Protein Engineering vol.7 no.3 pp.349–358, 1994

# Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?

I.N.Shindyalov[1], N.A.Kolchanov[1] and C.Sander[2]

[1]Institute of Cytology and Genetics, Russian Academy of Sciences, Siberian Department, Prospect Lavrentyeva 10, Novosibirsk 630090, Russia and [2]EMBL, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

**A method has been developed to detect pairs of positions with correlated mutations in protein multiple sequence alignments. The method is based on reconstruction of the phylogenetic tree for a set of sequences and statistical analysis of the distribution of mutations in the branches of the tree. The database of homology-derived protein structures (HSSP) is used as the source of multiple sequence alignments for proteins of known three-dimensional structure. We analyse pairs of positions with correlated mutations in 67 protein families and show quantitatively that the presence of such positions is a typical feature of protein families. A significant but weak tendency is observed for correlated residue pairs to be close in the three-dimensional structure. With further improvements, methods of this type may be useful for the prediction of residue–residue contacts and subsequent prediction of protein structure using distance geometry algorithms. In conclusion, we suggest a new experimental approach to protein structure determination in which selection of functional mutants after random mutagenesis and analysis of correlated mutations provide sufficient proximity constraints for calculation of the protein fold.**

*Key words:* distance geometry/homology/protein structure/residue contacts/sequence alignment

## Introduction

There is much experimental data demonstrating that the amino acid sequence contains the information necessary to determine the three-dimensional structure of a protein (Epstein *et al.*, 1963; Anfinsen, 1981; Creighton, 1983). However, the problem of protein structure prediction from the amino acid sequence remains unsolved in general. Information contained in the multiple sequences of an entire protein family, properly aligned, can be very useful to predict functionally important or surface residues (Lesser *et al.*, 1987; Valencia *et al.*, 1991) and to improve the prediction of protein secondary structure (Benner and Gerloff, 1991; Rost and Sander, 1993a,b).

Here we investigate the hypothesis that a sequence family contains information about residue–residue contacts in the three-dimensional structure. The main idea of the approach is the following. In the course of evolution, mutations in certain positions are fixed in a correlated manner, as the result of structural constraints imposed by the integrity of the three-dimensional structure (Godzik and Sander, 1989) or as a result of specific functional constraints. This idea was used by Altschuh *et al.* (1987) in their analysis of correlated amino acid substitutions in the coat protein structure of tobacco mosaic virus and seven related viruses. They showed that some pairs of positions with identical patterns of amino acid substitutions are close together in the three-dimensional structure of the coat protein. Later this approach was used for the analysis of three additional protein families (serine proteases, cysteine proteases and hemoglobins) (Altschuh *et al.*, 1988).

The approach is related to powerful methods for RNA secondary structure predictions based on the analysis of aligned sets of RNA molecules (Konigs and Hogeweg, 1989). Opposite positions in the functionally important hairpins of RNA molecules evolve in a correlated manner, with the conservation of the complementary nucleotide pairs. The rules of correlated evolution of base pairs are very simple, as each nucleotide can interact specifically only with one of three other nucleotides (A–U and C–G nucleotide pairs). In protein structures the situation is significantly more complicated, because there are no very simple rules for preferable residue–residue interactions in the three-dimensional structure.

In this paper, we present a new method for the identification of positions with correlated mutations in protein families. The method is based on the careful statistical analysis of correlation between accepted amino acid substitutions in different protein positions. Using this method, the phenomenon of correlated mutation is studied for a wide range of protein families. The analysis relies on the information in the database of protein structures (Protein Data Bank, Bernstein *et al.*, 1977) and of protein family alignments (HSSP, Sander and Schneider, 1991). Construction of phylogenetic trees was used to obtain amino acid substitutions which are fixed in the course of evolution of each family, rather than counting all possible pairs at one sequence position in the set of sequences, to avoid overcounting of mutations.

Applying stringent statistical criteria to filter out possible statistical noise, enough information remained to demonstrate some correlated mutations for 67 out of 91 considered protein families and to investigate the relation between correlated mutation and the position in the three-dimensional structure. Several factors providing for the appearance of such positions with correlated mutations are considered.

## Materials and methods

### *Sequence data*

Aligned amino acid sequences of 91 different families were taken from the HSSP database (Sander and Schneider, 1991). For each protein in the Protein Data Bank (PDB, Bernstein *et al.*, 1977), this database contains a multiple alignment of all proteins deemed homologous to the structurally known protein, i.e. those with sequence similarity above the threshold for structural homology. Initially a set of 105 non-homologous proteins was selected from the PDB, i.e. no pair in this set had significant sequence similarity (Hobohm *et al.*, 1992). Of these, 14 proteins were discarded because they had fewer than two aligned sequences in the HSSP family alignment. In each of the remaining 91 HSSP data sets, sequences near the twilight zone of sequence similarity were removed with a safety margin of plus 3 percentage points by requiring that sequence similarity to the master protein exceed

$$t(L) = 290.15L^{-0.562} + 3 \quad (1)$$

where $t(L)$ is the threshold of structural homology in percent identical residues (Sander and Schneider, 1991) and $L$ is the length of the alignment.

The PDB identifiers of the 91 protein families selected for analysis are (chain identifier in parenthesis): 1ACX, 1BP2, 1CBH, 1CCR, 1CD4, 1CRN, 1CSE(E), 1CSE(I), 1CTF, 1ECN, 1FC2(C), 1FXB, 1GCN, 1GCR, 1GD1(O), 1GOX, 1GP1(A), 1HOE, 1I1B, 1IL8(A), 1L13, 1LZ1, 1MBD, 1NXB, 1PAZ, 1PPT, 1PRC(H), 1PRC(L), 1R08(4), 1R69, 1SN3, 1TGS(I), 1TNF(A), 1UBQ, 1UTG, 1WSY(A), 1YPI(A), 2AZA(A), 2CA2, 2CCY(A), 2CDV, 2ER7(E), 2HLA(B), 2LIV, 2LTN(A), 2LTN(B), 2MEV(4), 2MHR, 2MLT(A), 2OVO, 2PAB(A), 2PLV(4), 2PLV(2), 2PLV(3), 2RNT, 2RSP(A), 2SGA, 2SNS, 2SOD(O), 2SSI, 2TMN(E), 2TS1, 351C, 3B5C, 3BLM, 3CLA, 3FXC, 3GAP(A), 3HMG(A), 3HMG(B), 3PGM, 4CPV, 4FD1, 4FXN, 4INS, 4INS(B), 4PTP, 4SGB(I), 4XIA(A), 5CPA, 5HIR, 5PTI, 5RXN, 6LDH, 7API(B), 7RSA, 8ADH, 8ATC(A), 8ATC(B), 8DFR, 9WGA(A). Unfortunately there is not enough room to cite individual references.

*Construction of phylogenetic trees*

For each set of aligned sequences, i.e. for each HSSP file, a phylogenetic tree was constructed by the UPGMA method (unweighted pair group method using arithmetic averages, Sokal and Sneath, 1963). Ancestral sequences for all vertices in the tree as well as the list of amino acid substitutions for each branch of the tree were determined by the method of Hartigan (1973).

*Statistical criterion for correlated mutations*

The main technical problem is to identify those pairs of positions which are significantly correlated. We do this by defining a matrix of correlated mutational events and estimating the probability to observe correlation in a particular pair in a random situation.

