Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions

D. A. Afonnikov*, D. Yu. Oshchepkov and N. A. Kolchanov

Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, 10 Lavrentyev Ave., Novosibirsk 630090, Russia

Received on December 10, 2000; revised on April 17 and June 20, 2001; accepted on August 2, 2001

ABSTRACT

Motivation: It is known that the physico-chemical characteristics of proteins underlying specific folding of the polypeptide chain and the protein function are evolutionary conserved. Detection of such characteristics while analyzing homologous sequences would expand essentially the knowledge on protein function, structure, and evolution. These characteristics are maintained constant, in particular, by co-ordinated substitutions. In this process, the destabilizing effect of a substitution may be compensated by another substitution at a different position within the same protein, making the overall change in this protein characteristic insignificant. Consequently, the patterns of co-ordinated substitutions contain important information on conserved physico-chemical properties of proteins, requiring their investigation and development of the corresponding methods and software for correlation analysis of protein sequences available to a wide range of users.

Results: A software package for analyzing correlated amino acid substitutions at different positions within aligned protein sequences was developed. The approach implies searching for evolutionary conserved physico-chemical characteristics of proteins based on the information on the pairwise correlations of amino acid substitutions at different protein positions. The software was applied to analyze DNA-binding domains of the homeodomain class. As a result, two conservative physico-chemical characteristics preserved due to the co-ordinated substitutions at certain groups of positions in the protein sequence. Possible functional roles of these characteristics are discussed.

Availability: The program package is available at http://wwwmgs.bionet.nsc.ru/programs/CRASP/

Contact: ada@bionet.nsc.ru

INTRODUCTION

The number of proteins with known sequences is rapidly increasing. In addition, the recent progress in multiple sequence alignment procedure favoured accumulation of multiple alignments in databases available via the Internet, for example BLOCKS (Henikoff and Henikoff, 1991), HSSP (Dodge et al., 1998), Pfam (Bateman et al., 1999), and PIR-ALN (Srinivasarao et al., 1999). Extraction of the novel information from the numerous data obtained so far requires development of new methods for analyzing sets of aligned protein sequences.

A promising approach for analyzing this kind of sequence data is correlation analysis of amino acid substitutions in protein sequences. It is based on the assumption that pairwise substitutions of the amino acid residues involved in the interactions that are important for the protein structure and function are fixed in a co-ordinated manner during the evolution (Lim and Ptitsyn, 1970; Altschuh et al., 1987; Kimura, 1991). Such an evolutionary mode may manifest itself in families of homologous sequences as a set of positions displaying statistically significant correlations between substitutions of amino acid residues. Correspondingly, the observed statistical relationships between amino acid substitutions at particular positions may reflect certain interactions between the residues located at these positions. In particular, it was suggested that some correlations between spatially close residues have a compensatory nature (Lim and Ptitsyn, 1970; Altschuh et al., 1987; Göbel et al., 1994; Shindyalov et al., 1994). Therefore, the information on co-ordinated amino acid substitutions may be useful for predicting the spatial structure of proteins.

Several approaches were suggested to detect statistical relationships between residue substitutions in proteins, including informational method (Korber et al., 1993; Clarke, 1995; Nagl et al., 1999; Atchley et al., 2000), analysis of correlations based on amino acid relatedness (Göbel et al., 1994), statistical evaluation of the pairwise amino acid substitution frequencies (Shindyalov et al.,

*To whom correspondence should be addressed.
1994; Chelvanayagam et al., 1997), analysis of correlations in terms of physico-chemical property values (Neher, 1994), and maximum likelihood approach (Pollock et al., 1999). However, the correlation between spatial proximity of a pair of residues and the co-ordinated substitutions at the corresponding positions appeared to be relatively weak (Göbel et al., 1994; Shindyalov et al., 1994; Clarke, 1995; Neher, 1994).

Nevertheless, certain examples demonstrate that correlation analysis is capable of extracting important information from sets of aligned sequences of homologous proteins. In particular, Benner and Gerloff predicted the secondary and tertiary structures of protein kinases from multiple alignment of their sequences. Analyzing the co-variation of amino acid residues, they demonstrated that two β-strands in the catalytic domain of the protein kinase molecule are antiparallel (Benner and Gerloff, 1991). Pazos et al. demonstrated that analysis of correlated substitutions of amino acid residues was sufficient to single out the right inter-domain docking solution amongst many wrong alternatives of two-domain proteins (Pazos et al., 1997a). Data on correlated changes in pairs of amino acid residues allowed the neural network predictions of residue–residue contacts to be improved (Fariselli and Casadio, 1999). Ortiz et al. applied the information on co-varying pairs of amino acid residues as tertiary restraints in ab initio structure prediction for small proteins and demonstrated that this approach was capable of assembling low-resolution tertiary structures for larger proteins than was ever possible (Ortiz et al., 1998). Data on conserved, correlated, and polar residues in protein sequences were used by Olmea et al. for filtering the incorrect threading solutions in protein fold predictions (Olmea et al., 1999). As a result, the ratio of correct versus incorrect folds was improved almost fourfold. Thus, it is beneficial to utilize the information on pairwise correlations of amino acid substitutions together with the data on protein secondary structure, accessibility of particular residues to the solvent, and other protein characteristics. Recently, such a unified approach was realized as a software package for protein sequence and structure analysis (Pazos et al., 1997b).

