Latent periodicity of serine–threonine and tyrosine protein kinases and other protein families

Andrew A. Laskin, Nikolay A. Kudryashov, Konstantin G. Skryabin, Eugene V. Korotkov

*Bioengineering Center of Russian Academy of Sciences, Prospect 60-tya Oktyabrya, 7/1, 117312 Moscow, Russia

Moscow Physical Engineering Institute, Kachirskoe Shosse 31, 115409 Moscow, Russia

Received 27 November 2004; received in revised form 18 April 2005; accepted 18 April 2005

Abstract

We identified latent periodicity in catalytic domains of approximately 85% of annotated serine–threonine and tyrosine protein kinases. Similar results were obtained for other 22 protein families and domains. We also designed the method of noise decomposition, which is aimed to distinguish between different periodicity types of the same period length. The method is to be used in conjunction with the method of cyclic profile alignment, and this combination is able to reveal structure-related or function-related patterns of latent periodicity. Possible origins of the periodic structure of protein kinase active sites are discussed. Summarizing, we presume that latent periodicity is the common property of many catalytic protein domains.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Periodicity; Alignment; Profile analysis; Protein kinase; Repeat; Domain substructure

1. Introduction

The development of mathematical techniques for investigation of symbolic sequences is now acquiring ever-growing importance, since nowadays large amounts of genetic information are being gathered (Benson et al., 2000; Stoeser et al., 2000; Adams et al., 2000; Venter et al., 2001). Certainly, there are hidden caches in those huge sets of sequences, but our present searching tools have limited sensitivity. The presence of novel ways to extract information from symbolic sequences would radically improve the ability to extract biologically significant knowledge from genetic texts, the understanding of gene evolution processes and evolutionary rearrangements of genomes, and also the ability to create dynamic models of cell’s genetic regulation and artificial proteins with predefined features.

One of the ways to investigate the features of a symbolic sequence is the investigation of its periodicity. The investigation of periodicity has reasonable biological meaning because multiple duplications of DNA sequence fragments followed by subsequent substitutions, insertions and deletions of symbols could serve as the ground for evolution of genes and genomes. This proposition may be validated by the fact that certain structure in genetic texts was previously revealed using various mathematical techniques applied to amino acid and nucleotide sequences (Shulman et al., 1981; Michel, 1986; Konopka et al., 1987; Tautz et al., 1986; Bell, 1996; Konopka and Martindale, 1995; Martindale and Konopka, 1996; Trifonov and Sussman, 1980; Zhurkin, 1983; Gabrielian and Bolshoy, 1999; Konopka and Chatterjee, 1988). The discovery of periodicity in active centers of enzymes could witness that, in the past, genes could be built up by simple repeating of certain relatively short DNA fragments. We may also suppose that such structure of protein active sites could mean possible participation of latent amino acid periodicity in choice and stabilization of the proper conformation of protein globule.

The techniques of dynamic programming (Heringa, 1994, 1998; Heringa and Argos, 1993; Benson, 1997, 1999; Heger...
Decomposition (ID) of symbolic sequences (Korotkov and searching for periodicity, which is based on the Information. We had developed our own mathematical approach to the decomposition of symbolic sequences (Korotkov and Korotkova, 1995; Korotkov et al., 1997, 2003; Chaley et al., 1999, Korotkova et al., 1999). The main idea of this approach is that information content of any symbolic sequence could be decomposed into mutually nonoverlapping constituents. Each of the constituents represents the mutual information between the investigated symbolic sequence and the artificial periodic sequence with some period length. The interdependence of mutual information and period length may be presented in the form of spectral graph that resembles Fourier power spectra, but it has substantially different properties (Korotkov et al., 2003). This decomposition allows us to eliminate the shortcomings peculiar to dynamic programming and Fourier transformation, and it allows us to detect so-called latent periodicity, that is, the periodicity that other techniques are powerless to detect.

However, like Fourier-based techniques, the method of information decomposition (in its current form) is not capable of finding statistically significant latent periodicity in presence of many insertions and deletions. This may lead us to the conclusion that substantial part of latent periodicity occurrences in genetic texts remains unseen by information decomposition-based techniques as well as by other known methods. The simplest method of searching for latent periodicity with insertions and deletions is the combination of information decomposition and modified profile analysis. In this combination, information decomposition can serve as the method that detects latent periodicity in some amino acid sequences and creates the periodicity matrix (Korotkov et al., 2003), which can be used to determine the weights for each amino acid at each position in the period. Then modified profile analysis allows us to identify periodicity of corresponding type (defined by cyclic position-weight matrix we have just constructed) in all the primary sequence data bank (such as Swiss-Prot) with possible insertions and deletions, and search results can be used to modify the profile matrix for increased sensitivity and specificity of the search.

The first goal of this publication is to present the method of noise decomposition. For many sequence families, perfect tandem periodicity is disrupted with indels; the cyclic alignment thereby is to expand our possibilities of identifying latent periodicities to the cases where sufficiently short indels are present. The method also allows us to distinguish between different periodicity types of the same period length because sometimes there are nearby but distinct types of periodicity. In this paper, we demonstrate that our decomposition technique is able to distinguish two latent repeat profiles as close as those present in serine-threonine and tyrosine kinases.

The second goal is to reveal the prevalence of latent periodicity in protein kinases. Latent periodicity we previously identified in catalytic domains of seven protein kinases turned out to be more common property of these proteins than one would expect. To achieve this, we applied modified iterative circular profile analysis and the method of noise decomposition; our efforts resulted in certain modification of the initial position-weight matrix and identification of latent periodicity in active sites of 1215 protein kinases presented in Swiss-Prot data bank. The data we gathered witness that latent periodicity is a property of at least a great majority of eukaryotic protein kinases.

The third goal of this publication is to show that there is latent periodicity in a number of different protein families. For this purpose, we applied the ID and noise decomposition methods to investigate some selected protein families from Swiss-prot data bank. We found out that amino acid sequences of many protein domains have latent periodicity with various period lengths. We discuss these results and propose that latent periodicity could reflect the origin of proteins from manifold ancient tandem duplications.

2. Methods and algorithms

Let us define at first, which kinds of periodicity we may call latent. Generally, the latent periodicity is periodicity that is hard to identify with proper level of statistical significance using internal homology search techniques. The homology between periods (repeats) is often determined using amino acid similarity matrices, such as PAM or BLOSUM (Benson, 1997, 1999; Heger and Holm, 2000, Andrade et al., 2000), where the weights for similar amino acids are higher than those for dissimilar ones, or using the autocorrelation function algorithm (Dodin et al., 2000).

Let us consider a set of sequences $S_1, S_2, \ldots, S^L$ of equal length $L$. We want to evaluate the overall similarity between these sequences; to do it, let us construct their (indel-free) multiple alignment:

$$
\begin{align*}
S^1_1 & \cdots S^1_n \\
S^2_1 & \cdots S^2_n \\
\vdots & \cdots \vdots \\
S^L_1 & \cdots S^L_n
\end{align*}
$$

