Acad. Sci. U.S.A. 85, 3329-3333.
Hartshorne, R. P., & Catterall, W. A. (1984) *J. Biol. Chem. 259*, 1667-1675.
Hjelmeland, L. M., & Chrambach, A. (1984) *Methods Enzymol. 104*, 305-318.
Isacoff, E. Y., Jan, Y. N., & Jan, L. Y. (1990) *Nature (London) 345*, 530-534.
Latorre, R., Oberhauser, A., Labarca, P., & Alvarez, O. (1989) *Annu. Rev. Physiol. 51*, 385-399.
Lu, L., Montrose-Rafizadeh, C., & Guggino, W. B. (1990) *J. Biol. Chem. 265*, 16190-16194.
MacKinnon, R., & Miller, C. (1989) *Biochemistry 28*, 8087-8092.
MacKinnon, R., Reinhart, P. H., & White, M. M. (1988) *Neuron 1*, 997-1001.
MacKinnon, R., Latorre, R., & Miller, C. (1989) *Biochemistry 28*, 8092-8099.
McCormack, K., Lin, J. W., Iverson, L. E., & Rudy, B. (1990) *Biochem. Biophys. Res. Commun. 171*, 1361-1371.
Miller, C. (1987) *Biophys. J. 52*, 123-126.
Miller, C., Moczydlowski, E., Latorre, R., & Phillips, M. (1985) *Nature (London) 313*, 316-318.
Nakayama, N., Kirley, T. L., Vaghy, P. L., McKenna, E., & Schwartz, A. (1987) *J. Biol. Chem. 262*, 6572-6576.
Newman, M. J., Foster, D. L., Wilson, T. H., & Kaback, H. R. (1982) *J. Biol. Chem. 256*, 11804-11808.
Neyton, J., & Miller, C. (1988a) *J. Gen. Physiol. 92*, 549-567.
Neyton, J., & Miller, C. (1988b) *J. Gen. Physiol. 92*, 569-586.
Oliva, C., Folander, K., & Smith, J. S. (1991) *Biophys. J. 59*, 450a.
Peterson, O. H., & Maruyama, Y., (1984) *Nature (London) 307*, 693-696.
Pragnell, M., Snay, K. J., Trimmer, J. S., MacLusky, N. J., Naftolin, F., Kaczmarek, L. K., & Boyle, M. B. (1990) *Neuron 4*, 807-812.
Rehm, H., & Lazdunski, M. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 4919-4923.
Ruppersberg, J. P., Schroter, K. H., Sakmann, B., Stocker, M., Sewing, S., & Pongs, O. (1990) *Nature (London) 345*, 535-537.
Slaughter, R. S., Shevell, J. L., Felix, J. P., Garcia, M. L., & Kaczorowski, G. J. (1989) *Biochemistry 28*, 3995-4002.
Smith, C., Phillips, M., & Miller, C. (1986) *J. Biol. Chem. 261*, 14607-14613.
Stoscheck, C. M. (1987) *Anal. Biochem. 160*, 301-305.
Sugg, E. E., Garcia, M. L., Reuben, J. P., Patchett, A. A., & Kaczorowski, G. J. (1990) *J. Biol. Chem. 265*, 18745-18748.
Takumi, T., Ohkubo, H., & Nakanishi, S. (1988) *Science 242*, 1042-1045.
Tanford, C., & Reynolds, J. A. (1976) *Biochim. Biophys. Acta 457*, 133-170.
VanDongen, A. M. J., Frech, G. C., Drewe, J. A., Joho, R. H., & Brown, A. M. (1990) *Neuron 5*, 433-443.
Vâzquez, J., Feigenbaum, P., Kaczorowski, G. J., & Garcia, M. L. (1988) *J. Cell Biol. 107*, 143a.
Vâzquez, J., Feigenbaum, P., Katz, G. M., King, V. F., Reuben, J. P., Roy-Contancin, L., Slaughter, R. S., Kaczorowski, G. J., & Garcia, M. L. (1989) *J. Biol. Chem. 264*, 20902-20909.
Vergara, C., & Latorre, R. (1983) *J. Gen. Physiol. 82*, 543-568.
Villarroel, A., Alvarez, O., Oberhauser, A., & Latorre, R. (1988) *Pflugers Arch. 413*, 118-126.

New Joint Prediction Algorithm (Q_7-JASEP) Improves the Prediction of Protein Secondary Structure

Vellarkad N. Viswanadhan,* Benjamin Denckla, and John N. Weinstein
Laboratory of Mathematical Biology, National Cancer Institute, Building 10, Room 4B-56, National Institutes of Health, Bethesda, Maryland 20892

Received January 28, 1991; Revised Manuscript Received August 26, 1991

ABSTRACT: The classical problem of secondary structure prediction is approached by a new joint algorithm (Q_7-JASEP) that combines the best aspects of six different methods. The algorithm includes the statistical methods of Chou-Fasman, Nagano, and Burgess-Ponnuswamy-Scheraga, the homology method of Nishikawa, the information theory method of Garnier-Osgurthope-Robson, and the artificial neural network approach of Qian-Sejnowski. Steps in the algorithm are (i) optimizing each individual method with respect to its correlation coefficient (Q_7) for assigning a structural type from the predictive score of the method, (ii) weighting each method, (iii) combining the scores from different methods, and (iv) comparing the scores for α-helix, β-strand, and coil conformational states to assign the secondary structure at each residue position. The present application to 45 globular proteins demonstrates good predictive power in cross-validation testing (with average correlation coefficients per test protein of $Q_{7,\alpha} = 0.41$, $Q_{7,\beta} = 0.47$, $Q_{7,c} = 0.41$ for α-helix, β-strand, and coil conformations). By the criterion of correlation coefficient (Q_7) for each type of secondary structure, Q_7-JASEP performs better than any of the component methods. When all protein classes are included for training and testing (by cross-validation), the results here equal the best in the literature, by the Q_7 criterion. More generally, the basic algorithm can be applied to any protein class and to any type of structure/sequence or function/sequence correlation for which multiple predictive methods exist.

Prediction of secondary and tertiary structures of a globular protein from the amino acid sequence remains a major unsolved problem in biology. This is the case despite considerable progress in several key areas, including energy minimization [e.g., Gibson and Scheraga (1986)], molecular dynamics [e.g., Karplus and McCammon (1983), Karplus and Weaver (1976), Levitt and Meirovitch (1983), and Rooman and Wodak (1988)], development of simplified protein potentials [e.g.,

This article not subject to U.S. Copyright. Published 1991 by the American Chemical Society

Crippen and Viswanadhan (1985), Gregoret and Cohen (1990), Miyazawa and Jernigan (1985), and Seetharamulu and Crippen (1991)], discrete enumeration (Covell & Jernigan, 1990), Monte Carlo simulations [e.g., Skolnick and Kolinski (1990)], database and rule-based approaches [e.g., Blundell et al. (1987)], pattern recognition methods [e.g., Cohen et al. (1983), Holley and Karplus (1989), Kneller et al. (1990), Qian and Sejnowski (1988), and Rose (1978)], and several other probabilistic predictive approaches [e.g., Burgess et al. (1974), Chou and Fasman (1978), and Garnier et al. (1978)]. The importance of this problem is underscored by the human genome initiative and the consequent need to make sense out of the enormous amounts of sequence information that are becoming available. All of the current methods for predicting secondary structure fall into one of three major classes: (i) those based on a theory of protein structure and folding [e.g., Lim 1974a,b and Schiffer and Edmundson (1967)]; (ii) those based on homology (or sequence similarity) with known structures (or substructures) [e.g., Levine and Garnier (1988), Levine et al. (1986), and Nishikawa and Ooi (1986)]; (iii) those based on statistical/empirical/mathematical principles [e.g., Burgess et al. (1974), Chou and Fasman (1978), Cohen et al. (1983), Garnier et al. (1978), Holley and Karplus (1989), and Kneller et al. (1990)]. Significant recent additions to the last class have resulted from the advent of computational neural networks, which embody rules learned heuristically by mapping a set of inputs to a set of outputs based on a training procedure [e.g., Bohr et al. (1988), Holley and Karplus (1989), Kneller et al. (1990), and Qian and Sejnowski (1988)].