Consider a phylogenetic tree for a set of amino acid sequences of length $L$ containing $K$ branches, i.e. $K-1$ proteins connected to an ancestor by mutational events (Figure 1). We describe the set of mutational events in the tree by a binary matrix

$$G = \{g_{lk}\},\ l = 1, \ldots, L;\ k = 1, \ldots, K \qquad (2)$$

Here $l$ is the residue position in the alignment and $k$ is the branch index in the phylogenetic tree. Each element of $g_{lk}$ is either 1 or 0 according to

$$g_{lk} = \begin{cases} 1, & \text{if a mutation occurred on branch } k \text{ at position } l \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

An example is given in Figure 1. A mutational event is simply defined by the fact that the residues at position $l$ are not equal in the two proteins connected by branch $k$. Note that this definition excludes consideration of back mutations or multiple mutations in a single branch. The mutational events considered are the simplest set of events connecting the two proteins at branch $k$, not an estimate of the actual historical mutations.

Two amino acid substitutions at positions $l_1$ and $l_2$ will be considered as correlated if they occur on the same $k$th branch of the phylogenetic tree, i.e.

$$g_{l_1k} \cdot g_{l_2k} = 1 \qquad (4)$$

for at least one branch $k$. So in a 'correlated mutation' residues change at two sequence positions 'simultaneously'.

The total number of correlated mutations in the phylogenetic tree for a pair of positions $l_1$ and $l_2$ is simply the sum over all branches

$$M_{l_1l_2} = \sum_{k=1}^{K} g_{l_1k} \cdot g_{l_2k} \qquad (5)$$

Given one or more correlated mutations at a pair of positions $l_1$, $l_2$, what is the criterion for concluding that the mutations at these two positions are generally correlated, so that the pair of positions can be called a 'correlated pair of positions'? The answer depends on a statistical estimate. If the probability $P(M_{l_1l_2})$ to obtain by chance $M_{l_1l_2}$ or more correlated mutations at positions $l_1$ and $l_2$ is small, then these positions can be considered as evolving in a correlated manner. For simplicity we will call such pairs $l_1$, $l_2$ 'correlated positions'.

The probability $P(M_{l_1l_2})$ assuming a random series of mutational events is estimated as follows. The total number of mutations in the tree is

$$N = \sum_{l=1}^{L} \sum_{k=1}^{K} g_{lk} \qquad (6)$$

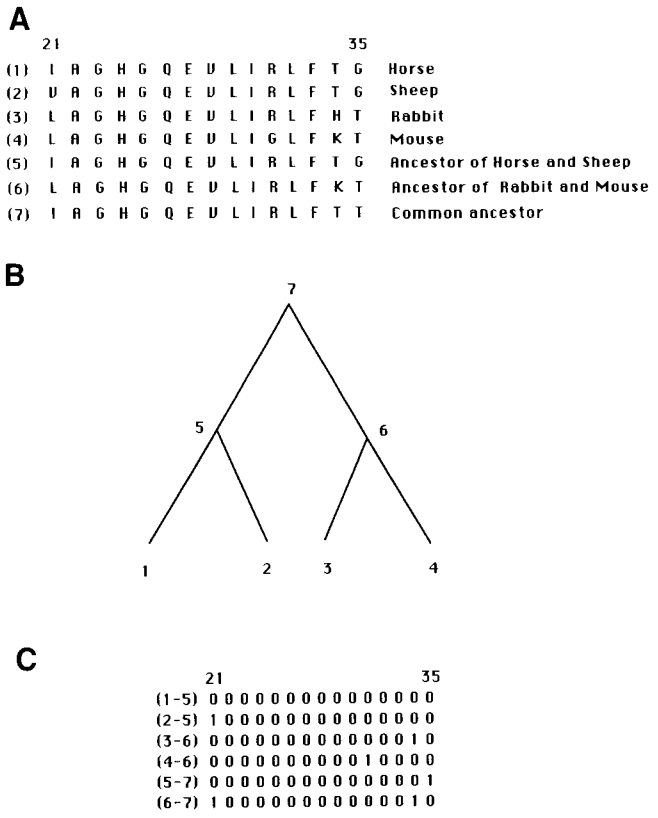


**Fig. 1.** Derivation of mutation matrix $G$ from the multiple sequence alignment via an evolutionary tree. Example from myoglobin, extracted from multiple sequence alignment data set 1mbd.hssp. (**A**) Amino acid sequences (residues 21–35 only) for four selected sequences and three reconstructed ancestral sequences. (**B**) Evolutionary tree for the seven sequence fragments; nodes labelled with the sequence number. (**C**) Matrix $g_{lk}$ of mutational events (equation 2); each row corresponds to one branch of the tree; the number 1 reflects a mutation in a branch (index $k$) at a particular position (index $l$); the number 0, no mutation. There is only one 'correlated' pair of mutations: in branch 6–7 at positions 21 (I → L) and 34 (T → K).

The average frequency of mutations on the $k$th branch, averaged over all $L$ positions is

$$G_k = \sum_{l=1}^{L} g_{lk}/N \tag{7}$$

and the average frequency of a mutation in the $l$th position, averaged over all branches, is

$$G_l = \sum_{k=1}^{K} g_{lk}/N \tag{8}$$

Interpreting the frequencies $G_l$ and $G_k$ as probabilities, a reasonable estimate of the probability of observing a mutation at the $l$th position on the $k$th branch of a tree that has a total number $N$ of mutations is

$$P_{lk} = 1 - (1 - G_k \cdot G_l)^N \tag{9}$$

where $(1 - G_k \cdot G_l)^N$ is the probability of seeing no mutation in spite of $N$ attempts and $(1 - G_k \cdot G_l)$ is the probability of seeing no mutation in a single attempt. Equation (9) is approximate and works if $N \ll L \cdot K$, i.e. if the matrix $G$ is only sparsely filled with mutations.

What then is the probability of obtaining by chance correlated mutations? More precisely, what is the probability $P(M_{l_1l_2})$ of obtaining $M_{l_1l_2}$ or more correlated mutations in positions $l_1$ and $l_2$? A reasonable estimate is

$$P(M_{l_1l_2}) = 1 - \sum_{j=0}^{M_{l_1l_2}-1} \sum_{i=0}^{C_K^j} \prod_{k=1}^{K} D_{l_1l_2k} \tag{10}$$

This expression is based on considering all possible ways of obtaining less than $M_{l_1l_2}$ correlated mutations in a tree with $K$ branches. The first sum reflects different total numbers $j$ of correlated mutations at positions $l_1$ and $l_2$ ($0 \le j \le M_{l_1l_2} - 1$). The second sum reflects different ways of placing $j$ correlated mutations in $K$ branches ($0 \le i \le C_K^j$), where the upper limit $C_K^j$ is given by the binomial coefficient.

The product over $K$ branches reflects the specification of one such way, i.e. of a particular tree. For each branch $k$ the elementary probability of seeing a mutation at $l_1$ and at $l_2$ is $P_{l_1k} \cdot P_{l_2k}$, that for not seeing a correlated mutation is $1 - P_{l_1k} \cdot P_{l_2k}$, using the single residue probabilities in equation (9). In order to conveniently take the product over all cases, correlated or not, we define a common symbol for the probability $D_{l_1l_2k}$ of a particular mutation state in branch $k$ at positions $l_1$ and at $l_2$

$$D_{l_1l_2k} = \begin{cases} P_{l_1k} \cdot P_{l_2k}, & \text{if } k\text{th branch contains correlated mutations in positions } l_1 \text{ and } l_2 \\ 1 - P_{l_1k} \cdot P_{l_2k}, & \text{otherwise} \end{cases} \tag{11}$$

How surprised should one be observing $M$ or more correlated mutations for a given pair of positions $l_1$ and at $l_2$? Statistical significance is expressed in terms of an empirically chosen probability threshold, $p_0$: if

$$P(M_{l_1l_2}) \le p_0 \tag{12}$$

then positions $l_1$ and $l_2$ are considered significantly correlated. We use $p_0 = 0.05$, corresponding to 1.96 standard deviations away from a random expectation or $p_0 = 0.20$, corresponding to 1.25 standard deviations (see Results).

