LATENT PERIODICITY OF PROTEIN KINASE AMINO ACID SEQUENCES

Andrew A. Laskin[1,2] , Nikolay A. Kudryashov[2], Eugene V. Korotkov[1,2,*]

[1] Bioengineering Center of Russian Academy of Sciences, Prospect 60-tya Oktyabrya, 7/1, 117312, Moscow, Russia.
[2] Moscow Physical Engineering Institute, Kashirskoe shosse 31, 115409, Moscow, Russia.

Contact: katrin22@mtu-net.ru; katrin2@biengi.ac.ru
Fax: 7-095-135-05-71
Tel: 7-095-135-21-61

* to whom correspondence should be addressed
Keywords: periodicity, alignment, profile analysis, protein kinase, repeat

Abstract:
In our previous studies, we found statistically significant latent periodicity in a huge amount of protein sequences and a proposition was made that many of periodicity patterns correspond to structural or functional features of protein families. However, it was unable to test since latent periodicity is feebly marked and often interrupted by insertions and deletions. A new method of noise decomposition for isolation of a genetic sequence pattern from its close analogs is introduced. When used in conjunction with cyclic profile alignment, it is able to isolate structure-related or function-related patterns of latent periodicity. In this study, we found latent periodicity in catalytic domains of about 85% of serine-threonine and tyrosine protein kinases. Therefore we presume that latent periodicity is a common property of these protein kinase catalytic domains. Possible origins of periodic structure of protein kinase catalytic domains are discussed. Detailed results may be viewed at http://periodicity.fromru.com

Introduction

The development of mathematical techniques for symbolic sequence studies is now acquiring ever-growing importance since large amounts of genetic information such as DNA base or amino acid sequences are being gathered [Benson et al., 2000; Stoesser et al., 2001; Adams et al., 2000; Venter et al., 2001]. Which information is able to be gained when one investigates symbolic sequences using today's mathematical achievements? Answer to this question determines 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 a dynamic model of cell's genetic regulation and artificial proteins with predefined features.

One of the methods of investigation of symbolic sequence organization is the investigation of its periodicity. The periodicity investigation has reasonable biological meaning because multiple duplications of DNA sequence fragments with subsequent substitutions, insertions and deletions of symbols could serve as the ground for evolution of genes and genomes. 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 active centers of proteins could mean possible participation of latent periodicity of amino acid sequences in choice and stabilization of the proper conformation of protein's globule.

To find periodicity today the techniques of dynamic programming [Heringa, 1994; Heringa and Argos, 1993; Heringa, 1998; Benson, 1997; Benson, 1999; Heger and Holm, 2000; Andrade et al., 2000] or Fourier transformaion [Taylor et al., 2002; Lobzin and Chechetkin, 2000; Dodin et al., 2000; Jackson et al., 2000; Rackovsky, 1998; Chechetkin and Lobzin, 1998; Coward and Drablos, 1998; Voss, 1992; Silverman and Linsker, 1996; McLahlan, 1993] are commonly used. The dynamic programming techniques allow one to effectively find repeats in symbolic sequence with

arbitrary number of insertions and deletions. But that approach is limited by symbol similarity matrix it uses (PAM and BLOSUM matrices are most commonly used for amino acid sequence studies). It inherently means that dynamic programming-based techniques are capable of finding periodicity with sufficiently high level of homology between distinct repeats. However, symbolic sequences may contain so called latent periodicity [Korotkov and Korotkova, 1995; Korotkov et al., 1997; Korotkov et al., 1999; Korotkova et al., 1999; Chaley et al., 1999], where significant homology between periods is not observed and periodicity can be only revealed as a property of a certain set of periods. These types of periodicity, previously found in many DNA and protein sequences, can be extensively omitted when using dynamic programming-based techniques.

Techniques that use Fourier transformation to search for periodicity in symbolic sequences are now capable of finding any kind of periodicity in absence of insertions and deletions of symbols. They represent symbolic sequence as a set of sequences of zeroes and ones, the size of this set equal to the size of used alphabet [Lobzin and Chechetkin, 2000]. However, using Fourier technique one can find statistically significant periodicity only if period length is way less than alphabet size and the investigated subsequence is not noticeably enriched by some symbol or group of symbols. Otherwise the statistical significance of this periodicity will diffuse over periods of shorter lengths and that makes it impossible to be found at proper significance level. Moreover, Fourier transformation does not allow finding periodicity of symbolic texts in presence of many insertions and deletions.

