Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 98–102, January 1994
Biochemistry

How frequent are correlated changes in families of protein sequences?

(compensatory mutations/protein structure/alignment/fluctuation/correlation coefficient)

ERWIN NEHER

Max-Planck-Institut für Biophysikalische Chemie, Abteilung Membranbiophysik, Am Fassberg, 37077 Göttingen, Federal Republic of Germany

Contributed by Erwin Neher, September 16, 1993

ABSTRACT **A loss-of-function point mutation in a protein is often rescued by an additional mutation that compensates for the original physical change. According to one hypothesis, such compensation would be most effective in maintaining a structural motif if the two mutated residues were spatial neighbors. If this hypothesis were correct, one would expect that many such compensatory mutations have occurred during evolution and that present-day protein families show some degree of correlation in the occurrence of amino acid residues at positions whose side chains are in contact. Here, a statistical theory is presented which allows evaluation of correlations in a family of aligned protein sequences by assigning a scalar metric (such as charge or side-chain volume) to each type of amino acid and calculating correlation coefficients of these quantities at different positions. For the family of myoglobins it is found that there is a high correlation between fluctuations in neighboring charges. The correlation is close to what would be expected for total conservation of local charge. For the metric side-chain volume, on the other hand, no correlation could be found.**

The dramatic increase in available data on protein sequences has made it clear that certain structural motifs of proteins are quite commonly represented in multiple forms (1). This is a consequence of the general feature that, in evolution, sequence changes more rapidly than structure (2). If this were so, then one would expect that, when comparing the various sequences of an alignment, some of the variations in sequences are compensatory. For example, if, in a given sequence, an amino acid side chain is particularly bulky with respect to the average at a given position, this might have been compensated in evolution by a particularly small side chain in a neighboring position, for preserving the general structural motif. Similar constraints might hold for other physical quantities such as amino acid charge or hydrogen bonding capacity.

Individual examples for such compensatory changes have been documented (3). If they were sufficiently frequent, one might be able to identify the mutually compensating partners by a statistical analysis, since there should be correlations in the frequency of occurrence of amino acids at the corresponding positions. Indeed, inferences from sequence variability and correlation analysis have been quite successful recently in elucidating structural features (4–8). This report deals with an attempt to quantify the frequency of compensatory changes in a given protein family.

Theory and Algorithms

The family of aligned sequences is considered to be a sample of a hypothetical population of all possible protein sequences that are able to form a certain three-dimensional motif. It is viewed as a set of random variables A_i, each of which represents one position i in the alignment and can assume as possible outcomes the 20 amino acids. The population is defined if all the probabilities f_i^ν (the probability of occurrence of amino acid type ν at position i) are known. For the general case f_i^ν should be considered as functions not only of the occurrence of an amino acid at position i but also of that at any other position in the family. In fact, we want to test whether the given distribution in the sample is significantly different from the null hypothesis, which is the assumption that amino acids at different positions in the alignment vary independently. In that case f_i^ν depends only on the given position i and type of amino acid ν and will be denoted p_i^ν. In the more general case the joint probability P of finding amino acid ν at position i and amino acid μ at position j is

$$P(A_i = \nu;\, A_j = \mu) = f_{i,j}^{\nu,\mu}, \qquad [1]$$

which again is a function of all the occupancies at other positions. For the null hypothesis it is a product of the distributions at positions i and j:

$$f_{i,j}^{\nu,\mu} = p_i^\nu p_j^\mu. \qquad [2]$$

Each amino acid ν has associated with it a physical quantity q_ν (such as a side-chain charge) such that we can define the random variables Q_i with

$$P(Q_i = q^\nu) = f_i^\nu \qquad [3]$$

$$P(Q_i = q^\nu;\, Q_j = q^\mu) = f_{i,j}^{\nu,\mu}. \qquad [4]$$

Then, the expected value $E\{Q_i\}$ is

$$E\{Q_i\} = \sum_\nu q^\nu f_i^\nu = \overline{Q}_i, \qquad [5]$$

and the expected variance σ_i^2 is

$$\sigma_i^2 = \sum_\nu (q^\nu - \overline{Q}_i)^2 f_i^\nu. \qquad [6]$$

Correlations among positions are conveniently expressed in terms of correlation coefficients,

$$\rho_{i,j} = \sum_{\nu,\mu} \frac{(q^\nu - \overline{Q}_i)(q^\mu - \overline{Q}_j)}{\sigma_i \sigma_j} f_{i,j}^{\nu,\mu}. \qquad [7]$$