In practice, the product over $k$ and sum over $i$ of equation (10) is evaluated approximately as

$$C_K^j \cdot p^j (1 - p)^{K-j}$$

where $p$ is a multiplicative average of $P_{l_1k} \cdot P_{l_2k}$ over all $l_1$, $l_2$ and $k$, i.e. $p$ is the average probability, per branch, of observing one simultaneous mutation. The average is such that in the special case $j = 0$ the approximate expression becomes exact—simultaneous mutations in any branch $k$ are independent events. So we have

$$<p> = 1 - \left( \prod_{k=1}^{K} (1 - P_{l_1k} \cdot P_{l_2k}) \right)^{1/K}$$

such that

$$(1 - <p>)^K = \prod_{k=1}^{K} (1 - P_{l_1k} \cdot P_{l_2k})$$

We have checked the validity of this approximation by explicit calculation.

With this precise definition (equation 12) of correlated positions, one can analyse the tree of a particular protein family and enumerate all $K^c$ correlated pairs of positions.

Given a set of correlated positions, what can be said about the physical reason behind the correlations? To answer this question, we next look into the relationship between spatial proximity of a residue pair in the three-dimensional structure and its tendency to evolve in a correlated fashion.

*Definition of side chain contacts*

The three-dimensional structures of proteins used for this analysis are from the Protein Data Bank (Bernstein *et al.*, 1977).

For each protein structure the distances $R(l_1,l_2)$ are calculated for all possible pairs of amino acids. The distance between pairs of amino acids is defined as the minimal distance between centres of atoms of the two side chains, excluding hydrogen atoms. For glycine, the $C_\alpha$ atom is used. Spatially close amino acids, i.e. residue pairs in contact, are defined by the condition

$$R(l_1,l_2) \le 6 \text{ Å} \tag{13}$$

*Statistical criterion for side chain contacts*

Given residue pairs with correlated mutations, we need to assess whether they have a significant tendency to be in spatial proximity (to have side chain−side chain contacts). To perform an explicit statistical estimate, we need the following variables:

$Q$, the total number of pairs;
$Q_r$, the number of pairs in contact;
$Q_{\bar{r}}$, the number of pairs not in contact;
$Q_{sr}$, the number of correlated pairs in contact;
$Q_{s\bar{r}}$, the number of correlated pairs not in contact;
$Q_{\bar{s}r}$, the number of non-correlated pairs in contact;
$Q_{\bar{s}\bar{r}}$, the number of non-correlated pairs not in contact.

The probability $P_r$ of two residues to be in contact is estimated as

$$P_r = (Q_{sr} + Q_{\bar{s}r})/Q$$

The probability $\tilde{P}$ to obtain by chance a particular combination of values $Q_{sr}$, $Q_{s\bar{r}}$, $Q_{\bar{s}r}$, $Q_{\bar{s}\bar{r}}$ is estimated using the binomial distribution

$$\tilde{P} = \begin{cases} \sum_{i=Q_{sr}}^{Q_s} C_{Q_s}^{i} (P_r)^i (1 - P_r)^{Q_s - i}, & \text{if } Q_{sr} \geq P_r \cdot Q_s \\ 1, & \text{otherwise} \end{cases} \quad (14)$$

where $Q_s = Q_{sr} + Q_{s\bar{r}}$.

If $\tilde{P} \leq \omega$ (where $\omega$ is a small value, here chosen empirically as $\omega = 0.05$, corresponding to two standard deviations away from random expectation) then there is a significant correlation between the fact that two residues evolve in a correlated manner and the fact that they are in contact, i.e. are spatially close in the three-dimensional structure. Below, this estimate is applied to pairs with a chain separation of $d > 10$, i.e. pairs involved in 'tertiary' structure contacts.

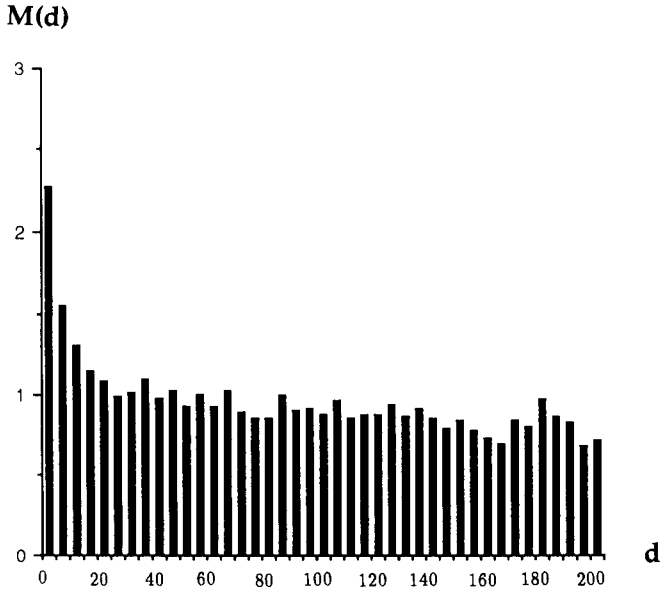


**Fig. 2.** Distribution $M(d)$ of the number of correlated pairs as a function of chain distance $d$, derived from multiple sequence alignments. For example, 1% of all correlated pairs are at a chain distance of 25–30 residues. Averaged over 67 proteins. The figure shows that most correlated pairs have small chain distance, but that some correlated pairs also occur at large chain distance.

## Results

### *Correlated positions in protein families*

Statistically significant correlated mutations were detected in 67 families of proteins of known three-dimensional structure (Table I), using a threshold of $p_0 = 0.05$ (equation 12), corresponding to two standard deviations away from random expectation. For example, 143 members of the trypsin family of serine proteases were analysed, aligned relative to $\beta$ trypsin (PDB code 4PTP, 223 residues). In this family, 2049 (8.4%) pairs of positions out of 24 531 possible pairs were significantly correlated in their mutational behaviour (24 531 = 223 × 222/2 − 222, not taking into account neighbouring positions). Twelve protein families had more than 2.0% of correlated pairs, but most had <1.0% of such pairs. Families with fewer family members also tended to have a lower number of correlated pairs, probably as a result of increased statistical noise in combination with a constant cut-off. It is reasonable to assume that more sequence information in each family would yield more correlated pairs.

So, a small but significant number of correlated pairs can be detected by our analysis in most protein familes. Which factors are responsible for their presence? Can they be exploited for structure prediction?

### *Relation between mutational correlation and one-dimensional chain separation of residue pairs*

Does the occurrence of correlated mutations depend on chain separation? Are chain neighbours more likely to mutate in unison? A histogram (Figure 2) of the number of correlated mutations $M(d)$ against chain distance $d$ for the 67 proteins of Table I provides some quantitative answers to these questions. Here, $M(d)$ is the fraction of pairs with chain separation $d$ that are mutationally correlated. The data, derived from multiple sequence information alone, show the following trends.

