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. of Mol. Biol. 147, 195–197 (1981)

3. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402 (1997)
4. Pearson, W.R.: Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132, 185–219 (2000)
5. Mirkin, B., Camargo, R., Fenner, T., Loizou, G., Kellam, P.: Aggregating homologous protein families in evolutionary reconstructions of herpesviruses. In: Ashlock, D. (ed.) Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 255–262 (2006)
6. Rocha, J., Rossello, F., Segura, J.: The Universal Similarity Metric does not detect domain similarity. Technical Report, Quantitative Methods, Q-bio QM (2006), http://arxiv.org/abs/q-bio/0603007
7. Vinga, S., Almeida, J.: Alignment-free sequence comparison – A review. Bioinformatics 19, 513–523 (2003)
8. Vert, J.-P., Saigo, H., Akutsu, T.: Local alignment kernels for biological sequences. In: Scholkopf, B., Tsuda, K., Vert, J.P. (eds.) Kernel Methods in Computational Biology. MIT Press, Cambridge (2004)
9. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
10. Rangwala, H., Karypis, G.: Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics 21(23), 4239–4247 (2005)
11. Qiu, J., Hue, M., Ben-Hur, A., Vert, J.-P., Noble, W.S.: A structural alignment kernel for protein structures. Bioinformatics 23(9), 1090–1098 (2007)
12. Sun, L., Ji, S., Ye, J.: Adaptive diffusion kernel learning from biological networks for protein function prediction. BMC Bioinformatics 9, 162 (2008)
13. Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S.: Mismatch string kernels for dis-criminative protein classification. Bioinformatics 20(4), 467–476 (2004)
14. Haussler, D.: Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, UC Santa Cruz (1999)
15. Cuturi, M., Vert, J.-P.: A mutual information kernel for sequences. In: Proc. of IEEE Int. Joint Conference on Neural Networks, vol. 3, pp. 1905–1910 (2004)
16. Thorne, J.L., Kishino, H., Felsenstein, J.: An evolutionary model for maximum likelihood alignment of DNA sequences. Journal of Molecular Evolution 33, 114–124 (1991)
17. Miklos, I., Lunter, G.A., Holmes, I.: A "long indel" model for evolutionary sequence alignment. Molecular Biology and Evolution 21(3), 529–540 (2004)
18. Miklos, I., Novak, A., Satija, R., Lyngso, R., Hein, J.: Stochastic models of sequence evolution including insertion-deletion events. Statistical methods in medical research 29 (2008)
19. Metzler, D.: Statistical alignment based on fragment insertion and deletion models. Bioinformatics 19, 490–499 (2003)
20. Dayhoff, M.O., Schwarts, R.M., Orcutt, B.C.: A model of evolutionary change in proteins. Atlas of Protein Sequences and Structures 5(suppl. 3), 345–352 (1978)
21. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89(22), 10915–10919 (1992)
22. Sulimova, V., Mottl, V., Kulikowski, C., Muchnik, I.: Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families. In: DIMACS Technical Report 2008-16. DIMACS Technical Report 2008-16, Rutgers University, 17 p. (2008), ftp://dimacs.rutgers.edu/pub/dimacs/TechicalReports/TechReports/2008/2008-16.pdf
23. Mercer, T.: Functions of positive and negative type and their connection with the theory of integral equations. Trans. London. Philos. Soc. A 209, 415–416 (1999)
24. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. Wiley, New York (1990)
25. Mirkin, B.: Clustering for Data Mining: A Data Recovery Approach. Chapman and Hall/CRC, Boca Raton (2005)