The total weight of this multiple alignment is generally a sum of position weights:

$$
W = \sum_{i=1}^{L} W_i. \tag{1}
$$

Here $W_i = W_i(S^1, \ldots, S^L)$ is some function that designates the degree of similarity between amino acids $S^1_i, \ldots, S^L_i$. Traditional homology search techniques cal-
calculate this quantity via pair-wise amino acid affinities:

$$W_i = \sum_{\alpha} \sum_{\beta > \alpha} M(S_\alpha, S_\beta),$$

(2)

where $M$ is some amino acid affinity matrix, such as PAM or BLOSUM. This expression may be rewritten in the form of sum by amino acid types:

$$W_i = \frac{1}{2} \sum_{j,k} n_{i,j} (n_{i,k} - \delta_{j,k}) M(j,k),$$

(3)

where $j$ and $k$ are amino acid types and $n_{i,j}$ amino acid frequencies, i.e. the numbers of amino acids of type $j$ at position $i$. We (Korotkov and Korotkova, 1995; Korotkov et al., 1997, 2003; Chaley et al., 1999; Korotkova et al., 1999) proposed another measure of similarity, which is based on concepts of information theory and called “information content”:

$$W_i = \sum_{j} n_{i,j} \frac{n_{i,j}}{N_{fj}} \ln \frac{n_{i,j}}{N_{fj}},$$

(4)

where $f_j$ is the frequency of occurrence of amino acids of type $j$ in the whole set of sequences. These measures are undoubtedly different, thus an alignment may score high using information-theoretic measure while scoring low using homology-based measure, and vice versa. But the phrase “high-scoring” does not mean anything, especially when comparing weights calculated with different measures. We have to make sure that the obtained value of $W$ is far above those calculated with sets of random, unrelated sequences. To do it, we shuffle initial sequences and calculate either p-value or Z-value of obtained alignment; low p-values, or high Z-values, witness for significant similarity between sequences $S_1, S_2, ... , S_N$. One usually sets up some threshold value, which exceeding is believed to mean that the similarity is not arisen by chance.

When $S_1, S_2, ... , S_N$ are consecutive equal-length slices of the sequence under investigation, we may say that significant similarity between them means significant periodicity in this sequence. As we said before, different similarity measures result in different weights and different significance values; in some cases, periodicity of a sequence may be apparent from information-theoretical viewpoint while omitted by homology searches. In our studies, we call this effect latent periodicity. In our previous studies (Korotkov and Korotkova, 1995; Korotkov et al., 1997, 2003; Chaley et al., 1999; Korotkova et al., 1999) we have shown that such latent periodicity is present in many sequences of biological importance.

Let us illustrate this with an example. Assume that the latent period is seven symbols long, and there is equal probability to encounter the amino acids shown in Table 1 for each position of the period, e.g. letters A, R, N, D, C, Q, E, Z, X, W, Y, B are equiprobable at positions $(1 + 7N)$, and other letters are absent at these positions. This table shows the artificial matrix used to represent the concept of latent periodicity. One of possible sequences satisfying the conditions in the table is:
Let us assess the probabilities \( \alpha \) (of meeting such, or better, homology between repeats by chance) and \( \beta \) (of obtaining the observed, or higher, value of mutual information by chance) for this sequence using the Monte-Carlo techniques. For this purpose we calculated the alignment score \( W \) from the Eqs. (1)–(3) using the BLOSUM50 affinity matrix, as well as the mutual information of the sequence (4) (Korotkov et al., 2003). Then we generated 200 random shuffles of the sequence with the same amino acid composition. For each of these sequences, the values of similarity score \( W \) from (1) and mutual information were obtained. Then we calculated the means and the variances of both \( W \) and mutual information, and we obtained \( Z \)-scores using the formula (15).

For our sample sequence, the calculated \( Z \)-values were equal to 2.96 and 6.5, respectively. That is, the probability \( \alpha \) is roughly estimated as 0.05 and the probability \( \beta \) is roughly estimated as \( 10^{-9} \). The shown values of \( \alpha \) and \( \beta \) illustrate that homology-based algorithms of searching for repeats are unable to identify such periodicities at a statistically significant level.

2.1. Analysis of the Swiss-Prot data bank

The concept of searching for latent periodicity with indels is graphically represented in Fig. 1. It should be noted that the initial periodic profiles for iterative searching are taken from our database of latent periodicities found in Swiss-Prot with ID (Korotkova et al., 1999). The information decomposition method had identified more than 12,000 subsequences with significant periodicity of various lengths and types (we refer to amino acids in protein sequences as symbols and to amino acid sequences as symbolic sequences because information decomposition technique is regardless of what type of symbolic information is to process—nucleotides, amino acid residues or even written text). Nearly 20% of the found amino acid sequences contained homologous periodicities, i.e. those periodicities had values of \( \alpha \) less than \( 10^{-6} \), and homology search techniques are thereby enough sensitive to identify them. We excluded those cases from our analysis because RADAR (Heger and Holm, 2000) and REPRO (George and Heringa, 2000) have been good tools to investigate them. We only consider a few cases of homologous periodicity in this paper in order to perform comparative analysis, while concentrating on non-homologous periodicities with \( \alpha > 5 \times 10^{-3} \) and \( \beta < 10^{-6} \).

Fig. 1. The framework of searching for latent periodicity using the iterated profile analysis.
If we create a new profile from the results of searching in the data bank and repeat this process, we may thereby perform the iterative refinement of the initial profile. To make this possible, the iteration method should be asymptotically stable, i.e. after a certain number of iterations the new profile should be a good match to the previous one. This condition virtually defines the choice of the formula to fill the position-weight matrix, since there is the theorem that specifies the connection between the asymptotic frequencies of symbols in high-scoring segments (namely, the results of searching a data bank) and the scoring scheme (Karlin et al., 1990). It may be written in the form:

\[ W_{i,j} = C \ln \frac{p_{i,j}}{\bar{r}_{i,j}} \]

(5)

where \( W_{i,j} \) is an element of the position-weight matrix for the symbol of type \( i \) at position \( j \), \( p_{i,j} \) is the fraction of the symbol of type \( i \) occurring at position \( j \) within the high-scoring segments, and \( \bar{r}_{i,j} \) is the frequency of occurrence of symbols of type \( i \) in scanned sequence, namely, in the data bank. Of course, pseudo-counts should be used in this formula where appropriate, in order to avoid logarithm errors. The scaling parameter \( C \) in the formula may be arbitrarily chosen (multipling all the weights and the scores by a factor changes neither the path of the alignment nor its statistical significance, provided that indel costs are multiplied by the same factor), so we can choose it large enough to round the values of weights to integers without substantial loss of precision, in order to speed up the computation.

From the viewpoint of the information theory, the problem of the calculation of new profile from the results of profile analysis is similar to that of the extraction of useful signal from a noisy channel (Schmidt, 1998; Wilbur and Neuwald, 2000). It is clear that certain patterns corresponding to frequent structural and functional features of proteins are over-represented (compared to what we could expect from their length and amino acid composition). Since profile searching is very sensitive, in most realistic cases some of the representatives of these structural units could be found even with a distantly related profile. If we include them in the multiple alignment to derive new profile, they will skew the profile into the direction of that over-represented group. It is likely to lead to a subsequent sharp increase of such representatives, and the initial motif would be completely changed (Altschul and Koonin, 1998). If we are searching for common profile for a family of proteins or domains, it may result in unwanted effects known as “profile wandering”. Moreover, it makes automation of successive iteration process virtually impossible—one has to choose “correct” scan results by hand in order to avoid profile “pollution”. When the number of results runs into the hundreds and thousands, “time-consuming” turns into “impossible”.

Information theory may be a way to circumvent this problem. The main idea of information-theoretic approach is to identify certain noise patterns in noisy signal, and then to separate signal from noise being aware of these patterns. For example, if we observe the excess of noise at some frequencies, these frequencies should be most suppressed. This principle is really working in noise reduction systems. To set apart from “genetic noise”, we propose the probabilistic consideration of sequence set that we call “noise decomposition” (ND). Let us represent this set as a mixture of “uncorrelated noise”, composed of sequences having no homology to the proteins under investigation and characterized by “background” symbol probabilities \( f_i \) and “correlated noise”, composed of sequences generally unwanted but having sufficient level of homology to be found with profile analysis together with the desired sequences. For our purposes, the main difference of correlated noise is that the distribution of the numbers of symbol occurrences over different positions of profile alignment is not casual because we propose the existence of a pattern (or patterns) of these false positives. The resulting noise will apparently be position-specific, let us denote its distribution as \( \pi_{i,j} \).