Previously we had developed the mathematical approach for searching for latent periodicity, which is based on information decomposition of symbolic sequence [Korotkov and Korotkova, 1995; Korotkov et al., 1997; Korotkov et al., 1999; Korotkova et al., 1999; Chaley 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 that constituent represents mutual information between the investigated symbolic sequence and artificial periodic sequence with period length equal to some prime number. For instance, in case of periodicity, which is 2 symbols long, we generate the artificial sequence 12121212121212... , where symbols are treated as numbers. This decomposition allows us to eliminate the shortcomings, peculiar to dynamic programming and Fourier transformation, and makes it possible to find the so called latent periodicity of symbolic texts. By latent periodicity we mean the periodicity, where significant homology between periods is not observed and periodicity can be only revealed as a property of a certain set of periods.

However, information decomposition at its present, just like Fourier-based techniques, is not capable of finding statistically significant latent periodicity in presence of many insertions and deletions. This may lead us to a conclusion that substantial part of latent periodicity occurrences in genetic texts remains unseen by information decomposition-based techniques, as well as other known methods.

The simplest method of searching for latent periodicity with insertions and deletions is a combination of information decomposition and modified profile analysis. At that information decomposition can serve as a method that, in a number of cases, reveals the latent periodicity of some amino acid sequences and results in periodicity matrix [Korotkova et al., 1999], which can be used to determine the weights for each amino acid at each position in a period. Then using of the modified profile analysis allows us to reveal periodicity of this given type (defined by cyclic position-weight matrix we have just constructed) in all the primary sequence data bank (such as Swiss-Prot), taking into account possible insertions and deletions, and also to modify the matrix with the object of increased sensitivity and specificity of the search.

In our present study for the first time we applied the modified profile analysis to latent amino acid periodicity, previously found in 7 protein kinases using information decomposition. Subsequent iterative use of the modified profile analysis resulted in noticeable modification of the initial position-weight matrix and discovery of latent periodicity in active sites of 1074 protein kinases from 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.

Methods and algorithms

We searched for latent periodicity of amino acid sequences using our previously developed technique of information decomposition of symbolic sequences. First all amino acid sequences in Swiss-Prot release 39 were scanned. The algorithm identified more than 12000 subsequences with statistically significant periodicity of various lengths and types, without insertions or deletions of symbols (we refer to amino acids in a protein sequence as symbols and to amino acid sequences as symbolic sequences, because information decomposition technique is regardless of what type of symbolic information it processes – nucleotides, amino acid residues or even written text). Then these identified periodicities provided the initial data for searching for given periodicities with the possibility of insertions and deletions. For this, we chose the cases with latent periodicity that have the same period length and reside within functionally equivalent domains of different proteins. For these cases we defined the latent periodicity matrices [Korotkova et al., 1999] that contain the numbers of amino acids at each position in the period. Finally, we merged all these matrices into one, which later served to introduce the position-weight matrix for subsequent modified profile analysis, allowing for insertions and deletions.

Profile analysis is now widely used in research in the fields of molecular biology and bioinformatics. A profile is a two-dimensional matrix, defining a pseudo-sequence. That is, while in a conventional comparison of symbolic sequences A={a_i} and B={b_j}, the value of affinity of the i-th symbol in A to the j-th symbol in B is determined using the mapping [a_i, b_j] -> **R** (with the mapping usually in the form of a matrix like PAM or BLOSUM), profile comparison uses the mapping [i, b_j] -> **R** [Gribskov et al., 1987]. This motivates us to call A a pseudo-sequence rather than a sequence. This technique is effective for data mining in genetic sequence databases, in contrast to conventional alignment [Smith and Waterman, 1981], appropriate for comparison of two protein sequences.

