Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 45 (2004) 525–546
polymer
www.elsevier.com/locate/polymer

On the use of secondary structure in protein structure prediction: a bioinformatic analysis

Armando D. Solis, S. Rackovsky*

Department of Biomathematical Sciences Box 1023, Mount Sinai Medical Center, One Gustave L. Levy Place, New York, NY 10029, USA

Received 10 May 2003; received in revised form 19 August 2003; accepted 19 August 2003

Abstract

The amount of structural information encoded in secondary structure can be measured by its ability to specify the correct peptide backbone conformation of protein chains. Using methodology derived from information theory, we generate optimized distributions of backbone phi–psi dihedral angle pairs given either correct or predicted three-state secondary structure. Entropy measurements on these distributions provide a means to determine the effect of secondary structure knowledge on identifying the actual 3D conformation of protein chains. We find that only a modest fraction of the total uncertainty in phi–psi conformation (from 14 to 38%, at 20–90° resolutions, respectively) is resolved even with perfect knowledge of secondary structure. We further show that prediction of secondary structures, because of an accuracy ceiling below 80%, degrades structural information substantially. If prediction accuracy is below 50%, virtually no advantage is gained from using the prediction. Moreover, even state-of-the-art prediction accuracy of 75% retains less than one-third of the structural information encoded in secondary structure. We demonstrate that the level of structural description affects the amount of information extracted. The effort to provide as much structural detail as possible, while faced with a limited structural data set, results in an optimum resolution in the vicinity of a 20°-partition of the (ϕ, ψ) plane. We show that structural information increases exponentially with prediction accuracy, revealing that even marginal gains in the performance of secondary structure prediction algorithms are important for the retention of structural information. We observe that different kinds of secondary structure prediction outputs (single-state prediction, single-state prediction with a confidence index, and three-state probability prediction) do not differ greatly in the amount of structural information they yield, so long as the methods formulated in this work to generate propensity distributions are applied appropriately. The optimal phi–psi probability distributions developed here may be useful in biasing searches in structure space. We discuss the sources of the degradation of information caused by errors in secondary structure prediction, and their consequences for the prediction of the 3D conformation of protein chains.

Keywords: Protein bioinformatics; Information theory; Secondary structure prediction

1. Introduction

The quality and resolution of structural features and patterns detected by statistical analysis of protein backbone chains is dependent on the descriptor used to specify structural data [1,2]. The most common backbone structural description, the assignment of each residue to one of the three types of secondary structure (2°), provides the simplest means to identify repeating backbone patterns [3]. However, since only three states (helix, extended, and coil) are used in this classification, potentially informative details of the local sequence dependence of backbone conformation are not efficiently recognized. Nonetheless, the 2° prediction problem has become a benchmark in the field of protein structure prediction, and has prompted the development of numerous prediction algorithms over the past two decades [4,5]. Higher resolution structural descriptors, such as the phi–psi dihedral angle description [1,6] and the C^{α} trace [2,7–9], are more successful in cataloguing nuances of the local sequence–structure relationship.

The goal of protein structure prediction is to assign the correct 3D conformation to a given amino acid sequence [10]. Because even complete knowledge of the secondary structure of a protein is not sufficient to identify its folded structure, 2° prediction schemes are only an intermediate step. Recent work has aimed to close the secondary–tertiary structure gap via homology modelling and other means [11–13]. As they assume the membership of *all* patterns of

* Corresponding author. Tel.: +1-212-241-5851; fax: +1-212-860-4630.
E-mail addresses: shelly@camelot.mssm.edu (S. Rackovsky), armando.solis@mssm.edu (A.D. Solis).

0032-3861/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.065

2° elements and fold motifs in the known database, such approaches may not always work. Complementary techniques are being developed to fold specified secondary structures via energy minimization and other computational procedures [14–17]. By holding helices and strands rigid in the course of the minimization, the problem is transformed into that of folding 2° elements into compact forms. This procedure is essentially the (non-trivial) 'coil' folding problem, which involves a search for the native structure of connective coil segments.

The difficulty increases substantially in cases when only *predicted* secondary structures are available. Efforts to increase the prediction accuracy in recent years have stalled at the 70–75% level [18–20]. The challenge in bridging the secondary–tertiary structure gap stems from a number of complications. Structurally homologous proteins show only 90% alignment in their 2° strings on average [21]. Moreover, some conformational irregularities are tolerated in the 'regular' helical [22] and sheet structures [23], and therefore their canonical forms are not always the most appropriate models. The 'coil' assignment (random coil or irregular) means only that the amino acid backbone does not occur within organized helical or sheet networks. (However, the actual backbone (ϕ, ψ) dihedral angle pair may still exist in helical or extended regions of the phi–psi space.) This classification provides little information on the actual backbone dihedral angles of coil residues. The most common output of prediction algorithms, a string of 2° states at each residue, glosses over the fact that, on average, about one in four assignments are wrong, and we do not know where in the chain they occur.

To address these difficulties, we have implemented an informatic analysis to explore the utility of 2° prediction schemes in the computational 3D structural solution of protein chains. The following questions have served as guides into this work:

(1) How much structural information is associated with knowledge of the correct secondary structure of a protein chain, and how much more is necessary to finally identify its native backbone conformation?
(2) How informative are limited accuracy outputs of 2° assignment algorithms?
(3) How can such defective outputs be exploited to assist in narrowing the search for the native conformation of the protein backbone?

Because of the significant level of uncertainty inherent in 2° prediction algorithms, it is important to sort out truly informative predictions from those that mislead.

To address the informatic relationship between secondary structure and the actual 3D backbone conformation of protein chains, we begin by formulating a procedure to generate optimal (ϕ, ψ) dihedral angle distributions for a specified 2° state. Using these distributions, we measure the amount of information actually resolved by full 2° knowledge. We then extend our procedures to situations where only predicted secondary structures (with limited accuracy) are known. Though some investigations have employed predicted secondary structure as 'biases' in simulations [24–28], no informatically optimized methods have been developed to preserve maximal structural information in these circumstances. We address this task here. We provide a method to evaluate the effectiveness of prediction algorithms of limited accuracy in extracting backbone structural information. Simultaneously, we generate distributions in (ϕ, ψ) space which salvage the information latent in predicted secondary structures.

Secondary structure prediction algorithms can be classified in two major categories with respect to the degree of detail they provide. Single-state prediction ($2°P$) algorithms give a single string composed of a best guess for the 2° state of each residue in the input sequence (e.g. [29–32]). Many of these algorithms include in the output a confidence index, a single value per prediction, which indicates the expected accuracy of the particular prediction. The other kind gives a more detailed output, a three-state probability prediction ($2°P3$) for each residue, embodied in a set of values which describe the probability of each of the three 2° types (e.g. [33–35]). Whereas $2°P$ algorithms output only a single letter (e.g. H), with or without a confidence index, $2°P3$ algorithms give a set of probabilities (e.g. 0.7 for H, 0.1 for E, and 0.2 for C) at each prediction site. In this work, we examine the informatic utility of these different outputs, and design methods to generate optimally informative (ϕ, ψ) structure propensities in each case.