From a given alignment, we would like to calculate at positions i and j the sample means m_i and m_j, as well as sample variances s_i and s_j and the sample correlation coefficients $r_{i,j}$. If the sample had independent sequences, the sample correlation coefficients would be given by:

$$r_{i,j} = \frac{1}{N} \sum_l \frac{(q_i^l - m_i)(q_j^l - m_j)}{s_i s_j}, \qquad [8]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

where the sum extends over all sequences indexed, l, and q_i^l and q_j^l represent the physical quantities associated with amino acids of sequence l at positions i and j, respectively. Likewise, the standard equations hold for m_i, m_j, s_i, and s_j (9). For the null hypothesis (independent positions) and for independent sequences the expectation value for $r_{i,j}$, as calculated according to Eq. **8**, is zero and the expected variance S_r^2 of calculated values $r_{i,j}$ values is given approximately by (9, 10)

$$S_r^2 \approx \frac{1}{N-4}, \qquad [9]$$

when N is larger than 4. The quantity $N-4$ can be considered the number of degrees of freedom for estimating S_r^2. Unfortunately, sequences within a family of homologous proteins (by definition of homology) are not statistically independent, so it is not easy to estimate S_r^2, when $r_{i,j}$ values are calculated according to Eq. **8**.

Here, an attempt is made to solve this problem partially by considering sequences pairwise. It is easily shown that for independent sequences an unbiased estimate for $r_{i,j}$ can be obtained by

$$r_{i,j} = \frac{1}{N}\frac{1}{2}\sum_l \frac{(q_i^{l1} - q_i^{l2})(q_j^{l1} - q_j^{l2})}{s_i s_j}, \qquad [10]$$

where the sum extends over N pairs (index l). The subscripts, as in Eq. **8**, refer to positions i and j along the sequence, and the superscripts refer to the two sequences forming a given pair. The sample variances s_i and s_j are, likewise, calculated by

$$s_i^2 = \frac{1}{2N}\sum_l (q_i^{l1} - q_i^{l2})^2. \qquad [11]$$

When sequences have a high degree of homology, then many of the terms in the sums of Eqs. **10** and **11** are zero, since the associated amino acids are identical. We interpret identities in amino acids in the sense that no mutation has occurred in evolution at the given position between sequences $l1$ and $l2$ and disregard the corresponding terms in the analysis. If, on the other hand, a mutation has occurred, we consider the two outcomes as independent. In practice this is implemented by dividing the sums in Eqs. **10** and **11** not by N, the number of pairs considered, but by $N_{i,j}$ and N_i, respectively, the number of terms at positions i (and j) without identities in the amino acids:

$$r_{i,j} = \frac{1}{N_{i,j}}\frac{1}{2}\sum_l \frac{(q_i^{l1} - q_i^{l2})(q_j^{l1} - q_j^{l2})}{s_i s_j} \qquad [10a]$$

$$s_i^2 = \frac{1}{2N_i}\sum_l (q_i^{l1} - q_i^{l2})^2. \qquad [11a]$$

This introduces an error, since identity in amino acids also can occur by multiple mutations. But this error is relatively small. In essence, the procedure considers not the original set of random variables A_i with probability distribution $f_{i,j}^{\nu,\mu}$, but a modified one (A_i^*) with joint probability distributions $f_{i,j}^{*\nu,\mu}$, where all values which satisfy $\nu = \mu$ are set to zero, and the remaining ones are renormalized. For the case of the null hypothesis the errors made by this replacement can be readily estimated when the amino acid profiles at positions involved are known. It is small, since errors partially cancel in determining $r_{i,j}$ values because they affect numerator and denominator of Eq. **10a** similarly.