Note that the positions with correlations between residue substitutions may not occur pairwise only, but also form networks. Networks of correlated protein positions were analyzed in DNA-binding domains of the homeodomain class (Clarke, 1995), ligand-binding pockets of nuclear receptors (Nagl et al., 1999), and domains of the helix–loop–helix class (Atchley et al., 2000). One of the problems arising while analyzing these networks is to discover the particular physico-chemical characteristics of the residues that are conserved through co-ordinated substitutions (Lim and Ptitsyn, 1970; Gerstein et al., 1994). For example, study of the sequences of GPI-modification sites suggested that compensating substitutions determined the constancy of the total volume of residues in the vicinity of o-site (Eisenhaber et al., 1998). This allowed the volume of the cleft in the active site of putative transaminase to be estimated and the recognition of the GPI-modification site sequences to be improved (Eisenhaber et al., 1999).

Thus, the analysis of co-ordinated substitutions in proteins would comprise the two major stages: (i) search for the pairs (or groups) of positions within a protein sequence where amino acid substitutions occur in a co-ordinated manner and their analysis and (ii) deduction and analysis of the corresponding conserved protein physico-chemical characteristics displayed by these groups of positions.

In this work, we are describing the software package CRASP, designed to detect clusters of correlated positions using multiple alignments of protein sequences and analysis of integral physico-chemical characteristics of such clusters. This package was used to study DNA-binding domains of the homeodomain class. We calculated the correlation coefficients and detected the pairs of positions within the homeodomain where the substitutions occurred in a co-ordinated manner. We discovered two clusters of positions where amino acid substitutions correlated with respect to the isoelectric point value. The evolutionary conserved physico-chemical characteristics of the residues involved were determined for these clusters.

SYSTEM

The software package CRASP is written in C, supplied with an HTML interface, and installed on the molecular biological server of the Institute of Cytology and Genetics, SB RAS (http://wwwmgs.bionet.nsc.ru/programs/CRASP). A schematic of the software package CRASP is shown in Figure 1. The package comprises five major modules detailed below.

(1) **User interface**, allowing the user to work in an interactive mode.

The input data for CRASP are multiple alignments of amino acid sequences in a FASTA format. To perform weighted calculations, the user can also input sequence weights or phylogenetic tree in a separate input window.

The output data of the package CRASP—a diagram of correlation coefficient matrices, diagram of a binary tree of the clusters of positions, and other types of information—could be represented in graphical and/or numerical forms.

(2) **Database on physico-chemical properties of amino acids.** We used information contained in the database AAindex (Nakai et al., 1988; Tomii and Kanehisa, 1996), describing over 400 properties.
The interface allows the user to select any property for analysis.

(3) Module for constructing correlation matrix calculates the matrix of correlations between values of a physico-chemical characteristic of amino acid residues in all the columns of a multiple alignment. In the case of homologous sequences, it is possible to use weighted calculation to take into account the evolutionary relatedness of protein sequences.

(4) Module for analyzing correlation matrix, designed to solve two problems: (a) detecting clusters of positions related with significant correlation coefficients; and (b) detecting local regions of the correlation matrix with a prevalence of significant correlation coefficients.

(5) Module for analyzing integral physico-chemical characteristics of clusters of positions. It estimates the constancy of the value of a characteristic considered for a cluster of positions and compares it with the value predicted by the model of uncorrelated residue substitutions.

In this work, the software package CRASP is applied to search for the protein integral physico-chemical characteristics that are maintained constant during the evolution through co-ordinated substitutions of amino acid residues. The approach proposed comprises the following stages. First, the pairwise correlation coefficients for the value of a physico-chemical property of the residues located at all the variable positions within a protein. Then, the groups of positions displaying significant correlations are determined. A set of integral physico-chemical properties of the protein that might be maintained constant by co-ordinated amino acid substitutions is inferred through analyzing the pairwise correlation coefficients within the detected groups of positions. Finally, the contribution of co-ordinated substitutions to the conservation of the properties inferred is estimated.

ALGORITHMS AND METHODS

Estimation of the pairwise correlation

Let us consider a sample of \( N \) aligned sequences of the length \( L \). Then, we consider a certain physico-chemical amino acid property \( f \) (for example, hydrophobicity, charge, side chain volume, etc.). A value of this property is attributed to every amino acid in the alignment. As a result, we obtain a matrix whose element \( f_{ki} \) is the value at the \( i \)th position of the \( k \)th sequence. To detect the co-ordinated substitutions, we calculate correlation coefficients between the amino acid physico-chemical property values, \( f_i \) and \( f_j \), at the positions \( i, j \) of the protein alignment (Afonnikov et al., 1997).

\[
\bar{f}_i = \frac{1}{N} \sum_{k=1}^{N} f_{ki}.
\]

(1)

In case of evolutionary unrelated sequences (for example, sequences with distant homology or obtained by \textit{in vitro} selection), the mean value of the variable \( f_i \) is estimated as follows:

\[
Covariance \ s_{ij} \ (if \ i \neq j) \ and \ variance \ (if \ i = j) \ are \ equal \ to
\]

\[
s_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (f_{ki} - \bar{f}_i)(f_{kj} - \bar{f}_j).
\]

(2)

To estimate the relation between the pair of variables \( f_i \) and \( f_j \), the linear correlation coefficient is calculated as

\[
r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} \cdot s_{jj}}},
\]

(3)

where \( s_{ij} \) are the elements of the co-variation matrix \( S \). If the correlation coefficient \( r_{ij} \) differs from zero significantly, then the values of the property in question at the pair of positions \( i, j \) are statistically related through the co-ordinated substitutions of amino acid residues at these positions.