2. Methodology

Our ultimate goal is to construct probability distributions which describe as accurately as possible the backbone structural propensities of a protein chain, given a specified, limited degree of knowledge. These distributions must extract the maximum information possible from the database of known structures [1]. They provide a means to measure informatic quantities of interest. If one takes a structural distribution as a measure of the relative likelihoods of the possible conformations, the width of the distribution should indicate the relative ease of locating the correct conformation using a prediction procedure.

A few words about notation are appropriate. We indicate the probability distribution of secondary structures as $P(\chi_{2°})$, with χ denoting the structural domain, and the subscript 2° referring to secondary structure. We shall restrict the 2° description to the three general states (H, E, and C), used in the standard DSSP assignment protocol [3].[1] Similarly, $P(\chi_{(\phi,\psi)})$ refers to the probability distribution of the phi–psi dihedral angle pair. We indicate probability

[1] The eight states defined by DSSP were collapsed into three major states via the following definitions: H = {G,H,I}: E = {E}; C = {B,C,S,T}.

distributions by (upper case) P and a specific probability by (lower case) p. We include a superscript, if necessary, to (ϕ, ψ), to indicate the level of resolution used in discretizing the structural domain. For instance, $(\phi, \psi)^{20^\circ}$ denotes an even partition of the phi–psi plane by an 18×18 grid system, which divides the space evenly into 324 bins of side 20°. In this work, we consider the following resolutions: 90°, 45°, 20°, 10°, and 5°. We use partitions which contain the axes of the (ϕ, ψ) plane (i.e. $\phi = 0$ and $\psi = 0$) as two of the specified boundaries.

2.1. Measuring residual structural entropy

The degree of uncertainty associated with the task of determining the correct backbone conformation of a protein chain can be measured by the discrete Shannon entropy of the universe of structures [36]

$$H(\chi) = -\sum_i p_i(\chi) \ln p_i(\chi) \tag{1}$$

where χ can represent any backbone structural descriptor, and i runs over all possible states of χ. This equation assumes that knowledge of the respective probabilities is complete; the effectiveness of the equation diminishes as approximations to the probabilities lose accuracy. The entropy unit is the 'nat' when the natural logarithm is used.

Larger partitions of the structural space (for instance, the three-state 2° scheme or various n-state phi–psi classifications, e.g. [37]) can be taken as intermediate, lower resolution steps toward identifying the actual native structure of the backbone. This is because more than one state in a fine partition can belong to a particular state in a coarse partition. For instance, a residue identified as being in one of the three 2° states can be in many possible (ϕ, ψ) states. We observe that the uncertainty in predicting the membership in a coarse partition cannot be greater than the uncertainty associated with a more detailed partition. For instance, $H(\chi_{2^\circ}) = 1.06$ nats, magnitudes smaller than $H(\chi_{(\phi,\psi)^{20^\circ}}) = 3.87$ nats. (Recall that entropy is a logarithmic measure.) These values formalize the fact that it is easier to find or 'guess' the correct secondary structure than the phi–psi conformation.

The desire to use finer structural discretization, in order to extract useful structural information, is mitigated by operational limits set by the size of the data set [2]. For frequencies to be statistically meaningful, one needs a substantial sampling of all regions of the structural space. In previous work, we developed computational methods to measure information gain as a function of both sequence and structural discretization, and incorporated them into an automatic information maximization procedure to generate optimum probability distributions [1]. These methods use as an underlying control a background distribution, which becomes more important as one uses increasingly fine partitions (for both sequence and structure), when the statistics become sparse. Here, we use the same principles to build approximations to probability distribution functions, and employ them to measure the associated entropies. The mechanics of the methodology will be elaborated in later sections, where concrete examples are discussed.

2.2. Structural propensities and residual entropies

We are interested in estimating the conditional probability distribution $P(\chi_k|\chi_K)$, which is the probability of a particular state in the small partition χ_k given membership in the large partition χ_K. For instance, we might wish to estimate the probability distribution of a residue in phi–psi space given its 2° class. From these conditional probabilities, the associated *residual entropies* $H(\chi_k|\chi_K)$ can be computed. We use the term 'residual' to emphasize the fact that these measure the amount of uncertainty remaining in the small partition χ_k after knowing the state in the large partition χ_K.

Specifically, we would like to estimate the conditional probability distribution $P(\chi_{(\phi,\psi)}|\chi_{2^\circ})$, and use it to calculate the residual entropy after the *correct* three-state secondary structure is known, or $H(\chi_{(\phi,\psi)}|\chi_{2^\circ})$. Clearly, $H(\chi_{(\phi,\psi)}|\chi_{2^\circ}) > 0$, because there is a distribution of (ϕ, ψ) values associated with any of the three standard 2° states.

We are also interested in incorporating sequence information, in the hope of further decreasing the residual entropy. We denote the corresponding entropy by $H(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{\text{seq}})$. The simplest sequence information, which we use here, comes from the identity of the amino acid at the site in question, or $Y_{\text{seq}} = Y_{\text{aa}}$. For example, the magnitude of the residual entropy $H(\chi_{(\phi,\psi)}|\chi_{2^\circ} = \text{H}, Y_{\text{aa}} = \text{G})$ is the uncertainty left to resolve in the phi–psi domain after knowing that the residue identity is glycine and its 2° state is helical.

Lastly, we develop a way to estimate the residual entropy if the 2° structure assignments provided are outputs of prediction algorithms of limited accuracy. We denote the corresponding entropy quantities of the two kinds of 2° prediction as $H(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{\text{aa}})$ and $H(\chi_{(\phi,\psi)}|\chi_{2^\circ P3}, Y_{\text{aa}})$, where the subscript $2^\circ P$ and $2^\circ P3$ indicate the single-state and three-state probability outputs, respectively.

Our objective is to build structural probability distribution functions which yield the minimum possible residual entropy [1]. The procedure to generate these optimal probability distributions is outlined in Section 2.3. We note that these probability distribution functions can be useful in a number of applications. In particular, structure prediction schemes can utilize the distributions to systematically bias the search for the native backbone conformation in 3D space (e.g. [38–40]).