The procedure described so far (using Eqs. **10a** and **11a**) allows for statistical dependence among the partners of a given pair and provides relatively unbiased estimates for the sample correlation coefficients. It should be pointed out that strongly biased results are obtained when Eqs. **10** and **11** are used on sequences with a high degree of homology. If the pairwise amino acid identity averaged over all positions and pairs is $\bar{p}$, then

$$N_i \approx (1-\bar{p})N \qquad [12]$$

$$N_{i,j} \approx (1-\bar{p})^2 N. \qquad [13]$$

Thus, by comparison of Eqs. **10** and **11** with Eqs. **10a** and **11a** it can be seen that the biased estimates of s_i^2 and $r_{i,j}$ values are only approximately $(1-\bar{p})$ times the unbiased ones.

Use of Eqs. **10a** and **11a** avoids these underestimates. Statistical dependence is reintroduced, however, when one averages the contributions from many sequence pairs, because the sequences of one pair are partially homologous to those of other pairs. This does not falsify estimates of s_i and $r_{i,j}$ values, because averaging of unbiased estimates does not introduce bias. However, it implies that there is not a simple relationship between the number of averages $N_{i,j}$, taken at positions i and j, and the number of degrees of freedom ($N-4$) required in Eq. **9** to estimate the expected variance S_r^2 of the $r_{i,j}$ values. We need to know S_r^2, however, in order to decide whether a given $r_{i,j}$ value is significantly different from zero.

It is found in the numerical analysis that S_r^2 depends strongly on the degree of conservation at the positions considered and that it decreases when more terms $N_{i,j}$ are available, as suggested by Eq. **9**. We therefore assume, in analogy to Eq. **9**, that position specific quantities $S_{r,i,j}^2$ can be defined

$$S_{r,i,j}^2 = \frac{1}{\alpha N_{i,j} - 4}, \qquad [14]$$

where α is a coefficient such that $(\alpha N_{i,j} - 4)$ represents the apparent number of degrees of freedom. This coefficient can be obtained from the scatter in the distribution of all $r_{i,j}$ values, since the overwhelming majority of these should represent independent positions (only very few out of all possible partners should be neighbors). For this purpose, the variance of all $r_{i,j}$ values S_r^2 and the mean $\bar{N}$ of $N_{i,j}$ values is determined. An approximation for α can be obtained by inverting Eq. **14**:

$$\alpha \approx (4 + 1/S_r^2)/\bar{N}. \qquad [15]$$

Strictly, α might also depend on the position considered. However, when all available $r_{i,j}$ values are subdivided into classes with different $N_{i,j}$ and the analysis is performed classwise, no major trend is observed.

The statistical analysis of a family of proteins is started by creating a table of pairs (the "pair list"). It was found (see below) that pairs for which the partners have overall amino acid identities between 60% and 95% are most appropriate for the analysis. Therefore, as a first step, a matrix of all pairwise overall identities is set up. Second, sequences are eliminated that have partners appreciably more similar than the 95% criterion. This leaves a number N_s of sequences for the analysis. Then, the subset of those pairs fulfilling the 60–95% criterion is created, and this is narrowed by random elimination to contain not more than 3 times N_s pairs in order to speed the calculation. Results are not improved substantially by including more pairs.

Once a pair list has been created, the analysis employs a "scale file" that assigns physical quantities to amino acids.

Variances according to Eq. **11a** are calculated for those positions where N_i in Eq. **11a** is larger than zero. Subsequently, all possible $r_{i,j}$ values are calculated according to Eq. **10a**. The $r_{i,j}$ values are accepted as valid if the variances associated with positions i and j are valid and if $N_{i,j}$ is larger than zero. Finally, the mean and variance of all $r_{i,j}$ values are calculated. In this calculation individual values are weighted according to the inverse of their expected variance (Eq. **14**)—i.e., by weight = $\alpha N_{i,j} - 4$. The analysis considers only those values with weights larger than zero. Thus, the proper weighting requires a knowledge of α, which, on the other hand, is determined, according to Eq. **15**, from the result of the statistical evaluation. This problem is solved iteratively. When 1 is used as a starting value for α, three or four iterations converge on a stable value which, in the case of the calculations described below, was in the range 0.2–0.4.

