Bayly, R. J., & Evans, E. A. (1966) *J. Labelled Compd. 2*, 1.
Chiorazzi, N., Eshhar, Z., & Katz, D. H. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 2091.
Cunningham, L. W., & Nuenke, B. J. (1961) *J. Biol. Chem. 236*, 1716.
David, G. S. (1972) *Biochem. Biophys. Res. Commun. 48*, 464.
David, G. S., & Reisfeld, R. A. (1974) *Biochemistry 13*, 1014.
Ellman, G. L. (1959) *Arch. Biochem. Biophys. 82*, 70.
Eshhar, Z., Benacerraf, B., & Katz, D. H. (1975) *J. Immunol. 114*, 872.
Gennis, R. B., & Cantor, C. R. (1972) *Biochemistry 11*, 2509.
Glover, J. S., Salter, D. N., & Shepherd, B. P. (1967) *Biochem. J. 103*, 120.
Green, N. M. (1970) *Methods Enzymol. 18A*, 418.
Green, N. M. (1975) *Adv. Protein Chem. 29*, 84.
Green, N. M., & Toms, E. J. (1973) *Biochem. J. 133*, 687.
Greenwood, F. C., Hunter, W. M., & Glover, J. S. (1963) *Biochem. J. 89*, 114.
Jasiewicz, M. L., Schoenberg, D. R., & Mueller, G. C. (1976) *Exp. Cell Res. 100*, 213.
Kato, K., Hamaguchi, Y., Fukui, H., & Ishikawa, E. (1975) *J. Biochem. (Tokyo) 78*, 235.
Kato, K., Hamaguchi, Y., Fukui, H., & Ishikawa, E. (1976) *Eur. J. Biochem. 62*, 285.
Katz, D. H. (1974) In *Immunological Tolerance: Mechanisms and Potential Therapeutic Applications* (Katz, D. H., & Benacerraf, B., Eds.) p 189, Academic Press, New York.
Katz, D. H., & Benacerraf, B. (1974) In *Immunological Tolerance: Mechanisms and Potential Therapeutic Applications* (Katz, D. H., & Benacerraf, B., Eds.) p 249, Academic Press, New York.
Kitagawa, T., & Aikawa, T. (1976) *J. Biochem. (Tokyo) 79*, 233.
Klotz, I. M., & Heiney, R. E. (1962) *Arch. Biochem. Biophys. 96*, 605.
Lindsay, D. G., & Shall, S. (1971) *Biochem. J. 121*, 737.
Liu, F.-T., & Leonard, N. J. (1979) *J. Am. Chem. Soc. 101* (in press).
Liu, F.-T., & Katz, D. H. (1979) *Proc. Natl. Acad. Sci. U.S.A.* (in press).
Liu, F.-T., Zinnecker, M., Hamoaka, T., & Katz, D. H. (1978) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 37*, 1375.
Lowry, O. H., Rosebrough, N. J., Farr, A. L., & Randall, R. J. (1951) *J. Biol. Chem. 193*, 265.
Maloy, W. L. (1977) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 36*, 873.
Smyth, D. G., Nagamatsu, A., & Fruton, J. S. (1960) *J. Am. Chem Soc. 82*, 4600.

Improvements in the Prediction of Protein Backbone Topography by Reduction of Statistical Errors†

Frederick R. Maxfield and Harold A. Scheraga*

ABSTRACT: We have simplified our previously published method for predicting the occurrence of residues in one of five conformational states [Maxfield, F. R., & Scheraga, H. A. (1976) *Biochemistry 15*, 5138] without sacrificing accuracy. An increase in the size of the data set from 3681 to 5082 residues led to a slight (1–2%) increase in the accuracy. In order to overcome the limitations in the accuracy due to statistical errors, we tested the usefulness of averaging the predictions for several homologous proteins at each position in the aligned sequence. When this procedure was used on 15 cytochrome *c* sequences and 24 globins, the accuracy of the predictions increased by 8 and 6%, respectively. Averaging did not improve the accuracy when only a few homologous sequences were available or when there was only a slight variability in the amino acid sequence. The improved accuracy from the use of homologous proteins and the slight improvement from an increase in the size of the data set are consistent with the hypothesis that statistical errors place a significant limitation on the accuracy of predictions which incorporate pairwise interactions between neighboring residues. Since a large increase in the size of the data set will be required to reduce the statistical errors significantly, the use of homologous sequences appears to be the most promising way to improve the predictions.

In a previous paper (Maxfield & Scheraga, 1976), which will be referred to as paper 1, we described and evaluated a new method for the prediction of the backbone conformational states of proteins. That method was used to assign each residue in a protein to one of five conformational states, based on intraresidue interactions and on the interactions with the four nearest neighbors on either side. The data used to estimate the effects of these intraresidue and medium-range interactions[1] were obtained from an analysis of 3681 residues in 20 proteins. By use of these data, residues were predicted to occur in one of the five conformational states with an accuracy of 56%. A two-state prediction (α helical or non-

† From the Department of Chemistry, Cornell University, Ithaca, New York 14853. *Received June 28, 1978; revised manuscript received October 2, 1978.* This work was supported by research grants from the National Science Foundation (PCM75-08691) and from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312).

[1] In this paper, medium-range interactions refer to the effects of neighboring residues, out to the fourth nearest neighbors, on the backbone conformation of a residue. Némethy & Scheraga (1977) have suggested another definition for medium-range interactions which is not adopted here in order to maintain consistency with our earlier paper (Maxfield & Scheraga, 1976).