Profiles are usually obtained as a result of multiple alignment of proteins, having some common property, and then used to search for homologues in data banks. However, we are also concerned with inverse problem: how to form a new profile from the results of profile analysis of data bank? It should have the following properties:

a) stability, i.e. it should be capable of finding at least a major part (ideally all) of the proteins it was formed from;

b) optimality, i.e. the score of these proteins should be as high as possible and statistical probability of casual finding of unrelated subsequence scoring this high should be as low as possible;

c) selectivity, i.e. we should find the least possible number (ideally no) proteins not having this property. The difference between b) and c) is that in the former case we deal with some statistical model of sequence data bank (usually in the form of a long sequence with given probabilities of symbols or groups of symbols), while in the latter case we we deal with real data bank. As we will see below, the difference is important.

If we create new profile from the results of searching a data bank and repeat this process, we can determine 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 formula to fill the position-weight matrix, since a theorem exists that specifies the connection between 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 can be written in this form:

$$W_{i,j} = C \ln \frac{p_{i,j}}{f_i}, \qquad (1)$$

where $W_{i,j}$ is an element of the position-weight matrix for symbol of type i at position j, $p_{i,j}$ is fraction of symbol of type i occurring at position j within the high-scoring segments, and f_i is the frequency of occurrence of symbols of type i in the sequence being scanned, namely, in the data bank. The scaling parameter C in the formula may be arbitrarily chosen (multiplying all weights and

scores by a factor does not change the path of the alignment or its statistical significance, provided that deletion costs are multiplied by the same factor), so we can choose it large enough to round the weight values to integers without substantial loss of precision in order to speed up the computation. Our numerical experiments showed that using other admissible variants of position-weight matrix calculation, successive iterations leaded to:

a) continuous increase in number of results and complete loss of selectivity, or
b) contrarily, decrease in number of results right down to zero, or
c) deformation of the initial profile from one iteration to another, as a rule, the results skewed to unrelated homologous repeats (we believe that it was due to irrelevant overestimation of difference between most frequent and less frequent, but valid, symbols).

These results conform with the principles of operation of the popular profile analysis package PSI-BLAST, whose authors came to similar conclusions [Altschul et al., 1997; Altschul and Koonin, 1998; Aravind and Koonin, 1999].

From the viewpoint of information theory, the problem is similar to that of useful signal extraction from a noisy channel [Schmidt, 1998; Wilbur and Neuwald, 2000]. However, it would not be correct to consider a set of real protein sequences to be like a mixture of uncorrelated noise and useful signals. 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 for determination of a new profile, they will skew it 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 will be completely changed [Altschul and Koonin, 1998]. This property is of a certain heuristic value, for example, for characterization of distant relationships between proteins and protein families; nevertheless, it may lead to unwanted effects when defining a common profile for a family of proteins or domains. Moreover, it makes automation of successive iteration process hardly possible – one has to choose "correct" results of scanning by hand to avoid profile "pollution". When the number of results runs into the hundreds and thousands, "time-consuming" turns into "impossible".

To this end in our present paper, we propose a different probabilistic consideration of a set of genetic sequences, which we call "noise decomposition". Let us represent it 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. The main difference of correlated noise for us 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 $\pi_{i.j}$.

It is clear that correlated noise may be not of a single type, but a composition of a few different types. We assign a distribution to each of these types and call it $q^a_{i,j}$, 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:

$$\pi_{i,j} = c_0 f_i + \sum_{k=1}^{N} c_k q^k_{i,j}, \quad \sum_{k=0}^{N} c_k = 1, \qquad (2)$$

according to it, we rewrite the expression for calculation of the elements of the position-weight matrix (1) with

$$W_{i.j} = C \ln \frac{p_{i,j}}{\pi_{i,j}}. \qquad (3)$$

Here c_k are "relative magnitudes" of different types of noise, in other words, their contribution into the resulting noise, with normalization, which is required for this condition to hold:

$$\forall j, \sum_{\text{all } i} \pi_{i,j} = 1., \quad (4)$$

since $\pi_{i,j}$ are probabilities.