Correlation coefficients are also displayed graphically in the form of dot matrix plots. In such "maps" the x axis and y axis represent amino acid positions i and j, respectively. The values of correlation coefficients are encoded by the sizes of the symbols. Under the assumption that the physical quantity under study (charge or side-chain volume) of neighboring positions is negatively correlated, such plots would display neighborhood relationships if only negative values significantly different from zero were plotted. To do so, signal-to-noise ratios were calculated for each point as the ratio of $r_{i,j}$ over $S_{r,i,j}$ (Eq. **14**). They were plotted as dots proportional in size to their values, in case they were negative, and if they exceeded a certain threshold. During the calculation of signal-to-noise ratios, values were discarded if, according to Eq. **14**, they would result in a negative $S^2_{r,i,j}$.

Results and Discussion

As a test case, the family of myoglobins was analyzed. An alignment containing 68 myoglobin sequences was obtained from A. Lesk (Medical Research Council Laboratory, Cambridge, U.K.). When the pair list (see above) was formed, 26 sequences were eliminated because they had >97% amino acid identities to other sequences of the family, so that the analysis was performed on a total of 42 myoglobins.

Initially, the side-chain volume was analyzed by using values as given by Klapper (11). It was possible to evaluate 2357 correlation coefficients. The overall mean $\overline{r_{i,j}}$ of these was close to zero:

$$\overline{r_{i,j}} = 0.028 \pm 0.53 \text{ (mean} \pm \text{SD)}. \quad [16]$$

This was expected, since for the vast majority of position pairs, the amino acids in question are not neighbors; therefore, they should be uncorrelated. From the standard deviation one can calculate (Eq. **9**) an apparent number of degrees of freedom of 3.6.

When the average correlation coefficient $\overline{r_{i,j}}^*$ was calculated for only those position pairs known to be neighbors from x-ray crystallography, the result was

$$\overline{r_{i,j}}^* = -0.001 \pm 0.53 \ (n = 93). \quad [17]$$

Here, neighbors were defined as the set of those position pairs whose side chains had at least one atomic contact.

This result is disappointing, since one would expect that fluctuations in side-chain volume should be compensated in the neighborhood for preservation of overall structure. However, Lesk and Chothia (12) have shown that such compensation may occur not at the nearest neighbor, but also by slight displacement of secondary structure elements or by compensation at places further away.

Side-chain charge, on the other hand, should require a more localized compensation. This was, indeed, found to be the case. When $r_{i,j}$ values were analyzed for the quantity side-chain charge (D and E, −1; K and R, +1; H, +0.5), again the mean of all $r_{i,j}$ values was indistinguishable from zero. The mean of all known neighbors (as defined above), however, was

$$\overline{r_{i,j}}^* = -0.405 \pm 0.49 \ (n = 40). \quad [18]$$

The standard error of the mean is 0.08, so there is highly significant negative correlation among neighboring charged residues. Significance of the result is corroborated by the finding that the value drops to close to zero if the list of neighbors is scrambled before the mean is calculated. The negative correlation stems partially from neighbors (in three-dimensional space) which are far apart along the chain, but partially also from chain neighbors. When the average $r_{i,j}$ was calculated from only those neighbors that were more than 4 residues apart from each other, a value of -0.25 ± 0.45 (n = 8) was obtained. The complement (neighbors closer than 5 residues) yielded an average of -0.57 ± 0.45 (n = 31).

The results obtained for side-chain charges show a remarkably high correlation, given the fact that a given fluctuation need not be compensated by a specific neighbor. In the list of