0006-2960/79/0418-0697$01.00/0

helical) yielded an accuracy of 76%. The accuracy of the two-state prediction was comparable to that obtained by other methods (Ptitsyn & Finkel'shtein, 1970; Lewis & Scheraga, 1971; Robson & Pain, 1971; Burgess et al., 1974; Chou & Fasman, 1974; Fasman et al., 1976; Lim, 1974; Tanaka & Scheraga, 1976; Nagano, 1977). In the five-state prediction, each conformational state was defined by a specific range of ϕ,ψ dihedral angles,[2] so that the predicted conformations could be used as the starting point for energy minimization or other procedures used to predict the three-dimensional structures of proteins (reviewed by Nemethy & Scheraga, 1977). Since the definition of conformational states in the five-state prediction was different from that used by others, it was not possible to compare directly the accuracy of this prediction with other methods. However, the percentage of residues correctly assigned to a specific region of conformational space (as opposed to predictions such as "coil" which contain little information) was higher with the five-state prediction than with other methods.

Since it was clear that the accuracy of prediction methods was not high enough to provide reliable starting conformations for protein folding algorithms, a detailed analysis of the statistical errors in the predictions was made. It was found that the data set of 3681 residues was not large enough to evaluate accurately the effects of pairwise interactions between neighboring residues, and the statistical error in evaluating these interactions was probably the major factor limiting the accuracy of the predictions. From this analysis, it became evident that the way to increase the accuracy of predictions, before further refinements in the method are undertaken, is by reducing the statistical error caused by the size of the data set.

In this paper, we explore the use of homologous protein sequences to improve the accuracy of predictions. The predictions at each position in an aligned sequence are averaged over all the proteins in the homologous family. The averaged prediction is then compared to the known structure of a member of the family. If statistical uncertainty is a major source of error in the predictions, then averaging over several different sequences should improve the accuracy.

Seven new protein structures have been added to our data set, increasing the number of residues from 3681 to 5082. The increased size of the data set reduces the statistical error, and we show that this allows us to simplify our prediction algorithm considerably without sacrificing accuracy.

It must be emphasized that the procedure discussed here, as well as all similar ones described in the literature, neglects long-range interactions. Therefore, it is theoretically impossible for such procedures to attain an accuracy of 100%. However, it remains to be seen how closely such approximate methods can approach 100% in accuracy and how accurate they have to be to serve as useful starting points in subsequent procedures that introduce the long-range interactions.

Methods

Backbone dihedral angles were calculated from X-ray coordinates which were provided by the Protein Data Bank, Brookhaven National Laboratories, Upton, NY. The 20 proteins used in our previous study (Maxfield & Scheraga, 1976) were used again, with any revisions entered into the Protein Data Bank files as of January, 1977. Coordinates for the following proteins were also used: trypsin (Chambers & Stroud, 1977), tuna cytochrome *c* (Swanson et al., 1977), the Fab′ fragment of the immunoglobulin IgG New (Poljak et al., 1974), ferredoxin (Adman et al., 1976), triose phosphate isomerase (Banner et al., 1976), the variable part of the Bence-Jones protein REI (Epp et al., 1975), and carbonic anhydrase B (Kannan et al., 1975). Structures for the following proteins were used to check the accuracy of predictions on the globin family: human fetal hemoglobin, γ chain (Frier & Perutz, 1977), and the α and β chains of horse hemoglobin (Bolton & Perutz, 1970).

Sequences and alignments of homologous proteins were obtained from Dayhoff (1972, 1973, 1976).

The details of the prediction method used in this paper have been described in paper 1, and only a brief outline will be given here. The (ϕ,ψ) conformational space is divided into five regions (see Figure 1 of paper 1). The symbols used for these regions, and the approximate (ϕ,ψ) values of the center of these regions, are as follows: ϵ (−90°,150°), α_R (−60°,−40°), ζ_R (−90°,60°), α_L (60°,60°), and ζ_L (60°,−90°). These five regions cover the *entire* (ϕ,ψ) space. In addition, when four or more consecutive residues have dihedral angles $-130° \leq \phi \leq -10°$ and $-90° \leq \psi \leq -10°$, these residues are considered to be part of a right-handed α helix, designated by α_h. When determining the accuracy of a prediction, we considered the α_h and α_R states to be identical, since the dihedral angles for these states are roughly the same.

Three classes of interactions are considered, and estimates of the conformational effects of these interactions are made from an analysis of proteins of known structure.

Intraresidue interactions include the interactions between the side-chain and backbone atoms within a residue. The conformational effects of these interactions are estimated by counting the number of times, n_{km}, that an amino acid, m, occurs in conformation k.

The effect of an amino acid on the conformation of its neighbors, independent of the identity of the neighbors, is considered to be the result of *nonspecific medium-range interactions*. For example, an amino acid with a strong tendency to be helical will generally increase the likelihood for its neighbors to be helical. This type of effect can be accounted for by nonspecific medium-range interactions. An estimate of the effect of these interactions is made by counting the number of times, n_{jkl}, that *any* residue at position i is in conformation state k when residue $i + j$ ($|j| \leq 4$) is a given amino acid l. These interactions have also been described by Robson & Pain (1974) and Robson & Suzuki (1976) and were included in a prediction algorithm by Garnier et al. (1978).

The conformational effects of pairwise interactions between two specific amino acids are defined to result from *specific medium-range interactions*. The influence of these interactions on conformation is estimated by counting the number of times, n_{jklm}, that amino acid m, at position i, is in conformation k, when residue $i + j$ ($|j| \leq 4$) is amino acid l.

A computer program provided in the supplementary material to paper 1 can be used to calculate the n_{km}, n_{jkl}, and n_{jklm}. The conformational states of residues, which are needed as input for this program, can easily be derived from the sequences and dihedral angles provided by the Protein Data Bank, using the definitions of conformational states given in paper 1.

The prediction of the conformation of an amino acid residue is based on the appropriate values of n_{km}, n_{jkl}, and n_{jklm} using eq 1–3, as described below. The tendency for a residue to be in a particular conformation is estimated by calculating the natural logarithm of the odds in favor of that conformation.

[2] The abbreviations and symbols for the description of the conformation of polypeptide chains conform to the rules adopted by an IUPAC–IUB Commission on Biochemical Nomenclature (1970) *Biochemistry 9*, 3471.