(i) Close neighbours in the protein chain have the highest probability to mutate in a correlated manner. An example of correlated evolution of contacting amino acids in chain proximity is given by positions 102 and 106 in the cytochrome $c$ family

**Table I.** Characteristics of correlated mutations in protein families

| ID | $L$ | $N$ | Protein name | $K^c$ | $K^t$ | $K^0$ |
|---|---|---|---|---|---|---|
| 4PTP | 223 | 143 | $\beta$-Trypsin | 2049 | 24 531 | 8.4% |
| 1CCR | 111 | 104 | Cytochrome $c$ | 268 | 5995 | 4.5% |
| 4CPV | 108 | 76 | Calcium-binding parvalbumin | 229 | 5671 | 4.0% |
| 1MBD | 153 | 102 | Myoglobin | 444 | 11 476 | 3.9% |
| 1CTF | 68 | 28 | L7/L12 50S ribosomal protein | 73 | 2211 | 3.3% |
| 3FXC | 98 | 62 | Ferredoxin | 146 | 4656 | 3.1% |
| 2OVO | 56 | 87 | Ovomucoid (third domain) | 45 | 1485 | 3.0% |
| 1NXB | 62 | 75 | Neurotoxin B | 52 | 1830 | 2.8% |
| 7RSA | 124 | 52 | Ribonuclease A | 171 | 7503 | 2.3% |
| 1CSE (E) | 274 | 21 | Subtilisin | 773 | 37 128 | 2.1% |
| 1ECN | 136 | 19 | Haemoglobin | 191 | 9045 | 2.1% |
| 1BP2 | 123 | 84 | Phospholipase A2 | 151 | 7381 | 2.0% |
| 1GD1 (O) | 334 | 46 | D-Glyceraldehyde-3-phosphate dehydrogenase | 1028 | 55 278 | 1.9% |
| 1LZ1 | 130 | 42 | Lysozyme | 153 | 8256 | 1.9% |
| 2CA2 | 256 | 18 | Carbonic anhydrase II | 545 | 32 385 | 1.7% |
| 2ER7 (E) | 330 | 17 | Endothiapepsin | 931 | 53 956 | 1.7% |

**Table I.** *Continued*

| ID | $L$ | $N$ | Protein name | $K^c$ | $K^t$ | $K^0$ |
|---|---|---|---|---|---|---|
| 2SOD (O) | 151 | 28 | Cu,Zn superoxide dismutase | 187 | 11 175 | 1.7% |
| 5PTI | 58 | 42 | Trypsin inhibitor | 27 | 1596 | 1.7% |
| 9WGA (A) | 171 | 109 | Wheat germ agglutinin (isolectin 2) | 245 | 14 365 | 1.7% |
| 2PLV (2) | 268 | 25 | Poliovirus (type 1, mahoney strain) | 555 | 35 511 | 1.6% |
| 3HMG (A) | 328 | 70 | Haemagglutinin | 857 | 53 301 | 1.6% |
| 8DFR | 186 | 19 | Dihydrofolate reductase | 256 | 17 020 | 1.5% |
| 6LDH | 329 | 28 | Lactate dehydrogenase | 793 | 53 628 | 1.5% |
| 8ADH | 374 | 31 | Apo-liver alcohol dehydrogenase | 877 | 69 378 | 1.3% |
| 3HMG (B) | 175 | 69 | Haemagglutinin | 189 | 15 051 | 1.3% |
| 2PLV (3) | 235 | 24 | Poliovirus (type 1, mahoney strain) | 362 | 27 261 | 1.3% |
| 2HLA (B) | 99 | 30 | Human class I histocompatibility antigen | 60 | 4753 | 1.3% |
| 2LTN (A) | 181 | 23 | Pea lectin | 208 | 16 110 | 1.3% |
| 2LTN (B) | 47 | 31 | Pea lectin | 13 | 1035 | 1.3% |
| 3BLM | 257 | 15 | $\beta$-Lactamase | 383 | 32 640 | 1.2% |
| 1TGS (I) | 56 | 25 | Trypsin inhibitor | 18 | 1485 | 1.2% |
| 1GCR | 174 | 105 | $\gamma$-II Crystallin | 182 | 14 878 | 1.2% |
| 1GCN | 29 | 41 | Glucagon | 4 | 378 | 1.1% |
| 4FD1 | 106 | 24 | Ferredoxin | 53 | 5460 | 1.0% |
| 4FXN | 138 | 11 | Flavodoxin | 87 | 9316 | 0.9% |
| 1IL8 (A) | 71 | 17 | Interleukin 8 | 21 | 2415 | 0.9% |
| 3B5C | 85 | 21 | Cytochrome *b5* | 30 | 3486 | 0.9% |
| 1CD4 | 173 | 35 | Cd4 | 121 | 14 706 | 0.8% |
| 1UBQ | 76 | 17 | Ubiquitin | 21 | 2775 | 0.8% |
| 1YPI (A) | 247 | 13 | Triose phosphate isomerase | 246 | 30 135 | 0.8% |
| 8ATC (A) | 310 | 10 | Aspartate carbamoyltransferase | 379 | 47 586 | 0.8% |
| 1ACX | 107 | 8 | Actinoxanthin | 38 | 5565 | 0.7% |
| 2CCY (A) | 127 | 8 | Cytochrome *c'* | 52 | 7875 | 0.7% |
| 5RXN | 54 | 14 | Rubredoxin | 9 | 1378 | 0.7% |
| 2RNT | 104 | 11 | Lys 25-ribonuclease t1 | 37 | 5253 | 0.7% |
| 5CPA | 307 | 9 | Carboxypeptidase A$\alpha$ | 265 | 46 665 | 0.6% |
| 1PAZ | 120 | 14 | Pseudoazurin | 45 | 7021 | 0.6% |
| 1PRC (L) | 273 | 10 | Photosynthetic reaction centre | 232 | 36 856 | 0.6% |
| 3CLA | 213 | 9 | Type III chloramphenicol acetyltransferase | 119 | 22 366 | 0.5% |
| 1I1B | 151 | 7 | Interleukin-1$\beta$ | 54 | 11 175 | 0.5% |
| 1SN3 | 65 | 18 | Scorpion neurotoxin (variant 3) | 11 | 2016 | 0.5% |
| 2CDV | 107 | 7 | Cytochrome c3 | 21 | 5565 | 0.4% |
| 351C | 82 | 9 | Cytochrome c551 | 12 | 3240 | 0.4% |
| 1CSE (I) | 63 | 10 | Eglin-c | 5 | 1891 | 0.3% |
| 2AZA (A) | 129 | 13 | Azurin | 23 | 8128 | 0.3% |
| 4INS (B) | 30 | 64 | Insulin | 1 | 406 | 0.2% |
| 1CBH | 36 | 7 | C-terminal domain of cellobiohydrolase I | 1 | 595 | 0.2% |
| 1TNF (A) | 152 | 10 | Tumour necrosis factor-$\alpha$ (cachectin) | 25 | 11 325 | 0.2% |
| 1R08 (4) | 40 | 16 | Rhinovirus 14 | 1 | 741 | 0.1% |
| 1CRN | 46 | 12 | Crambin | 1 | 990 | 0.1% |
| 1GP1 (A) | 183 | 8 | Clutathione peroxidase | 15 | 16 471 | 0.0% |
| 1L13 | 164 | 5 | Lysozyme | 1 | 13 203 | 0.0% |
| 2PLV (4) | 62 | 16 | Poliovirus (type 1, mahoney strain) | 1 | 1830 | 0.0% |
| 2TMN (E) | 316 | 6 | Thermolysin | 46 | 49 455 | 0.0% |
| 3GAP (A) | 208 | 6 | Catabolite gene activator protein | 1 | 21 321 | 0.0% |
| 3PGM | 230 | 7 | Phosphoglycerate mutase | 7 | 26 106 | 0.0% |
| 4SGB (I) | 51 | 9 | Potato inhibitor | 1 | 1225 | 0.0% |

ID, Protein Data Bank identifier, protein subunit in parentheses; $L$, chain length; $N$, number of sequences in multiple sequence alignment; $K^c$, number of pairs of correlated positions; $K^t$, total number of pairs of positions; $K^0$, $K^c/K^t$ in percent.