To check the significance of the correlation coefficient, the threshold value \( r_c \) is calculated. If \( |r_{ij}| < r_c \), then the substitutions of residues at positions \( i, j \) are statistically independent. For the unrelated protein sequences, the critical value \( r_c \) for the linear correlation coefficient could
be estimated using the equation:

\[ |r_{ij}| = \frac{t^2_p}{\sqrt{t^2_p + m}}, \tag{4} \]

where \( t_p \) is the value of the Student’s statistics at the significance level \( P \) with \( m = N - 2 \) degrees of freedom (Kendall and Stuart, 1967).

Note that the significant correlation coefficient (equation 3) for a pair of residues does not necessarily require that these residues interact directly (Lapedes et al., 1997). A long-range correlation may occur if the interaction is mediated by other residues.

Therefore, we use the partial correlation coefficient to estimate the dependence between values of the physicochemical characteristic in a pair of positions:

\[ r_{ij:g} = \frac{-a_{ij}}{\sqrt{a_{ii} a_{jj}}} \tag{5} \]

Parameters \( a_{ij} \) are the elements \( i, j \) of the matrix \( A \) calculated as

\[ A = S^{-1}. \]

where \( S = \{s_{ij}\} \) is a co-variation matrix (to invert the co-variation matrix \( S \), we decompose it into eigenvectors and eigenvalues \( \lambda_i \) and calculate the reciprocal transformation with eigenvalues of \( 1/\lambda_i \)). The partial coefficient (equation 5) reflects the correlation between the variables \( f_i \) and \( f_j \), provided that all the rest \( g = L - 2 \) variables are fixed (Kendall and Stuart, 1967). This allows a direct relationship between positions \( i \) and \( j \) of the protein to be estimated. Note that for the partial correlation coefficient (equation 5), the number of degrees of freedom in equation (4) is \( m = N - L \) (Kendall and Stuart, 1967).

We have demonstrated that the partial correlation coefficient (equation 5) allows the effect of remote correlations to be decreased significantly while estimating the correlations between amino acid substitutions in Boltzmann type models of amino acid substitutions (Afonnikov, 2000b).

Data weighting

Evolutionary relatedness of the sequences analyzed is another source of errors while estimating the significance of correlation coefficients (equations 3 and 5). It is known that amino acid sequences within homologous families had common ancestors at certain stages of their evolution and, therefore, are statistically related. Conventionally, topology of phylogenetic trees reflects evolutionary relationships between sequences (the tree is star-shaped if the sequences are unrelated). It has been shown that the evolutionary relatedness of protein sequences has a strong effect on the estimates of correlation measures of various types (Pollock and Taylor, 1997; Lapedes et al., 1997). If the evolutionary aspects are ignored, then the critical value of the correlation coefficient \( r_c \) is underestimated. This results from the overestimation of the parameter \( m \) (number of degrees of freedom) in equation (4) for the protein sequences whose evolutionary tree has other than a star-shaped topology (Felsenstein, 1985).

In this work, we studied the effect of evolutionary relatedness of sequences on correlation coefficient estimates by numerical simulations. A Poisson model with equal rates of independent amino acid substitutions was used. Four types of phylogenetic tree topologies—star-shaped, random, balanced, and maximally unbalanced—were simulated. The overall evolution time was equal (0.4) for all the topology types; length of all the sequences, 500. The number of sequences in the sample \( N \) varied from 8 to 100; 1000 samples were generated for each topology type and sample size. We studied the behaviour of variance of correlation coefficient estimates depending on the number of sequences and phylogenetic tree topology. The effective sample size \( N_{\text{eff}} = \frac{\pi^2}{4} + 1 \) was considered as the parameter that characterizes the variance. It is known that for a sample of \( N \) independent observations over two independent Gaussian variables, the variance \( D_r \) of the correlation coefficient estimates depends on the sample size asymptotically as \( D_r = 1/(N - 1) \) (Anderson, 1958). Therefore \( N_{\text{eff}} \) is expected to be close to \( N \) for the star-shaped tree topology. Its decrease would indicate an increase in variance of correlation coefficients compared with the variance for an independent sample. We estimated the value \( D_r \) for distribution of pairwise correlation coefficients for the simulated sequences and averaged it over 1000 samples. The results obtained are shown in Figure 2 (empty circles correspond to the star-shaped topology; filled symbols, to the remaining topologies). It is evident that the value of \( N_{\text{eff}} \) is close to \( N \) in the case of the star-shaped topology, as it was actually expected. In the cases of other tree topologies, \( N_{\text{eff}} < N \), and the differences are close to twofold. Consequently, the variance of estimates of the correlation coefficients (equation 3) for related sequences is essentially higher than for the independent sequences. In the case of related sequences, the sequences are assumed independent and significance of the correlation coefficient is estimated using equation (4), the threshold \( r_c \) appears too low; consequently, the fraction of the pairs of positions where the substitutions are independent and \( |r_{ij}| > r_c \) exceeds the specified error level 1 - \( P \).

Different approaches are used for a more accurate estimation of the critical value \( r_c \). Some authors simulate numerically the evolution of proteins with independent residue substitutions to estimate this threshold (Lapedes et al., 1997; Nagl et al., 1999; Pollock et al., 1999; Atchley et al., 2000); however, this method is very time-consuming.