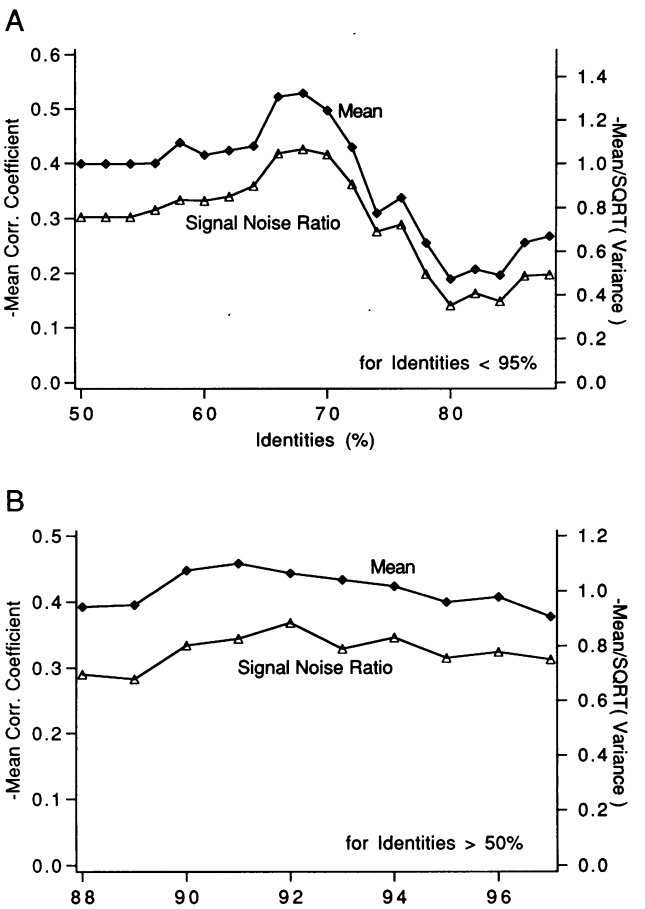

FIG. 1. Mean correlation coefficient between neighboring charged residues and signal-to-noise ratio as a function of analysis parameters. Negative mean correlation coefficient, $\overline{r_{i,j}}^*$ (see Eq. **18**), between charges of neighboring residues was calculated for different sets of sequence pairs. Signal-to-noise ratio is $\overline{r_{i,j}}^*/S_r$, as calculated numerically from all $r_{i,j}$ values. In *A*, sequence pairs were selected with overall amino acid identities larger 5% and smaller than the value given by the abscissa. In *B*, sequence pairs were selected with overall amino acid identities <50% and larger than the value given by the abscissa. The number of such position pairs, over which averages were taken, ranged from 23 (90% identity) (*A*) to 90. See text for further explanation. SQRT, square root.

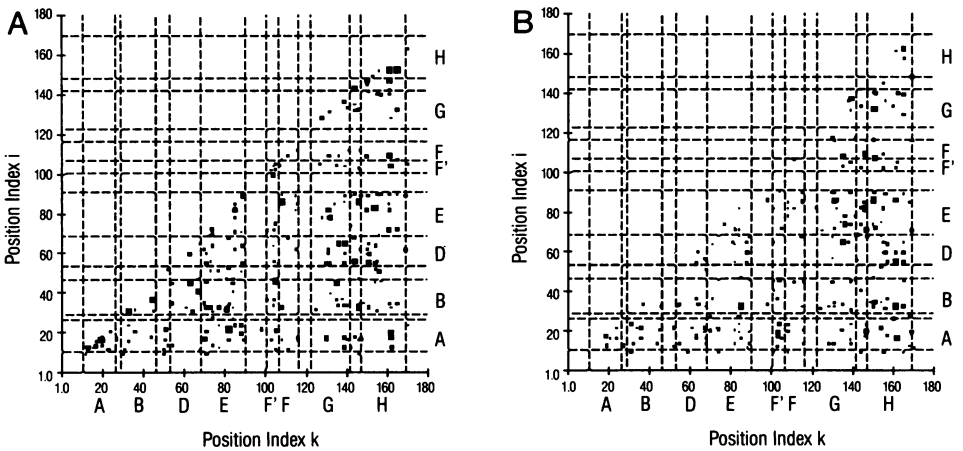

FIG. 2. Dot matrix plot of correlation coefficients among side-chain charges. A dot was plotted whenever the absolute value of the signal-to-noise ratio (see text for explanation) of a correlation coefficient between positions index i and j exceeded the value of 1.7. The size of the symbol is proportional to the absolute magnitude of the signal-to-noise ratio. *A* displays negative values, which are expected to indicate compensatory correlations between neighbors. *B* shows positive values, for which no such correlation is expected. For this and the following maps the amino acid scale considered both full charges and also partial charges for side chains with a pK close to 7. Relative values: 1 for H, K, and R; −1 for D and E; −0.5 for C and Y; 0 for all other ones. The alignment included not only myoglobins but also hemoglobin α chains and hemoglobin β chains (a total of 212 sequences). The grid and the lettering indicate secondary structure elements according to Lesk and Chothia (12).