It is clear that correlated noise may be not of single type, but it may be some composition of a few different types. We assign a distribution to each of these types and call it \( q_{c_k} \), where the upper index corresponds to different constituents of the correlated noise, composed of different types of false positives. Then the correlated noise can be expressed by this formula:

\[ \pi_{i,j} = c_0 f_i + \sum_{k=1}^{N} c_k q_{c_k}^{t_{i,j}} \sum_{k=0}^{N} c_k = 1, \]

(6)

according to it, we rewrite the expression for calculation of the elements of the position-weight matrix (5) as:

\[ W_{i,j} = C \ln \frac{p_{i,j}}{\bar{r}_{i,j}} \]

(7)

Here \( c_k \) are “relative magnitudes” of different types of noise, in other words, their contribution into the resulting noise. The normalization in (6) is required to hold this condition:

\[ \sum_{i \in I} \pi_{i,j} = 1 \]

(8)

since \( \pi_{i,j} \) re probabilities. In order to define \( c_k \), we will weight different types of noise as follows. Weighting coefficients \( c_k \) should be proportional to the relative fraction of the corresponding type of noise in the total input signal, i.e. if the source bank contains \( M_1 \) sequences with correlated noise type 1, \( M_2 \) sequences with correlated noise type 2 and so on, and also \( M_0 \) sequences without correlation with our pattern, then the relative contribution of these noises into the resulting noise will be proportional
to these numbers, i.e.

$$\frac{c_0}{M_0} = \frac{c_1}{M_1} = \frac{c_2}{M_2} = \ldots$$ \hspace{1cm} (9)

However, in fact, the numbers $M_0, M_1, M_2, \ldots$ are unknown. We will thereby estimate them from the results of initial search. Since the insignificant subsequences should be treated as noise, we will take into account only statistically significant results (see Eq. (15)). If the search results contain $N_1$ sequences with correlated noise type 1, $N_2$ sequences with correlated noise type 2 and so on, we assume that

$$\frac{c_0}{N_0} = \frac{c_1}{N_1} = \frac{c_2}{N_2} = \ldots$$ \hspace{1cm} (10)

Notice that $N_0$ could not be estimated this way (if we do that, $c_0$ will be set too low, and the new profile (7) will be distorted; our experiments demonstrated that the iterative process diverges under this condition). Our experiments show that 0.75 is a good estimate for $c_0$ in most cases (see later).

We calculated $r_{ij}$ via pair-wise global alignment of all sequences in true positive set; let the alignment score of sequences $i$ and $j$ be $S(i, j)$. These values were utilized to calculate $T(k)$, which stands for prevalence of $k$-like sequences in true positive set:

$$T(k) = \sum_j \max \left( 0, S(i, k) \right)$$ \hspace{1cm} (11)

Index $j$ runs through all the set of true positives. The meanings of terms in the sum (11) are: the term with $j = k$ always equals 1 (any sequence is self-similar), terms from unrelated sequences are zeroes, and terms from similar sequences range from 0 to 1. That is, we get $T(k) = 1$ in case that there are no sequences similar to $k$; we get $T(k) = N$ if all $N$ sequences in the true positive set are identical; we get a value from 1 to $N$ if the sequences are similar, depending on the level of similarity. Then we calculate the values of $r_{ij}$:

$$r_{ij} = \frac{\rho_{ij}}{T(k)}$$ \hspace{1cm} (12)

where $\rho_{ij}$ are the frequencies of occurrence of symbol $i$ at position $j$ in the cyclic alignment of sequence $k$. Thus we eliminate possible over-representation of some sorts of sequences in Swiss-Prot data bank.

2.3. Iterated profile analysis

Iterations were performed as follows. First of all, we used the initial matrix of periodicity determined via the algorithm described in Section 2.1, and we calculated the $W_{ij}$ using (5).

Then we used this position-weight matrix in the procedure of cyclic alignment, which will be described in Section 2.4 of the paper. Cyclic alignment of all amino acid sequences in Swiss-Prot was carried out, and all statistically significant results ($Z > 6.0$) were gathered.

After that, we divided these results into classes, namely, true positives and false positives of different classes, for the formulae (6) and (7) to be applied. Two ways of division into classes were tried out, namely, keyword analysis and clustering.

We worked with the Swiss-Prot data bank, which is curated and contains extensive information about protein entries. We extracted information about identified proteins from their descriptions (DE field), keywords (KW field) and feature tables (FT field) (Junker et al., 1999). We divided the results into three classes using keywords. Class 1 comprised the proteins we wanted to identify, and classes 2, 3, 4, \ldots comprised the groups of proteins we wanted to filter off. Then we considered the class 1 as true positives and other classes as different types of correlated noise. This approach is applicable when proper information is presented in Swiss-Prot or any other database.

If such information is absent in a data bank, clustering may be used to split the proteins into classes. We performed the clustering experiment for protein kinases as follows. First, we made the pair-wise comparisons of identified subsequences using the global alignment technique (Needleman and Wunsch, 1970). Then we built the distance matrix between those sequences using the formula:

$$\text{Distance}(A, B) = \frac{\text{AlignmentScore}(A, A) + \text{AlignmentScore}(B, B))}{2 - \text{AlignmentScore}(A, B)}$$

This distance matrix was used in single linkage cluster analysis, and the threshold of merging was adjusted to output two large classes. Then we checked whether these clusters are related to the two types of protein kinases and found incomplete correlation (about 90%), i.e. there were both serine-threonine and tyrosine kinases in each cluster, but the separation was about 90%. We concluded that the information in Swiss-Prot is more reliable than clustering, and it should be used when possible.

After the formation of the classes we calculated the values of $p_{ij}$ and $q_{ij}$, which are simply positional residue frequencies for the corresponding classes, as well as the values of $N_i$, the cardinalities of the classes. We let $f_i$ equal to the residue frequencies in Swiss-Prot. Then we were able to apply the formulae (11), (10), (6), and (7) consecutively in order to obtain the new values of $W_{ij}$, i.e. the new cyclic profile matrix.

Using the new $W_{ij}$ we repeated the search of the Swiss-Prot data bank according to Section 2.4 and obtained the new set of amino acid sequences with $Z > 6.0$. After that, we repeated the procedure of selection of true and false positives as shown above in this section and recalculated $W_{ij}$ again. The iteration process was carried out until the set of results after some iteration was virtually the same as the set before that iteration, that is, the identity of two sets was more 95%.

Our experiments showed that 3–5 iterations were enough in most cases.
2.4. Cyclic alignment and statistical significance

It is clear that the noise decomposition technique is quite applicable to conventional (linear) profiles (Gribskov et al., 1987), but in the present study it was used for periodic profile training aimed to investigate latent periodicity of protein sequences (Korolkov et al., 1999; Korolkov et al., 2003). The reasoning given above is valid, and the expressions (5)–(12) hold in both cases. The problem of finding a good cyclic profile is much more sophisticated than the problem of finding a linear profile, since there is a bunch of diverged copies (repeats) of some pattern within a sequence, instead of one copy. Internal divergence of repeats superimposes on the divergence between sequences; hence, cyclic patterns are rather feebly marked. We have made a modification of the algorithm presented in (Fischetti et al., 1992), which is called locally optimal cyclic alignment (Laskin et al., 2003).

What is cyclic alignment? The conventional alignment is matching of a sequence to some pattern, for example, “QWERTY”. The cyclic alignment is then matching of that sequence to virtual periodically elongated pattern “…QWERTYQWERTYQWERTY…” of course, in most cases only a part of the sequence will be matched (i.e. cyclic alignment will be local). Our main idea is to present cyclic alignment in the form of a path that connects the nodes of a two-dimensional cylindrical lattice, where one of the coordinates corresponds to position in (linear) sequence, and another (cyclic) corresponds to position in the cyclic profile (compare to conventional sequence alignment, which can be presented in the form of a path between the nodes of a flat two-dimensional lattice, coordinates being the positions in the compared sequences).