Data weighting is another possible approach to take into account the evolutionary relatedness of the sequences analyzed (Göbel et al., 1994; Chelvanayagam et al., 1997).
Revealing clusters of correlated positions

In this work, we are using the data weighting approach based on Felsenstein’s method (Felsenstein, 1985), based on the assumption that the less is the time of divergence of two organisms from their common ancestor, the stronger is the relation between homologous protein sequences of these organisms (Felsenstein, 1973; Altschul et al., 1989).

To verify the efficacy of this approach, we used a numerical experiment. The conditions of the experiment were the same as above, but Weighted Estimates (WE) for $r_{ij}$ were implemented. The results obtained are shown in Figure 2 (empty symbols). It is evident that $N_{\text{eff}}$ is close to $N$ for the WE s, similar to the case of independent sequences. This means that the variances of WEs of correlation coefficients for different phylogenetic tree topologies are similar, thereby allowing equation (4) to be used for estimating $r_c$ with $m = N - 2$ degrees of freedom, similarly to the case of unrelated sequences. It may be expected here that the fraction of independent pairs of positions displaying $|\alpha_{ij}| > r_c$ is close to the specified error $1 - P$.

Detecting matrix regions with high density of correlated pairs

One of the options provided by CRASP is detection of the matrix local regions with significant prevalence of correlated elements. These regions are chosen so that the number $l$ of pairs with $r_{ij} > r_c$ within the square window of a user-defined size is significantly greater than the number of such pairs $K_{\text{rand}}$ expected randomly within the window in question. To determine the statistical significance of the prevalence of the correlated pairs in this window, we used an approximation of the binomial distribution assuming that the pairs were randomly distributed over the entire matrix and that the expected fraction of the pairs was typically less than 10%. In this case, the probability $p$ of the number of pairs with $r_{ij} > r_c$ in this window to exceed the observed value $l$ equals

$$p = 1 - \sum_{k=0}^{n-l} \binom{n}{k} q^k (1 - q)^{n-k},$$

where $q$ is a fraction of the pairs with $r_{ij} > r_c$ for the entire matrix; $n$, the number of matrix elements in the window. A region is considered significant at level $\alpha$ if $p < p_{\alpha}$, that is, this corresponds to the significance level $p_{\alpha} = 100\%(1 - p_{\alpha})$. The regions with a higher prevalence of significant correlations are shown in a separate matrix diagram.

Integral physico-chemical characteristics of proteins

Let us consider a cluster $\alpha$ of the correlated protein positions. Let us select an $f$-th physico-chemical characteristic. The integral physico-chemical characteristic of this
cluster $F_{ak}$ for the $k$th sequence of the protein family is determined by the following equation:

$$F_{ak} = \sum_{i \in \alpha} c_i \cdot f_{ki}.$$  

(6)

where $f_{ki}$ is the value of this physico-chemical property of the amino acid residue located at the $i$th position of the $k$th sequence; $c_i$, an arbitrary real number determining the contribution of the $i$th position to the value of the integral characteristic $F_{ak}$. Equation (6) allows a wide range of integral physico-chemical characteristics of the protein to be described depending on the values of coefficients $c_i$ associated with individual positions of the protein sequence. In particular, such characteristics as the total charge of a protein ($c_i = 1$; $\alpha$, all the positions of the protein sequence; and $f$, the charge of a residue), volume of its hydrophobic core ($c_i = 1$; $\alpha$, positions of all the residues forming the core; and $f$, the volume of a residue), projection of the helix hydrophobic moment ($c_i$, coefficient of the helix hydrophobic moment projection at the $i$th position; $\alpha$, positions of alpha-helix; and $f$, the hydrophobicity of a residue), etc., belong to this class.

The mean value of the integral physico-chemical characteristic of the cluster alpha $F_{\alpha}$ in the case of unrelated sequences is calculated as

$$F_{\alpha} = \frac{1}{N} \sum_{k=1}^{N} F_{ak}. \tag{7}$$

Variance of the integral characteristic $D(F_{\alpha})$ is used as the measure of conservation of the value $F_{\alpha}$ for a set of its variability (Lim and Ptitsyn, 1970; Gerstein et al., 1994) for the sample of protein sequences:

$$D(F_{\alpha}) = \frac{1}{N-1} \sum_{k=1}^{N} (F_{ak} - F_{\alpha})^2. \tag{8}$$

In the case of evolutionary related sequences, we apply the Felsenstein’s method (Felsenstein, 1973, 1985) for estimating $F_{\alpha}$ and $D(F_{\alpha})$. Variance of the integral characteristics $D(F_{\alpha})$ may be expressed through variances $D(f_i)$ of the physico-chemical property at all the positions within the cluster $\alpha$ and the correlation coefficient $r_{ij}$. Thus, it comprises two components—$D_{var}$, formed by independent substitutions, and $D_{cov}$, depending on co-ordinated residue substitutions:

$$D(F_{\alpha}) = \sum_{i, j \in \alpha} c_i c_j r_{ij} D(f_i) D(f_j) = \sum_{i \in \alpha} c_i^2 D(f_i) + \sum_{i, j \in \alpha} c_i c_j r_{ij} D(f_i) D(f_j) = D_{var} + D_{cov}. \tag{9}$$

Note that $D_{var}$ is always positive or equal to zero, while $D_{cov}$ can be positive, negative, and equal to zero. The last case may be considered as a null hypothesis for testing the significance of the contribution of co-ordinated substitutions to $D(F_{\alpha})$. If $r_{ij} = 0$, the expected value of variance of the physico-chemical property $D_{exp}(F_{\alpha})$ coincides with $D_{var}$:

$$D_{exp}(F_{\alpha}) = D_{var} = \sum_{i \in \alpha} c_i^2 D(f_i). \tag{10}$$

If any co-ordinated substitution is detected in the cluster of positions $\alpha$, then

$$D(F_{\alpha}) \neq D_{exp}(F_{\alpha}), \tag{11}$$

and the null hypothesis is rejected.