(Figure 4). These residues are located on an $\alpha$ helix that participates in a hydrophobic cluster. According to criterion (13) these positions evolve in a correlated manner [$P(M_{l_1 l_2}) = 0.04$]. All pairs of substituted amino acids in these positions conserve the hydrophobic cluster.

(ii) The longer the chain distance between two positions, the lower the probability of correlated mutation.

(iii) Even positions which are very distant along the chain can mutate in a correlated manner.

*Evaluation of the spatial proximity of correlated positions*

To make the analysis of correlated mutations relevant for the prediction of three-dimensional structure, it is useful to concentrate on residue pairs that bridge secondary structure segments, i.e. pairs for which the chain separation is not small. So we concentrate on pairs with a chain separation of $d > 10$ and ask if there is a connection between the tendency of some pairs of positions to evolve in a correlated manner and their spatial proximity in the three-dimensional structure of proteins. The

**Table II.** Relation between the correlated evolution of positions and their spatial proximity in the three-dimensional structure: contact prediction, all positions

| Protein family | $Q_{sr}$ | $Q_{s\bar{r}}$ | $Q_{\bar{s}r}$ | $Q_{\bar{s}\bar{r}}$ | $H$ | $\hat{H}$ | $\tilde{P}$ |
|---|---|---|---|---|---|---|---|
| Trypsin inhibitor | 6 | 10 | 160 | 952 | 37.5% | 14.7% | 0.0215 |
| Phospholipase A2 | 9 | 101 | 160 | 6058 | 8.2% | 2.7% | 0.0028 |
| Cytochrome *c* | 15 | 185 | 134 | 4716 | 7.5% | 3.04% | 0.0009 |
| Haemoglobin | 8 | 144 | 160 | 7563 | 5.3% | 2.1% | 0.0169 |
| γ-II Crystallin | 11 | 140 | 330 | 12 885 | 7.3% | 2.6% | 0.0018 |
| Myoglobin | 16 | 355 | 173 | 9609 | 4.3% | 1.9% | 0.0018 |
| Photosynthetic reaction centre | 7 | 201 | 248 | 33 997 | 3.4% | 0.7% | 0.0010 |
| Carbonic anhydrase II | 18 | 486 | 508 | 29 123 | 3.6% | 1.7% | 0.0039 |
| Endothiapepsin | 19 | 821 | 636 | 49 564 | 2.3% | 1.3% | 0.0141 |
| Pea lectin | 7 | 167 | 248 | 14 113 | 4.0% | 1.8% | 0.0348 |
| Cu,Zn superoxide dismutase | 10 | 143 | 281 | 9436 | 6.5% | 2.9% | 0.0158 |
| β-Trypsin | 62 | 1666 | 395 | 20 455 | 3.6% | 2.0% | 0.00001 |
| Lactate dehydrogenase | 17 | 714 | 523 | 49 467 | 2.3% | 1.1% | 0.0027 |
| Apo-liver alcohol dehydrogenase | 15 | 799 | 733 | 64 519 | 1.8% | 1.1% | 0.0447 |
| Aspartate carbamoyltransferase | 15 | 318 | 574 | 43 934 | 4.5% | 1.3% | 0.00004 |

$Q_{sr}$, the number of correlated and spatially close positions; $Q_{s\bar{r}}$, the number of correlated and spatially distant positions; $Q_{\bar{s}r}$, the number of non-correlated and spatially close positions; $Q_{\bar{s}\bar{r}}$, the number of non-correlated and spatially distant positions; $H$, fraction of pairs of correlated positions which are spatially close ('correctly predicted'); $\hat{H}$, fraction of pairs of positions which are spatially close; $\tilde{P}$, probability obtaining the set of numbers $\{Q\}$ by chance.

estimate of statistical significance of the connection is performed using equation (14).

We find that of the 65 proteins which contain correlated positions at a chain separation of $d > 10$, 15 proteins have a significant correlation (Table II). An example is trypsin inhibitor, with 16 correlated pairs of positions at a chain distance of $d > 10$ Å. Of these (see Figure 5), six pairs have side chain contacts (true positives), while 10 make no contact (false positives), corresponding to 37% correct prediction. This is a non-trivial result, as the probability to obtain by chance six close and correlated positions is estimated as 0.02. This and other examples (Table II) establish a statistical link between correlated evolution and spatial proximity for some sequence positions. In these cases specific residue–residue interactions appear to be responsible for the evolutionary pressure that leads to correlated mutations. The number of such correctly predicted contact pairs ($Q_{sr}$) ranges from six for trypsin inhibitor to 62 for trypsin. However, in this part of the analysis, the average accuracy of contact prediction ($H$) is low, rarely exceeding 5% (Table II). These levels are too low for general prediction of contacts between side chains. Note, however, that our statistical criteria are very strict and may be relaxed in prediction applications.

*Prediction of hydrophobic contact between secondary structure elements*

In a refined phase of the analysis, the following modified criteria were applied in order to improve the predictive value of the correlations.

(i) All proteins of length $L > 200$ were excluded. In addition, only pairs at a chain separation $d < 70$ were considered. These conditions were chosen so as to restrict the analysis to single domain proteins and intradomain contacts.

(ii) Only pairs with a chain separation of $d > 10$ were considered. This condition eliminates from consideration all residue pairs where spatial proximity is simply the result of closeness along the protein sequence.

(iii) Only positions with a conservative hydrophobic pattern of amino acid substitutions were considered, i.e. positions in which more than 90% of all residue replacements in the multiple sequence alignment remain in the hydrophobic class (V, L, I, W, F, C). This condition reflects the crucial role of the hydrophobic core in the determination of protein structure.

(iv) In this part of the analysis, a more relaxed threshold level of $p_0 = 0.20$, corresponding to 1.25 standard deviations away from random expectation, was used for the selection of correlated positions. The choice of cut-off is empirical, and reflects a trade-off between a larger number of pairs and higher significance.

**Table III.** Relation between the correlated evolution of positions and their spatial proximity in the three-dimensional structure: contact prediction for hydrophobic positions

| Protein family | $Q_{sr}$ | $Q_{s\bar{r}}$ | $Q_{\bar{s}r}$ | $Q_{\bar{s}\bar{r}}$ | $H$ | $\hat{H}$ |
|---|---|---|---|---|---|---|
| Cytochrome *c* | 6 | 11 | 14 | 76 | 35.3% | 18.7% |
| Subtilisin | 4 | 4 | 16 | 56 | 50.0% | 25.0% |
| Myoglobin | 10 | 29 | 14 | 139 | 25.6% | 12.5% |
| Interleukin 8 | 4 | 3 | 16 | 35 | 57.1% | 34.5% |
| Tumour necrosis factor-α | 2 | 4 | 22 | 269 | 33.4% | 8.1% |
| Calcium-binding parvalbumin | 3 | 1 | 2 | 12 | 75.0% | 27.8% |
| Flavodoxin | 2 | 1 | 7 | 31 | 66.7% | 22.7% |
| Ribonuclease A | 4 | 9 | 25 | 99 | 30.8% | 21.2% |

For notation see Table II. Hydrophobic positions are defined in the text.