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.

$$\frac{c_0}{M_0} = \frac{c_1}{M_1} = \frac{c_2}{M_2} = ... \quad (5)$$

However, in fact, the numbers M_0, M_1, M_2... are unknown. Therefore, we will estimate them from the results of an initial search. Moreover, since the insignificant subsequences should be treated as noise, we will take into account only statistically significant results. At that the number of results of any type (hence, the quantitative composition of noise) will depend on the chosen significance threshold, however, our experiments showed that this assumption is valid for our problem. In other words, 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

$$\frac{c_0}{N_0} = \frac{c_1}{N_1} = \frac{c_2}{N_2} = ... \quad (6)$$

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 (3) will be distorted; our experiments demonstrated that the iterative process diverges under this condition). So we set N_0 equal to the total number of false positives in the initial search results; experiments showed that iteration process converges in this case.

To obtain values of $q^a_{i,j}$ and $p_{i,j}$ we have to divide the initial search results into true positives and false positives of different types. We have tested two methods to accomplish this: using keyword analysis and using clustering. As we determined, the keyword analysis technique provides a more precise decomposition of results, but it is applicable only when the data bank we work with is already curated with appropriate keywords; the clustering technique works on uncharacterized set of sequences but requires sequence similarity within each class. Details on our implementation of these techniques will be given below.

It is clear that noise decomposition is quite applicable to conventional (linear) profiles [Gribskov et al., 1987], but in the present study it was used for periodic profile training with the aim of latent periodicity investigation of protein sequences [Korotkov et al., 1999; Korotkova et al., 1999]. The reasoning given above is valid and expressions (1)-(6) hold in both cases. The problem of finding a good cyclic profile is much more sophisticated than the linear problem, because in cyclic alignment we have a number of diverged copies (repeats) of a pattern within a sequence instead of one. Internal divergence of repeats superimposes on divergence of different sequences; hence, cyclic patterns are much more feebly marked and hard to assign to some structural or functional unit. For the cyclic profile search problem we have made a modification of the algorithm of [Fischetti et al., 1992], called locally-optimal cyclic alignment. 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 a position in the linear sequence and another (cyclic one) is for a position in the cyclic profile (compared to conventional sequence alignment, which can be presented in the form of a path between nodes of the two-dimensional lattice, coordinates being the positions in 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 such path has a total score, which is the sum of gap penalties and weights of symbol-to-position matches.

The optimal cyclic alignment is the path with the highest possible total score. We have shown that, just like finding of the best linear alignment using Smith-Waterman formula [Smith and Waterman, 1981], it can be found when we fill cell-by-cell the similarity matrix $S_{i,j}$, having a cyclic

(wrapped) index i, namely, $S_{i-L,j} \equiv S_{i+L,j} \equiv S_{i+2L,j} \equiv \ldots \equiv S_{i,j}$, where L is length of the period. The formulae for recursive filling of $S_{i,j}$ are:

$$S_{i,j} = \max\{S'_{i,j}, \max_{1\le k\le L-1}[S'_{(i-k),j} - d_k]\}, \text{ where} \quad (7)$$

$$S'_{i,j} = \max\left\{0, S_{i-1,j-1} + w_{i,j}, \max_{1\le k\le j}\left[S_{i,j-k} - d_k\right]\right\} \quad (8)$$

Here $w_{i,j}$ is the weight of the jth symbol in the sequence at the ith position in the profile, d_k is the gap penalty for insertion/deletion of k successive symbols. As usual, to find the optimal local alignment we find the highest element of S-matrix and recreate the path down to the first zero element. The value of the highest element is the total score of optimal local alignment; this value was used to check whether the alignment is statistically significant.

A Monte-Carlo method was used to assess the statistical significance of alignments. The assessment was performed separately for each sequence, taking into account its length and composition. To assess the statistical significance of an alignment in this study we aligned the appropriate random sequences of the same length and composition as the real sequence (to avoid composition bias effects) using the same cyclic profile. Then, considering the distribution of obtained weights to be normal [Webber and Barton, 2001; Korotkov et al., 2000], we could estimate the Z-value of the obtained alignment from the mean and the variance of these random scores:

$$Z = \frac{S_{real} - E(S)}{\sqrt{D(S)}} \quad (9)$$

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 when analyzing a random test sequence with the same number of symbols as the total number of symbols in Swiss-Prot.

The preliminaries for cyclic profile alignment were given in [Laskin et al., 2002]; proofs and implementation details will be discussed elsewhere. The described algorithms were implemented in C++, and are available upon request by e-mail (for details visit our Web site http://periodicity.fromru.com).