Good secondary structure prediction is an important starting point for modeling supersecondary and tertiary aspects of protein structure (Taylor & Thornton, 1983). Since secondary structure predictions by any single algorithm seem to have reached a level of saturation, we wondered whether an appropriate joint approach would improve predictions. Although joint approaches for secondary structure prediction have been used earlier (Argos et al. 1976; Biou et al., 1988: Lenstra, 1977; Mr'azek & Kypr, 1988; Viswanadhan et al., 1990b), the rules for combining different methods have remained arbitrary and ad hoc, as also have the ways to evaluate them. Argos et al. (1976) combined five different methods and observed that their joint prediction quality remained at the same level as that of the best method included. Later, Nishikawa and co-workers (Nishikawa, 1983; Nishikawa & Ooi, 1986b) used a joint approach based on a combination of three different methods and reported no substantial improvement in the prediction quality. There have also been a number of applications to particular proteins or a homologous set, using ad hoc rules (Bourgeois et al., 1979; Crawford et al., 1987; Schulz et al., 1974; Viswanadhan et al. 1990b). All of these joint algorithms combine "final category" predictions of different methods, rather than working directly with the numerical prediction scores used to arrive at the final categorizations. Information is thus wasted; prediction scores indicate the relative certainty of prediction and should be used if we are interested in more than a "majority vote" for a joint prediction method.

The objective of the present work is to present a joint approach using the numerical prediction scores of each method. An important consideration in this regard is the criterion to use for the fine tuning of each method and for optimizing the performance of the joint algorithm. In the present work, we demonstrate that a prediction quality index known as "Q_7" (defined as the correlation coefficient for predicted and observed secondary structure in a two-state prediction model) shows a well-defined maximum with respect to selection of a threshold for categorization. In contrast, another popular index known as "Q_3" (defined as the percentage of correct predictions) often does not exhibit a well-defined maximum in that regard. Hence, we use the Q_7 index for fine tuning the various methods as well as for the joint algorithm.

Traditionally, proteins have been divided into structural classes (Levitt & Chothia 1976) on the basis of patterns in secondary structure. In the present work, we consider a database of 45 proteins containing *all* important protein classes for initial application of this joint prediction algorithm. We show here that a joint algorithm (Q_7-JASEP) based on a combination of weighted prediction scores consistently improves prediction quality (Q_7) over the component methods and over any other single method applicable to all globular proteins.

METHODS

(A) The Protein Secondary Structure Database

High-resolution crystal structures of a set of 45 globular proteins, including all important protein classes, are identified (Table I) from the most recent version of the Brookhaven Protein Data Bank (Bernstein et al., 1977). The secondary structure category of each amino acid residue in these proteins is objectively assigned from X-ray crystallographic coordinates using the DSSP algorithm (Kabsch & Sander, 1983). The DSSP secondary structure classification into eight types is regrouped into a three-state classification (α-helix, β-strand, and coil). The sets of secondary structure assignments form the "targets", which are used in parametrization of the algorithm (discussed below) or for comparing with predictions.

(B) The Q_7-JASEP Algorithm

The basis of the Q_7 Joint Algorithm for Secondary Structure Prediction (Q_7-JASEP) of proteins is summarized schematically in Figure 1. The algorithm includes four stages: I. RUN-PRED, II. THRESHER, III. JASEP-INIT, and IV. JASEP-FIN. These are briefly described below.

(*I*) RUN-PRED. In this first step, a sequence is analyzed by running each of six prediction methods (described in section C of Methods). Quantitative scores, obtained from each of the six individual predictive methods, are passed on to the THRESHER stage of the algorithm.

(*II*) THRESHER. Rather than consider the "final category" predictions of each method, we preferred to use the numerical predictive potentials, hereafter denoted $p_{i,j}(k)$, where k denotes the predictive method, i the residue position, and j the secondary structure type ($j = \alpha$ for helical, $j = \beta$ for extended, and j = c for coil). This approach maximizes the amount of information extracted from each predictive method, eliminating the less important (often subjective) rules for transforming the numerical prediction scores into category predictions.

First, the numerical scores (which are expressed originally in the quantitative range and units characteristic of each method) are transformed by range scaling into real numbers between −1 and +1. This transformation is achieved for each method by steps a and b.

(*a*). Let $t_j(k)$ represent the threshold for a method k and structural type j that classifies all residue positions i into j or non-j type on the basis of the following criterion: if $p_{i,j}(k) > t_j(k)$, then position i is predicted to be in state j and vice versa. To find the *optimal* threshold $T_j(k)$, the set of predictive scores (for all proteins) is scanned at 1000 equally spaced points [possible thresholds, $t_j(k)$] along the entire numerical range of predictive scores. $T_j(k)$ is thus found by determining the quality of predictions (overall Q_3 or Q_7 for the database) as