As a result of these restrictions, the number of pairs analysed is much smaller, as is the number of proteins, but the average level of correct prediction of residue–residue contacts increases to 23%. The prediction success depends sensitively on the distance cut-off used in the definition of residue–residue contacts. Note that a more relaxed cut-off on distance (here 6 Å between the closest two atoms) would lead to a higher level of accuracy and may still be useful in distance geometry calculations. Proteins with a particularly high fraction of correctly predicted hydrophobic contacts are given in Table III.

An alternative way to assess the connection between mutational correlation and spatial distance is inspection of the distribution of interresidue distances. Figure 3 quantifies the relationship between mutational correlation and residue contacts in terms of separate histograms for correlated and non-correlated residue pairs, as a function of residue–residue spatial distance. A similar figure is in Altschuh *et al.* (1988). Correlated pairs tend to occur

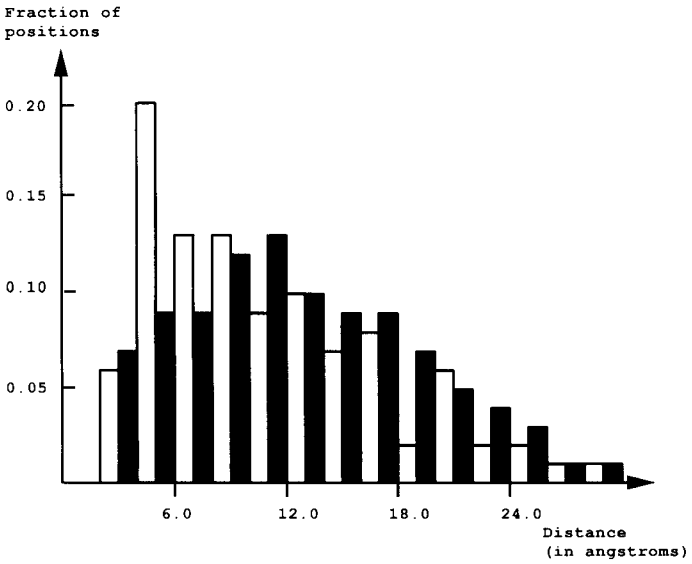


**Fig. 3.** Evidence for weak information about spatial closeness in correlated residue pairs. Histogram of the fraction of residue pairs as a function of spatial distance. The subset of mutationally correlated pairs (white) tends to occur at smaller spatial distances than do all pairs (black). Mutational correlation for particular residue pairs is assessed using equation (12), with a cut-off of $p_0$ of 0.20. Residue distance is the distance of closest interatomic approach between side chains. The numbers are accumulated for the eight proteins in Table III.

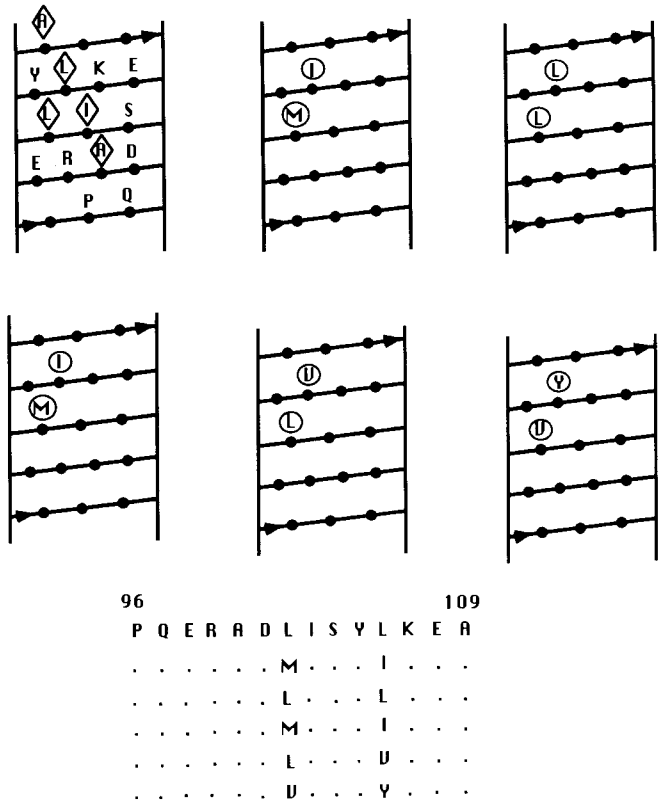


**Fig. 4.** Example of correlated mutations at positions 102 and 106 of the C-terminal $\alpha$-helix of cytochrome *c* (residues 96–109 of data set 1CCR). Helical net diagram with the sequence from bottom left to top right. Note the hydrophobic patch on the surface of the $\alpha$-helix (diamonds). In different homologous sequences, positions 102 and 106 (circled) are occupied by the residue pairs LL, MI, LL, MI, LV and VY. The correlation between positions 102 and 106 is understandable in terms of the spatial proximity of the two residues.

at smaller interresidue distances, but are also present at larger distances, limiting the predictive value.

*Prediction of residue contacts in λ repressor from mutation–selection experiments*

In a series of elegant experiments, Lim and Sauer (1989) modified genes for the small λ repressor protein by randomizing some sequence positions using cassette mutagenesis and then selecting functional proteins from among a very large number of possible sequences. The interpretation of the effect of mutations on repressor structure is straightforward for single mutations, but more complicated for multiple substitutions. Here, we use the data of Lim and Sauer (1989) to analyse the occurrence of correlated pairs of mutations.

The selection experiments of Lim and Sauer (1989) are analogous to a process of natural evolution in which the relationship between ancestor and descendant is a particularly simple one. Accordingly, their functional sequences can be arranged in a star-like tree, different from hierarchical phylogenetic trees, with branches radiating outward from the wild-type protein. From this tree, we derived a list of mutations in each branch. The accuracy of this data is higher than that normally obtained from multiple sequence alignments, for which the mutational history cannot be reconstructed in full detail. The statistical criterion (12) was then applied to the data for the seven randomized positions in the hydrophobic core of λ repressor in order to detect pairs of positions with simultaneously accepted residue substitutions. Thirty mutant amino acid sequences with from one to four mutations relative to the wild-type were considered. The distance matrix between $C_\alpha$ atoms of these seven positions is given in Table IV.

Among the $7 \times 6/2 = 21$ possible pairs of positions, four pairs have significantly correlated mutations. These pairs are underlined in Table IV and the corresponding $P(M_{l_1 l_2})$ values according to equation (12) given in the caption. Strikingly, all four correlated pairs of residues have a mutual distance between $C_\alpha$ atoms of $< 5.0$ Å. According to the exact Fisher criterion the probability of obtaining by chance four correlated and spatially close positions is equal to $P = 0.04$.

*A new experimental approach to protein structure determination?*

It is tempting to suggest a new experimental approach to protein structure determination. First, the sequence of a protein would be partially randomized. Second, molecules which are catalytically active or capable of specific binding would be selected, e.g. by a phage display system or by selective growth advantage in cell culture, and their sequences determined. Assuming that functional molecules have an intact fold, an analysis of correlated sequence changes may yield sufficient constraints for the calculation of the three-dimensional protein structure. Such constraints could be combined with information from other, e.g. spectroscopic, experiments and from secondary structure prediction methods. The evolution experiments of Lim and Sauer (1989) on λ repressor may mark the beginning of a new series of selection experiments aiming at protein structure determination.