2.3. Estimating residual entropy with known 2° assignment

Procedures to generate optimal probability distributions and calculate the associated residual entropies are described here. The data from which estimates for $p(\chi_{(\phi,\psi)_m}|\chi_{2^\circ})$ are made are illustrated in Fig. 1(A)–(C). These figures show

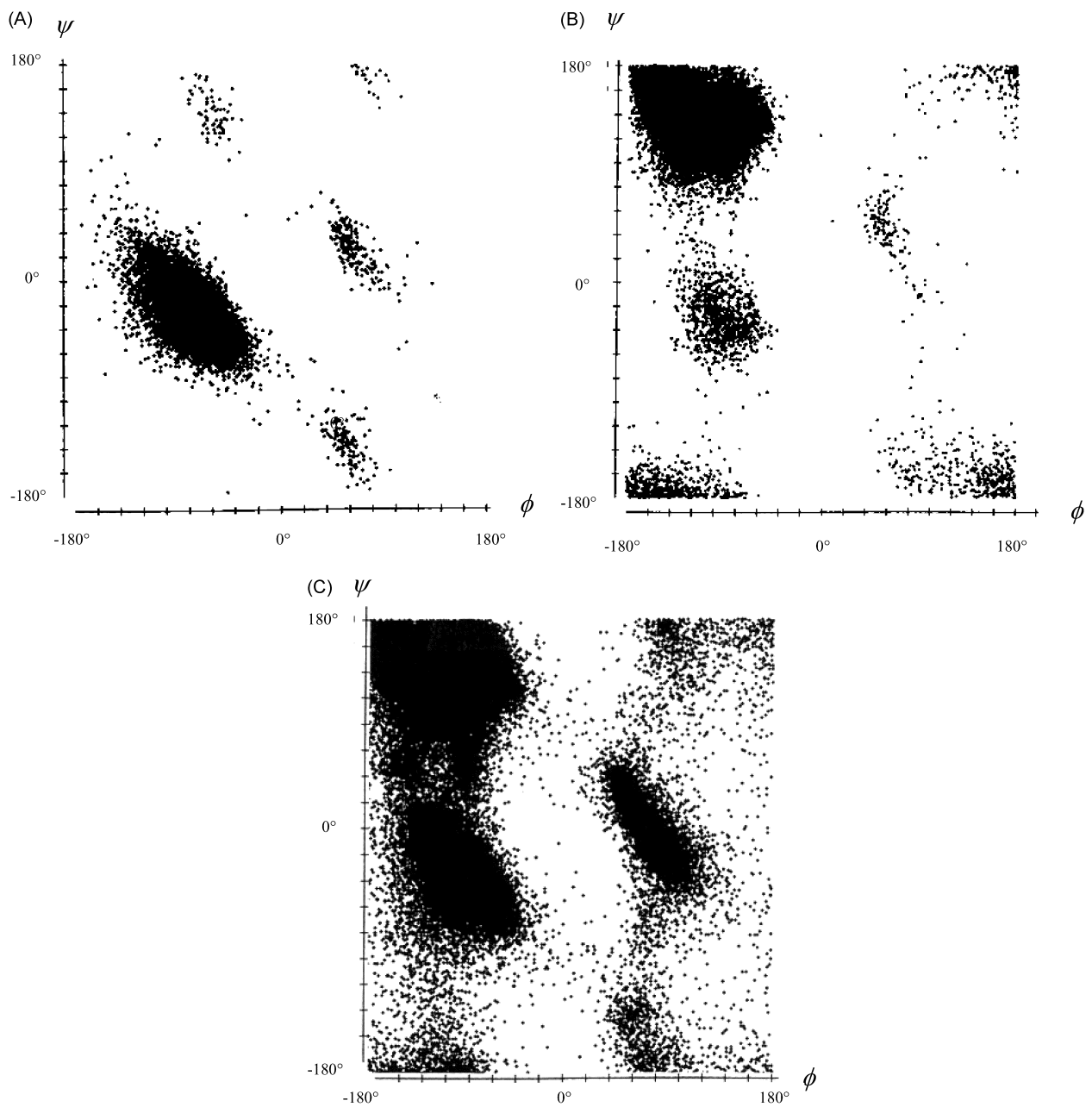


Fig. 1. The distribution of backbone structures from a representative set of protein chains. These illustrations represent the diversity in (ϕ, ψ) dihedral angle pairs of (A) helical, (B) extended, and (C) coil residues in the data set.

the distribution of backbone structures from a representative set of protein chains.[2] It is important to buffer raw frequencies against fluctuation effects brought about by sparseness of data [1]. We use a hybrid coefficient γ and a background distribution $B(\chi)$ as follows

$$p(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_i}) = [\gamma B(\chi_{(\phi,\psi)_m}) + n(\chi_{2^\circ_i})\,\pi(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_i})] / [\gamma + n(\chi_{2^\circ_i})] \tag{2a}$$

where the quantity $n(\chi_{2^\circ_i})$ denotes the total number of occurrences in the ith 2° state, π denotes a raw frequency, and $\pi(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_i})$ is given by

$$\pi(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_i}) = n(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_i}) / n(\chi_{2^\circ_i}) \tag{2b}$$

where $n(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_i})$ is the number of times a residue with

[2] Construction of the data set of protein structures used in this work is based on the algorithm for generating a representative set of protein chains by Hobohm and Sander [41]. We use the list containing proteins up to 25% pairwise sequence-homologous. We use 1045 protein chains, with a total of 207,834 residues. The list of protein chains can be obtained from http://www.mssm.edu/biomath/papers/proteindataset.html.

specified secondary structure i falls into the phi–psi structure bin m. Use of the background distribution compensates for the fact that some conformations (especially in fine structural partitions) do not occur in the data set used to estimate the probability distribution. If raw frequencies (i.e. Eq. (2b)) are used exclusively, bins or states that are not populated in the data set have probability estimates equal to zero. If such a probability distribution is employed to bias searches for the native structure, these states, some of which may actually occur, will never be accessed. The hybrid probability estimate reflects the reality that we have incomplete knowledge, arising from a limited data set.

The hybrid coefficient γ is an adjustable parameter designed to balance the contribution of raw frequencies and the background distribution to the final conditional probabilities. Our objective is to find the set of structural distributions with lowest entropy by searching over a range of γ. If there is a substantial number of data points per sequence and structure partition, the optimization procedure for γ will produce a hybrid distribution weighting the raw frequency counts more heavily than the background. In this situation, the well-populated raw frequency counts are taken as an adequate approximation of the underlying true probability distribution. On the other hand, if the number of unique states that subdivide the data is too large for the number of data points, the best hybrid coefficient will generate a distribution which favors the background over the raw frequency counts. As an extreme example, if one subdivides the phi–psi space into 360×360 bins (i.e. 1° resolution), the resulting number of unique states, 129,600, is of the same magnitude as the number of data points in our data set (207,834). Raw frequency distributions generated from such a partition will be sparse, making them misleading as representations of the true probability distributions. A natural solution is to buffer them with a background distribution.