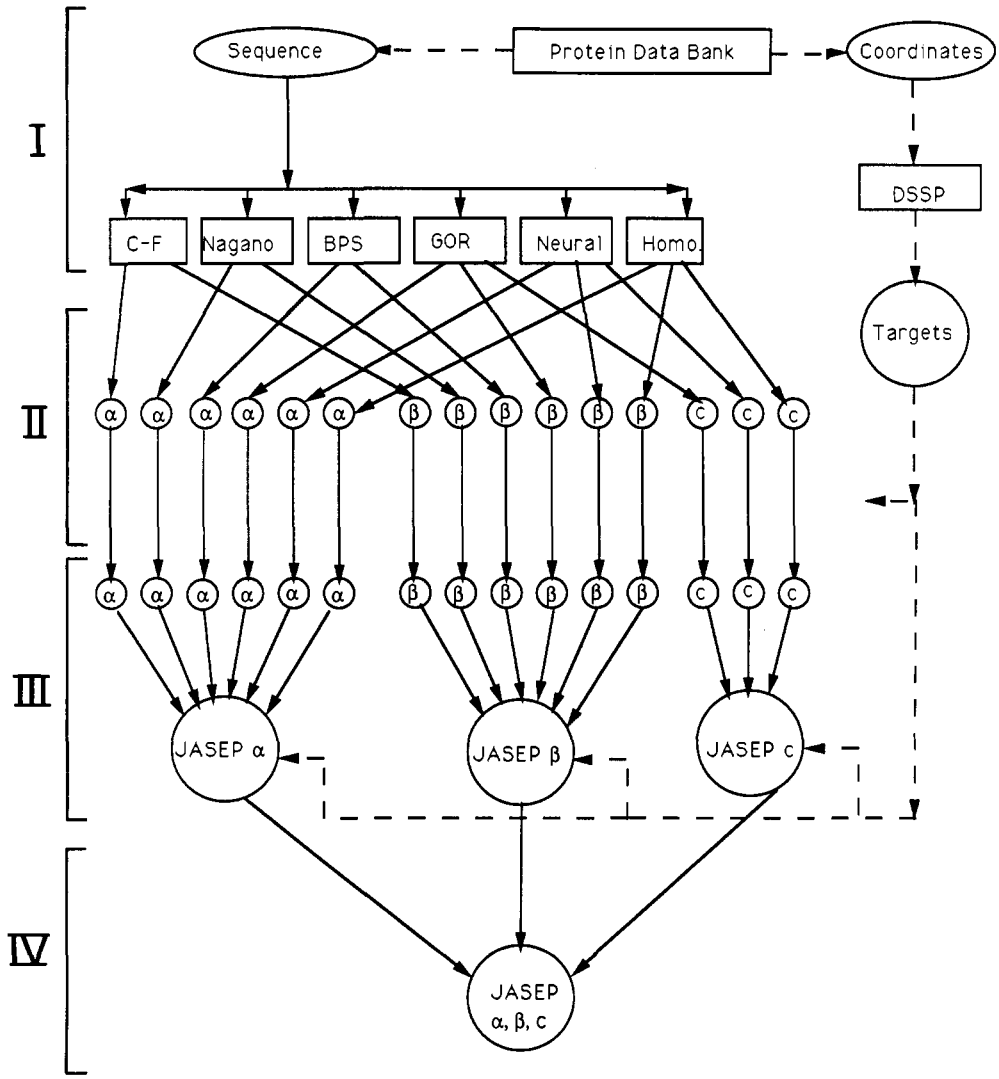

FIGURE 1: Schematic view of the Q_7-JASEP algorithm showing the pattern of information flow. The input sequence is obtained either from the PDB (the Brookhaven Protein Data Bank) or from the user. In brief, the stages of the algorithm are as follows. (I) RUN-PRED: Predictions are performed using six different methods (see text for details): C–F (Chou and Fasman), Nagano, BPS (Burgess, Ponnuswamy, and Scheraga), GOR (an information theory approach of Garnier, Osgurthope, and Robson), Neural (a neural network approach by Qian and Seznowski), and Homo (a homology method by Nishikawa). From the results for each individual method, numerical prediction scores are obtained for helical (α), extended (β), and coil (c) states (as applicable). For training purposes, targets are derived from the DSSP algorithm as described under Methods. (II) THRESHER: These targets are then used to determine the optimal threshold (see the text and Figure 2) for each structure–method pair. (III) JASEP-INIT: Quantitative prediction scores for each structure–method pair are then combined with appropriate weights to obtain Q_7-JASEP-α, Q_7-JASEP-β, and Q_7-JASEP-coil predictions (two-state predictions). In the training mode, the targets are then used again to optimize the threshold for each combination of scores. (IV) JASEP-FIN: In the final step, the state with the highest Q_7-JASEP potential in the sequence is picked for each residue position. Steps indicated with dashed lines do not apply when making predictions on a new, or "test", protein.

a function of threshold value, $t_j(k)$.

(*b*). $P_{i,j}(k)$'s at or above the threshold $T_j(k)$ are then scaled between 0 and 1, and those at or below $T_j(k)$ are scaled between 0 and −1. A scaled potential above zero indicates positive prediction for structure type j.

(*III*) JASEP-INIT Let $P_{i,\alpha}(k)$, $P_{i,\beta}(k)$ and $P_{i,c}(k)$ represent the *scaled* potentials for α, β, and coil structures, respectively, at the residue position i for method (k). For each method, three weights $\omega_\alpha(k)$, $\omega_\beta(k)$, and $\omega_c(k)$ are associated with the three secondary structure types. For methods that specify only α and β, the weights for "coil" are set to zero. If n methods are used in the prediction of a given type of secondary structure j, then the joint potential $P_{i,j}$ at position i of the residue sequence is given by

$$P_{i,j} = \sum_{k=1}^{n} \omega_j(k) P_{i,j}(k) \qquad (1)$$

The weight $\omega_j(k)$ for each structure–method pair (i,k) in eq 1 was set equal to $Q_{7,j}$ for method k. At this stage, the joint potentials are further range scaled by picking the zero point at an appropriate threshold using the same procedure (THRESHER) as described above for individual predictors. The target residue structure assignments derived from DSSP for the training protein set are used in optimizing the thresholds and estimating the weights. Parameters in the individual methods are based on data sets that include all 4 classes of proteins, as assembled by the respective authors. The sequence of operations shown with broken lines in Figure 1 is omitted when making predictions on a "test" database (or a new protein sequence).

(*IV*) JASEP-FIN Final predictions of Q_7-JASEP are obtained by comparing the three Q_7-JASEP potentials for α-helix, β-strand, and coil structures and assigning the state with the highest combined potential at each position along the sequence. In the present version of the algorithm, no smoothing is performed, and intrasegment cooperativity effects are not included.

Table I: Protein Database Used in the Present Study

PDB code	protein	ref
2AAT	aspartate aminotransferase	Smith et al. (1986)
3LZM	lysozyme	Weaver et al. (1989)
5CPA	carboxypeptidase A	Rees et al. (1983)
3APR	acid protease	Suguna et al. (1987)
1FX1	flavodoxin	Watenpaugh et al. (1973)
1HIP	high potential iron protein	Carter et al. (1974)
3GRS	glutathione reductase	Karplus and Schulz (1987)
7TLN	thermolysin	Holmes et al. (1983)
1PHH	*p*-hydroxybenzoate hydrolase	Schreuder et al. (1988)
2GCR	γ-crystallin	White et al. (1989)
2LDX	apolactate dehydrogenase	Hogrefe et al. (1987)
6PTI	pancreatic trypsin inhibitor	Wlodower et al. (1987)
7ATC	carbamoyltransferase	Kim et al. (1987)
1CYC	ferrocytochrome	Tanaka et al. (1975)
8CAT	catalase	Fita et al. (1986)
1ACX	actinoxanthin	Pletnev et al. (1982)
4MDH	malate dehydrogenase	Birktoft et al. (1987)
3CNA	concanavalin	Hardman and Ainsworth (1972)
3PGK	phosphoglycerate kinase	Bryant et al. (1974)
2B5C	cytochrome b_5	Matthews et al. (1972)
1RHD	rhodanase	Borkakoti et al. (1982)
3RP2	rat mast cell protein	Remington et al. (1988)
3PGM	phosphoglycerate mutase	Winn et al. (1982)[a]
2ALP	α-lytic protein	Fujinaga et al. (1985)
4FXN	flavodoxin (semiquinone)	Smith et al. (1977)
4APE	endothiapepsin	Pearl and Blundell (1984)
1TIM	triose phosphate isomerase	Banner et al. (1976)
1HMQ	haemerythrin	Stenkamp et al. (1984)
1ABP	L-arabinose binding protein	Gilliland and Quiocho (1981)
1TGS	trypsinogen	Bolognesi (1982)
8LDH	lactate dehydrogenase	Abad et al. (1987)
3EST	elastase	Meyer et al. (1988)
8DFR	dihydrofolate reductase	Mathews et al. (1985)
2SNS	staphylococcal nuclease	Cotton et al. (1979)
3ADK	adenylate kinase	Dreusicke et al. (1988)
2CCY	cytochrome *c*	Finzel et al. (1985)
8ADH	apo-liver alcohol dehydrogenase	Colonna et al. (1986)
5MBN	myoglobin	Takano (1984)
4GPD	D-gdp-dehydrogenase	Murthy et al. (1980)
3ICB	Ca-binding protein	Szebenyi and Moffat (1986)
3BP2	phospholipase A2	Dijkstra et al. (1984)
7PCY	plastocyanin	Collyer (1990)
9PAP	papain	Kamphuis et al. (1984)
1CCR	rice cytochrome	Ochi et al. (1983)
2SOD	superoxide dismutase	Tainer et al. (1982)

[a] Unpublished data from the Brookhaven Protein Data Bank.