In this case, the co-ordinated substitutions contribute additionally to the variance of the integral characteristics $F_{\alpha}$, namely, aiding the conservation of the characteristics $F_{\alpha}$, if

$$D(F_{\alpha}) < D_{exp}(F_{\alpha}). \tag{12}$$

On the other hand, if the inequality

$$D(F_{\alpha}) > D_{exp}(F_{\alpha}) \tag{13}$$

is fulfilled, this means that the characteristics $F_{\alpha}$ is hypervariable due to co-ordinated substitutions.

For testing inequality (equation (12)), we may use the ratio of variances $D_{exp}(F_{\alpha})/D(F_{\alpha})$, which, under the above null hypothesis follows $F$ distribution with $L_\alpha(N-1)$ and $N-1$ degrees of freedom, respectively, where $L_\alpha$ is the number of positions in the cluster. For verifying (equation (13)), we use the ratio $D(F_{\alpha})/D_{exp}(F_{\alpha})$ with $N-1$ and $L_\alpha(N-1)$ degrees of freedom (Selvin, 1998), respectively.

Additionally, we applied a Monte Carlo technique to estimate the statistical significance of the observed deviation of $D(F_{\alpha})$ from $D_{exp}(F_{\alpha})$. Let us consider a group of positions of the cluster $\alpha$ with a known mean value $f_i$ and variance $D(f_i)$ of the physico-chemical characteristic in question for each position. We generate a sample with a size $N$ of independent variables whose means and variances equal the estimates of their means and variances at each position within the cluster $\alpha$. Then, we estimate the variance $D_{rand}(F_{\alpha})$ for each simulated sample using equation (8). We repeat this procedure $M$ times, and construct the distribution of $D_{rand}(F_{\alpha})$.

When we test the hypothesis on conservation of the integral characteristic $F_{\alpha}$ we estimate the fraction $p(D_{rand}(F_{\alpha}) < D(F_{\alpha}))$. On the contrary, if $F_{\alpha}$, is hypervariable, we estimate the fraction $p(D_{rand}(F_{\alpha}) > D(F_{\alpha}))$. If $p$ is lower than a specified significance level $\omega$, we consider the deviation of $D(F_{\alpha})$ from $D_{exp}(F_{\alpha})$ statistically significant at a significance level of $P_\omega = 100\%(1 - p_\omega)$. 

1040
IMPLEMENTATION

Protein sequences and tertiary structures

In this work, we investigated DNA-binding domains of the homeodomain class (Bürglin, 1994). Many transcription factors involved in gene regulation contain these domains binding specifically to the corresponding DNA site (Wingender, 1997). The spatial structure of homeodomain is resolved (Kissinger et al., 1990). Namely, the homeodomain forms three α-helices, designated as H1 (positions 10–22), H2 (positions 28–37), and H3 (positions 42–58), joined by short loops. Helix H3 odomain forms three isoelectric point as an example for analysis (White are available at the CRASP www-site.

The sample of homeodomain sequences was extracted from the Pfam database (Bateman et al., 1999), entry PF00046. The alignment was not manually validated, however, we have removed identical sequences and the sequences with many deletions or insertions. The resulting sample contained \( N = 372 \) sequences. The positions of the multiple alignment where less than five amino acids were found were discarded too. Thus, the number of positions analyzed was \( L = 47 \). The phylogenetic tree necessary for obtaining the weighted variance and covariance estimates was constructed using the program CLUSTALW (Thompson et al., 1994) by the neighbor-joining method. The data on sequence samples are available at the CRASP www-site.

In this work, we selected the value of amino acid isoelectric point as an example for analysis (White et al., 1978) due to the following reasons. It is known that values of amino acid isoelectric point characterize the charge of residues (Nakai et al., 1988), and electrostatic interactions play an important role in formation of DNA–protein complexes, determining non-specific interactions between DNA and proteins (von Hippe and Berg, 1986). Thus, this property is likely to be important for the structure and/or function of homeodomain. Previously, only one type of integral characteristics connected with volume of the side groups of residues—total volume of protein hydrophobic core (Lim and Ptitsyn, 1970; Lim and Sauer, 1989; Gerstein et al., 1994; Clarke, 1995) and volume of the group of residues in the protein active site (Eisenhaber et al., 1998) was studied. However, certain experimental data suggest that the total charge of a group of amino acid residues might follow a similar conservation pattern (Jamieson et al., 1994). We obtained similar data on zinc-finger domains (Afonnikov and Wingender, 1998), DnaJ domains of heat shock proteins (Afonnikov, 2000a) and DNA-binding domains of CREB, and AP-1 proteins (Afonnikov et al., 1997). This suggested us to search for similar patterns in the homeodomain family proteins.