The choice of background distribution is dictated by the conditional probability distribution function to be estimated and the variable being optimized. The background distribution must not involve the variable of interest. For instance, the background distribution we use to estimate the effect of 2° knowledge on backbone conformation, as embodied in $P(\chi_{(\phi,\psi)}|\chi_{2^\circ})$, is $P(\chi_{(\phi,\psi)})$. Of course, the latter has to be estimated as well. A large data set may be able to support moderate partitions of (ϕ,ψ); however, a background distribution must still be used to buffer the estimate for $P(\chi_{(\phi,\psi)})$ in extreme cases (i.e. small data set size or fine structural partitions). The most conservative background distribution possible is the uniform distribution, which assumes no prior knowledge of (ϕ,ψ) propensities. If the structural partition is chosen judiciously, with the size of the data set in mind, then the optimal γ will be found to have a low value.

To approximate the entropies $H(\chi_{(\phi,\psi)}|\chi_{2^\circ})$, we rewrite Eq. (1) as an expectation:

$$H(\chi_{(\phi,\psi)}|\chi_{2^\circ}) = E[-\ln P(\chi_{(\phi,\psi)}|\chi_{2^\circ})] \tag{3a}$$

With our non-redundant set of protein structures, we can estimate this expectation as

$$E[-\ln P(\chi_{(\phi,\psi)}|\chi_{2^\circ})] = -(1/n_{\mathrm{tot}}) \sum_{j}^{n_{\mathrm{tot}}} \ln p(\chi_{(\phi,\psi)_j}|\chi_{2^\circ}) \tag{3b}$$

where the summation is over all instances in the data set of size n_{tot}. The probability estimate on the right hand side of Eq. (3b) is computed from the observed frequencies, excluding the data point being considered from the computation. This is accomplished by subtracting 1 from the numerator and the denominator of Eq. (2a), or

$$p(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_j}) = [\gamma\, p(\chi_{(\phi,\psi)_m}) + n(\chi_{2^\circ_j})\, \pi(\chi_{(\phi,\psi)_m}|\chi_{2^\circ_j}) - 1] / [\gamma + n(\chi_{2^\circ_j}) - 1] \tag{4}$$

The effect of sequence information on the residual entropy is also of interest. We compute the residual entropy of the distribution from estimates of $P(\chi_{(\phi,\psi)_m}|\chi_{2^\circ}, Y_{\mathrm{aa}})$, where Y_{aa} is the specified amino acid. As a preliminary step, we only include single site sequence, as the residue of the site of interest contains the greatest amount of structural information. The background distribution used to estimate this probability is $P(\chi_{(\phi,\psi)_m}|Y_{\mathrm{aa}})$.

2.4. Conditional probability distributions for prediction schemes of limited accuracy

The methods we have developed thus far to generate the optimized structural probability distributions $P(\chi_{(\phi,\psi)_m}|\chi_{2^\circ})$ and $P(\chi_{(\phi,\psi)_m}|\chi_{2^\circ}, Y_{\mathrm{aa}})$ only apply when we have complete knowledge of the three-state 2° structure. In reality, the 2° assignment from prediction is only about 70–75% accurate, and the locations of mis-assigned residues are unknown.

It is not obvious how ab initio schemes can utilize this defective output to bias a computational search. We illustrate the problem with an example. If a residue in a protein chain predicted as H is actually helical, the probability distribution derived from Fig. 1(A) should suffice. However, if this residue is actually observed in a β-sheet with dihedral angles $(-90, 150^\circ)$, the same probability distribution will not be able to locate the native backbone conformation as easily, as this region has a low sampling probability. If the prediction algorithm is only 70% accurate, the misprediction of an E residue as an H residue is likely to occur more than once within a protein chain of average length.

Algorithms which output parameters to indicate prediction confidence are more useful. For instance, the PHD algorithm [34] outputs a 'reliability index,' a single number per prediction site, between 0 and 9, which apparently correlates to the accuracy in prediction. Though such information may be incorporated as user-controlled biases

in the conformational search, a systematic and automatic way to integrate this information with probability distributions is necessary.

Clearly, structural distributions developed from data which use correct 2° prediction will result in loss of information when applied to cases with 2° assignment errors. We, therefore, ask whether it is possible to develop structure distributions which are optimized to give the best possible predictions when error is known to be present. These distributions will be constructed, as were those above, from observed and background distributions. However, the introduction of error in 2° prediction will change the relative weighting of the two components. We seek to reoptimize this weighting in the presence of 2° prediction errors.

We, therefore, outline a procedure to build structural probability distributions that use as input three-state 2° predictions of specified accuracy Q_3. This procedure can also use another input f_{bad}, the fraction of residues in helices mispredicted as extended structures and vice versa (called 'bad' predictions) [4,42] to supplement Q_3. We denote these optimized distributions as $P(\chi_{(\phi,\psi)_m}|\chi_{2^\circ P}, Y_{\mathrm{aa}})$ to emphasize that the input 2° classification is a prediction.

The width of the probability distribution can be measured by the conditional entropy

$$H(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{\mathrm{aa}}) = \sum_{i,j} H(\chi_{(\phi,\psi)}|\chi_{2^\circ P_i}, Y_{\mathrm{aa}_j})\, p(\chi_{2^\circ P_i}, Y_{\mathrm{aa}_j})$$

$$= -\sum_{i,j} p(\chi_{2^\circ P_i}, Y_{\mathrm{aa}_j}) \sum_k p(\chi_{(\phi,\psi)_k}|\chi_{2^\circ P_i}, Y_{\mathrm{aa}_j})$$

$$\ln[p(\chi_{(\phi,\psi)_k}|\chi_{2^\circ P_i}, Y_{\mathrm{aa}_j})] \tag{5a}$$

Like Eq. (3a), the expression can be re-written as an expectation

$$H(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{\mathrm{aa}}) = E\{-\ln[P(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{\mathrm{aa}})]\} \tag{5b}$$

As before, an estimate of the entropy can be made by approximating the expectation:

$$H(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{\mathrm{aa}}) = -(1/n_{\mathrm{tot}}) \sum_k^{n_{\mathrm{tot}}} \ln[p(\chi_{(\phi,\psi)_k}|\chi_{2^\circ P}, Y_{\mathrm{aa}})] \tag{5c}$$

There are two major components to this procedure. First, the probability distribution $P(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{\mathrm{aa}})$ is estimated using $P(\chi_{(\phi,\psi)_j}|Y_{\mathrm{aa}})$ as the background distribution:

$$p(\chi_{(\phi,\psi)_k}|\chi_{2^\circ P}, Y_{\mathrm{aa}})$$

$$= [\gamma\, p(\chi_{(\phi,\psi)_k}|Y_{\mathrm{aa}}) + n(\chi_{2^\circ P}, Y_{\mathrm{aa}})\, \pi(\chi_{(\phi,\psi)_k}|\chi_{2^\circ P}, Y_{\mathrm{aa}})$$

$$- 1] / [\gamma + n(\chi_{2^\circ P}|Y_{\mathrm{aa}}) - 1] \tag{5d}$$

Second, the fact that the prediction accuracy is less than 100% is incorporated in the estimate. This is accomplished by a Monte Carlo procedure which randomly generates the 2° structure at each prediction site in order to simulate the output of an algorithm of given accuracy Q_3. The procedure also simulates the fraction of bad mispredictions, f_{bad}.