Results

In previous studies [Korotkov et al., 1999; Korotkova et al., 1999] we identified 7 protein kinase amino acid sequences with latent periodicity of length 18 amino acids, in the absence of insertions and deletions (Table 1). The example of latent periodicity is shown in Fig.1.

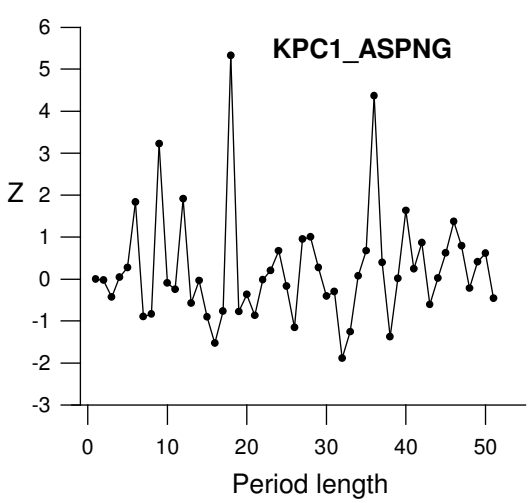

Fig. 1. Information decomposition spectrum for amino acid sequence
WWAFGVLIYQ MLLQQSPFRG
EDEDEIYDAI LADEPLYPIH
MPRDSVSILQ KLLTREPELR
LGSGPTDAQE VMSHAFFRNI
NWDDIYHKRV PPPFLPQISS PTD
from KPC1_ASPNG (residues 954 – 1056). Full protein kinase domain of this protein contains residues 770-1030.

All of them were serine-threonine protein kinases except M3KA_HUMAN, having dual specificity. The periodicity was 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). This suggested that the periodicity of 18 amino acids is a characteristic property of protein kinase active sites. We set a problem to find

out, first, to what extent this periodicity is peculiar to protein kinase active sites if insertions and deletions are permitted, and, second, do serine-threonine and tyrosine protein kinases share the same periodicity pattern and can we improve sensitivity and specificity of search by using separate periodicity profiles for different kinase types instead of one.

We averaged amino acid occurrence frequencies in different profile positions of all 7 cases mentioned above and made the initial position-weight matrix using formula (1). This matrix was used to scan Swiss-Prot release 40 [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. Values of 3.5 for gap opening and 0.7 for gap extension were found to be optimal and used thereafter.

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 results into classes and formed two new position-weight matrices using (2)-(6), 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. Two ways of division into classes were tried out, keyword analysis and clustering.

Table 1. Proteins where latent periodicity without insertions and deletions was previously found.

Swiss-Prot ID	Start position	End position	Protein description in Swiss-Prot
KDBE_SCHPO	400	478	PUTATIVE SERINE/THREONINE-PROTEIN KINASE C22E12.14C (EC 2.7.1.-).
KEMK_MOUSE	85	181	PUTATIVE SERINE/THREONINE-PROTEIN KINASE EMK (EC 2.7.1.-).
KPC1_ASPNG	954	1056	PROTEIN KINASE C-LIKE (EC 2.7.1.-).
KPCL_RAT	526	565	PROTEIN KINASE C, ETA TYPE (EC 2.7.1.-) (NPKC-ETA) (PKC-L).
M3KA_HUMAN	97	181	MITOGEN-ACTIVATED PROTEIN KINASE KINASE 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 PROTEINKINASE).
CC2_CARAU	88	160	CELL DIVISION CONTROL PROTEIN 2 HOMOLOG (EC 2.7.1.-) (P34 PROTEINKINASE) (CYCLIN-DEPENDENT KINASE 1) (CDK1).

We worked with the Swiss-Prot data bank, which is curated and contains relevant information about protein kinase type [Junker et al., 1999]. We extracted information about proteins under investigation from their descriptions (DE field), keywords (KW field) and feature tables (FT field). We divided the results into three classes depending by keywords. Class 1 comprised proteins, which did not contain the keyword “protein kinase”. Class 2 comprised proteins, which were identified as a tyrosine protein kinase, while class 3 comprised all other proteins. Then we consider one of the classes (2 or 3) to be true positives and another 2 classes to be different types of correlated noise. This approach is applicable when relevant information is presented in Swiss-Prot or any other database (one can use SQL to query a number of databases at a time).