This path contains diagonal steps, which describe matching of a symbol from the sequence and a position of the profile, as well as steps along the axes, which describe insertions or deletions. Every path of such kind has a total score, which is the sum of gap costs and weights of symbol-to-position matches (see Laskin et al., 2003) for details and figures).

The optimal cyclic alignment is the path with the highest possible total score. We have shown that it can be found by means of cell-by-cell filling of the similarity matrix $S_{ij}$, in which one of the indices (for instance, $i$) is cyclic or wrapped, namely, $S_{i+Lj} \equiv S_{i-Lj} \equiv \ldots \equiv S_{i+j}$, where $L$ is the period length (Laskin et al., 2003), just as we find the best linear alignment using the Smith–Waterman formula (Smith and Waterman, 1981). The formulae for recursive filling of $S_{ij}$ are:

$$S_0 = \max \{S_{0,j}, \max_{1 \leq a < L} [S_{a,j} - d_i]\}, \quad (13)$$

$$S_j = \max \{0, S_{j-1} + u_0, \max_{1 \leq a < L} [S_{a+j-k} - d_i]\} \quad (14)$$

Here $u_0$ is the weight of the jth symbol in the sequence, at the jth position in the profile, it is the element of matrix $W$ from (5) or (7)), $d_i$ is the gap penalty for insertion/deletion of i successive symbols. As usual, to find the optimal local alignment we have to identify the highest element of $S$-matrix and recreate the path from it down to the first zero element. The value of the highest element is the total score of the optimal local alignment; this is the value we utilized to check whether the alignment is statistically significant.

The Monte-Carlo technique was used to assess the statistical significance of alignments (Chaley et al., 2003). The assessment was performed separately for each sequence, taking into account its length and composition. To assess statistical significance of an alignment in this study, we aligned a number of random sequences with the same length and composition as the real sequence (in order to avoid composition bias effects) against the same cyclic profile. Then we could calculate the mean and the variance of those random scores and estimate the $Z$-value of the obtained alignment, assuming that the distribution of the obtained weights is normal (Webber and Barton, 2001; Chaley et al., 2003):

$$Z = \frac{S_{\text{real}} - E(S)}{\sqrt{D(S)}} \quad (15)$$

The threshold value of $Z$ was chosen to be equal to 6.0. Our numerical experiments showed that we are unlikely to observe $Z$-values greater than 6.0 in a random test sequence set with the total number of symbols equal to the total number of symbols in Swiss-Prot.

The described algorithms were implemented in C++, and they are available upon request by e-mail (for details visit our Web site http://bioinf.narod.ru/periodicity).

3. Results

First of all, we would like to demonstrate the consistency reliability and the usefulness of our techniques in searching for known types of tandem repeats. Ankyrin and leucine-rich repeats were chosen for this purpose. The initial profiles of these repeats were obtained using ID with the period lengths equal to 33 and 24 residues, correspondingly.

We identified 146 of 150 sequences containing at least three marked ankyrin repeats (three times the period length was chosen to be minimal length of latenly-periodic subsequence). Needless to say, identified periodic subsequences covered the marked repeats. Scanning the rest of the Swiss-Prot data bank, we identified additional 57 sequences, in which less than three ankyrin repeats were marked by conventional techniques. Hence, we identify more repeats than conventional techniques are able to identify. The only false positive, hypothetical protein P50938 was identified because of perfect tandem periodicity of another type in it.

We also identified 261 of 270 sequences with leucine-rich repeats with no false positives. As in the previous case, we identified many additional leucine-rich repeats in those sequences. For example, in the protein P09661 (U2 small nuclear ribonucleoprotein) we identified the additional fourth repeat, which is unseen by conventional techniques.
Table 2

The protein kinases where latent periodicity without insertions and deletions was initially found

| Swiss-Prot ID | First position | Last position | Protein description in Swiss-Prot |
|---------------|----------------|---------------|----------------------------------|
| KDBE          | 400            | 478           | Putative serine-threonine-protein kinase C22E12.14C (EC 2.7.1.). |
| KEMK          | 85             | 181           | Putative serine-threonine-protein kinase EMK (EC 2.7.1.). |
| KPC1_ASPNG    | 954            | 1056          | Protein kinase C-LIKE (EC 2.7.1.). |
| KPC1_RA       | 526            | 565           | Protein kinase C-ETA type 2.7.1.- (NPCK-ETA) (PKC-4). |
| MSKA_HUMAN    | 97             | 181           | Mitogen-activated protein kinase 10 (EC 2.7.1.) (mixed lineage kinase 2) (protein kinase MST). |
| CC22_XENLA    | 85             | 148           | Cell division control protein 2 homolog 2 (EC 2.7.1.-) (P34 protein kinase). |
| CC2_CARAU     | 88             | 160           | Cell division control protein 2 homolog (EC 2.7.1.-) (P34 protein kinase) (cyclin-dependent kinase 1) (CDK1). |

We checked it with Pfam, SMART, as well as with dedicated repeat-finding software, namely, REP (Andrade et al., 2000), REPRO (George and Heringa, 2000), and RADAR (Heger and Holm, 2000). The 3D-structure of this protein is known, and the presence of the fourth repeat is evident. Another example is the protein P16473 (thyroid stimulating hormone receptor), where we identified 10 leucine-rich repeats, while other techniques identify no more than six repeats. The modeled 3D-structure of this protein testifies that the identified region consists completely of LRR repeats. Hence, the usage of ID and cyclic alignment techniques may lead to new discoveries even in well-known repeat families. We also identified additional reliable repeats in many other known repeat families, but we were not concentrated on those studies (the reason is explained in Section 2.1).

In our present study we initially identified seven protein kinase amino acid subsequences having latent periodicity with the period length equal to 18 residues in the absence of insertions and deletions (Table 2 and Fig. 2). All of them were serine–threonine protein kinases except MSKA_HUMAN having dual specificity. The periodicities were located in the protein kinase catalytic domains (in a number of cases, we also encountered periodicity with period lengths of multiples or divisors of 18), and the periodicity profiles were similar. This allowed us to suggest that the periodicity of 18 amino acids is a characteristic property of protein kinase active sites. We had set a problem to find out, first, to which extent this periodicity is peculiar to protein kinase active sites if insertions and deletions are permitted, and, second, whether serine–threonine and tyrosine protein kinases share the same periodicity pattern, and whether we can improve the sensitivity and the specificity of search using separate periodicity profiles for different kinase types instead of one.

We averaged the amino acid occurrence frequencies in different profile positions of all the seven cases mentioned above and made the initial position-weight matrix using the formula (2). This matrix was used to scan Swiss-Prot release 41 (Bairoch and Apweiler, 2000) using cyclic profile alignment. A number of values of gap-opening and extension costs were tried out, taking into account sensitivity and specificity.
Total number of protein kinases having latent periodicity. In this study we demonstrate the presence of latent periodicity in 22 families in addition to protein kinases. Latent periodicity is identified in more than 75% of members of each family. The list of these families is shown in Table 6. The corresponding profiles and search results are presented in Table 3. The authors’ Web site at http://bioinf.narod.ru/periodicity.

Further application of the cyclic profile analysis to the latent periodicities identified in protein sequences with the ID technique allowed us to identify other protein families having latent periodicity. In this study we demonstrate the presence of latent periodicity in 22 families in addition to protein kinases. Latent periodicity is identified in more than 75% of members of each family. The list of these families is shown in Table 6. The corresponding profiles and search results are presented at http://bioinf.narod.ru/periodicity/new.