The simulation is implemented as follows. (1) Both the 2° and the (ϕ, ψ) structures of each residue in the data set are noted, and their respective frequencies are counted. (2) A proportion, given by f_{bad}, of residues of states H or an E is randomly selected, and mispredictions of E or H are substituted. (3) Another set of mispredictions are made involving sites where the initial or final 2° state is C, to generate a total proportion of wrong predictions corresponding to $1 - Q_3$. (4) The probability $p(\chi_{(\phi,\psi)}|\chi_{2^\circ P})$ is calculated using Eq. (5d), setting $\chi_{(\phi,\psi)}$ equal to the actual $(\phi, \psi)^{20^\circ}$ structure of the particular residue. (5) Steps 1–4 are applied to every residue in the data set, and the entropy is then calculated via Eq. (5c). (6) To strengthen the estimate for the entropy, the procedure is applied to the whole data set as many times as necessary for the average value to converge. (We find that reasonable convergence is achieved when the procedure is repeated 50 times.)

2.5. Secondary structure prediction algorithms outputting three-state probability distributions

More sophisticated 2° prediction algorithms output a probability distribution to describe the relative probabilities of all possible states. For instance, a prediction output for a residue can take the following form $p_Q(\mathrm{H}) = 0.66$, $p_Q(\mathrm{E}) = 0.10$, $p_Q(\mathrm{C}) = 0.24$, instead of the single-state output of H, which carries less information about the prediction. In this work, we denote this three-state probability output as $P_Q(\chi_{2^\circ})$, and specify its values by $\{p(\mathrm{H}), p(\mathrm{E}), p(\mathrm{C})\}$ (e.g. $P_Q(\chi_{2^\circ}) = \{0.66, 0.10, 0.24\}$). The subscript Q is a reminder that these probabilities are generated by a prediction algorithm.

One would like to interpret such probability distributions as follows: given the incomplete information available to the algorithm to specify the correct 2° state (e.g. inputs of local sequence, multiple sequence alignments, sequence/structural environments, etc.), the distribution is the relative likelihood of finding a set of conditions consistent with each of the three 2° forms. The biophysical underpinnings of the prediction may be obscure, but the distributions are essentially informatic in nature. They reflect our state of ignorance—the residue in question does not actually exist in three 2° states in the given proportion, but folds into one, and only one, state.

All prediction algorithms can, in principle, output an index resembling strength of prediction for *each* type of secondary structure. Even those which output a single state work implicitly with a scoring system, from which a majority-rules decision is implemented to arrive at the single-state output. Such indices can be transformed into pseudo-probabilities by a simple normalization. Such probabilities, however, do not necessarily describe the true, underlying distribution. For instance, a particular residue in the helical state may be correctly predicted by an

algorithm outputting $P_Q(\chi_{2^\circ}) = \{0.80, 0.08, 0.12\}$ as well as another with $P_Q(\chi_{2^\circ}) = \{0.34, 0.33, 0.33\}$, if a majority-rules decision is enforced. Clearly, these distributions cannot both satisfy the underlying distribution. Black-box algorithms, such as neural nets, may be able to output a distribution, but their relevance may be limited to the majority-rules condition from which they are trained.

This underlying distribution, which we denote as $P_{\text{true}}(\chi_{2^\circ})$, is unknowable. The algorithms can only attempt to approximate it. We will see that successful approximations to the underlying probability distribution should assist in extracting the maximal amount of information from the data set. On the other hand, inaccurate estimates can mislead, and, therefore, may lead to a degradation in our ability to extract information. It is thus necessary to differentiate between these two kinds of outputs. In this work, we measure the ability of both types of three-state probability outputs to extract information, and thereby gauge the added benefit of more detailed output as compared to the simpler single-state output.

Given a probability distribution $P_Q(\chi_{2^\circ})$ which describes the relative probabilities of finding the particular residue in each of the three 2° states, a sequence-specific hybrid distribution in the (ϕ, ψ) space can be generated by the total probability rule

$$P_Q(\chi_{(\phi,\psi)}|\chi_{2^\circ P3}, Y_{\text{aa}}) = \sum_i P(\chi_{(\phi,\psi)}|\chi_{2^\circ_i}, Y_{\text{aa}})\, p_Q(\chi_{2^\circ_i}) \qquad (6)$$

where the summation (indexed by i) goes through the three 2° states, and $\chi_{2^\circ P3}$ denotes the 2° prediction with the three-state probability output (as compared to $\chi_{2^\circ P}$, which refers to the single-state prediction output) (e.g. used in Ref. [39]). The probability distribution $P(\chi_{(\phi,\psi)}|\chi_{2^\circ_i, Y_{\text{aa}}})$, which describes the contour on the (ϕ, ψ) plane for each of the 20 amino acids in the 2° state i, is generated using methods described in previous sections. (For instance, Fig. 7(A) shows the distribution for helical alanine residues, $P(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ} = \text{H}, Y_{\text{aa}} = \text{Ala})$.) The term $p_Q(\chi_{2^\circ_i})$ is the probability of state i, which modulates the features of the resulting hybrid distribution, which is a weighted average of the three canonical contour maps for the specific amino acid Y_{aa}. It is easy to see that if the prediction is perfect (i.e. $p_Q(\chi_{2^\circ_c}) = 1$ for the correct state c, and 0 for the two incorrect states), then $P_Q(\chi_{(\phi,\psi)}|Y_{\text{aa}})$ reduces to $P(\chi_{(\phi,\psi)}|Y_{\text{aa}}, \chi_{2^\circ_c})$, the (ϕ, ψ) distribution of the amino acid Y_{aa} in state c. We note that in order to produce the hybrid distribution which best describes the (ϕ, ψ) propensities, the term $P_Q(\chi_{2^\circ})$ should, in principle, approximate the underlying probabilities $P_{\text{true}}(\chi_{2^\circ})$. Otherwise, the resulting hybrid distribution will be unnecessarily broad.

To measure the informatic quantities involved in this kind of prediction, we simulate the prediction process, characterized by Q_3 and f_{bad}, by a Monte Carlo procedure. Note that, in contrast to the Monte Carlo procedure outlined in Section 2.4, we now work with a three-state 2° probability distribution. We first describe the case when the three-state probability output is consistent with the underlying distribution (i.e. $P_Q(\chi_{2^\circ}) = P_{\text{true}}(\chi_{2^\circ})$). We wish to randomly generate 'true' three-state distributions for each residue, consistent with pre-selected values of Q_3 and f_{bad}. We will then ask whether an alternative prediction algorithm, which give the same prediction under a majority-rules selection scheme, but which produces different output probabilities, extracts the same amount of structural information from the data set.