If such information is absent in a data bank, clustering can be used to split the proteins into classes. We performed the protein kinase clustering experiment as follows. First, we made pairwise comparisons of identified subsequences using a global alignment method [Needleman and Wunsch, 1970]. Then we built the matrix of distances between subsequences using the formula:

Distance(A,B)=(AlignmentScore(A,A)+AlignmentScore(B,B))/2–AlignmentScore(A,B).

This distance matrix was used in cluster analysis with the single linkage method, and merge threshold adjusted to output 2 large classes. Then we checked if these clusters related to 2 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 prevalence was about 90%. We conclude that the information in Swiss-Prot is more reliable than clustering and should be used when possible. From this point on we used the keywords method.

Class separation resulted in 2 position-weight matrices, which were later optimized (trained) with iterative searches (described above) to find the highest number of serine-threonine or tyrosine protein kinases, respectively, while keeping the specificity. We performed iterations until the result set was nearly stable (generally 3-5 iterations turned out to be adequate).

Table 2. Summary of Swiss-Prot scanning using two profiles described in text.

Matrix type	Serine/threonine-protein kinases	Tyrosine-protein kinases
Total number of protein kinases present in Swiss-Prot release 40	963 (43 – dual specificity kinases)	369 (43 – dual specificity kinases)
Total number of protein kinases having the Z more than 6.0	774	301
False positives	55	5
Another type protein kinases found among false positives	47 (tyrosine kinases)	5 (serine/threonine kinases)

The summary of final results is presented in Table 2. In both cases, we found latent periodicity in more than 80% of the target protein kinases and reached kinase separation levels higher than 94%, most errors being due to dual-specific kinases (usually found using serine-threonine profile rather than tyrosine). To obtain the latest versions of latent periodicity profiles and results sets for these and other investigated instances, visit the authors' Web site at http://periodicity.fromru.com.

Discussion

The latent periodicity notion and search technique was initially presented in [Korotkov and Korotkova, 1995] and refined in subsequent works [Korotkov and Korotkova, 1996; Korotkov et al., 1997; Korotkov et al., 1999; Korotkova et al., 1999; Chaley et al., 1999]. As a result of performed studies, we revealed the existence of various types of latent periodicity in numerous amino acid sequences [Korotkov et al., 1999; Korotkova et al., 1999]. These found latently periodic sequences belonged to various proteins, with various period lengths and periodicity patterns. However, the question of functional significance of identified latent periodicities and its correlation with protein structures remained open. To a great extent this resulted from the information decomposition method being incapable of revealing latent periodicity interrupted with insertions and deletions, so this approach omitted a substantial subset of proteins with certain functional domains, so that 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 type that is common for a whole protein superfamily. We have achieved this result by means of the development of novel mathematical methods and software, capable of finding statistically significant latent periodicity of a predefined type in presence of insertions and deletions.

Protein kinases, i.e. enzymes whose function is to transfer phosphate residues from ATP to other proteins, are known to provide a 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 kinases and tyrosine kinases (according to the phosphorylated residue). In addition, there also exist protein kinases with dual specificity [Kentrup et al., 1996]. While there are also other types of protein kinases structurally dissimilar to these kinases, which phosphorylate other residues. They do not share the types of periodicity described in this paper, nor they were found to have any other common periodicity pattern. In 1999, we found a few latent periodicity instances in histidine protein kinases. We applied the same techniques to these results, however, iterated cyclic profile searches found only a few close homologues, and so latent periodicity was not found to be common for the whole histidine protein kinase superfamily.

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 3D structure elements [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 asparatic acid residue mentioned above, the valine residue that interacts with ATP adenine, 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 2 times more frequently in tyrosine kinases than in serine-threonine ones (on average, 5.96 vs 3.05 indels per site), i.e. we observe larger deviations from perfect periodicity.

Indubitably, the crucial outcome of our studies is the very fact that there is (as we suppose, far from unique) very important class of proteins whose active center turned out to be latently periodic. Our hypothesis about existence of latent periodicity types, common for whole protein classes, has been verified.