## Discussion

We have attempted to show quantitatively that correlated mutation of residue pairs is a characteristic feature of protein structure that can be extracted from multiple sequence data. In addition,

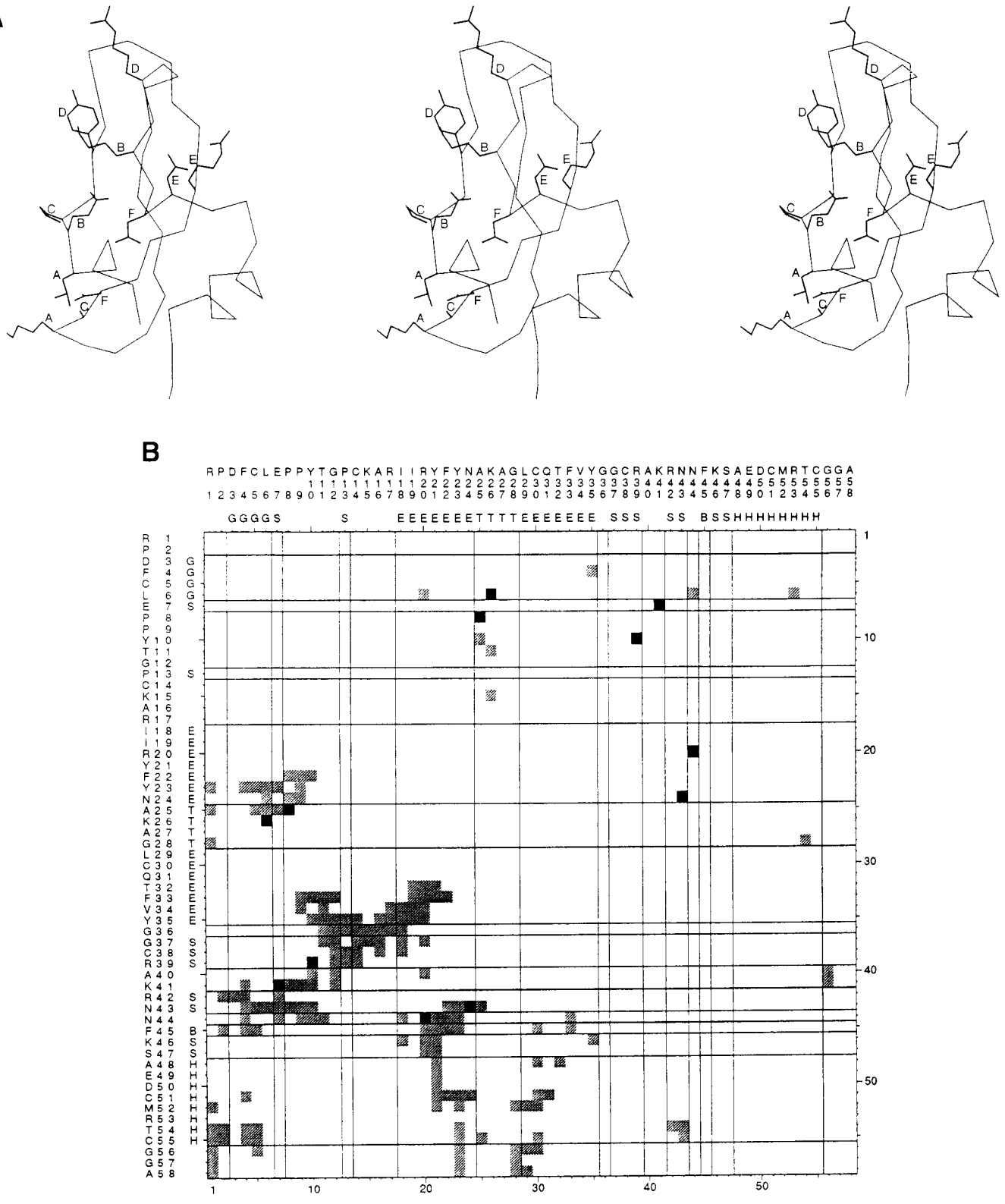


**Fig. 5.** Predicted and observed chain-distant contacts in the structure of trypsin inhibitor (data set 5PTI). (**A**) Stereo view. Residue pairs involved in correctly predicted contacts. Each residue in a pair is labelled with the same letter. For example, Arg20 and Asn44 (middle, right) are labelled with the letter E; the contact between these two residues is predicted on the basis of correlated mutations and is also observed in the crystal structure. The six residue pairs shown correspond to the six black squares in the top right (and bottom left) of the contact map (B). (**B**) Predicted (top right) and observed (bottom left) contact map. Only chain-distant contacts ($|i - j| > 10$, where $i$ and $j$ are residue numbers) are included in the analysis and in this map. Top right: all predicted contacts (significantly correlated pairs) are indicated, correctly predicted ones by black squares and incorrectly predicted ones by gray squares. Bottom left: all contacts observed in the crystal structure are shown, correctly predicted ones by black squares and those not predicted by grey squares. Note that six out of the 16 predicted contacts are correctly predicted, two in each of the important chain-distant contact regions. However, the number of correctly predicted contact pairs is relatively small compared to the set of all observed contacts.

**Table IV.** Distance matrix for the hydrophobic core of λ repressor

| | | | | | | |
|---|---|---|---|---|---|---|
| 36V | 7.3 | | | | | |
| 40M | 7.8 | 5.1 | | | | |
| 47V | 7.1 | 4.0 | 4.6 | | | |
| 51F | 2.3 | 4.8 | 7.1 | 4.1 | | |
| 57L | 5.5 | 10.1 | 7.0 | 9.6 | 7.8 | |
| 65L | 3.9 | 4.8 | 4.1 | 6.2 | 4.1 | 3.2 |
| | 18L | 36V | 40M | 47V | 51F | 57L |

The four pairs with significantly correlated mutations are underlined. Their spatial proximity can be reliably predicted from sequence information alone. Probabilities are $P(M_{18,15}) = 0.037$, $P(M_{36,47}) = 0.016$, $P(M_{36,51}) = 0.0048$, $P(M_{47,51}) = 0.0070$, calculated according to equation 10. Distances between residues (Å) are distances of closest approach of side chains atoms ($C_\alpha$ for Gly) calculated from model coordinates derived from the $C_\alpha$ trace as deposited in the PDB by the algorithm of Holm and Sander (1991).

we have detected a tendency for amino acid substitutions to be accepted in a correlated fashion in evolution for pairs of residues that are in physical contact, i.e. form structurally specific interactions.

One type of residue–residue interaction, S–S bonds, provides particularly strong constraints on protein tertiary structure. According to our results, pairs of positions with Cys residues frequently evolve in the correlated manner. Another kind of specific interaction are contacts between oppositely charged side chains that can form H-bonds or interact electrostatically. We found in several proteins correlated pairs of oppositely charged or polar side chains.

Interactions between side chains in the hydrophobic core is a key factor in determining a protein's structure (Richmond and Richards, 1978; Cohen *et al.*, 1979; Lesk and Chothia, 1980). Residues in the cores of proteins are very closely packed, occupying almost all available interior space, at a density resembling that observed in crystals of small inorganic molecules (Chothia, 1976). The hydrophobic core volume of proteins is essentially constant in the course of evolution. For example, in the globin family the total volume of the protein core remains roughly constant with an r.m.s. deviation of 15 $Å^3$ (Lim and Ptitsyn, 1970).

Similar estimates have been made for the λ repressor protein (Lim and Sauer, 1989). The core volume of functionally intact mutants is centred around the wild-type value with an r.m.s. deviation of 20 $Å^3$. The constancy of hydrophobic core volume in the course of evolution is the result of compensating mutations (substitution of small–large by medium–medium or large–small pairs) (Lim and Ptitsyn, 1970).