A prediction is generated for every residue in the data set. The probability of a correct prediction, $p_{\text{true}}(\chi_{2^\circ} = \text{correct})$ is designed to fluctuate around the given Q_3 *at each site*, with a set standard deviation: $Q_3 = E\{p_{\text{true}}(\chi_{2^\circ} = \text{correct})\}$. If the prediction is correct, then the probabilities for the remaining two 2° states are generated by randomly dividing the balance $1 - Q_3$. For instance, if we are given $p_{\text{true}}(\chi_{2^\circ} = \text{correct}) = 0.80$ for a particular residue in the coil state, we generate the correct prediction with probability 0.80. The probabilities for the helical and extended states are then generated by the random division of the remainder, 0.20, into two quantities, e.g. 0.13 and 0.07. Since we assume that $P_Q(\chi_{2^\circ}) = P_{\text{true}}(\chi_{2^\circ})$, the distribution for this case is $P_Q(\chi_{2^\circ}) = \{0.13, 0.07, 0.80\}$.

The chance that the prediction is incorrect is $1 - p_{\text{true}}(\chi_{2^\circ} = \text{correct})$. If a residue is determined by the Monte Carlo procedure to be incorrectly predicted, the incorrect state must be selected following another constraint, f_{bad}. This constraint affects only helical and extended structures: if the residue in question falls in either of the two structures, then the probability of generating a bad prediction (an H mispredicted as E, and vice versa) follows the probability[3] $f_{\text{bad}}/[(1 - Q_3) f_{\text{H+E}}]$, where $f_{\text{H+E}}$ is the fraction of non-coil structures (in our data set, $f_{\text{H+E}} = 0.59$). For example, if the residue in question is in the helical state, and the Monte Carlo procedure, with $p_{\text{true}}(\chi_{2^\circ} = \text{correct}) = 0.64$, determines that it be mispredicted, the probability of generating E as a prediction is $f_{\text{bad}}/[(1 - Q_3) f_{\text{H+E}}]$. If the simulation generates an E, then the balance $1 - p_{\text{true}}(\chi_{2^\circ} = \text{correct}) = 0.36$ is divided between the two other structural states H and C. The corresponding probability for the bad (H ↔ E) prediction, in this case $p_{\text{true}}(\chi_{2^\circ} = \text{H})$, was designed to fluctuate around the generating probability $f_{\text{bad}}/[(1 - Q_3) f_{\text{H+E}}]$. A possible output distribution for this helical residue, after the enforced misprediction, is $P_Q(\chi_{2^\circ}) = \{0.22, 0.64, 0.14\}$. One can see that this distribution would mispredict the helical residue as E (because its probability 0.64 is highest), under a majority-rules regime.

This point-per-point generation of prediction across the data set, using all the relevant probabilities, produces a

[3] What we would like in this case is the probability of a bad prediction (H ↔ E) given that it is a misprediction and the residue is either an H or E. Therefore, by simple probability rules, f_{bad}, the global probability of bad predictions, must be transformed by dividing by the probability of conditions added (i.e. $(1 - Q_3)$ and $f_{\text{H+E}}$).

single set of predictions when a majority-rules scheme with global characteristics conforming to the two prior constraints Q_3 and f_{bad}. (To confirm, we explicitly counted the number of correct predictions and the number of bad mispredictions in every simulation, and found that these two counts conform to Q_3 and f_{bad} with 99.9% accuracy.)

The simulation above describes the situation where the three-state probability output is the underlying distribution. This was assured because the distribution used to generate guesses was the same as that used to simulate the three-state probability distribution output. We also looked at the informatic behavior of algorithms which output three-state probability distributions not consistent with the underlying distribution established by the Monte Carlo simulation, $P_Q(\chi_{2^\circ}) \neq P_{true}(\chi_{2^\circ})$. Two extreme cases were examined. The first is a small perturbation from the underlying distribution $P_{true}(\chi_{2^\circ})$ to give $P_Q(\chi_{2^\circ}) = P_{true}(\chi_{2^\circ}) \pm \delta$. The other case is that in which the underlying distribution is completely ignored, and $P_Q(\chi_{2^\circ})$ is generated randomly. Both situations, however, are designed so that their majority-rules predictions are consistent with the prediction of each residue in the data set, arising from the original, underlying (Monte Carlo generated) distribution.

To ensure this, we used the same Monte Carlo simulation of the prediction process described above, but with some important alterations. For the small perturbation case, we systematically perturbed the $P_{true}(\chi_{2^\circ})$ by a small value (by 0.05 and 0.10) in either direction to give $P_Q(\chi_{2^\circ})$. (As a reminder, the former quantity, $P_{true}(\chi_{2^\circ})$, governed the generation of the prediction, while the latter, $P_Q(\chi_{2^\circ})$, is taken as the simulated output of the prediction algorithm.) For the randomly generated $P_Q(\chi_{2^\circ})$ case, we simply conjured $P_Q(\chi_{2^\circ})$ consistent with the real as well as the predicted secondary structures generated by the Monte Carlo simulation. For instance, if a particular helical residue was mispredicted as extended (with $P_{true}(\chi_{2^\circ}) = \{0.20, 0.50, 0.30\}$) then the randomly generated $P_Q(\chi_{2^\circ})$ must have $P_Q(\mathrm{E}) > P_Q(\mathrm{H})$ and $P_Q(\mathrm{E}) > P_Q(\mathrm{C})$ (e.g. $P_Q(\chi_{2^\circ}) = \{0.12, 0.75, 0.13\}$).

With the definition of the conditional probability (Eq. (6)), the computation of the residual entropy is straightforward. The residual entropy is computed using an equation analogous to Eq. (5b),

$$H(\chi_{(\phi,\psi)}|\chi_{2^\circ P3}, Y_{aa}) = E\{ -\ln[P_Q(\chi_{(\phi,\psi)}|\chi_{2^\circ P3}, Y_{aa})]\} \quad (7a)$$

Again, an estimate of the entropy can be made by approximating the expectation by a summation across the entire data set, or

$$H(\chi_{(\phi,\psi)}|\chi_{2^\circ P3}, Y_{aa}) = -(1/n_{tot}) \sum_{k}^{n_{tot}} \ln[p(\chi_{(\phi,\psi)_k}|\chi_{2^\circ P3}, Y_{aa})] \quad (7b)$$

We find that even a single Monte Carlo pass through the entire data set results in a reliable estimate for the various entropic quantities of interest. Nonetheless, to ensure reliability, we average quantities from 20 independent passes through the entire data set.