At this time the origin of observed latent periodicity is unclear, 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 duplications were properly arranged to form even more catalytically active domain. It is a fact that repeats in DNA sequences 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 structure stability increased. Subsequent mutations formed even bettter packed structure of these domains and fine-tuned the functionality, at the same time the mutations resulted in periodicity diffusion, loss of homology between distinct repeats. That means that we may call the residual internal similarities "echoes" of the ancient protein formation processes.

Latent periodicity may be also involved in stabilization of protein structure and proper folding. It is well known that protein folding is supervised by chaperone proteins that bind to growing polypeptide chain [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 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. For many cases we observe structure-related periodicity, 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-structure preference. At that, the periodic motif itself determines a supersecondary structure peculiar to a type of protein domains or single-domain proteins [Laskin et al., 2002].

To what extent is latent periodicity a common property of structural and functional protein units? This question can be answered only after building of complete database of latent periodicity profiles and their structural and functional features. Its creation is likely to demand great processing power and development of new methods for latent periodicity detection and characterization. It is now under consideration.

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., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F. et.al (2000) The Genome Sequence of Drosophila melanogaster. Science, 287:2185-2195.

Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. TIBS, 23, 444-447.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389–3402

Andrade, M.A., Ponting, C.P., Gibson, T.J. and Bork, P. (2000) Homology-based Method for Identification of Protein Repeats Using Statistical Significance Estimates. J. Mol. Biol., 298, 521-537

Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J. Mol. Biol., 287, 1023–1040.

Bairoch, A. and Apweiler, R. (2000) The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res., 25, 45-48.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L. (2000) GenBank. Nucl Acids Res., 28:15-18.

Benson, G. (1997) Sequence alignment with tandem duplication. J Comput Biol. 4(3):351-67.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. Jan 15;27(2):573-80.

Chaley, M.B., Korotkov, E.V. and Skryabin, K.G. (1999) Method revealing latent periodicity of the nucleotide sequences for a case of small samples. DNA Res., 6, 153-163.

Chechetkin, V.R., Lobzin, V.V. (1998) Nucleosome units and hidden periodicities in DNA sequences. J Biomol Struct Dyn, 15:937-947.

Coward, E. and Drablos, F. (1998) Detecting periodic patterns in biological sequences. Bioinformatics, 14, 498-507.

Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C., Marcourt, L. (2000) Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. J Theor Biol 206:323-326.

Goldsmith, E.J., Cobb, M.H. (1994) Protein kinases. Curr Opin Struct Biol. 4(6):833-40

Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Profile analysis: Detection of distantly related proteins. Proc. Natl Acad. Sci. USA, 84, 4355–4358.

Hanks, S.K., Quinn, A.M., Hunter, T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science. 241(4861):42-52. Review.

Heger, A. and Holm, 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(3):233-43.

Heringa, J. (1998) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Curr Opin Struct Biol, 8:338-345.

Heringa, J., Argos, P. A method to recognize distant repeats in protein sequences. Proteins. 1993 Dec;17(4):391-41

Hunter, T. (1991) Protein kinase classification. Methods Enzymol. 200:3-37.

Jackson, J.H., George, R., Herring, P.A. (2000) Vectors of Shannon information from Fourier signals characterizing base periodicity in genes and genomes. Biochem. Biophys. Res. Commun, 268:289-292

Junker, V.L., Apweiler, R. and Bairoch, A. (1999) Representation of functional information in the SWISS-PROT data bank. Bioinformatics, 15, 1066-1067.

Karlin, S., Dembo, A. and Kawabata, T. (1990) Statistical composition of high-scoring segments from molecular sequences. Ann. Stat. 18, 571-581.

Kentrup, H., Becker, W., Heukelbach, J., Wilmes, A., Schurmann, A., Huppertz, C., Kainulainen, H., Joost, H.G. (1996) Dyrk, 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(7):3488-95.

Knarr, G., Modrow, S., Todd, A., Gething, M.J., and Buchner J. (1999) BiP-binding Sequences in HIV gp160. J Biol Chem, 274, 29850–29857

Korotkov, E.V. and Korotkova, M.A. (1995) DNA regions with latent periodicity in some human clones. DNA Seq., 5, 353-358.