(C) Methods Used in Q_7-JASEP

We focused on a set of generically different and well-known methods implemented on a VAX 8350 computer [Kanehisa, 1987; Viswanadhan et al. 1990b): (i) a neural network model (Qian & Sejnowski, 1988); (ii) a homology method (Nishikawa & Ooi, 1986); (iii) the GOR information theory approach (Garnier et al., 1978; Gibrat et al., 1987); (iv) the Chou–Fasman method (Chou & Fasman, 1978); (v) the Burgess–Ponnuswamy–Scheraga (BPS) method (Burgess et al., 1974); and (vi) the Nagano method (Nagano, 1973, 1974). We briefly highlight the main points of each approach here. The original references should be consulted for additional information.

(*a*) *Neural Network Model.* A neural network predictive scheme consists of a set of nonlinear processing units connected with a specific topology, input encoding and output decoding functions, and the set of weights and biases produced by network training. The model we use is that of Rumelhart et al. (1986), which uses no hidden layers. Parameters are taken from Tables 13–16 of Qian and Sejnowski (1988). The input is a window of 13 amino acids from a given sequence centered at a given residue position. Outputs are the predictive scores for each structural state (α-helix, β-strand, or coil conformations).

(*b*) *Information Theory.* The GOR information theory method (Garnier et al., 1978; Gibrat et al., 1987) is a Bayesian approach for secondary structure prediction. Information carried by an event y about event x is given by

$$I(x;y) = \log [p(x/y)/p(x)] \quad (2)$$

where $p(x/y)$ is the conditional probability of x given y, and $p(x)$ is the a priori probability of x. The information carried by the local sequence R_{j-8}, ..., R_{j+8} (event y) about the secondary structure in the set (helix, turn, strand, coil) assumed by residue R_j (event x) is determined, and the residue is assigned the secondary structure with the highest information value (after an estimated "decision constant" is added to each prediction score). For present purposes, however, we take into account the information measure of each type of structure.

(*c*) *Homology (Similarity) Modeling.* Nishikawa and Ooi (1986) assembled a reference database of proteins with known crystal structures and used it to develop a method for assigning secondary structure based on sequence similarity. For each residue position in the test protein represented by a window of 11 residues centered at the position of interest, they computed "homology scores" with respect to all 11-residue peptides in the reference protein database. These scores were ranked, summed up, and weighted for each structure type. This was followed by the smoothing of these scores to eliminate short helices and β-strands before assigning structural categories. The homology score for each pair of peptides was taken as the sum of correlation coefficients for six physical–chemical properties of the amino acid residues in the two peptides. It is worth noting that even a high homology score of this kind does not necessarily represent any evolutionary kinship or common ancestry for the two peptides in question.

(*d*) *Chou–Fasman Method.* In this study, we use the parameters P_α and P_β, expressing the potential to form α-helix and β-strand, respectively, as defined by Chou and Fasman (1978) and deduced from a set of 39 globular proteins of known structure. We did not employ the complete prediction protocol as specified by Chou and Fasman (1978) but instead considered separately the prediction score for each type of structure at each residue position, as described earlier.

(*e*) *Nagano Method.* Nagano's method (Nagano 1973, 1974) employs a linear combination of weights representing measures of statistical constraint at different residue separations (up to six-residue spacings) along the sequence to assign local structure on the basis of short-range interactions. The parameter values are based on a survey of globular protein crystal structures. In this method, a high negative score represents a position prediction. We have reversed the sign of all numerical prediction scores for the sake of uniformity with other methods.

(*f*) *Burgess–Ponnuswamy–Scheraga Method.* Empirical rules that assign each residue in a sequential nonapeptide to one of the four states of the set (α-helical, extended, turn, coil) are derived in a probabilistic formulation of the prediction problem (Burgess et al., 1974). Statistical parameters are gathered from a survey of protein crystal structures, that is, from the distribution of each residue type among regions of

Table II: Comparison of Q_7-JASEP with Its Component Methods, for Two-State Prediction Models Using a Cross-Validation Scheme

	$Q_{7,\alpha}$		$Q_{7,\beta}$		$Q_{7,c}$	
method	train	test	train	test	train	test
homology	0.424	0.408	0.395	0.390	0.393	0.388
GOR III	0.402	0.383	0.422	0.393	0.416	0.410
neural networks	0.343	0.335	0.378	0.370	0.382	0.374
BPS	0.285	0.272	0.113	0.111		
Chou–Fasman	0.227	0.216	0.196	0.183		
Nagano	0.346	0.320	0.235	0.230		
Q_7-JASEP (phase III)	0.448	0.427	0.436	0.414	0.425	0.417

(ϕ–φ)-space characteristic of different secondary structures. Here, we combine the turn and coil states into a single state (coil) and consider the prediction scores for each type.

(D) Assessment of Predictive Information

Seven different quality indices have been proposed (Schulz & Schirmer, 1979) for assesment of prediction quality. The most commonly used, Q_3, is simply the percentage of correct predictions. For a two-state prediction model (e.g., helix–coil prediction) this index can be quite misleading. High Q_3 may result from a trivial prediction such as "all coil" (see Results). We have, therefore, focused on the correlation coefficient (Matthews, 1975) $Q_{7,i}$ for each two-state prediction model (i or non-i):

$$Q_{7,i} = \frac{(p_i n_i) - (u_i o_i)}{[(n_i + u_i)(n_i + o_i)(p_i + u_i)(p_i + o_i)]^{1/2}} \quad (3)$$

Here, i is the secondary structure state in question, p_i the number of residues correctly predicted to belong to type i, and n_i the number of residues correctly predicted *not* to belong to type i; u_i and o_i refer to the number of residues underpredicted and overpredicted, respectively.

(E) Cross-Validation of Q_7-JASEP

For training and testing of the algorithm, we divide the set of 45 proteins shown in Table I into training subsets of 35 proteins and corresponding test subsets of 10 proteins. Each training subset is chosen to minimize homologies with any protein in the corresponding test subset. Nine such distinct training–testing experiments are conducted, so that each protein is represented twice in these nine test subsets. Thus, the total number of test results (9 × 10) equals twice the size of the original data set (2 × 45). In each training–testing experiment, the optimal threshold, $T_j(k)$, and the correlation coefficient (eq 3) are evaluated for each structure–method pair (i,k) from the training set. The weight for each structure–method pair (eq 1) is set equal to the corresponding correlation coefficient obtained from each of the training sets. Then this weight and threshold are used to perform predictions on corresponding test subsets. Independent of the above cross-validation, all 45 proteins were used for training and the corresponding correlations on the data set are evaluated for comparison.