The logarithm of the odds was used instead of the more familiar probability because of some mathematically convenient properties of the logarithm of the odds. If Θ_k is the probability for a residue to be in conformation k, then the logarithm of the odds in favor of k is $\ln [\Theta_k/(1 - \Theta_k)]$. The predicted conformation is the one with the largest value of the logarithm of the odds. A computer program used to calculate the logarithm of the odds in favor of each conformation for each residue in a protein is provided as part of the supplementary material to paper 1. A complete user's guide is provided with the program.

The logarithm of the odds in favor of conformation k, for an amino acid m, based on intraresidue interactions and medium-range interactions (see eq A-16 of paper 1), is given by

log of odds in favor of k =

$$\ln [(n_{km} + 0.1)/ \sum_{\substack{k'=1 \\ k' \neq k}}^{6} (n_{k'm} + 0.1)] + \sum_{\substack{j=-4 \\ j \neq 0}}^{4} I_j \quad (1)$$

where the first term represents the effect of intraresidue interactions, and the I_j's, which represent the effect of medium-range interactions, are calculated using eq 2 and 3. The index k' varies from 1 to 6 to include six possible conformational states. The 0.1 which appears in terms in eq 1–3 was added to avoid the computational difficulties which would occur if one of the terms were equal to zero and has a negligible effect on the predictions. Equations 1–3 have been changed from their equivalent formulations in paper 1 (eq A-16, A-17b, and A-15, respectively) by rearranging some of the terms; the content of the equations is not changed by these rearrangements.

The calculation of I_j is somewhat more complex than the calculation of the effects of intraresidue interactions in eq 1. Ideally, only the specific medium-range interactions would be used, since the effects of nonspecific medium-range interactions are included implicitly in the specific interactions. As described in paper 1, the current size of the data set is not sufficiently large for accurate estimates of the effects of specific medium-range interactions, while nonspecific medium-range interactions can be estimated fairly reliably. Therefore, the prediction method described here and more fully in paper 1 uses a mixture of specific and nonspecific interactions to estimate the effects of neighboring residues. The balance between specific and nonspecific interactions is determined by an adjustable parameter, S. For $S = 0$, the effect of neighboring residues is determined exclusively by specific interactions, and for $S = \infty$, these effects are determined exclusively by nonspecific interactions. When S is equal to the number of occurrences of a specific pairwise interaction in the data set (i.e., $S = \sum_{k=1}^{6} n_{jklm}$ for a pairwise interaction described by particular values of j, l, and m), the specific and nonspecific interactions are weighted equally in the prediction. On the average, values of S near 15 would give an even balance between the two types of interactions with the enlarged data base used in this paper. The accuracy of predictions using various values of S is tested to determine the best value for this parameter.

Specifically, the I_j's in eq 1 are calculated, as described in paper 1 (eq A-17b), as

$$I_j = \ln (n_{jklm} + a_{jklm} + 0.1) - \ln [\sum_{\substack{k'=1 \\ k' \neq k}}^{6} (n_{jk'lm} + 0.1) + b_{jklm}] - \ln (n_{km} + 0.1) + \ln [\sum_{\substack{k'=1 \\ k' \neq k}}^{6} (n_{k'm} + 0.1)] \quad (2)$$

where $a_{jklm} + b_{jklm} = S$, and (see paper 1, eq A-15)

$$a_{jklm}/b_{jklm} = (n_{km} + 0.1)(n_{jkl} + 0.1)[\sum_{\substack{k'=1 \\ k' \neq k}}^{6} \sum_{l=1}^{21} (n_{jk'l} + 0.1)][\sum_{\substack{k'=1 \\ k' \neq k}}^{6} (n_{k'm} + 0.1)]^{-1}[\sum_{\substack{k'=1 \\ k' \neq k}}^{6} (n_{jk'l} + 0.1)]^{-1}[\sum_{l=1}^{21} (n_{jkl} + 0.1)]^{-1} \quad (3)$$

The index l varies from 1 to 21 to include each of the 20 amino acids as neighbors, with the value 21 used to indicate the end of the polypeptide chain. The use of the terms a_{jklm} and b_{jklm} in eq 2 is mathematically equivalent to adding S additional occurrences to the data set for each specific medium-range interaction. Using eq 3, these additional "occurrences" are distributed among the six conformational states in the ratio which would be expected on the basis of intraresidue interactions and nonspecific medium-range interactions. A medium-range interaction calculated using eq 2 and 3 is, thus, a weighted mixture of specific and nonspecific medium-range interactions, with the weighting determined by the value of S. Larger values of S will increase the weighting given to nonspecific medium-range interactions.

It may be helpful to provide the justification for adding S additional occurrences to each specific medium-range interaction in the manner done in eq 2 and 3. In statistical problems, one is often required to make an estimate of some parameter based on information obtained from a small sample. In a variety of applications, it has been shown that these estimates can be improved by using information which is not contained in the sample (Efron, 1975; Efron & Morris, 1973; Lindley, 1965). This additional information may be regarded as prior knowledge about the current sample. In our case, the small sample is the occurrences of a specific pairwise interaction in the data set, and the prior knowledge is the information we have about intraresidue and nonspecific medium-range interactions. Bayesian methods, which are used in our prediction procedure (see paper 1), provide a particularly simple method for incorporating this prior knowledge; one simply adds "occurrences" to each of the conformational states in proportion to the number expected on the basis of the prior information. In order to see if this procedure is useful in our application, we used the total number of "occurrences" to be added to each specific medium-range interaction as an adjustable parameter, S. When $S = 0$, the prior knowledge is not used at all. If the procedure is not useful, then $S = 0$ should give the best predictions. We emphasize that S is the only adjustable parameter in eq 1–3. Once S is chosen, a_{jklm} and b_{jklm} are completely determined by the data.