We also counted the latently-periodic Swiss-Prot sequences that did not belong to the corresponding families. Their numbers did not exceed 7% of the numbers of proteins in corresponding families, and they were equal to zero for two thirds of them. We propose that the presence of such sequences may be caused by imperfect annotation of the Swiss-Prot data bank and possible additional functions of protein kinases. Latent periodicity is not peculiar to the families they were found in.

As a result of this initial scanning, we obtained about 100 statistically significant periodic subsequences from both serine–threonine and tyrosine protein kinases. It is a fact that catalytic domains of serine–threonine and tyrosine protein kinases have fairly homologous primary sequences as well as similar 3D structure (Taylor et al., 1995). Hence, we decided to form two periodic profiles according to these two types of protein kinase catalytic cores. To do it we divided the results into classes and formed two new position-weight matrices using (6)–(12), so that in one case serine–threonine protein kinases were considered to be true positives, and tyrosine protein kinases were considered to be a kind of correlated noise, and vice versa for the other matrix. Noise decomposition resulted in two position-weight matrices, which were later optimized (trained) with iterative searches in order to find the highest number of members of each family. The list of these families is shown in Table 6. The corresponding profiles and search results are presented at http://bioinf.narod.ru/periodicity/new.

Further application of the cyclic profile analysis to the latent periodicities identified in protein sequences with the ID technique allowed us to identify other protein families having latent periodicity. In this study we demonstrate the presence of latent periodicity in 22 families in addition to protein kinases. Latent periodicity is identified in more than 75% of members of each family. The list of these families is shown in Table 6. The corresponding profiles and search results are presented at http://bioinf.narod.ru/periodicity/new.

The authors’ Web site at http://bioinf.narod.ru/periodicity.

We also counted the latently-periodic Swiss-Prot sequences that did not belong to the corresponding families. Their numbers did not exceed 7% of the numbers of proteins in corresponding families, and they were equal to zero for two thirds of them. We propose that the presence of such sequences may be caused by imperfect annotation of the Swiss-Prot data bank and possible additional functions of hypothetical proteins.

Table 3
Noise decomposition of the protein kinase superfamily

| Profile | S/T kinases | Y kinases | Other proteins |
|---------|-------------|-----------|----------------|
| Initial profile with ID | 63 | 41 | 18 |
| Initial profile, after five iterations with test set of S/T kinases, no noise decomposition performed | 615 | 109 | 58 |

Results of the analysis of the Swiss-prot data bank (release 41) for the presence of amino sequences with latent periodicities of the corresponding types

Table 4
Results of the analysis of the Swiss-prot data bank (release 41) for the presence of amino sequences with latent periodicities of the corresponding types

| Type of periodicity profile | Serine–threonine-protein kinases | Tyrosine-protein kinases |
|----------------------------|----------------------------------|--------------------------|
| Total number of protein kinases present in Swiss-Prot release 41 | 1116 (62 of them are dual-specificity kinases) | 348 |
| Total number of protein kinases having Z more than 6.0 | 903 | 312 |
| False positives | 37 | 11 |
| Another type protein kinases found among false positives | 4 (tyrosine kinases) + 1 (unknown kinase) | 4 (serine–threonine kinases) + 6 (tyr kinase-like) |
Table 5

| Accession number |  
|------------------|------------------|
| **Position range** | **Alignment** |
|  
| Serine–threonine-protein kinase CBK1 | P53894 | 10.7 | 449–610 |
| Serine–threonine-protein kinase, cGMP-dependent protein kinase, isozyme 1 | Q03042 | 16.4 | 505–655 |
| NT-3 growth factor receptor precursor, TrkC tyrosine DE kinase | Q16288 | 16.3 | 663–764 |
| Tyrosine-protein kinase transforming protein YES | P00527 | 16.4 | 370–457 |

The upper sequences of the alignments show the positions in the 18 amino acid period, and the lower sequences are amino acid subsequences with latent periodicity. Corresponding periodicity matrices are shown on the web site http://bioinf.narod.ru/periodicity.

4. Discussion

The notion of latent periodicity and the technique of searching for it was initially presented in (Korotkov and Korotkova, 1995) and refined in subsequent investigations (Korotkov et al., 1997, 2003; Chaley et al., 1999; Korotkova et al., 2003). These identified latently-periodic sequences belong to proteins of various types, with various period lengths and periodicity patterns. However, the question of functional significance of the identified latent periodicities and its correlation with protein structures remained open. To a great extent, this resulted from the information decomposition technique, which is incapable of revealing latent periodicity interrupted with insertions and deletions; so, this approach omitted a substantial subset of proteins with certain functional domains, and no inference about the relationship between latent periodicity and protein functionality could be made. This paper presents a pioneer work that shows the existence of latent periodicity in whole protein families. In fact, achieving this result required the development of the iterative mathematical method and the software capable of finding statistically significant latent periodicities of a predefined type (found by the ID technique) in presence of insertions and deletions.

The applied techniques of noise decomposition and iterative analysis do not mean that the periodicity we identified in protein kinases and other protein families is not a latent periodicity anymore. This can be seen from Fig. 3, where information decompositions of a few protein kinase amino acid sequences after their cyclic alignment to the corresponding periodicity profiles (i.e., with proper indels in these sequences) are presented. The periodicity of 18 residues is evident from the spectrograms of these information decompositions (graphs on the left), though it is absent in the original subsequences (graphs on the right). Similar decomposition spectra may be obtained from all the other protein kinase amino acid sequences (as well as in sequences from other protein families aligned with corresponding profiles) after noise decomposition and iterative analysis.

The existence of latent periodicity in protein families conforms with the results of other authors who studied the structure of genetic sequences by means of information-theoretic (Salamon and Konopka, 1992; Salamon et al., 1993; Konopka, 2001; Allison et al., 2000; Konopka and Owens, 1990) and statistical (Konopka, 1994, 1997, 2003b; Konopka and Smythers, 1987; Guibas and Odlyzko, 1981; Gentleman and Mullin, 1989; Breen et al., 1985) approaches. Those investigations have demonstrated the prevalence of certain nucleotide or amino acid patterns and distributions in sets of DNA and protein sequences, and this serves as a ground for our abilities both to do the Noise Decomposition and to reveal subtle periodic signals in genetic texts.
Table 6
Protein families where latent periodicities with insertions and deletions were found

|   | Protein family                      | Period length, aa | Number of found proteins with latent periodicity |
|---|-------------------------------------|-------------------|--------------------------|
| 1 | Ankyrin repeats                     | 33                | 203                      |
| 2 | Leucine-rich repeats                 | 24                | 261                      |
| 3 | Acetoacetate synthases               | 25                | 28                       |
| 4 | Granulo-bound glycogen synthase      | 41                | 11                       |
| 5 | Tra synthetases                     | 17                | 71                       |
| 6 | Acyl and adler transforases, EIF     | 6                 | 136                      |
| 7 | Potassium channels                  | 25                | 55                       |
| 8 | Acyl transferring synthases, active site | 25                  | 124                      |
| 9 | Dithiobostin synthases               | 30                | 15                       |
| 10| Various GTP-binding proteins        | 23                | 270                      |
| 11| Tryptophan 2-monooxygenases         | 28                | 6                        |
| 12| Various CoA-related proteins, maybe CoA-binding site | 30            | 42                       |
| 13| Asparyl proteases                   | 41                | 54                       |
| 14| MHC-I Antigens                      | 12                | 148                      |
| 15| Methyltransferases                  | 31                | 30                       |
| 16| ACC-oxidases                        | 32                | 17                       |
| 17| Polymerase core                     | 13                | 103                      |
| 18| Interleukin 12, growth factors      | 14                | 12                       |
| 19| ATP synthases, active site          | 14                | 82                       |
| 20| Hemoglobin                          | 15                | 92                       |
| 21| Hodgepodge proteins                 | 16                | 59                       |
| 22| Actins                              | 17                | 256                      |
| 23| Lyases                              | 10                | 428                      |
| 24| Pyridal phosphate binding proteins  | 16                | 76                       |

The first two cases are the examples of homologous periodicity identified with our techniques. Cyclic alignments against the corresponding cyclic profiles are shown at the web site http://bioinf.narod.ru/periodicity/new.