Results and Discussion

As described under Methods, scores for α, β, and coil structures were obtained from the six individual approaches for each of the 45 proteins shown in Table I. For five of the six methods, prediction scores were then rescaled between −1 and +1. An exception to this precedure was the homology method. Since smoothing and indirect string comparison (by correlation of physical–chemical properties) are essential components of this algorithm, it was not possible to extract a potential for each residue position as we could for other methods. Hence, we assigned a score of +1 if a given residue position as "predicted" in a two-state model and −1 otherwise.

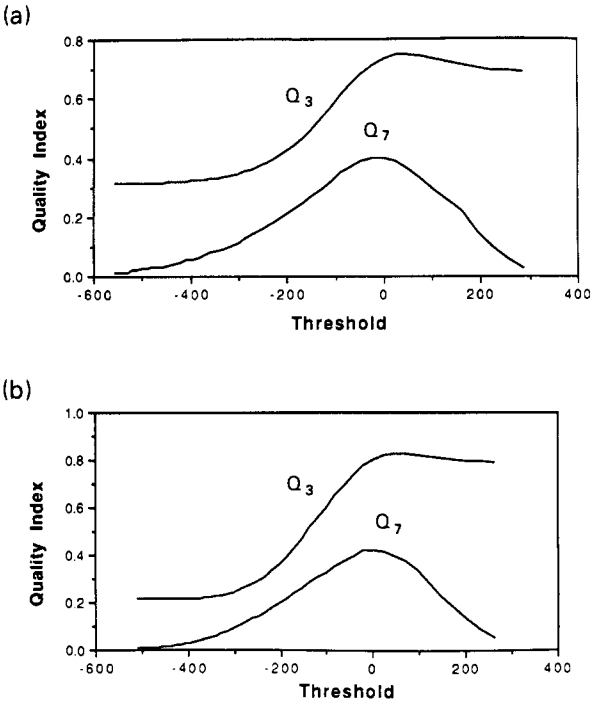

FIGURE 2: Variation of Q_3 and Q_7 as a function of threshold for the two-state predictions using the entire data set of Table I. Panels a and b show the variation of a quality index with threshold for α and β structure predictions using the GOR III method.

Although seven different types of quality index are available (Schulz & Schirmer, 1979), only two, viz., Q_3 and Q_7, are used with any frequency. The correlation coefficient (Q_7) is of general applicability and widespread use in protein structural theory [e.g., Viswanadhan (1987)], binding site modeling [e.g., Viswanadhan et al. (1990a)], etc., but it has been used less often than Q_3 for secondary structure prediction. Figure 2a,b shows the variation of Q_3 and Q_7 as a function of the threshold applied to convert the numerical prediction scores to secondary structure assignments. Results displayed here are those for the GOR III method, based on the protein database shown in Table I. Data are not shown for other methods, but the plots are essentially similar. The most important feature in both panels is the presence of one clear maximum for the Q_7 profiles, indicating an optimal threshold. The Q_3 profiles, to the contrary, reach a maximum at a certain threshold beyond which there is practically no change. It is easily seen that the end points of Q_3 profiles sum to unity and correspond to fractions in each of the two states. In most cases, Q_3 can be tuned to an almost optimal value simply by assigning all residues to the state with the greater frequency. This behavior may yield deceptively high values, even for a trivial prediction such as "all coil". The threshold values used in the present study for making predictions have been taken from the profiles for Q_7 for each method used here.

Table II shows the quality indices (Q_7) of different methods for the three secondary structural states. The "training set"

results in that table are derived from the entire database of proteins shown in Table I. A cross-validation study of Q_7-JASEP (and its component methods) was carried out as described under Methods (see section E under Methods). In this cross-validation, each protein was represented twice in the test sets, and the average from the two test results for each protein is reported ("test" scores in Table II). The quality indices we obtained for these methods are different from the original authors' because (i) we used DSSP assignments, whereas the original authors used other criteria for classification, (ii) we considered each set of scores (α, β, or coil) separately, and (iii) we optimized each individual threshold using the Q_7 index. Line 7 shows the results at the end of the third phase of Q_7-JASEP. These entries are the most appropriate to compare with values for the individual methods. It is clear that these scores are better than those obtained from any of the individual component methods. Given the differences in databases and criteria used to discern secondary structures, it is obviously difficult to make direct comparisons among methods. However, the purpose of these comparisons is to point out the value of using this combination strategy. It will be unfair to rate any given method on the basis of these scores alone unless the complete prediction protocol specified in each method was used. Such an exercise is obviously beyond the scope of the present work.

In Table III, the average test scores $Q_{7,\alpha}$, $Q_{7,\beta}$, and $Q_{7,c}$ are reported for each protein. The improvement obtained by the joint approach (average test scores per protein: $Q_{7,\alpha}$, = 0.41, $Q_{7,\beta}$ = 0.47, and $Q_{7,c}$ = 0.41) over the individual methods is shared by all types of structure. The average Q_7 for all three structures (0.43) is just equal to the best result reported in the literature from another joint prediction approach (Biou et al., 1988; Garnier & Levin, 1991). However, the latter method has steps in the algorithm to account for cooperativity effects. The scores obtained with Q_7-JASEP will presumably be enhanced when such correlative corrections are added to it. On average, Q_7-JASEP yielded 64% (Q_3) correct predictions, even though the thresholds were optimized for Q_7, not Q_3. This was not an improvement over the best of existing methods; however, this figure is really the consequence of a trade-off between Q_3 and Q_7 when selecting an appropriate threshold for each set of predictive scores (see the discussion relating to Figure 2 above). As an example of the application of Q_7-JASEP, Figure 3a–c shows profile maps of the final predictive scores for helix, β-strand, and coil conformations in *p*-hydroxybenzoate hydrolase. Also shown in the middle of each profile map are the DSSP secondary structure assignments from crystallographic coordinates.

It should be noted that a jackknife analysis (leaving one protein out and making predictions on that protein, and repeating the procedure for all proteins in the data set) can sometimes lead to misleading results. For example, if the data set has some homologous groups, a jackknife method of testing can lead to higher scores for a test protein when one or more homologous proteins are present in the training data set. Hence, we kept the training and test sets as independent as possible by minimizing the homology between the two.

Concatenation of independent predictions for α, β, and coil structures in the final phase (phase IV) of Q_7-JASEP *lowers* the prediction quality index (Q_7) because of the additional constraint that the sum of predicted α, β, and coil residues must be equal to the total number of residues in each protein. We have not obtained concatenated results for each method independently, but only for Q_7-JASEP for each test described above. After concatenation, we calculated two types of averages, the arithmetic mean of scores for each protein in the data set (average 1) and the global average (average 2) computed from the quantities n_i, o_i, p_i, and u_i (eq 3) for the entire data set. In that equation, for any single protein, if the denominator (or the numerator and the denominator) goes to zero, the Q_7 index is undefined: in such cases, we assigned the maximum value of 1.0 for Q_7-JASEP as well as for the individual methods. This assignment has no influence on average 2 (global average). Results obtained before and after concatenation (testing and training averages) are shown in Table IV. It is worth mentioning that the improvements in Q_7 seen with Q_7-JASEP are shared by all classes of proteins. That is, this method does not favor a particular protein class. If the class of the protein is known with 100% accuracy, one might obtain better scores by using different parameter sets (Kneller et al., 1990); however, uncertainity in class predictions limits the utility of such an approach, and this may cause erroneous interpretation of the secondary structure for a protein sequence assigned to the wrong class.