3. Results and discussion

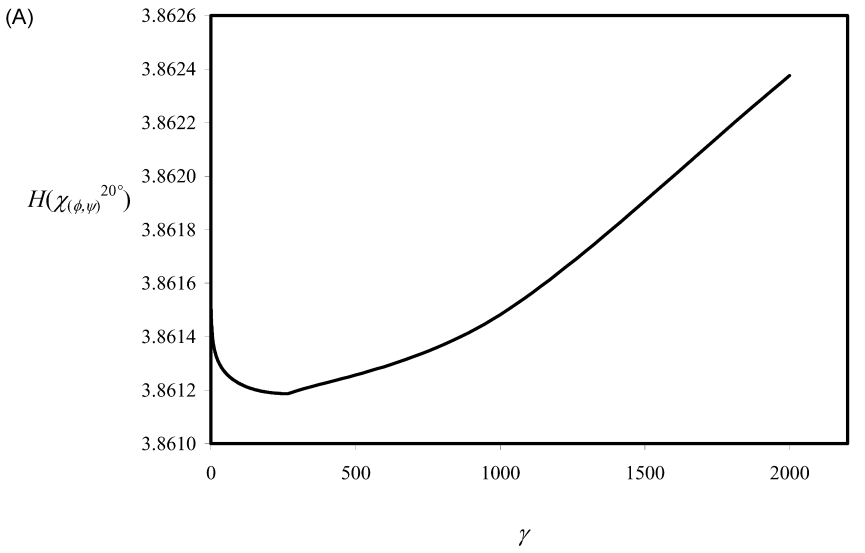
3.1. Uncertainty in determining the (ϕ, ψ) conformation of protein chains with known secondary structure

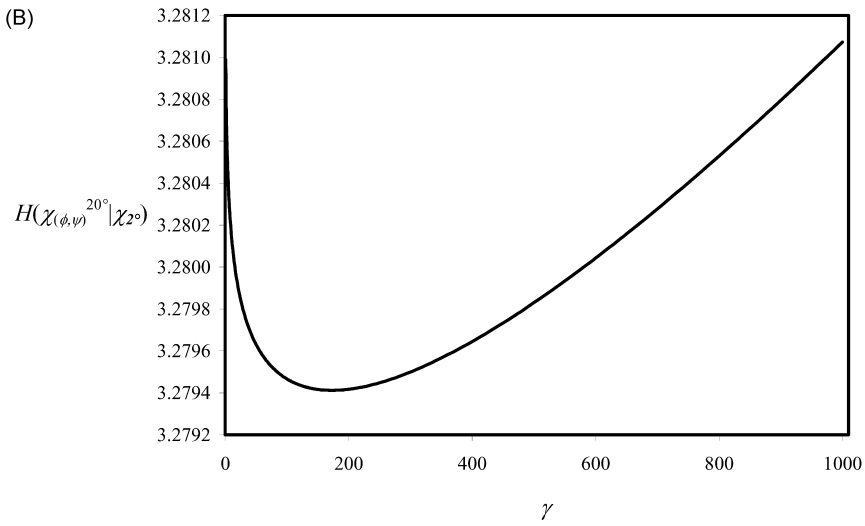
Entropic quantities relating to the (ϕ, ψ) dihedral angle space are calculated from structural probability distributions, generated by the weighted combination of raw frequencies and a properly chosen background distribution. This combination, embodied in Eq. (4), is characterized by a hybrid distribution coefficient γ, chosen to return the lowest residual entropy. Generating structural distributions via the hybrid method protects against over-zealous partition of the sequence and structure domains, and ensures that enough data are available to provide meaningful statistics.

The over-all uncertainty in determining the $(\phi, \psi)^{20^\circ}$ conformation is given by the entropy $H(\chi_{(\phi,\psi)^{20^\circ}})$, computed with the uniform distribution as background. The dependence of the entropy on the hybrid coefficient, shown in Fig. 2(A), exhibits a clear single minimum of 3.861 nats at the optimum hybrid coefficient $\gamma = 266$.

We use the resulting hybrid distribution $P(\chi_{(\phi,\psi)^{20^\circ}})$ of the universe of structures as the background distribution to generate $P(\chi_{(\phi,\psi)}|\chi_{2^\circ})$, the distribution of backbone phi–psi conformation given the correct three-state 2° conformation. The optimal probability distribution (at $\gamma = 173$) yields a residual entropy $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ})$ of 3.279 nats. (The dependence of $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ})$ on γ is shown in Fig. 2(B).) The same procedure was implemented to search for the optimum hybrid coefficient associated with $H(\chi_{(\phi,\psi)^{5^\circ}}|\chi_{2^\circ})$, the case where the phi–psi space is subdivided into 72×72 equally sized bins. This search, summarized in Fig. 2(C), identifies the hybrid coefficient ($\gamma = 1855$) which yields the lowest residual entropy of 5.845 nats. The latter case, in which the

Fig. 2. The dependence of the residual entropy on the hybrid coefficient γ. (A) Measuring $H(\chi_{(\phi,\psi)^{20^\circ}})$. The uniform distribution was chosen as the background distribution to build the distribution of (ϕ, ψ) dihedral angles discretized into an 18×18 grid (20° resolution). The plot exhibits a clear single minimum of 3.861 nats at the optimum hybrid coefficient $\gamma = 266$. (B) Measuring $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ})$. We use the resulting hybrid distribution $P(\chi_{(\phi,\psi)^{20^\circ}})$ of the universe of structures as the background distribution to generate $P(\chi_{(\phi,\psi)}|\chi_{2^\circ})$, the distribution of backbone phi–psi conformation given the correct three-state 2° conformation. The optimal probability distribution (at $\gamma = 173$) yields a residual entropy $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ})$ of 3.279 nats. (C) Measuring $H(\chi_{(\phi,\psi)^{5^\circ}}|\chi_{2^\circ})$. The search for the optimum hybrid coefficient associated with $H(\chi_{(\phi,\psi)^{5^\circ}}|\chi_{2^\circ})$, the case where the phi–psi space is subdivided into 72×72 equally sized bins identifies the hybrid coefficient ($\gamma = 1855$) which yields the lowest residual entropy at 5.845 nats.