Robustness of the correlation coefficient estimates

To evaluate the robustness of correlation coefficients obtained, we used the initial sample of homeodomain sequences to generate 300 random samples with less size. The size \( N_r \) of the reduced sample was also determined randomly in the range of \( 0.8N < N_r < N \). For each reduced sample, the pairwise partial correlation coefficients of isoelectric point values of amino acids at all the positions of the sequences were estimated. The variance \( D(\rho_{r_{ij}}) \) of the normalized correlation coefficient \( \rho_{r_{ij}} = \frac{r_{ij}}{\sqrt{1/(N_r - 1)}} \) was calculated for each pair of positions \( i, j \). The higher is the variance \( D(\rho_{r_{ij}}) \), the less reliable is the correlation coefficient estimate for the pair in question. The estimate of the correlation coefficient for a pair of positions was considered robust if either its significance was the same in all runs or its value \( D(\rho_{r_{ij}}) \) is lower than 95% of the \( D(\rho_{r_{ij}}) \) estimates (5% of the pairs displaying highest \( D(\rho_{r_{ij}}) \) values were rejected).

Detecting clusters of correlated positions in homeodomain

The method described above was used for calculating the matrix of robust partial correlation coefficients between the isoelectric point values at the considered positions of homeodomain sequence. Using this matrix, we constructed a dendrogram to group the positions within the homeodomain displaying correlated substitutions (Figure 3).

The dendrogram displays two big clusters of positions with pairwise correlation coefficients at a significance level over 99.9% (\( r_{ij} \) values are listed in Table 1). The residues forming cluster I are located in spatial vicinity in the region where helix H1 (R15, R18, E19, and N23) contacts helix H2 (R30, Q33, S36, and E37; Figure 4). The residues forming cluster II are located in the N-terminal loop (position F8), the loop between helices H1 and H2 (Y25), and within helices H1 (L13 and K17) and H3 (E42 and K52). Except for residues at positions K17 and K52, the residues of this cluster have no contacts, and some of them (F8, L13, Y25, and E42) are located in the vicinity of DNA phosphate groups.

Studying integral characteristics of the clusters

Table 1 demonstrates that typical of cluster I is negative values of all the significant correlation coefficients between its positions. This means that the residues within the cluster are substituted in a compensatory manner with respect to the isoelectric point value, that is, the higher is this value at one position, the lower it is at the other. Therefore, we propose that the net \( Q_I \) value, which is a sum of the \( pI \) values of the amino acids forming cluster
where $pI_i$ is the isoelectric point value of the amino acid located at $i$th position. Variance of this value $D(Q_I)$ for homeodomains equals 80.72, while the expected variance calculated for independent substitutions according to equation (8) $D_{exp}(Q_I) = 127.74$. In this case, $D_{exp}(Q_I)/D(Q_I) \approx 1.58$. Thus, the condition (12) is fulfilled, thereby indicating that the integral characteristic $Q_I$ is conserved during the evolution of homeodomains. The statistical significance of equation (12) for $Q_I$ estimated by a Monte Carlo technique at $M = 10^5$ amounts to $P > 99.999\%$ (Table 2, Figure 5a). Note that the value $D_{rand}(Q_I)$ for neither of the random samples was less than the variances calculated for real protein sequences. The critical values of the ratio of variances estimated by $F$ distribution at 95 and 99% significance levels equal 1.13 and 1.19, respectively. These values were significantly less than the ratio $D_{exp}(Q_I)/D(Q_I)$.

The position pairs contained in cluster II may be divided into two groups: one with negative significant correlation coefficients (K52–K17 and K52–E42), the other with positive correlation coefficients (L13–F8, Y25–F8, and E42–F8). Interestingly, all the three pairs of the second group include position 8.

The positive correlation coefficients in the pairs involving position 8 are likely to suggest the compensatory nature of substitutions with respect to the difference in $pI$ values. This means that the co-ordinated substitutions in this case occur in such a way that the higher is the $pI$ value at one position, the higher it is at the other. In this process, the differences between the values of isoelectric points remain approximately constant. Hence, we have suggested that the characteristic

$$Q_{II} = pI_{13} + pI_{17} + pI_{25} + pI_{42} + pI_{52} - pI_8$$  \hspace{1cm} (15)$$
is conserved by the co-ordinated substitutions within cluster II. The $pI_8$ is considered with a negative sign, as position 8 correlates positively with positions 13, 25, and 42, while correlations with positions 17 and 52 are insignificant. The analysis performed demonstrated that the variance $D(Q_{II})$ for homeodomains is lower than the expected value for independent substitutions, that is, $D_{exp}(Q_{II})/D(Q_{II}) \approx 1.17$ (Table 2, Figure 5). Thus, inequality (12) is fulfilled, and consequently, the integral characteristic $Q_{II}$ is conservative. The statistical significance of equation (12) for $Q_{II}$ estimated by a Monte Carlo technique with $M = 10^5$ amounted to $P > 95\%$ (Table 2). The critical values of the ratio of variances, estimated by $F$ distribution at 95 and 99% significance levels, equal 1.14 and 1.20, respectively. These values are also consistent with the estimates of the $Q_{II}$ conservation obtained in the numerical experiments (Table 2).

Thus, the analysis performed suggests that the detected integral characteristics of clusters I and II, $Q_I$ and $Q_{II}$, determined by equations (14) and (15) are conservative.