With the appearance of a large number of prediction methods, it seems natural to try combining them to produce more accurate and robust predictions. The present joint al-

Table III: Results of Cross-Validation Analysis for the Set of 45 Proteins of Table I[a]

PDB code	Q_7,α	Q_7,β	Q_7,c
2AAT	0.44	0.42	0.42
3LZM	0.47	0.81	0.45
5CPA	0.46	0.45	0.52
3APR	0.29	0.38	0.29
1FX1	0.55	0.37	0.31
1HIP	0.25	0.12	0.23
3GRS	0.39	0.33	0.33
7TLN	0.66	0.27	0.43
1PHH	0.40	0.41	0.44
2GCR	−0.08	0.44	0.25
2LDX	0.31	0.34	0.31
6PTI	0.29	0.60	0.67
7ATC	0.23	0.22	0.29
1CYC	0.73	1.00	0.55
8CAT	0.45	0.52	0.36
1ACX	1.00	0.42	0.41
4MDH	0.31	0.42	0.33
3CNA	0.64	0.38	0.33
3PGK	0.49	0.43	0.47
2B5C	−0.12	0.24	0.17
1RHD	0.45	0.54	0.49
3RP2	0.24	0.04	0.14
3PGM	0.51	0.28	0.40
2ALP	0.27	0.29	0.37
4FXN	0.28	0.60	0.33
4APE	0.42	0.43	0.38
1TIM	0.56	0.65	0.40
1HMQ	0.55	1.00	0.59
1ABP	0.27	0.28	0.22
1TGS	0.39	0.44	0.49
8LDH	0.35	0.29	0.33
3EST	0.40	0.44	0.49
8DFR	0.17	0.27	0.40
2SNS	0.62	0.46	0.43
3ADK	0.37	0.58	0.31
2CCY	0.55	1.00	0.55
8ADH	0.34	0.36	0.47
5MBN	0.44	1.00	0.44
4GPD	0.49	0.49	0.48
3ICB	0.84	1.00	0.77
3BP2	0.29	0.06	0.15
7PCY	0.44	0.36	0.43
9PAP	0.46	0.25	0.61
1CCR	0.60	1.00	0.51
2SOD	−0.05	0.50	0.53

[a] Each protein (identified by its PDB code) was tested twice, and the average value of Q_7 is shown for each type of secondary structure. See Table IV for overall average values.

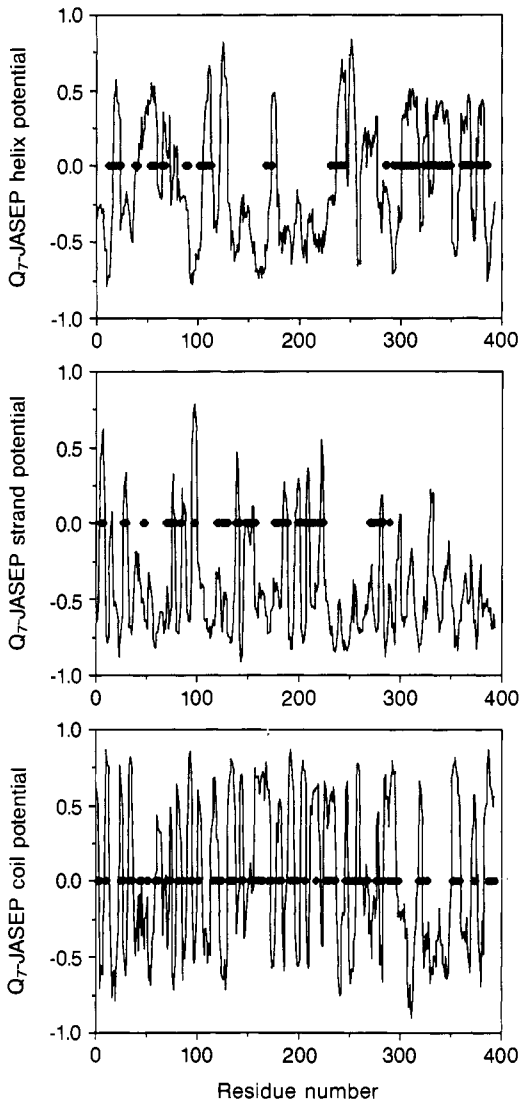

FIGURE 3: Predictions of Q_7-JASEP for *p*-hydroxybenzoate hydrolase. (a, top) Prediction score profile for the α-helix structure. DSSP assignments for α-helix structure (i.e., target structures obtained from crystallographic coordinates) are shown in the middle of the profile map (i.e., at potential = 0) as diamonds. Residues with Q_7-JASEP α-helix potentials greater than zero are assigned the α-helix structure; those below zero are assigned to the non-α-helix category. No smoothing of the edges of the predicted structures was done. (b, middle) Analogous plot for the β-strand/non-β-strand categorization; (c, bottom) Analogous plot for the coil/noncoil categorization.

gorithm is an attempt in that direction. In formulating Q_7-JASEP, we chose to optimize the Q_7 index, which has well-defined optimum values for the individual methods. We used the DSSP assignments as targets because they are objective. Some other widely used methods have been based on crystallographers' assignments, which are often subjective. Refinement of the individual methods by optimization of their parameter values using the DSSP targets may improve their individual, as well as collective, performance. Inclusion of cooperativity effects (intrasegment correlations) (Jernigan & Szu, 1979) and smoothing of predictions are expected to enhance the predictive quality of the scores further. We have determined that the weights used in Q_7-JASEP for individual methods could be optimized using such techniques and that better prediction scores would result. Our principal purpose here was to obtain a direct comparison between Q_7-JASEP and its component methods, rather than to optimize the absolute scores by including cooperativity effects and weight optimization. Such extensions of the method and applications to new

Table IV: Summary of Results of the Application of Q_7-JASEP to 45 Proteins

expt	$Q_{7,\alpha}$	$Q_{7,\beta}$	$Q_{7,c}$
training			
phase III	0.45	0.44	0.43
phase IV			
(a) av 1[a]	0.40	0.48	0.42
(b) av 2[b]	0.45	0.41	0.41
testing			
phase III	0.43	0.41	0.42
phase IV			
(a) av 1[a]	0.41	0.47	0.41
(b) av 2[b]	0.43	0.40	0.40

[a] Av 1 is the average over n_p proteins for each type of structure:

$$\frac{\sum_{i=1}^{n_p} Q_{7,i}}{n_p}$$

[b] Av 2 is calculated by (i) applying eq 3 to each test set (using n_i, etc., for the total number of correct predictions, etc., in that set) and (ii) calculating the weighted average for all test sets, the weight being the total number of residues in that test set. This calculation gives the average value weighted by the number of residues in each protein.

proteins will be discussed elsewhere (manuscript in preparation). We have also constructed artificial neural networks as a complimentary approach to weight optimization and nonlinear combination of different methods. This approach will be discussed elsewhere (manuscript in preparation).

The present approach includes six well-known methods, each of which is general in its scope; hence Q_7-JASEP is also general in its scope. That is, parametrization for any of the component methods does not depend on structural particulars of any one class of proteins. Combination of specific models for one class (Cohen et al., 1983; Kneller et al., 1990) would improve the predictions further when predictions are restricted to that class. However, that would limit the applicability of the algorithm to one class and also would require prior knowledge of the class into which a test protein falls.