A second parameter, P_{max}, with values between 0 and 1, was also tested in paper 1 as a second method to adjust the balance between the two types of short-range interactions. P_{max} was used to screen specific medium-range interactions for their statistical significance. With $P_{max} = 0$, only nonspecific medium-range interactions were used, and with $P_{max} = 1$, only specific interactions were used. For intermediate values, specific medium-range interactions which were statistically significant at the level of P_{max} were used. This screening procedure was used to remove from the prediction specific medium-range interactions which were not statistically significant at a chosen level (P_{max}). As shown in the Results section, the screening procedure does not lead to increased accuracy with the new, enlarged data set. Hence, the screening procedure may be eliminated, and all medium-range interactions may be calculated using eq 2 and 3. Elimination of the screening procedure is formally equivalent to setting P_{max}

Table I: Occurrence of Six Conformational States in New and Old Data Sets

	conformational state[a]						
	ϵ	α_h	ζ_R	α_L	ζ_L	α_R	total
old data,[b] 20 proteins	1408 38.3%	1112 30.2%	274 7.4%	172 4.7%	91 2.5%	624 17.0%	3681
new data, 27 proteins	2144 42.2%	1334 26.2%	391 7.7%	227 4.5%	155 3.0%	833 16.4%	5084

[a] As defined in the text. [b] This is the data set used previously (Maxfield & Scheraga, 1976).

= 1.

In addition to calculating an estimate for the logarithm of the odds in favor of a conformation, the variance in that estimate can also be calculated, as described in paper 1 (eq A-18 and A-19).

The prediction method used in paper 1 was used without modification to determine the effects of increasing the size of the data set on the accuracy of the predictions.

For the averaged predictions using homologous families of proteins, the prediction program described above was used to estimate the logarithm of the odds favoring each conformation and the variance in this estimate for each residue in the homologous proteins. The sequences for the homologous proteins were then aligned using the diagrams of Dayhoff (1972, 1973, 1976). For each position, the weighted average of the logarithm of the odds for each conformation was calculated, using the variances as the weighting factors. The predicted conformation was the one with the largest value for the weighted-average logarithm of the odds.

Results

Since seven new protein structures have been added to the data set since our initial analysis, we felt that it was necessary to examine the effects of these new data on the predictions before proceeding with the use of homologous sequences to improve the predictions. Data obtained from the crystal structures of trypsin, cytochrome *c*, the Fab′ fragment of IgG New, ferredoxin, triose phosphate isomerase, the variable part of the Bence-Jones protein REI, and carbonic anhydrase B were added to the data set obtained from an analysis of 20 proteins in paper 1 (Maxfield & Scheraga, 1976). Several effects of the incorporation of these new data were noted. As shown in Table I, the fraction of residues in the extended (ϵ) conformation was much higher in the seven new proteins than in the previous data set, and the fraction of α-helical residues (α_h) was much lower. The original data set contained four globin chains, and the occurrence of α-helical structures in a "typical" globular protein may have been overestimated.

The new data set contains nearly 40% more residues than the data set used in paper 1. This increase in the number of residues is expected to increase the statistical reliability of the data. The expected results of this increased reliability are (1) a decreased need to rely on nonspecific (as opposed to specific) medium-range interactions and (2) an increase in the accuracy of the predictions.

As described in the Methods section (and more fully in paper 1), the prediction method uses a combination of specific and nonspecific medium-range interactions. Two adjustable parameters, P_{max} and S, were used to determine the balance between the use of specific and nonspecific interactions, with large values of P_{max} (i.e., values near 1) and small values of S (i.e., values near 0) increasing the reliance on specific medium-range interactions. In paper 1, it was found that the most accurate predictions were obtained when $P_{max} = 0.1$ and $S = 40$. This combination of adjustable parameters meant that more than half of the medium-range interactions were estimated exclusively by nonspecific interactions (due to P_{max}), and those interactions which included some contribution from specific interactions were still heavily weighted by the nonspecific interactions (due to S).

Table II: Number of Residues Predicted Correctly for Several Values of the Adjustable Parameters[a]

	new data set					old data set
P_{max}	0.1	1.0	1.0	1.0	1.0	1.0[b]
S	40	0	30	40	60	40
no. correct	2820	2515	2818	2857	2848	2802
% correct	55.5	49.5	55.4	56.2	56.0	55.1

[a] From a sample of 5084 residues in 27 proteins; i.e., the predictions (using both the new and old data sets) were made on 27 proteins. [b] This value of P_{max} is used only for comparison with the best prediction made with the new data set.

Several values of P_{max} and S were also tested using the new data set, as summarized in Table II. The most accurate predictions are now obtained with $P_{max} = 1$ and $S = 40$. This means that the screening of specific medium-range interactions using P_{max} no longer leads to improved accuracy with the larger data set. Without this screening, all medium-range interactions will now have at least some contribution from specific pairwise interactions. While the optimal value of S remained constant, this actually represents a further decrease in the reliance on nonspecific interactions, since the relative value of S compared to the number of occurrences per pairwise interaction (which is, of course, larger in the new data set) determines the weighting of specific interactions in the prediction.

As shown in Table II, the best value for S was determined by optimizing the predictions for all 27 proteins as a group. In examining the predictions for individual proteins, we found that one prediction was most accurate when $S = 0$, six when $S = 30$, and 15 when $S = 40$ or $S = 60$ (there were several ties). This indicates that the best value of S is between 40 and 60 for most proteins. Thus, we could have optimized S using a small group of test proteins rather than the whole data set, and the optimal value of S would have been nearly the same ($\sim$40) for most random choices of test proteins.

A comparison of the number of residues predicted correctly (see Table II), using the old and new data sets, demonstrates that the incorporation of the new proteins increases the number of correct predictions by 55 residues, or slightly more than 1% of the total number of residues.