Let us discuss the existing limitations of our current abilities to identify periodicities. First, it should be noted that in the present study cyclic alignment and noise decomposition techniques were used in conjunction with the information decomposition technique that provided the initial periodicity profiles (see Section 2.1). It is still not capable of handling indels, so we may omit a number of (maybe homologous) periodicity families with indels being present in all sequences. Nevertheless, searching for homologous tandem repeats is rather investigated problem, and it was not the aim of our investigations. We concentrated on searching for feebly marked periodicities, and the conjunction of the ID and ND techniques is the only method to study them so far. One may also notice that the noise decomposition method itself can be applied to study any profile, not only cyclic one. From this viewpoint, it may be considered as an independent technique.

Second, the necessity to find initial periodicity types without insertions and deletions also affected our choice of maximal possible size of indels; it was limited to the period length. If we allow for larger deletions, we would find many linked dispersed repeats in different parts of sequences instead of periodicity in domains, which is of our interest. It is possible that there are proteins in some investigated families, in which such large indels occurred nevertheless; in that case, we were unable to identify them. But we did that for the sake of specificity.

Third, our numerical experiments have demonstrated that the used estimation of statistical significance (15) is valid only if the length of alignment is greater than 50 (Chaley et al., 2003). It means that the cyclic profile search technique is capable of identifying periodicity ranging along 50 or more amino acid residues; short dispersed repeats could not be identified. But it was abovementioned that we avoided dispersed repeats in our searches. Moreover, the ID technique can be good at finding periodicities only if at least three consecutive repeats are present, and we developed the ND technique with this limitation of minimal length of latently-periodic subsequences (i.e. 54 residues for protein kinases) in mind. That is, if only a part of the protein kinase catalytic domain is conserved, our software would not identify it; however, such scrap of the domain is unlikely to be functioning.

We have made the comparative analysis of the power of our method, in which we tried to reveal the periodicity of protein kinases and other latently-periodic protein families (Table 6) with RADAR (Heger and Holm, 2000) and REPRO (George and Heringa, 2000) software. Both failed to do that.

In this paper we did not pose a problem to reveal all existing cases of latent periodicity in protein families. It is a rather laborious problem, and it requires great computational power. We focused our efforts to demonstrate that we are able to identify periodicities in most of proteins in some families, in which only a few members were initially found to be latently-periodic. The periodic structure of all these proteins and domains was not described before.

Let us discuss specifically the latent periodicity of protein kinases, its possible meaning, and its relationship with the spatial organization of these proteins. Protein kinases, i.e. enzymes whose function is to transfer phosphate residues
Fig. 3. Information decompositions of serine–threonine and tyrosine protein kinases after cyclic alignment against the corresponding profile (denoted by ones) and before it (denoted by twos). The cyclic alignments of these sequences are shown in Table 5. Accession numbers are: A, P53894; B, Q03042; C, Q16288; D, P00527.
from ATP to other proteins, are known to provide an important role in cell signaling. There are many subfamilies of protein kinases with internal homology of 90% and higher. Mutual homology between subfamilies is much weaker, usually about 30%. Two classes of protein kinases are structurally very similar, serine–threonine and tyrosine kinases (according to the phosphorylated residue). In addition, there are also exist protein kinases with dual specificity (Kemptrup et al., 1996). There are also other types of protein kinases structurally dissimilar to these kinases, and they phosphorylate other residues. They do not share the types of periodicity described in this paper, nor were they found to have any other common periodicity pattern. We think they could have another type of latent periodicity, or the presence of substantial deletions and insertions in their amino acid sequences makes it impossible to find latent periodicity in protein kinases of those types using the noise decomposition technique.

As is well-known (Hanks et al., 1988; Hunter, 1991), the catalytic domain of protein kinases, where our periodic subsequences reside, could be divided into 12 subdomains, on the one hand, highly evolutionarily conservative and, on the other hand, related to the elements of 3D-structure (Taylor et al., 1992; Goldsmith and Cobb, 1994). Subdomains I-IV are responsive for ATP binding and form antiparallel β-sheets. Subdomains VIa-XI bind the substrate and initiate the phosphate ion transfer. They are a bit more variable (presumably to provide substrate specificity) and composed of mostly α-helices.

Subdomains alternate with less conservative sites that usually form loops, with the period of these alternations being close to 18 residues. Using our protein kinase periodicity profiles, in various protein kinases we find periodic sites, about 100 residues long, located in subdomains VIb, VII, VIII, and IX. These subdomains contain functionally important features such as the catalytically active asparatic acid residue in subdomain VIb and the activation loop between subdomains VII and VIII. Many amino acid residues within these subdomains are of critical importance for proper folding and functioning of the active center. These are the aspartic acid residue mentioned above, the valine residue that interacts with ATP adeinne, the lysine residue that interacts with phosphate ion, the asparagine and asparatic acid residues in subdomain VII that retain inhibiting and activating Mg ions, the asparatic acid residue in subdomain IX that stabilizes the catalytic loop, and also a few other residues that provide ionic bonds and regulate enzyme activity, being subject for phosphorylation or autophosphorylation (Taylor and Radzio-Andzelm, 1994). Notice that, for example, the aforementioned asparatic acid residues are separated from each other by 18 and 36 residues (the numbers are given for cAMP-dependent mammalian protein kinase A, which is usually a model enzyme for studying kinase structure; we obtained similar values for other proteins), so they are both placed at the same position in the periodic repeat (namely, position 2), although their functions are different. Therefore, we see that the period length is close to a subdomain. To make sure we compared cyclic profile alignment of protein kinase A (Swiss-Prot accession P05132) with its subdomain structure. It turned out that subdomain borders are located at period positions 14 and 15. Hence, there is a clear relationship between periods and subdomains.

It was previously proposed (Kruse et al., 1997; Muller et al., 1999) that tyrosine protein kinases were evolutionarily derived from serine–threonine ones by means of isolation of catalytic domain nucleotide sequences by insertion of introns and subsequent pasting of these mobile elements, slightly altered with mutations, into some other proteins with kinase activity, called “ancestral kinases”. This would result in greater variability of catalytic domain lengths in tyrosine kinases, because there are no gap restrictions for mobile elements. Our results favor that hypothesis, because we found out that insertions and deletions occur almost two times more frequently in tyrosine kinases than in serine–threonine ones (on average, 5.96 versus 3.05 insertions and deletions per site), i.e. we observe larger deviations from perfect periodicity.

At this time the exact origin of the observed latent periodicities presented in this paper is not clear, however, we can propose a few possible explanations of this phenomenon. First, we may suppose that catalytic domains were initially much smaller than what we observe now. However, they were able to duplicate, and the duplications were properly arranged to form even more catalytically active domain. It is a fact that DNA sequence repeats facilitate replication errors at their location, thereby promoting new tandem repeats. We suppose that, as the number of repeats grew, the ancestor protein benefited, i.e. its catalytic activity and structural stability increased. Subsequent mutations formed even better packed structure of these domains and fine-tuned the functionality, at the same time the mutations resulted in the diffusion of periodicity, the loss of homology between distinct repeats. That means that we may call the residual internal similarities “echoes” of the ancient processes of protein formation (Ohno, 1970, 1984; Ohno and Epplen, 1983).