It should be obvious that, in principle, the joint algorithm described here can be applied to include the prediction of features other than secondary structure. It can be applied to any instance in which there exist multiple methods for assigning protein residues (or nucleic acid bases) in a data set to categories. Examples to be considered include the identification of promoter sites, transmembrane domains, and antigenic sites (B- and T-cell).

CONCLUSION

In this paper, we present a new joint algorithm for predicting protein secondary structure. Unlike previous joint approaches, this algorithm (i) directly uses numerical scores from the individual component methods and (ii) weights the predictions of various methods on the basis of information from the database. More generally, the basic algorithm can be applied to any type of structure/sequence or function/sequence correlation for which there are multiple predictive methods.

ACKNOWLEDGMENTS

We are grateful to Dr. H. R. Guy and Dr. R. L. Jernigan for their insightful comments on this work. Parts of computation were carried out using CRAY XMP and VAX 8350 computers at The Advanced Scientific Computation Laboratory (ASCL), Frederick, MD. We thank the ASCL for facilities and staff support.

Registry No. Aspartate aminotransferase, 9000-97-9; lysozyme, 9001-63-2; carboxypeptidase A, 11075-17-5; acid protease, 9074-09-3;

glutathione reductase, 9001-48-3; thermolysin, 9073-78-3; *p*-hydroxybenzoate hydrolase, 9059-23-8; apolactate dehydrogenase, 9001-60-9; pancreatic trypsin inhibitor, 9087-70-1; carbamoyltransferase, 66304-42-5; catalase, 9001-05-2; actinoxathin, 59680-34-1; malate dehydrogenase, 9001-64-3; concanavalin, 11028-71-0; phosphoglycerate kinase, 9001-83-6; cytochrome b_5, 9035-39-6; rhodanase, 9026-04-4; phosphoglycerate mutase, 9032-62-6; endothiapepsin, 37205-60-0; triosephosphate isomerase, 9023-78-3; trypsinogen, 9002-08-8; lactate dehydrogenase, 9001-60-9; elastase, 9004-06-2; dihydrofolate reductase, 9002-03-3; staphylococcal nuclease, 9013-53-0; adenylate kinase, 9013-02-9; cytochrome *c*, 9007-43-6; alcohol dehydrogenase, 9031-72-5; D-*GDP*-dehydrogenase, 9001-50-7; phospholipase A2, 9001-84-7; papain, 9001-73-4; superoxide dismutase, 9054-89-1.

References

Abad-Zapatero, C., Griffith, J. P., Sussman, J. L., & Rossmann, M. G. (1987) *J. Mol. Biol. 198*, 445–467.

Argos, P., & Levine, M. (1972) *Cold Spring Harbor Symp. Quant. Biol. 36*, 387.

Argos, P., Schwarz, J., & Schwarz, J. (1976) *Biochim. Biophys. Acta 439*, 261–273.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., & Wilson, I. A. (1976) *Biochem. Biophys. Res. Commun. 72*, 146–155.

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

Biou, V., Gibrat, J. F., Levin, J. M., Robson, B., & Garnier, J. (1988) *Protein Eng. 2*, 185–191.

Birktoft, J. J., Bradshaw, R. A., & Banaszak, L. J. (1987) *Biochemistry 26*, 2722–2734.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., & Thornton, J. M. (1987) *Nature (London) 326*, 347–352.

Bohr, H., Bohr, J., Brunak, S., Cotteril, R. M. J., Lautrup, B., Norskov, L., Olsen, O. H., & Peterson, S. B. (1988) *FEBS Lett. 241*, 223–228.

Bolognesi, M., Gatti, G., Menegatti, E., Guarneri, M., Marquart, M., Papamokos, E., & Huber, R. (1982) *J. Mol. Biol. 162*, 839–868.

Borkakoti, N., Moss, D. S., & Palmer, R. A. (1982) *Acta Crystallogr. Sect. B 38*, 2210–2217.

Bourgeois, S., Jernigan, R. L., Szu, S. C., Kabat, E. A., & Wu, T. T. (1979) *Biopolymers 18*, 2625–2643.

Bryant, T. N., Watson, H. C., & Wendell, P. L. (1974) *Nature (London) 247*, 14–17.

Burgess, A. W., Ponnuswamy, P. K., & Scheraga, H. A. (1974) *Isr. J. Chem. 12*, 482–494.

Carter, C. W., Kraut, J. J., Freer, S. T., Xuong, N., Alden, R. A., & Bartsch, R. G. (1974) *J. Biol. Chem. 249*, 4212–4225.

Chou, P. Y., & Fasman, G. D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol. 47*, 45–148.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., & Fletterick, R. J. (1983) *Biochemistry 22*, 4894–4904.

Collyer, C. A., Guss, J. M., Sugimura, Y., Yoshizaki, F., & Freeman, H. C. (1990) *J. Mol. Biol. 211*, 617–632.

Colonna, F., Perahia, D., Karplus, M., Eklund, H., Branden, C. I., & Tapia, O. (1986) *J. Biol. Chem. 261*, 15273.

Cotton, F. A., Hazen, E. E., Jr., & Legg, M. J. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 2551–2555.

Covell, D. G., & Jernigan, R. L. (1990) *Biochemistry 29*, 3287–3294.

Crawford, I. P., Niermann, T., & Kirschner, K. (1987) *Proteins: Struct., Funct., Genet. 2*, 118–129.

Crippen, G. M., & Viswanadhan, V. N. (1985) *Int. J. Pept. Protein Res. 25*, 487–509.

Dijkstra, B. W., Kalk, K. H., Drenth, J., de Haas, G. H., Egmond, M. R., & Slotboom, A. J. (1984) *Biochemistry 23*, 2759–2766.

Dreusicke, D., Karplus, P. A., & Schulz, G. E. (1988) *J. Mol. Biol. 199*, 359–371.

Finzel, B. C., Weber, P. C., Hardman, K. D., & Salamme, F. R. (1985) *J. Mol. Biol. 186*, 627–643.

Fita, I., Silva, A. M., Murthy, M. R. N., & Rossmann, M. G. (1986) *Acta Crystallogr. Sect. B. 42*, 497–515.

Fujinaga, M., Delbaere, L. T. J., Brayer, G. D., & James, M. N. G. (1985) *J. Mol. Biol. 184*, 479–502.

Garnier, J., & Levin, J. M. (1991) *Comput. Appl. Biosci. 7*, 133–142.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.

Gibrat, J. F., Garnier, J., & Robson, B. (1987) *J. Mol. Biol. 198*, 425–443.

Gibson, K. D., & Scheraga, H. A. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 5649–5633.

Gilliland, G. L., & Quiocho, F. A. (1981) *J. Mol. Biol. 146*, 341–362.

Gregoret, L. M., & Cohen, F. E. (1990) *J. Mol. Biol. 211*, 959–974.

Hardman, K. D., & Ainsworth, C. F. (1972) *Biochemistry 11*, 4910–4919.

Hogrefe, H. H., Griffith, J. P., Rossmann, M. G., & Goldberg, E. (1987) *J. Biol. Chem. 262*, 13155–13162.

Holley, L. H., & Karplus, M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 152–156.