crystallographic contacts a given amino acid had 7.6 neighbors. However, many of these did not carry a charge in any of the sequences available. Among those positions displaying fluctuations in charge, there were on average 3.6 neighbors which also did so. Thus, if a change in charge at a given position were required to be locally compensated by one of the fluctuating neighbors, one would expect an average probability of 0.28, not very different from the correlation coefficient found.

The analysis presented so far was performed on the basis of a pair list that comprised sequence pairs with overall amino acid identities between 60% and 95%. Fig. 1 shows that this appears to be the optimal range of sequence similarities. Here the boundaries of the range in amino acid identities used for the analysis were varied. Both the mean correlation coefficient of neighbors $\overline{r_{i,j}}^*$ (see Eq. **18**) and the signal-to-noise ratio $(\overline{r_{i,j}}^*/S_r)$ were plotted. In Fig. 1*A* the upper boundary of pairwise identities was fixed to 95% and the lower boundary was varied between 50% and 90%. It is seen that the signal-to-noise ratio rises slightly between 50% and 70% values but drops precipitously when the analysis is restricted to pairs more similar than 70%. In Fig. 1*B* the lower boundary was fixed to 50%, and the upper boundary was varied between 88% and 98%. The signal-to-noise ratio is relatively constant between 92% and 98%, which shows that including pairs with partners more similar than 95% does not improve the result.

In spite of the high correlation coefficient (Eq. **18**), it is not possible to identify with reasonable confidence the neighbors of a given position. Such identification would require a signal-to-noise ratio $(r_{i,j}/S_{r,i,j})$ of ≈ 2. To achieve this a set of sequences ≈ 6 times larger (about 250 sequences) than the present one would be required.

Nevertheless, structural information may be obtained from the correlation coefficients for cases when it is possible to average over elements of secondary structure, such as parallel or antiparallel helices, where periodicities in the correlation coefficients are expected (see below). Also, a structural model that provides a list of neighbors could be tested by calculating the average correlation coefficient among all predicted neighbors. If conditions were otherwise similar to those of the case studied for Eq. **18**, a mean very much different from zero would be expected for a correct model.