The accuracy of predictions for the new proteins, using the old and new data sets, is summarized in Table III. These predictions were made using $S = 40$ and $P_{max} = 1$. When the new proteins were included in the data set, the accuracy of the predictions for the new proteins increased from 56.5 to 58.5%. These predictions (and all other predictions in this paper) were made with the residue being predicted removed from the data set; thus, the increased accuracy is not due to a bias in the data set caused by including the residue being predicted. (See paper 1 for a further discussion of this point.) The accuracy of predictions for the new proteins using the old

Table III: Percentage of Residues Predicted Correctly by Use of the Old and New Data Sets

protein	% correct: old data	new data
trypsin	60.0	60.9
cytochrome *c*	54.5	55.4
Fab′ of IgG New		
light chain	56.3	61.7
heavy chain	52.8	59.6
Bence-Jones protein REI	71.5	73.3
ferredoxin	44.2	44.2
triose phosphate isomerase	62.4	60.4
carbonic anhydrase B	48.8	48.8
total	56.5	58.5

data set (56.5%) is slightly higher than the accuracy (54.5%) reported in paper 1 for proteins contained in the old data set for $P_{max} = 1$ and $S = 40$. The fact that predictions of proteins outside the old data set are roughly as accurate as predictions of proteins in the data set demonstrates that the method is not biased towards proteins in the data set.

Using the expanded data set and our simplified prediction method (with $S = 40$), we tested the use of homologous protein sequences to improve the accuracy of predictions. Predictions were made for each protein in a homologous family, and the predictions for each position in the aligned sequence were averaged over the entire family. This procedure was tested on the cytochrome *c* proteins and the globins since a large number of complete sequences were available for these proteins with considerable variation in the amino acid sequence. Similar tests on smaller families of proteins, or families with little variability in the amino acid sequence, showed little difference between the predictions for a single protein and the averaged prediction. Among the families tested were ribonuclease (six sequences), lysozyme-lactalbumin (six sequences), ferredoxin (six sequences), and cytochrome b_5 (four sequences).

The averaged prediction for 15 cytochrome *c* molecules was compared to the observed conformational states of tuna ferricytochrome *c* (inner molecule) as reported by Swanson et al. (1977). The cytochrome *c* sequences used were the following: tuna, human, pig, chicken, snapping turtle, rabbit, kangaroo, dogfish, lamprey, fruit fly, silkworm moth, wheat, *Candida krusei*, baker's yeast, and *Neurospora crassa*. The results of the averaged prediction are presented in Table IV. The comparison between the averaged prediction and a prediction based only on the tuna cytochrome *c* sequence is summarized in Table V. For the predictions reported in Tables IV and V, both tuna cytochrome *c* and cytochrome c_2 of *Rhodospirillum rubrum* were removed from the data set in order to avoid biasing the predictions in regions where one of the cytochrome *c* sequences would be homologous to one of these two proteins. [There is some sequence homology between the cytochrome *c* sequences and cytochrome c_2 (Dayhoff, 1972).] The averaged prediction has an accuracy of 60% (61 correct out of 101 residues) in predicting the tuna cytochrome *c* conformations as compared to 52% (53/101) for the prediction based solely on the tuna sequence; i.e., the accuracy of the predictions has been improved by 8%. As shown in Table V, the improvement in the accuracy comes from more of both the ϵ and ($\alpha_R + \alpha_h$) conformations being predicted correctly.

The predictions for globins were also generally improved by averaging over 24 sequences, as summarized in Table VI. The globin chains used for the averaging were the following: the α chains of human, dog, horse, kangaroo, echidna, chicken, viper, newt, carp, and desert sucker hemoglobins; the β chains of human, dog, kangaroo, echidna, chicken, horse, potoroo, and frog hemoglobins; the γ chain of human fetal hemoglobin;

Table IV: Prediction of Tuna Cytochrome *c* Conformation by Use of 15 Sequences

residue no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
obsd conformation[a]		ϵ	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_R	α_R	α_R	ϵ
predicted conformation[b]		ϵ	ϵ	α_h	α_h	α_h	α_h	α_h	ϵ	ϵ	ϵ	α_h	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ
sequence	G	D	V	A	K	G	K	K	T	F	V	Q	K	C	A	Q	C	H
residue no.	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
obsd conformation[a]	ϵ	α_R	ϵ	ϵ	ζ_L	ϵ	ϵ	ϵ	ϵ	α_R	ϵ	ϵ	ϵ	α_R	ϵ	ϵ	α_R	ϵ
predicted conformation[b]	ϵ	ϵ	α_h	α_h	α_h	α_h	α_h	ϵ	ϵ	α_R	ϵ	ϵ	ζ_R	α_R	ϵ	α_L	α_R	ϵ
sequence	T	V	E	N	G	G	K	H	K	V	G	P	N	L	W	G	L	F
residue no.	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
obsd conformation[a]	α_L	ϵ	ϵ	ϵ	α_L	α_R	ϵ	ϵ	ϵ	ζ_L	ζ_R	ϵ	ϵ	α_h	α_h	α_h	α_h	α_h
predicted conformation[b]	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	α_R	α_R	ζ_R	α_h	α_h
sequence	G	R	K	T	G	Q	A	E	G	Y	S	Y	T	D	A	N	K	S
residue no.	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
obsd conformation[a]	α_h	ζ_L	ϵ	ϵ	ϵ	ϵ	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	ζ_R	α_h	α_h
predicted conformation[b]	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	α_h	α_h	α_h	α_h	α_h	α_h	ϵ	α_h	α_h	ϵ	ϵ	α_h
sequence	K	G	I	V	W	N	N	D	T	L	M	E	Y	L	E	N	P	K
residue no.	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
obsd conformation[a]	α_h	α_h	ζ_R	ϵ	ζ_L	ϵ	ζ_R	ζ_R	ζ_R	ϵ	α_R	ϵ	ϵ	α_R	ϵ	α_h	α_h	α_h
predicted conformation[b]	ϵ	ϵ	ϵ	ϵ	α_L	α_R	ϵ	ϵ	ϵ	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h
sequence	K	Y	I	P	G	T	K	M	I	F	A	G	I	K	K	K	G	E
residue no.	91	92	93	94	95	96	97	98	99	100	101	102	103					
obsd conformation[a]	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h						
predicted conformation[b]	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_h	α_R	α_R						
sequence	R	Q	D	L	V	A	Y	L	K	S	A	T	S					

[a] Based on the structure of the oxidized inner molecule of cytochrome *c* as reported by Swanson et al. (1977). Conformational states which were the same in the oxidized inner, oxidized outer, and reduced molecules are underlined (see Discussion). [b] The predictions based on 15 homologous cytochrome *c* sequences were averaged, as described in the text.