Latent periodicity may be also involved in the stabilization of protein structure and in its proper folding. It is well known that protein folding is supervised by the chaperone proteins that bind to growing polypeptide chains (Ruddon and Bedows, 1997; Thulasiraman et al., 1999). This binding is not strictly specific, but there are certain binding preferences, the main factors being charge and hydrophobicity of amino acid sequence sites (Takenaka et al., 1995; Knarr et al., 1999). We suppose that the periodic distribution of these properties along the sequence facilitates uniform distribution of chaperones and such uniformity is required (or desirable) for fast and proper folding. In many cases the periodicity we observe is structure-related, that is, different positions in a period correspond to different secondary structure preferences. For example, a period may consist of two parts, one showing α-helix preference and another showing β-sheet preference (Laskin et al., 2003). At that, the periodic motif
itself determines a supersecondary structure peculiar to a type of protein domains or single-domain proteins. What is the prevalence of latent periodicity, is it a common property of structural and functional protein units? This question could be answered only after the building of complete database of latent periodicity profiles correlated with the structural and functional features of the matching proteins. Its creation is likely to demand great processing power and the development of new methods for latent periodicity detection and characterization. Our up-to-date results show that 25 different protein families have latent periodicity, and the number of these families is increasing each week as our investigations go on. We can propose that latent periodicity is common phenomenon for great number of protein families.

P.S. At the time of writing this paper, the new homology-based repeat-finding tool TRUST (Szklarczyk and Heringa, 2004) has appeared on the scene. It is claimed to outperform the above-mentioned RADAR and REPRO in sensitivity and accuracy of repeat detection. The tool is available as a Web service as well as downloadable for local execution. We have run the downloaded package against our database of latently-periodic protein kinase active sites using default parameters; again, no significant repeats were found. This confirms our proposition that the usage of homology-based repeat-finding techniques is insufficient for the investigation of regular structures in genetic sequences.

Acknowledgements

We would like to thank Dr. Michael Ochs from Fox Chase Cancer Center (Philadelphia, PA) for helpful comments and suggestions on the manuscript.

References

Adams, M.D., et al., 2000. The genome sequence of Drosophila melanogaster. Science 287, 2183–2195.
Allison, L., Stern, L., Edgcape, T., Dix, T.I., 2000. Sequence complexity for biological sequence analysis. Comput. Chem. 24, 43–55.
Altschul, S.F., Koonin, E.V., 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem. Sci. 23, 444–447.
Andrade, M.A., Ponting, C.P., Gibson, T.J., Bork, P., 2000. Homology-based method for identification of protein repeats using statistical significance estimates. J. Mol. Biol. 298, 521–537.
Bai, A., Neuwelt, E.A., 2000. The Swiss-Prot protein sequence database and its supplement THMBL. Nucleic Acids Res. 28, 45–48.
Bell, G., 1996. Evolution of simple sequence repeats. Comput. Chem. 20, 41–48.
Benso, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L., 2000. GenBank. Nucleic Acids Res. 28, 15–18.
Benso, G., 1997. Sequence alignment with tandem duplication. J. Comput. Biol. 4, 351–367.
Benso, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.
Brenner, S., Waterman, M.S., Zhang, N., 1985. Renewal theory for several patterns. J. Appl. Prob. 22, 226–234.
Chaley, M.B., Korotkov, E.V., Kudryashov, N.A., 2003. Latent periodicity of 21 bases typical for mcp gene is widely present in various bacterial genera. DNA Seq. 14, 37–52.
Chaley, M.B., Korotkov, E.V., Skyabin, K.G., 1999. Method revealing latent periodicity of the nucleotide sequences for a case of small samples. DNA Res. 6, 153–163.
Chechik, V.R., Lefkin, V.V., 1998. Nucleosome units and hidden periodicities in DNA sequences. J. Biomol. Struct. Dyn. 15, 937–947.
Coward, E., Drablos, F., 1998. Detecting periodic patterns in biological sequences. Bioinformatics 14, 496–507.
Dolgin, G., Vגודלבייסט, P., Levine, P., Cordier, C., Marcout, P., 2000. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. J. Theor. Biol. 206, 323–326.
Fischetti, V., Landau, G., Schmidt, J., Sellers, P., 1992. Identifying periodic occurrences of a template with applications to a protein structure. In: Apostolico, A., Crochemore, M., Gùdd, Z., Manber, U. (Eds.), Proceedings of the Third Annual Symposium on Combinatorial Pattern Matching. Lecture Notes in Computer Science, vol. 644, Springer-Verlag, pp. 111–120.
Gabrielian, A., Bolshoy, A., 1999. Sequence complexity and DNA curvature. Comput. Chem. 23, 263–274.
Gianfcrea, J.F., Mullin, R.C., 1989. The distribution of the frequency of occurrence of nucleotide subsequences based on their overlap capability. Bioinformatics 45, 35–52.
George, R.A., Heringa, J., 2000. The REPRO server: finding protein internal sequence repeats through the Web. Trends Biochem. Sci. 25, 515–517.
Goldsmith, E.J., Cobb, M.H., 1994. Protein kinases.Curr. Opin. Struct. Biol. 4, 835–840.
Grishov, M., McLachlan, A.D., Eisenberg, D.B., 1987. Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84, 4355–4358.
Guibas, L.J., Odlyzko, A.M., 1981. String overlaps, pattern matching and nonamotves. J. Comb. Theory (A) 30, 183–208.
Hanks, S.K., Quinn, A.M., Hunter, T., 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241, 42–52.
Heger, A., Helm, L., 2000. Rapid automatic detection and alignment of repeats in protein sequences. Proteins 41, 224–237.
Heringa, J., 1994. The evolution and recognition of protein sequence repeats. Comput. Chem. 18, 215–243.
Henstra, J., 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Curr. Opin. Struct. Biol. 8, 336–343.
Heringa, J., Argos, P., 1993. A method to recognize distant repeats in protein sequences. Proteins 17, 391–410.
Hunter, T., 1981. Protein kinase classification. Meth. Enzymol. 200, 37–37.
Jackson, J.H., George, R., Heringa, P.A., 2000. Vectors of Shannon information from Fourier signals characterizing base periodicity in genes and genomes. Biochem. Biophys. Res. Commun. 286, 289–292.
Junker, V.L., Apweiler, R., Bairoch, A., 1999. Representation of functional information in the SWISS-PROT data bank. Bioinformatics 15, 1066–1067.
Karlin, S., Dembo, A., Korostova, T., 1990. Statistical composition of high-scoring segments from molecular sequences. Ann. Stat. 18, 571–583.
Kontopou, H., Becker, W., Hunkelbusch, J., Wilms, A., Schumman, A., Hupertz, C., Kammladen, H., Jost, H.G., 1996. Dyk, a dual specificity protein kinase with unique structural features whose activity is dependent on tyrosine residues between subdomains VII and VIII. J. Biol. Chem. 271, 3486–3495.
Kearney, G., Modrow, S., Todd, A., Gehring, M.J., Buchner, J., 1999. Bp-binding Sequences in HIV gp160. J. Biol. Chem. 274, 29850–29857.
Konopka, A.K., 1994. Sequences and codes: fundamentals of biomolecular cryptology. In: Smith, D. (Ed.), Biocomputing: Informatics and Genome Projects. Academic Press, San Diego, pp. 119–174.
Korotkov, A.K., 1997. Theoretical molecular biology. In: Meyers, R.A. (Ed.), Encyclopedia of Molecular Biology and Medical Genetics. In: Cooper, D.N. (Ed.), Nature Encyclopedia of the Human Genome, vol. 3. Nature Publishing Group Reference, London, pp. 464–469.

Korotkov, A.K., 2003b. Bicosmolecular sequence analysis: pattern acquisition and frequency counts. In: Cooper, D.N. (Ed.), Nature Encyclopedia of the Human Genome, vol. 3. Nature Publishing Group Reference, London, pp. 311–322.