Unfortunately, the data for correlated mutations does not only contain residue pairs that are in direct physical contact. Other factors apparently have a role, complicating the analysis. Among possible explanations for correlated evolution of residue positions distant in the chain we consider the following.

(i) Residues which are distant in the three-dimensional structure of native protein could be involved in specific interactions that are important for stabilization of some intermediate structures along the protein folding pathway (e.g. Creighton, 1980).

(ii) Another important factor may be protein stability. In many proteins amino acid substitutions have been detected that reduce protein stability (Alber, 1989). Most of these are substitutions in the hydrophobic core and their effects can easily be explained, but there are exceptions. In some cases, protein stability can be restored not only by neighbouring but also by distant suppressor substitutions. For example in the λ Cro protein, the detrimental effect of the mutation Ile30 → Leu is suppressed by several mutations in distant surface positions (Gln27 → Pro, Tyr26 → Cys, etc.) (Pakula and Sauer, 1989). The reasons for this 'effective interaction' is not clear. Possibly some compensating mutations of spatially distant residue pairs maintain a constant level of protein stability (not too stable, not too unstable).

(iii) Conformational change during protein function may have a role. It is well known that reversible conformational rearrangements for some proteins are essential for biological function (Perutz, 1983). Residues which are distant in the crystal or solution structure may be brought together at a definite stage of protein function. Evolution of such residue pairs would be correlated, but the pair would not be close in the native structure.

These additional effects complicate the prediction of residue contacts in the native structure. Perhaps, however, they can be exploited positively and provide information on spatial rearrangements during protein folding or function.

The approach used here could be useful in the prediction of protein three-dimensional structure as it produces information about possible contacts between specific residue positions. Once a list of potential residue–residue contacts has been derived, distance geometry approaches can be used to calculate explicit three-dimensional coordinates (Goel and Ycas, 1979; Crippen, 1981; Crippen and Havel, 1987). Contact prediction can also be useful in the context of combinatorial approaches (Cohen *et al.*, 1982), by providing additional constraints that can be used in the combinatorial filtering procedure.

There are several possible ways of improving the efficiency of the present approach.

(i) Using more effective algorithms for alignment of amino acid sequences (C.Sander and R.Schneider, in preparation).

(ii) Using more effective methods for the construction of phylogenetic trees.

(iii) Using nucleotide sequences of genes coding for proteins to increase the accuracy of reconstruction of mutation events in the phylogenetic tree.

(iv) Explicitly using physico-chemical characteristics of amino acids (charge, hydrophobicity, volume of side chains, etc.) in the statistical analysis.

(v) Refinement of the formal criteria for recognition of predicted contacts in the total set of correlated pairs.

Information about correlated pairs of residues may also be directly useful in protein engineering. Experiments could be designed in which certain positions are systematically mutated and protein function assayed, with the specific goal of obtaining information about spatial proximity or functional interaction of residue pairs, followed by prediction of the three-dimensional structure based on the experimental results.

We have begun to quantitatively exploit the wealth of information contained in multiple sequence alignments for the analysis of correlated sequence positions, using tree construction and careful statistical estimates. The results are not yet sufficient to predict protein tertiary structure, but the observed effects are encouraging and point the way toward protein structure determination by a combination of mutational experiments, analysis of correlated mutations and distance geometry calculation of explicit coordinates.

## Acknowledgements

We thank M.Scharf for help with CONAN contact analysis, R.Schneider for help with the HSSP database of protein families, G.Vriend and C.Ouzounis for help with the protein engineering package WHAT IF, all members of the Protein Design Group for helpful discussions and the Computer Group at EMBL for system support. C.S. is grateful for support from the Human Frontiers Science Program. I.N.S. and N.H.K. are grateful for financial support from the EMBL visitors' program and the Russian State Human Genome Project.

## References


Alber,T. (1989) *Annu. Rev. Biochem.*, **58**, 765–798.

Altschuh,D., Lesk,A.M., Bloomer,A.C. and Klug,A. (1987) *J. Mol. Biol.*, **193**, 693–707.

Altschuh,D., Verner,T., Berti,P., Moras,D. and Nagai,K. (1988) *Protein Engng*, **2**, 193–199.

Anfinsen,C.B. (1973) *Science*, **1981**, 223.

Benner,S.A. and Gerloff,D. (1990) *Adv. Enzyme Regul.*, **31**, 121–181.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Chothia,C. (1976) *J. Mol. Biol.*, **105**, 1–14.

Cohen,F.E., Richmond,T.J. and Richards,F.M. (1979) *J. Mol. Biol.*, **132**, 275–288.

Cohen,F.E., Sternberg,M.J.E. and Taylor,W.R. (1982) *J. Mol. Biol.*, **156**, 821–862.

Creighton,T.E. (1980) In Jaenicke,R. (ed.), *Protein Folding*. Elsevier/North-Holland Biomedical Press, Amsterdam, pp. 427–446.

Creighton,T.E. (1983) *Proteins: Structure and Molecular Properties*. W.H.Freeman, New York.

Crippen,G.M. (1981) *Distance Geometry and Conformational Calculations*. Wiley, New York.

Crippen,G.M. and Havel,T.F. (1988) *Distance Geometry and Molecular Conformation*. Research Studies Press, Taunton, UK.

Epstein,C.J., Goldberger,R.F. and Anfinsen,C.B. (1963) *Cold Spring Harbor Symp. Quant. Biol.*, **28**, 439.

Goel,N.S. and Ycas,M. (1979) *J. Theor. Biol.*, **77**, 253–305.

Hartigan,I.A. (1973) *Biometrics*, **29**, 53–65.

Godzik,A. and Sander,C. (1989) *Protein Engng*, **2**, 589–596.

Hobohm,U., Sander,C., Scharf,M. and Schneider,R. (1992) *Protein Sci.*, **1**, 409–417.

Holm,L. and Sander,C. (1991) *J. Mol. Biol.*, **218**, 183–194.

Konigs,D.A.M. and Hogeweg,P. (1989) *J. Mol. Biol.*, **207**, 597–614.

Lesk,A. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.

Lesser,G.J., Lee,R.H., Zehfus,M.H. and Rose,G.D. (1987) In Oxender,D.L. and Fox,C.F. (eds), *Protein Engineering*. Alan R.Liss Inc., New York, pp. 175–179.

Lim,V.I. and Ptitsyn,O.B. (1970) *Mol. Biol.*, **4**, 373–382.

Lim,W.A. and Sauer,R.T. (1989) *Nature*, **339**, 31–36.

Pakula,A.A. and Sauer,R.T. (1989) *Proteins: Struct. Funct. Genet.*, **5**, 202–210.

Perutz,M.F. (1983) *Mol. Biol. Evol.*, **1**, 1–28.

Richmond,T.J. and Richards,F.M. (1978) *J. Mol. Biol.*, **119**, 537–555.

Rost,B. and Sander,C. (1993a) *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.

Rost,B. and Sander,C. (1993b) *J. Mol. Biol.*, **232**, 584–599.

Sander,C. and Schneider,R. (1991) *Proteins: Struct. Funct. Genet.*, **9**, 56–68.

Sokal,R.R. and Sneath,P.H.A. (1963) *Principles of Numerical Taxonomy*. W.H.Freeman, San Francisco.

Valencia,A., Kjeldgaard,M., Pai,E.F. and Sander,C. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 5443–5447.


*Received October 20, 1992; revised October 26, 1993; accepted November 5, 1993*