Table V: Observed and Predicted Conformational States for Tuna Cytochrome *c*

	ϵ	$\alpha_h + \alpha_R$	ζ_R	α_L	ζ_L	total
no. observed	32	56	6	3	4	101
no. predicted using the tuna sequence	44	49	5	2	1	
no. correct	19	34	0	0	0	53
no. predicted by averaging predictions from 15 sequences	49	49	2	1	0	
no. correct	22	39	0	0	0	61

myoglobin of human, sperm whale, and red kangaroo; globin V of sea lamprey; and lamprey globin.

The new and old data sets contain the α and β chains of human hemoglobin, sperm whale myoglobin, and sea lamprey globin V. In addition, crystal structures were available for the γ chain of human fetal hemoglobin and for the α and β chains of horse hemoglobin. The averaged prediction of globin chains was made using a data set which had all globin chains removed in order to avoid biasing the predictions in regions where one of the 24 globin sequences would be homologous to one of the chains in the complete data set. The removal of all globin chains from the data set reduces the occurrence of helical residues in the data set, and, as seen in Table VI, the accuracy of predictions of globins is reduced as a consequence. However, for testing the usefulness of making averaged predictions, it is important to have a data set which is not biased towards the sequences used for the test. When the last two columns of Table VI are compared, it can be seen that the averaged prediction is generally more accurate than the predictions based only on the sequence for a single globin chain. The 6% average improvement in the accuracy is roughly the same as the 8% improvement noted for the tuna cytochrome *c* prediction. These two cases indicate that an averaged prediction, based on the sequences of several homologous proteins with considerable amino acid substitutions, can be significantly more accurate than a prediction based only on the sequence of the protein of interest.

Discussion

We have been developing a prediction method which could be used to generate starting conformations for procedures designed to predict the three-dimensional structures of proteins (see review by Némethy & Scheraga, 1977). Other possible applications of prediction methods were discussed by Chou & Fasman (1978). Recognizing that the usefulness of prediction methods is a subject of controversy (Matthews, 1975; Burgess & Scheraga, 1975; Maxfield & Scheraga, 1976), we designed our method with a firm statistical basis which allowed us to determine the role of statistical errors in limiting the accuracy of predictions (Maxfield & Scheraga, 1976). The method is simple to use and provides unambiguous predictions. By use of a computer program, the predicted conformation is determined using eq 1–3.

The prediction procedure used in this paper is similar to the method described by Robson (1974). Robson & Pain (1971) described a prediction method based on intraresidue interactions and specific medium-range interactions, and Garnier et al. (1978) have recently described a prediction algorithm based on intraresidue interactions and nonspecific medium-range interactions. Aside from the definitions of conformational states, these methods are nearly equivalent to our method with $S = 0$ or $S = \infty$, respectively. Additionally, Robson (1974) and Garnier et al. (1978) have described the use of decision constants to improve the predictions. These decision constants are used to change the relative number of residues predicted in each conformational state, and the use of these parameters increases the accuracy of predictions (Garnier et al., 1978).

The main improvement in our method over that of Robson and his colleagues is that we have included *both* specific and nonspecific medium-range interactions in the same prediction. As discussed in the Methods section and in paper 1, the Bayesian methods provide a very natural way for including both types of interactions. The results presented in Table II and in paper 1 demonstrate that predictions based on a mixture of the two types of interactions are superior to predictions based on only one type of medium-range interaction.

Effect of Increasing the Data Set. In paper 1, we concluded that the size of the data set (3681 residues) was sufficient for good estimates of the effects of intraresidue and nonspecific medium-range interactions, but the specific medium-range interactions could not be estimated reliably. Similar conclusions were reached by Robson & Suzuki (1976). Therefore, we were interested to see if a 40% increase in the size of the data set would increase the accuracy of the predictions. As shown in Tables II and III, a 1–2% increase in accuracy could be obtained by increasing the data set from 3681 to 5082 residues. This increase in accuracy is rather small. However, it must be remembered that there are still only about 15 occurrences in the data set per pairwise interaction, and the statistical error in estimating the effects of specific medium-range interactions remains very large. Obviously, it would take a very large increase in the size of the data set to obtain good estimates for the effects of specific medium-range interactions.

Although the accuracy of the predictions was not greatly improved, the increased size of the data set did allow us to

Table VI: Effect of Averaging on the Accuracy of Predictions of Globins

protein	no. of residues	no. correct		
		A[a]	B[b]	C[c]
human hemoglobin				
α chain	139	87	61	73
β chain	144	80	58	79
fetal γ chain	144	86	60	80
horse hemoglobin				
α chain	139	92	50	73
β chain	144	100	63	75
sperm whale myoglobin	151	114	98	74
sea lamprey globin V	146	91	76	72
total	1007	650 (64.5%)	466 (46.3%)	526 (52.2%)

[a] A, by use of the new data set. [b] B, with all globins removed from the new data set. [c] C, with globins removed from the new data set and the average prediction from 24 globins used.

simplify our prediction procedure (by eliminating the use of the parameter P_{max}) without sacrificing accuracy. This change in the procedure also means that specific medium-range interactions play a larger role in the predictions.