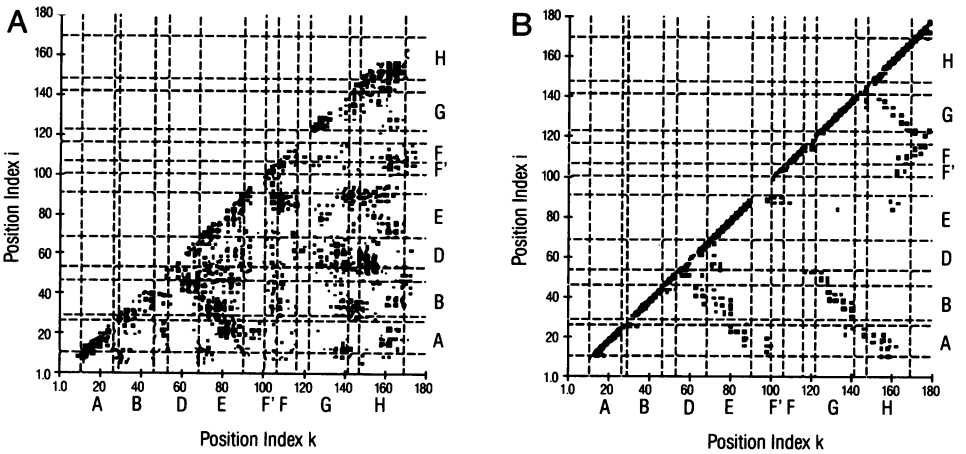

FIG. 3. (*A*) Maps of Fig. 2 after application of a filtering procedure designed to emphasize periodicities which are expected to occur if two helices are in contact in either parallel or antiparallel fashion. Each correlation coefficient was replaced by the weighted average of itself (weight 1) and those map neighbors (weight, 0.5) which were 3 and 4 positions apart on both the x and the y axis. In addition neighbors with the following index offsets were included with weight 0.25: (3, 3); (3, 4); (3, −4); (−3, −3); (−3, 4); (−3, −4); (4, 3); (4, 4); (4, −3); (−4, 3); (−4, −4). (*B*) "Contact map," in which a dot is plotted for each position pair that are known from crystallography to be in contact.

Another way of averaging is by visual inspection of the result. This can be done by viewing dot matrix plots like those of Fig. 2. For the following a larger set of sequences and a more complete metric was used (see Fig. 2 legend for details). As detailed in the *Theory and Algorithms* section, the plots show only those correlation coefficients considered to be significantly negative, at positions according to their indices. The sizes of the symbols encode an estimate for signal-to-noise ratio. In such a plot neighbors along the polypeptide chain are located close to the diagonal. Fig. 2*A* shows that there is a somewhat higher density of symbols along the diagonal. This pattern represents the neighborhood relationships along the polypeptide chain. No such density is seen in Fig. 2*B*, which shows, as an exception, positive correlations. The globins contain seven helical segments, some of which form antiparallel contacts. Depending on the angle between the helices, side-chain contacts with a periodicity of 3 or 4 are expected (12). Such contacts should show up as densities of points along lines perpendicular to the diagonal. Some indication of this can be seen in Fig. 2*A*. The features gain some clarity, however, when appropriate filtering procedures, exploiting the expected periodicities, are applied. In Fig. 3*A* each point in the map is replaced by a weighted average among its neighbors. The weights, as indicated in the legend to Fig. 3, emphasize periodicities of 3 and 4. With this procedure, the diagonal (representing contacts along the chain and across helical turns) becomes much clearer. Also, some streaks perpendicular to the diagonal appear. Not all of these are significant (since the filtering procedure induces streak-like distortions); however, a comparison with a contact map displaying all crystallographic contacts (Fig. 3*B*) shows clearly that certain features of the contact map are reproduced by the map of correlation coefficients.

The analysis presented here shows that there is significant correlation among residues in the family of myoglobin sequences. It also indicates that these correlations may be useful in identifying neighborhood relationships among sequence position without prior knowledge of three-dimensional structure.

I thank Arthur Lesk and Henry Lester for their advice and encouragement. I thank Frank Würriehausen for software development for many aspects of this work and Fred Sigworth, David Colquhoun, and Georg Casari for advice on the manuscript. This work was initiated during a sabbatical at the California Institute of Technology, supported by the Fairchild Foundation. An alignment of globin protein sequences was provided by Arthur Lesk, Medical Research Council, Cambridge.

1. Sander, C. & Schneider, R. (1991) *Proteins Struct. Funct. Genet.* **9**, 56–68.
2. Chothia, C. & Lesk, A. M. (1986) *EMBO J.* **5**, 823–826.
3. Oosawa, K. & Simon, M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 6930–6934.
4. Lesk, A. M. & Boswell, D. R. (1992) *Curr. Opin. Struct. Biol.* **2**, 242–247.
5. Benner, S. A. & Gerloff, D. (1990) in *Advances in Enzyme Regulation*, ed. Weber, G. (Pergamon, Oxford), Vol. 31, pp. 121–181.
6. Russell, R. B., Breed, J. & Barton, G. J. (1992) *FEBS Lett.* **304**, 15–20.
7. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. (1987) *J. Mol. Biol.* **193**, 693–707.
8. Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. (1988) *Protein Eng.* **2**, 193–199.
9. Spiegel, M. R. (1975) *Probability and Statistics* (McGraw-Hill, New York), p. 267.
10. Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York), p. 122.
11. Klapper, M. H. (1971) *Biochim. Biophys. Acta* **229**, 557–566.
12. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136**, 225–270.