---

**Table 1. Significant partial pairwise correlation coefficients of the isoelectric point values of amino acids at different positions of clusters I and II of homeodomain**

<table>
<thead>
<tr>
<th>Positions</th>
<th>$r_{ij}/g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R18–R15</td>
<td>-0.23</td>
</tr>
<tr>
<td>E19–R15</td>
<td>-0.22</td>
</tr>
<tr>
<td>N23–E19</td>
<td>-0.21</td>
</tr>
<tr>
<td>R30–E19*</td>
<td>-0.33</td>
</tr>
<tr>
<td>R30–N23</td>
<td>-0.19</td>
</tr>
<tr>
<td>Q33–E19</td>
<td>-0.19</td>
</tr>
<tr>
<td>Q33–R30</td>
<td>-0.20</td>
</tr>
<tr>
<td>S36–Q33</td>
<td>-0.20</td>
</tr>
<tr>
<td>E37–R15</td>
<td>-0.27</td>
</tr>
<tr>
<td>E37–E19</td>
<td>-0.19</td>
</tr>
<tr>
<td>E37–Q33</td>
<td>-0.19</td>
</tr>
<tr>
<td>L13–F8</td>
<td>0.18</td>
</tr>
<tr>
<td>Y25–F8*</td>
<td>0.33</td>
</tr>
<tr>
<td>E42–F8</td>
<td>0.20</td>
</tr>
<tr>
<td>K52–K17*</td>
<td>-0.27</td>
</tr>
<tr>
<td>K52–E42</td>
<td>-0.23</td>
</tr>
</tbody>
</table>

The threshold value $r_c = 0.182$ at the significance level $P = 99.9\%$. The types of residues listed correspond to the sequence of *engrailed* homeodomain (Kissinger et al., 1990). Asterisk marks the residues displaying the highest correlations according to Clarke (1995).
engrailed (1983), certain residues of the homeodomain two helices contact. According to criterion of Barlow and certain optimal value of the net charge at the site where substitutions within cluster I provide for maintaining a particular, we may hypothesize that the charge-compensating residue substitutions by electrostatic interactions. In par-

fl

lations discovered re

and the co-ordinated substitutions are involved in their conservation.

**DISCUSSION**

The performed analysis of DNA-binding domains of the homeodomain class has detected numerous pairs of positions where the amino acid substitutions are compensatory with respect to the isoelectric point value. It appeared that these positions formed two clusters, thus, the question arises as to what is the functional role of such clusters of correlated positions. The isoelectric point values of amino acids reflect their total charge (Nakai et al., 1988), therefore, the correlations discovered reflect the constraints imposed onto residue substitutions by electrostatic interactions. In particular, we may hypothesize that the charge-compensating substitutions within cluster I provide for maintaining a certain optimal value of the net charge at the site where two helices contact. According to criterion of Barlow and Thornton (1983), certain residues of the homeodomain *engrailed* (Kissing et al., 1990) belonging to cluster I (R30–E19, E37–R15, and E19–R15) form salt bridges (Figure 4). It is likely that the discovered compensatory substitutions maintain these interactions during the protein evolution. These interactions may be important for stabilising packing of helices H1 and H2. On the other hand, charges of the rest residues from the same cluster may also contribute to the helix packing stability. Consequently, the net charge of the residues forming cluster I is a critical parameter in question that determines the stability of helix packing. This parameter is conserved through co-ordinated amino acid substitutions.

The analysis performed has demonstrated that the majority of residues from cluster II (with the exception of positions 17 and 52) are located in the vicinity of phosphate groups in the DNA backbone. Thus, we may hypothesize that the correlations of pI values within cluster II reflect interactions of the corresponding residues with the negatively charged sugar–phosphate backbone of the DNA double helix. It is likely that these interactions assist correct orientation of homeodomain and DNA. In this process, specific charge distribution may appear functionally important and conserved during the evolution through co-ordinated substitutions of amino acid residues. Note that the dependence of substitutions at three pairs of positions belonging to clusters I and II with highest absolute values of correlation coefficients—F8–Y25, K52–K17, and E30–R19—was detected earlier (Clarke, 1995) using information theory. As is noted above, the pairs K52–K17 and E30–R19 contact directly in the spatial structure of the homeodomain *engrailed* (Kissing et al., 1990), while the residues at positions F8–Y25 are located in the vicinity of the sugar–phosphate backbone.

---

**Table 2.** Comparison of variances of characteristics $Q_I$ and $Q_{II}$ in the sample of homeodomain sequences and random samples

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>$D(F)$</th>
<th>$D_{exp}(F)$</th>
<th>$D_{rand}(F)$</th>
<th>$m(D_{rand}(F) &gt; D(F))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_I$</td>
<td>80.72</td>
<td>127.74</td>
<td>128.09</td>
<td>100,000</td>
</tr>
<tr>
<td>$Q_{II}$</td>
<td>16.17</td>
<td>18.94</td>
<td>18.99</td>
<td>98,339</td>
</tr>
</tbody>
</table>

$D(F)$, variance of each characteristic in the sample of homeodomain sequences; $D_{exp}(F)$, expected variance in the case of independent substitutions; $D_{rand}(F)$, mean variance in random samples; and $m$, the number of random samples where the expected variance value exceeded the variances of $D(F)$.

---

**Fig. 4.** Spatial structure of the homeodomain–DNA complex (Kissing et al., 1990). Spheres represent side groups of residues forming clusters I (dark spheres) and II (light spheres). Corresponding cluster numbers are shown in parentheses. Alpha helices and the DNA backbone are indicated.