Predictions Based on Homologous Sequences. The use of averaged predictions for homologous families of proteins led to an 8% increase in the accuracy of the tuna cytochrome *c* prediction and a 6% increase in the accuracy of the prediction of globins. These results indicate that, when a large family of homologous proteins with substantial variability in the amino acid sequence is available, the prediction obtained by averaging over the entire family is likely to be considerably more accurate than a prediction based on a single sequence. For smaller families of proteins, or families with few amino acid substitutions, we found that there was little or no advantage in using the averaged prediction. The sequences used in the cytochrome *c* and globin family predictions were chosen in order to include a large number of amino acid substitutions. By averaging over several protein sequences, we may reduce some of the errors which occur because of statistical uncertainty in the data set. The improvements noted as a result of an enlarged data set or the use of homologous sequences are both consistent with the hypothesis that improvements in the accuracy of predictions can be achieved by reducing the statistical error which results from the small size of the data set. Recently, Garnier et al. (1978) have reported preliminary results which indicated that improvements of 5–10% could be obtained by using homologous sequences.

The use of homologous families of proteins to make averaged predictions depends on the assumption that the structures of the homologous proteins are the same. For our purposes, this means that the residues in the same positions in the aligned sequence should be in the same conformational state. This is generally a good assumption. For example, using the alignment of Dayhoff (1972), the serine proteases trypsin and elastase have identical amino acids in 40% of the aligned positions, and we find that 80% of the aligned positions is in the same conformational state. In pairwise comparisons between the globins listed in Table VI, we find an average sequence homology of 42% and an average conformational state homology of 87%. The percentage of conformational homology between different members of a homologous family is similar to the conformational homologies, which are discussed below, among tuna ferrocytochrome *c* and two independent molecules of tuna ferricytochrome *c*. Since the amount of conformational homology is considerably higher than the accuracy of the predictions, it seems likely that no substantial error is introduced in an averaged prediction by assuming that homologous sequences are directing the aligned residues to be in the same conformational states in each member of a homologous family.

In paper 1, we discussed several of the factors that limit the accuracy of predictions of backbone topography. In that paper, we concluded that the quantity of data currently available is a major source of error for prediction methods which include the effects of specific pairwise medium-range interactions. The results presented in this paper support that conclusion. The quality of X-ray crystal structures could also affect the accuracy of predictions, and the recent publication of the coordinates of tuna ferrocytochrome *c* and two symmetry-independent molecules of tuna ferricytochrome *c* (which were designated inner and outer) by Dickerson's group (Swanson et al., 1977; Takano et al., 1977; Mandel et al., 1977) provides some insight into the limits which the quality of X-ray data might place on the accuracy of predictions. The X-ray data for the three cytochrome *c* structures were analyzed independently, and the structures of the molecules were compared using several measures of the similarities or differences between the deduced structures (Mandel et al., 1977). These authors reported that there were no unambiguous differences in the main-chain conformations of the three molecules. Nevertheless, there were some individual residues with large differences in their backbone dihedral angles in the three structures. These discrepancies were attributed to the difficulty in finding a unique fit of the polypeptide chain to the observed electron density. As a result, the conformatinal states assigned to the reduced molecule differed from the state assigned to the oxidized inner and outer molecules in 22 and 19% of the residues, respectively. The two oxidized molecules differed in 11% of their conformations. It was found that 75% of the residues are in the same conformational state in all three structures (see Table IV), and only two residues were in different conformational states in all three structures. If the cytochrome *c* structures are typical of the coordinates used for our data set, then the discrepancies mentioned above would suggest that roughly 10% of the residues in the data set are assigned to the wrong conformational state. These errors will affect the parameters used to estimate the effects of intraresidue and medium-range interactions, and they will also appear when comparing the predicted conformations to the X-ray structures. It is interesting to note that, for the residues where all three experimental structures agree, the accuracy of the averaged cytochrome *c* prediction is 66%, vs. 60% accuracy for the same prediction compared to the entire oxidized inner molecule (which is the molecule recommended by the crystallographers), 57% accuracy for the oxidized outer molecule, and 54% accuracy for the reduced molecule. This indicates that the residues which are poorly determined in the X-ray structure are predicted with lower accuracy, possibly because of incorrect dihedral angles in the crystal structure.

The current accuracy of our prediction method, even with the improvements described in this paper, is still far from the accuracy of the experimental determinations of the conformations of residues in proteins The improvements noted in this paper when the data set was increased by 40% indicate that reduction of statistical errors can partially reduce the gap between the accuracy of the predictions and the accuracy of crystal structures. Further improvements in the accuracy of predictions can be expected as new structures, and more accurate determinations of old structures, become available. In the meanwhile, procedures such as the use of homologous proteins, as described in this paper, can provide a significant increase in the accuracy of predictions.

References

Adman, E. T., Sieker, L. C., & Jensen, L. H. (1976) *J. Biol. Chem. 251*, 3801.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., & Wilson, I. A. (1976) *Biochem. Biophys. Res. Commun. 72*, 146.

Bolton, W., & Perutz, M. F. (1970) *Nature (London) 228*, 551.

Burgess, A. W., & Scheraga, H. A. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 1221.

Burgess, A. W., Ponnuswamy, P. K., & Scheraga, H. A. (1974) *Isr. J. Chem. 12*, 239.

Chambers, J. L., & Stroud, R. M. (1977) *Acta Crystallogr., Sect. B 33*, 1824.

Chou, P. Y., & Fasman, G. D. (1974) *Biochemistry 13*, 222.

Chou, P. Y., & Fasman, G. D. (1978) *Annu. Rev. Biochem. 47*, 251.