Holmes, M. A., Tronrud, D. E., & Matthews, B. W. (1983) *Biochemistry 22*, 236–240.

Jernigan, R. L., & Szu, S. C. (1979) *Macromolecules 12*, 1156–1159.

Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.

Kamphuis, I. G., Kalk, K. H., Swarte, M. B. A., & Drenth, J. (1984) *J. Mol. Biol. 179*, 233–256.

Kanehisa, M. (1987) *IDEAS-91*, Distributed by Advanced Scientific Computing Laboratory, National Cancer Institute, Frederick, MD.

Karplus, M., & Weaver, D. L. (1976) *Nature (London) 260*, 404–406.

Karplus, M., & McCammon, J. A. (1983) *Annu. Rev. Biochem. 53*, 263–300.

Karplus, M., & Schulz, G. E. (1987) *J. Mol. Biol. 195*, 701–729.

Kim, K. H., Pan, Z., Honzatko, R. B., Ke, H., & Lipscomb, W. N. (1987) *J. Mol. Biol. 196*, 853–875.

Kneller, D. G., Cohen, F. E., & Langridge, R. (1990) *J. Mol. Biol. 214*, 171–182.

Lenstra, J. A. (1977) *Biochim. Biophys. Acta 491*, 333–338.

Levine, J. M., & Garnier, J. (1988) *Biochim. Biophys. Acta 955*, 283–295.

Levine, J. M., Robson, B., & Garnier, J. (1986) *FEBS Lett. 205*, 303–308.

Levitt, M., & Chothia, C. (1976) *Nature (London) 261*, 552–557.

Levitt, M., & Meirovitch, H. (1983) *J. Mol. Biol. 168*, 595–620.

Lijk, L. J., Kalk, K. H., Bradenbury, N. P., & Hol. W. G. J. (1983) *Biochemistry 22*, 2952–2957.

Lim, V. I. (1974a) *J. Mol. Biol. 88*, 857–872.

Lim, V. I. (1974b) *J. Mol. Biol. 88*, 873–894.

Matthews, B. W. (1975) *Biochim. Biophys. Acta 405*, 442–451.

Matthews, D. A., Bolin, J. T., Burridge, J. M., Filman, D. J., Volz, K. W., Kaufman, B. T., Beddel, C. R., Champness, J. N., Stammers, D. K., & Kraut, J. (1985) *J. Biol. Chem. 260*, 381.

Meyer, E. F., Radhakrishnan, R., & Epp, O. (1988) *Acta Crystallogr. Sect. 44*, 26–38.

Miyazawa, S., & Jernigan, R. L. (1985) *Macromolecules 18*, 534–552.

Mr'azek, J., & Kypr, J. (1988) *Comput. Appl. Biosci. 4*, 297–302.

Murthy, M. R. N., Garavito, R. M., Johnson, J. E., & Rossmann, M. G. (1980) *J. Mol. Biol. 138*, 859–872.

Nagano, K. (1973) *J. Mol. Biol. 75*, 401–421.

Nagano, K. (1974) *J. Mol. Biol. 84*, 337–372.

Nishikawa, K. (1983) *Biochim. Biophys. Acta* 285–299.

Nishikawa, K., & Ooi, T. (1986) *Biochim. Biophys. Acta 871*, 45–54.

Ochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, H., & Morita, Y. (1983) *J. Mol. Biol. 166*, 407–418.

Pearl, L., & Blundell, T. (1984) *FEBS Lett. 174*, 96–101.

Pletnev, V. Z., Kuzin, A. P., & Malinina, L. V. (1982) *Bioorg. Khim. 8*, 1637.

Qian, N., & Sejnowski, T. J. (1988) *J. Mol. Biol. 202*, 865–884.

Rees, D. C., Lewis, M., & Lipscomb, W. N. (1983) *J. Mol. Biol. 168*, 367–387.

Remington, S. J., Woodbury, R. G., Reynolds, R. A., Matthews, B. W., & Neurath, H. (1988) *Biochemistry 27*, 8097–8105.

Rooman, M. J., & Wodak, S. J. (1988) *Nature (London) 335*, 45–49.

Rose, G. D. (1978) *Nature (London) 272*, 586–590.

Rumelhart, D., Hinton, G., & Williams, R. (1986) *Nature (London) 323*, 533–536.

Schiffer, M., & Edmundson, A. B. (1967) *Biophys. J. 7*, 121–135.

Schreuder, H. A., Van der Laan, J. M., Hol, W. G. J., & Drenth, J. (1988) *J. Mol. Biol. 199*, 637.

Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B., & Nagano, K. (1974) *Nature (London) 250*, 140–142.

Seetharamulu, P., & Crippen, G. M. (1991) *J. Math. Chem. 6*, 91–110.

Skolnick, J., & Kolinski, A. (1990) *Science (N.Y.) 23*, 1121–1125.

Smith, D. L., Ringe, D., Finlayson, W. L., & Kirsch, J. F. (1986) *J. Mol. Biol. 191*, 301–302.

Smith, W. W., Burnett, R. M., Darling, G. D., & Ludwig, M. L. (1977) *J. Mol. Biol. 117*, 195–225.

Stemkamp, R. E., Sieker, L. C., & Jensen, L. H. (1984) *J. Am. Chem. Soc. 106*, 618–622.

Suguna, K., Padlan, E. A., Smith, C. W., Carlson, W. D., & Davies, D. R. (1987) *J. Mol. Biol. 196*, 877–900.

Szebenyi, D. M. E., & Moffat, K. (1986) *J. Biol. Chem. 261*, 8761–8777.

Tainer, J. A., Getzoff, E. D., Richardson, J. S., & Richardson, D. C. (1982) *J. Mol. Biol. 160*, 181–217.

Tanaka, N., Yamane, T., Tsuhihara, T., Ashida, T., & Kakudo, M. (1975) *J. Biochem. (Tokyo) 77*, 147–162.

Takano, T. (1984) in *Methods and Applications in Crystallographic Computing*, Oxford University Press, Oxford, U.K.

Taylor, W. R., & Thornton, J. M. (1983) *Nature (London) 301*, 540–542.

Viswanadhan, V. N. (1987) *Int. J. Biol. Macromol. 9*, 39–48.

Viswanadhan, V. N., Ghose, A. K., & Weinstein, J. N. (1990a) *Biochim. Biophys. Acta 1039*, 356–366.

Viswanadhan, V. N., Weinstein, J. N., & Elwood, P. C. (1990b) *J. Biomol. Struct. Dyn. 7*, 985–1001.

Wallace, B. A., Cascio, M., & Mielke, D. L. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 9423–9427.

Watenpaugh, K. D., Sieker, L. C., & Jensen, L. H. (1973) *Proc. Natl. Acad. Sci. U.S.A. 77*, 3857–3860.

Weaver, L. H., Gray, T. M., Gruetter, M. G., Anderson, D. E., Wozniak, J. A., Dahlquist, F. W., & Matthews, B. W. (1989) *Biochemistry 28*, 3793–3797.

White, H. E., Driessen, H. P. C., Slingby, C., Moss, D. S., Turnell, W. G., & Lindley, P. F. (1989) *J. Mol. Biol. 207*, 217–235.

Wlodower, A., Nachman, J., Gilliland, G. L., Gallagher, W., & Woodward, C. (1987) *J. Mol. Biol. 198*, 469–480.