**Fig. 5.** Distributions of the variance values of the characteristics (a) $Q_I$ and (b) $Q_{II}$ in simulated random samples with independent changes in pI values at individual positions within homeodomain: the x-axis, $D_{rand}$; the y-axis, fraction of samples with the corresponding $D_{rand}$ value; arrows indicate the variance values at homeodomain positions.
Statistical relationships between substitutions of residues are in no way unique of homeodomains. Similar patterns of amino acid pair preferences were discovered both in primary and tertiary structures of various proteins (Naor et al., 1996; Azarya-Sprinzak et al., 1997; Gugolya et al., 1997; Cootes et al., 1998; Rani and Mitra, 1996; Weiss and Herzel, 1998). It is assumed that co-ordinated substitutions appear due to structural or functional constraints of different physical nature that are imposed on proteins during their evolution. Such constraints include the necessity of preserving steric contacts (Göbel et al., 1994; Shindyalov et al., 1994), disulphide bonds (Kreisberg et al., 1995), ionic interactions (Clarke, 1995), or specific contacts with other molecules (Clarke, 1995; Pazos et al., 1997a; Nagl et al., 1999).

It is well known that homologous proteins display various conserved properties, that is, the physico-chemical characteristics that are maintained constant during the evolution of a protein family. As a rule, these characteristics depend on the properties of residues located at a number of positions within a protein, that is, they are integral. Protein hydrophobic core volume, net charge of a spatially close group of residues in an active site, and hydrophobic moment of an alpha helix exemplify such integral characteristics. Three mechanisms are proposed to underlie the conservation of such characteristics (Lim and Ptitsyn, 1970; Gerstein et al., 1994).

The first mechanism stems from the absolute invariance of positions within a protein important for its structure or function (for example, active centres, sites of contacts with subunits, or the regions interacting with ligands). It implies that any mutations at these positions impair the function of the protein. Analysis of homologous proteins confirms the absence of amino acid substitutions at these positions within the protein family.

The second mechanism involves conservative amino acid substitutions at certain positions in the protein sequence. These substitutions occur in such a way that the corresponding physico-chemical properties remain virtually unchanged (for example, the substitution of isoleucine with leucine in the hydrophobic core of a globule).

The third mechanism implies the existence of co-ordinated amino acid substitutions in the protein sequence. The actual occurrence of the third mechanism was investigated earlier in several works on the estimation of the volume constancy of protein hydrophobic core (Lim and Ptitsyn, 1970; Gerstein et al., 1994) and conservation of the total volume of residues in the vicinity of ω-site (Eisenhaber et al., 1998). Data on physico-chemical characteristics presumably conserved due to the third mechanism may be useful for research into the protein structure and function. On the other hand, this knowledge may find application in recognition procedures, being an additional criterion for filtering false predictions (Eisenhaber et al., 1999).

Characteristic of our approach is its ability not only to detect correlated substitution amino acid residues and clusters of correlating positions, but also to deduce previously unknown physico-chemical characteristics of proteins presumably conserved due to co-ordinated substitutions and analyze them. The approach we are proposing enables detection of co-ordinated substitutions and infer conserved protein characteristics with respect to particular physico-chemical properties of amino acid residues (volume, polarity, hydrophobicity, etc.). Each conserved characteristic reflects a particular type of interaction between amino acid residues. For example, the volume of residues determines their steric interactions; isoelectric point value characterizes the charge; while hydrophobicity, the interaction of a residue with the surrounding water molecules. Obviously, the co-ordinated substitutions may appear due to a complex interplay between a number of physico-chemical properties. For example, certain co-varying positions in the homeodomain detected using informational approach (Clarke, 1995) escaped our analysis. Involvement of the interactions other than electrostatic may explain this inaccuracy. However, the method developed at this stage is not intended to cover the totality of the possible interacting protein characteristics. We envision the interactions of amino acid residues as a hierarchical system ordered with respect to the impact of particular interaction types on the structure and function of a protein. It is likely that the most stringent constraints on amino acid substitutions are imposed by the top interactions with this hierarchy. As our method detects the most pronounced characteristics, it may be used as a first approximation to studying the complex interplay of amino acid properties. However, detection of all the integral protein characteristics and assessment of their particular contributions to the protein structure and function requires further investigation.

In this work, we are discussing an example of the integral protein characteristic and the capability of co-ordinated amino acid substitutions in conserving its value. Since the value of these integral characteristics—total charge of detected functional clusters of amino acid residues—is formed by a set of the residues, it should not imply necessarily direct interactions of residue. The same is true for other integral protein characteristics, such as distribution of charge on the protein surface, specific packing patterns of secondary structure elements, and relative location of active sites. Detection and analysis of remote interactions between both individual residues and larger elements of the protein require further investigation.
CONCLUSIONS
In this work, we are describing a software package CRASP, designed for analysis of co-ordinated amino acid substitutions in families of protein sequences. This package is available via the Internet and comprises three major modules.

The first module allows the pairwise correlation between substitutions of residues at individual positions within a protein sequence to be determined using linear and partial correlation coefficients between the values of certain physico-chemical characteristics of an amino acid. The second module allows the clusters of positions exhibiting significant correlation coefficients to be detected. The third module allows the conservation of the integral physico-chemical characteristics in question to be evaluated, comparing its constancy (variability) with the value expected in the random model of independent residue substitutions.

The CRASP package may be useful for theoretical investigation of protein structural and functional characteristics as well as for the design of mutational experiments.

ACKNOWLEDGEMENTS
The work is supported by the Russian Foundation for Basic Research (grant no. 98-07-91078, 98-07-90126, 99-07-90203, 99-04-49879, 00-07-90337), Russian National Human Genome Program, Integration Project of SB RAS IG2000/65, Integration grant of SB RAS No 66, The Committee of Science and Technology of the Russian Federation. The authors are grateful to Dmitry A.Grigorovich for providing the www-interface of the programs and Galina B.Chirikova for translation of the manuscript into English.

REFERENCES