Korotkov, E.V. and Korotkova, M.A. (1996) Enlarged similarity of nucleic acid sequences. DNA Res., 3, 157-164.

Korotkov, E.V., Korotkova, M.A., Rudenko, V.M. (2000) MIR--family of repeats common for vertebrate genomes. Mol Biol (Mosk). 34(4):553-9. Russian.

Korotkov, E.V., Korotkova, M.A., Rudenko, V.M. and Skryabin, K.G. (1999) Latent periodicity regions in amino acid sequences. Mol. Biol., 33, 611-617.

Korotkov, E.V., Korotkova, M.A., Tulko, J.S. (1997) Latent sequence periodicity of some oncogenes and DNA-binding protein genes. CABIOS, 13:37-44.

Korotkova, M.A., Korotkov, E.V. and Rudenko, V.M. (1999) Latent periodicity of protein sequences. J. Mol. Model., 5, 103-115.

Kruse, M., Muller, I.M., Muller, W.E. (1997) Early evolution of metazoan serine/threonine and tyrosine kinases: identification of selected kinases in marine sponges. Mol Biol Evol. 14(12):1326-34.

Laskin A., Korotkov E., Kudryashov N. (2002) New method of latent periodicity detection may determine structurally related proteins and protein families. Proc. 3rd Int. Conf. On Bioinformatics of Genome Regulation and Structure, 3:97-99 Novosibirsk

Lobzin, V.V., Chechetkin, V.R. (2000) Order and correlations in genomic DNA sequences. The spectral approach. Uspekhi Fizicheskikh Nauk, 170:57-81

McLachlan, A.D. (1993) Multichannel Fourier analysis of patterns in protein sequences. J.Phys.Chem, 97:3000-3006.

Muller, W.E., Kruse, M., Blumbach, B., Skorokhod, A., Muller, I.M. (1999) Gene structure and function of tyrosine kinases in the marine sponge Geodia cydonium: autapomorphic characters in Metazoa. Gene. 238(1):179-93.

Needleman, S.B. and 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.

Rackovsky, S. (1998) Hidden sequence periodicities and protein architecture. Proc. Nat. Acad. Sci., 95:8580-8584.

Ruddon, R.W. and Bedows, E. (1997) Assisted Protein Folding. J Biol Chem,. 272, 3125–3128,

Schmidt, J.P. (1998) An information theoretic view of gapped and other alignments. Pac Symp Biocomput 561-72

Silverman, B.D., Linsker, R. (1996) A measure of DNA periodicity. J.Theor.Biol., 118:295-300.

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

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Lombard, V., Lopez, R., Parkinson, H., Redaschi, N., Sterk, P., Stoehr, P., Tuli, M.A. (2001) The EMBL nucleotide sequence database. Nucl. Acids Res, 29:17-21.

Takenaka, I.M., Leung, S.M., McAndrew, S.J., Brown, J.P., Hightower, L.E. (1995) Hsc70-binding Peptides Selected from a Phage Display Peptide Library that Resemble Organellar Targeting Sequences. J Biol Chem, 270, 19839-19844

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. 8:429-62.

Taylor, S.S., Radzio-Andzelm E. (1994) Three protein kinase structures define a common motif. Structure 2(5):345-55.

Taylor, S.S., Radzio-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(13):1255-66

Taylor, W.R., Heringa, J., Baud, F., Flores, T.P. (2002) A Fourier analysis of symmetry in protein structure. Protein Eng. 15(2):79-89.

Thulasiraman,V., Yang, C.F. and Frydman, J. (1999) In vivo newly translated polypeptides are sequestered in a protected folding environment. EMBO J 18,.85–95,

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. et.al (2001) The sequence of the human genome. Science, 291:1304-1351.

Voss, R.F. (1992) Evolution of long range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett., 25:3805-3808.

Webber, C. and Barton, G.J. (2001) Estimation of P-values for global alignment of protein sequences. Bioinformatics, 17, 1158-1167.

Wilbur, W.J., Neuwald, A.F. (2000) A theory of information with special application to search problems. Comp Chem 24 33–42