Korotkov, A.K., Chatterjee, D., 1988. Distance Analysis and Sequence Properties of Functional Domains in Nuclear Acids and Proteins. Gene Anal. Technolo 5, 87–93.

Korotkov, A.K., Martinuda, C., 1995. Noncoding DNA Ziff’s law and language. Science 68, 789.

Korotkov, A.K., Owens, J., 1990. Complexity charts can be used to map functional domains in DNA. Gene Anal. Technolo. Appl. 7, 35–38.

Korotkov, A.K., Smithers, G.W., 1987. DISTAN-a program which detects significant distances between short oligonucleotides. Comput. Appl. Biosci. 3, 193–201.

Korotkov, A.K., Smithers, G.W., Owens, J., Maatel Jr., J.V., 1987. Distance analysis helps to establish characteristic motifs in intron sequences. Gene Anal. Technol. 4, 63–74.

Korotkov, A.K., Korotkova, M.A., 1995. DNA regions with latent periodicity in some human clones. DNA Seq. 5, 353–358.

Korotkov, A.K., Korotkova, M.A., Kudryashov, N.A., 2003. Informa-
tion decomposition of symbolic sequences. Phys. Lett. A 312, 198–210.

Korotkov, A.K., Korotkova, M.A., Tulko, J.S., 1997. Latent periodicity of some oncogenes and DNA-binding protein genes. CARBON 13, 37–44.

Korotkov, M.A., Korotkov, E.V., Redzhi, V.M., 1999. Latent periodicity of protein sequences. J. Mol. Model. 5, 103–115.

Krause, M., Mullen, I.M., Mullen, W.E., 1997. Early evolution of meta-
morphisms in tRNA families. Identification of selected kinases in marine sponges. Mol. Biol. E vol.14, 1326–1334.

Laskin, A.A., Chaley, M.B., Korotkov, E.V., Kudryashov, N.A., 2003. The locally optimal method of cyclic alignment to reveal latent periodicities in genetic texts. NAD-binding protein sites. Mol. Biol. 37, 463–473.

Lapin, V.V., Chernikov, V.R., 2000. Order and correlations in genomic DNA sequences. The spectral approach. Uspekhi Fizicheskikh Nauk 170, 57–81.

Martinuda, C., Korotkov, A.K., 1996. Oligonucleotide frequencies in DNA follow a Yule distribution. Comput. Chem. 20, 35–38.

McLachlan, A.D., 1993. Multichannel Fourier analysis of patterns in pro-
tein sequences. J. Phys. Chem. 97, 3000–3008.

Michel, C.J., 1986. New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its application. J. Theor Biol. 120, 223–236.

Mullen, W.E., Krause, M., Krumm, B., Skorokhod, A., Mullen, M.I., 1999. Gene structure and function of tRNA families. Identification of selected kinases in marine sponges. Mol. Biol. E vol.14, 1326–1334.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Ohno, S., 1970. Evolution by gene duplication. Springer-Verlag, Berlin.

Ohno, S., 1984. Repeats of base oligomers as the primordial protein-encoding sequence. Proc. Natl. Acad. Sci. U.S.A. 80, 3391–3395.

Ohno, S., 1993. Multichannel Fourier analysis of patterns in protein sequences. Proc. Natl. Acad. Sci. U.S.A. 90, 4850–4854.

Ridlon, R.W., Bedows, E., 1997. Assisted protein folding. J. Biol. Chem. 272, 3125–3128.

Salamon, F., Korotkova, A.K., 1992. A maximum entropy principle for distribution of local complexity in naturally occurring nucleotide sequences. Comput. Chem. 16, 177–174.

Salamon, F., Wootton, J.C., Korotkova, A.K., Hansen, L.K., 1993. On the robustness of maximum entropy relationships for complexity distributions of nucleotide sequences. Comput. Chem. 17, 135–146.

Schmidt, J.P., 1998. An information theoretic view of gapped and other alignments. Pac. Symp. Biocomput., 561–572.

Sheilan, M.J., Siebenberg, C.M., Wooton, J.C., 1981. The coding function of nucleotide sequences can be discerned by statistical analysis. J. Theor. Biol. 88, 409–420.

Silverman, B.D., Lusker, R., 1996. A measure of DNA periodicity. J. Theor. Biol. 118, 295–301.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Stoesser, G., Baker, W., van den Brink, A., Cannon, E., Garcia-Pastor, M., Kane, C., Kukilova, T., Lombard, V., Lopez, R., Parkinson, H., Redochni, S., Steck, P., Stiefel, P., Tuli, M.A., 2001. The EMBL nucleotide sequence database. Nucleic Acids Res. 29, 17–21.

Szkarczyk, R., Herina, J., 2004. Tracking repeats using significance and transitivity. Bioinformatics 20, 1311–1317.

Takamura, I.M., Leung, S.M., McAndrew, S.J., Brown, J.P., Higham, L.E., 1995. Hsc70-binding peptides selected from a phage display peptide library that resemble organelar targeting sequences. J. Biol. Chem. 270, 10839–10844.

Tautz, D., Trink, M., Dover, G.A., 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature 322, 452–455.

Taylor, S.S., Knighton, D.R., Zheng, J., Ten Eyck, L.F., Sowadski, J.M., 1992. Structural framework for the protein kinase family. Annu. Rev. Cell Biol. 8, 429–462.

Taylor, S.S., Radziej-Andzelm, E., 1994. Three protein kinase structures define a common motif. Structure 2, 345–355.

Taylor, S.S., Radziej-Andzelm, E., Hunter, T., 1995. How do protein kinases discriminate between serine-threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. FASEB J. 9, 1255–1266.

Taylor, W.R., Heringa, J., Bard, F., Flores, T.P., 2002. A Fourier analysis of symmetry in protein structure. Protein Eng. 15, 79–89.

Thulasiram, V., Yang, C.F., Boyland, J., 1999. In vivo newly translated polyepitides are sequestered in a protected folding environment. EMBO J. 18, 85–95.

Thomas, E.N., Sussman, J.L., 1980. The pitch of chromatin DNA is 170, 57–81.

Tschumper, H., 1998. Hidden sequence periodicities and protein architec-
ture. Proc. Natl. Acad. Sci. U.S.A. 95, 8580–8584.

Taylor, S.S., Radziej-Andzelm, E., Hunter, T., 1995. How do protein kinases discriminate between serine-threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. FASEB J. 9, 1255–1266.

Taylor, W.R., Heringa, J., Bard, F., Flores, T.P., 2002. A Fourier analysis of symmetry in protein structure. Protein Eng. 15, 79–89.

Thulasiram, V., Yang, C.F., Boyland, J., 1999. In vivo newly translated polyepitides are sequestered in a protected folding environment. EMBO J. 18, 85–95.

Trifonov, E.N., Sussman, J.L., 1980. The pitch of chromatin DNA is 170, 57–81.

Taylor, S.S., Ten Eyck, L.F., Sowadski, J.M., 1992. Structural framework for the protein kinase family. Annu. Rev. Cell Biol. 8, 429–462.

Taylor, S.S., Radziej-Andzelm, E., 1994. Three protein kinase structures define a common motif. Structure 2, 345–355.

Taylor, S.S., Radziej-Andzelm, E., Hunter, T., 1995. How do protein kinases discriminate between serine-threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. FASEB J. 9, 1255–1266.

Taylor, W.R., Heringa, J., Bard, F., Flores, T.P., 2002. A Fourier analysis of symmetry in protein structure. Protein Eng. 15, 79–89.

Thulasiram, V., Yang, C.F., Boyland, J., 1999. In vivo newly translated polyepitides are sequestered in a protected folding environment. EMBO J. 18, 85–95.