Dayhoff, M. O., Ed. (1972) *The Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Silver Spring, MD.
Dayhoff, M. E., Ed. (1973) *The Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 1, National Biomedical Research Foundation, Silver Spring, MD.
Dayhoff, M. O., Ed. (1976) *The Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 2, National Biomedical Research Foundation, Silver Spring, MD.
Efron, B. (1975) *Adv. Math. 16*, 259.
Efron, B., & Morris, C. (1973) *J. Am. Stat. Assoc. 68*, 117.
Epp, O., Lattman, E. E., Schiffer, M., Huber, R., & Palm, W. (1975) *Biochemistry 14*, 4943.
Fasman, G. D., Chou, P. Y., & Adler, A. J. (1976) *Biophys. J. 16*, 1201.
Frier, J. A., & Perutz, M. F. (1977) *J. Mol. Biol. 112*, 97.
Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97.
Kannan, K. K., Notstrand, B., Fridborg, K., Lövgren, S., Ohlsson, A., & Petef, M. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 51.
Lewis, P. N., & Scheraga, H. A. (1971) *Arch. Biochem. Biophys. 144*, 576.
Lim, V. I. (1974) *J. Mol. Biol. 88*, 873.
Lindley, D. V. (1965) *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Part 2, pp 141–153, Cambridge University Press, Inference, London.
Mandel, N., Mandel, G., Trus, B. L., Rosenberg, J., Carlson, G., & Dickerson, R. E. (1977) *J. Biol. Chem. 252*, 4619.
Matthews, B. W. (1975) *Biochem. Biophys. Acta 405*, 442.
Maxfield, F. R., & Scheraga, H. A. (1976) *Biochemistry 15*, 5138.
Nagano, K. (1977) *J. Mol. Biol. 109*, 251.
Némethy, G., & Scheraga, H. A. (1977) *Q. Rev. Biophys. 10*, 239.
Poljak, R. J., Amzel, L. M., Chen, B. L., Phizackerley, R. P., & Saul, F. (1974) *Proc. Natl. Acad. Sci. U.S.A. 71*, 3440.
Ptitsyn, O. B., & Finkel'shtein, A. V. (1970) *Biofizika 15*, 757.
Robson, B. (1974) *Biochem. J. 141*, 853.
Robson, B., & Pain, R. H. (1971) *J. Mol. Biol. 58*, 237.
Robson, B., & Pain, R. H. (1974) *Biochem. J. 141*, 883.
Robson, B., & Suzuki, E. (1976) *J. Mol. Biol. 107*, 327.
Swanson, R., Trus, B. L., Mandel, N., Mandel, G., Kallai, O. B., & Dickerson, R. E. (1977) *J. Biol. Chem. 252*, 759.
Takano, T., Trus, B. L., Mandel, N., Mandel, G., Kallai, O. B., Swanson, R., & Dickerson, R. E.. (1977) *J. Biol. Chem. 252*, 776.
Tanaka, S., & Scheraga, H. A. (1976) *Macromolecules 9*, 168.

Comparative in Vivo Nitrogen-15 Nuclear Magnetic Resonance Study of the Cell Wall Components of Five Gram-Positive Bacteria†

Aviva Lapidot* and Charles S. Irving

ABSTRACT: The proton-decoupled 9.12-MHz ^{15}N NMR spectra of ^{15}N-labeled *Bacillus subtilis*, *Bacillus licheniformis*, *Staphylococcus aureus*, *Streptococcus faecalis*, and *Micrococcus lysodeikticus* intact cells, isolated cells walls, and cell wall digests have been examined. The general characteristics of Gram-positive bacteria ^{15}N NMR spectra are described and spectral assignments are provided, which allow in vivo ^{15}N NMR to be applied to a wide range of problems in bacterial cell wall research. The qualitative similarity of the intact cell and cell wall spectra found in each bacteria allowed the ^{15}N resonances observed in the proton broad-band noise-decoupled ^{15}N NMR spectra of intact cells to be assigned to cell wall components. Each of the five Gram-positive bacteria displayed a unique set of cell wall ^{15}N resonances, which reflected variations in the primary structure of peptidoglycans and the amounts of teichoic acid and teichuronic acid in the cell wall, as well as the dynamic properties of the cell wall polymers. Spectral assignments of cell wall ^{15}N resonances assigned to teichoic D-Ala residues, teichuronic acid and acetamido groups, and peptidoglycan acetamido, amide, peptide, and free amino groups have been made on the basis of specific isotopic labeling and dilution experiments, comparison of chemical shifts to literature values, determination of pH titration shifts, cell wall fractionation experiments, and comparative analysis of the cell wall lysozyme digest spectra in terms of the known primary sequences of peptide chains. All the peptidoglycan ^{15}N peptide resonances observed in the intact cells and isolated cell walls could be accounted for by residues in the bridge or crossbar regions of the peptide chains, which indicated that only the cross-linking groups had a high degree of motional freedom. Thermal- and pH-induced conformational changes around the cross-linking D-Ala residues were detected in the *B. licheniformis* cell wall lysozyme digest products. Comparison of the proton broad-band noise-decoupled and gated decoupled intact cell and cell wall ^{15}N spectra indicated that broad-band proton decoupling resulted in nulling of cytoplasmic resonances and enhancement of the cell wall resonances by the $^{15}N\{^1H\}$ nuclear Overhauser effect.

The chemical compositions of the Gram-positive bacterial cell walls and the primary structures of cell wall peptidoglycans, teichoic acids, and teichuronic acids have been thoroughly characterized (Ghuysen & Shockman, 1973; Tipper, 1970; Hughes, 1968). Much less is known about the physical and structural properties of bacterial cell wall polymers (Rogers, 1974) due to difficulties in making conformational and dynamic measurements on peptidoglycans. This has prevented the adequate testing of various three-dimensional structural models proposed for bacterial cell wall peptidoglycan (Tipper, 1970; Keleman & Rogers, 1971; Braun et al., 1973; Oldmixon et al., 1974; Formanek et al., 1974).

Nuclear magnetic resonance spectroscopy is a powerful tool for probing the conformations, noncovalent bonding interactions, and molecular motion of polymeric molecules.

† From the Department of Isotope Research, Weizmann Institute of Science, Rehovot, Israel. *Received July 25, 1978.*

0006-2960/79/0418-0704$01.00/0