(A)
3.8626
3.8624
3.8622
3.8620
3.8618
3.8616
3.8614
3.8612
3.8610
$H(\chi_{(\phi,\psi)}^{20^\circ})$
0
500
1000
1500
2000
γ


(B)
3.2812
3.2810
3.2808
3.2806
3.2804
3.2802
3.2800
3.2798
3.2796
3.2794
3.2792
$H(\chi_{(\phi,\psi)}^{20^\circ}|\chi_{2^\circ})$
0
200
400
600
800
1000
γ

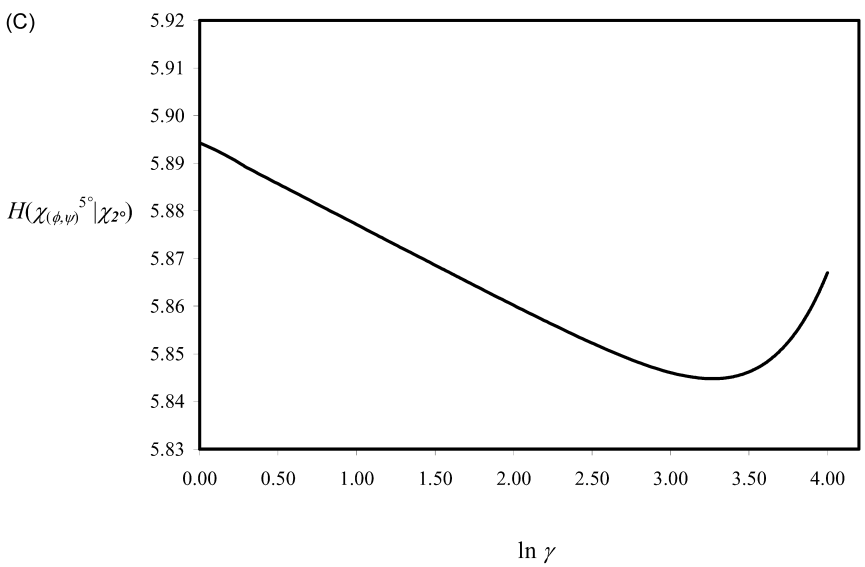

(C)
5.92
5.91
5.90
5.89
5.88
5.87
5.86
5.85
5.84
5.83
$H(\chi_{(\phi,\psi)}^{5^\circ}|\chi_{2^\circ})$
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00
ln γ

Table 1
Entropies and residual entropies of interest

Entropy term[a] Resolution[b]	Background distribution[c]	Optimal γ[d]	Entropy value (nats)
$H(\chi_{(\phi,\psi)})$			
90°	Uniform: $B = 1/m^2$	18	1.703
45°		65	2.567
20°		266	3.860
10°		784	5.065
5°		2124	6.402
$H(\chi_{(\phi,\psi)}\|\chi_{2^\circ})$			
90°	Universe: $B = p(\chi_{(\phi,\psi)})$	3	1.182
45°		55	2.031
20°		173	3.279
10°		568	4.482
5°		1855	5.845
$H(\chi_{(\phi,\psi)}\|Y_{aa})$			
90°	Universe: $B = p(\chi_{(\phi,\psi)})$	90	1.542
45°		290	2.371
20°		859	3.622
10°		2063	4.803
5°		4863	6.106
$H(\chi_{(\phi,\psi)}\|\chi_{2^\circ}, Y_{aa})$			
90°	Singlet: $B = p(\chi_{(\phi,\psi)}\|Y_{aa})$	9	1.057
45°		32	1.881
20°		98	3.118
10°		258	4.345
5°		810	5.755
$H(\chi_{(\phi,\psi)}\|\chi_{2^\circ P}, Y_{aa})$	Singlet: $B = p(\chi_{(\phi,\psi)}\|Y_{aa})$	see Fig. 8 for values at different levels of Q_3 and f_{bad}	

[a] The form of the equations, a weighted combination of raw frequencies and the chosen background distribution, used to calculate these terms are discussed in the text.
[b] Resolution refers to the size of square bins used to evenly divide the (ϕ, ψ) plane.
[c] Background component used to calculate the associated entropy. These terms must also be calculated accordingly, using the same strategy of optimizing the weighted combination of a raw frequency and a proper background distribution.
[d] Hybrid coefficients which give the lowest entropy.

phi–psi space is subdivided into more unique states (5184 vs. 324), leads to a larger hybrid coefficient, which is required to buffer sparser raw frequencies. One expects the value of the optimum hybrid coefficient to continue to rise as finer partitions are made to the structure domain. Conversely, the computation for $H(\chi_{(\phi,\psi)^{90^\circ}}|\chi_{2^\circ})$, from a coarse partition of the phi–psi space, generates a low hybrid coefficient, $\gamma = 3$. Since the same data set is subdivided into fewer unique structural states, the resulting raw frequency distribution is a more adequate estimate of the true probability distribution $P(\chi_{(\phi,\psi)^{90^\circ}}|\chi_{2^\circ})$, and the contribution of the background is diminished.

Table 1 summarizes the results of calculating various entropic quantities at different levels of structural discretization. Major results and observations are as follows.

3.1.1. The residual entropy associated with backbone conformation when the three-state secondary structure is known fully is still very high (i.e. $H(\chi_{(\phi,\psi)}|\chi_{2^\circ}) \gg 0$).

This observation highlights the major theoretical and computational difficulty in attempting to 'fold' 2° elements into their native 3D structure. The high residual entropy value, $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ}) = 3.279$ nats, explains the modest performance of computational schemes which attempt to fold known segments of helices and sheets of a given protein chain. Moreover, even within the organized segments, such as helices and sheets, the residual entropy $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ} = \{\mathrm{H}, \mathrm{E}\})$ is still large: $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ} = \mathrm{H}) = 2.003$ nats and $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ} = \mathrm{E}) = 3.218$ nats.

The major challenge lies in finding the backbone conformation of residues which are in the coil state. The average residual entropy of a given residue in the coil state, $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ} = \mathrm{C})$, is 4.362 nats, a value which emphasizes the wide variability of phi–psi dihedral angles in coil residues. Coil segments are the 'flexible' hinges connecting helices and sheets in a packed state, and folding secondary structures to form native tertiary interactions is equivalent to searching the backbone (ϕ, ψ) of coil segments for the correct packing of helices and sheets. Even if canonical helical and extended structures for the structured segments are assumed, the multitude of ways they can be organized into a tertiary domain is reflected in the high residual

entropy. Furthermore, the complexity of folding rigid 2° segments depends on the fraction of coil residues in the protein chain, since 4.364 nats is the average residual entropy for each (ϕ, ψ) dihedral angle pair in the coil state.

3.1.2. The uncertainty in determining backbone conformation even after complete knowledge of three-state secondary structure is much greater than the uncertainty resolved by an accurate assignment of secondary structure (i.e. $H(\chi_{(\phi,\psi)}|\chi_{2^\circ}) > H(\chi_{2^\circ})$).

The entropy associated with determining the correct 2° assignment, $H(\chi_{2^\circ})$, in our data set, composed of 41.15% coil, 36.42% helical, and 22.43% extended/sheet, is 1.069 nats per residue, from the Shannon entropy equation (Eq. (1)). (Because there are only three possible states, and there is a large amount of data to fill these states with observations, the calculation does not require using a hybrid distribution approximation. Instead, Eq. (1) can be used directly.)

Informatically, it is more difficult to find the correct $(\phi, \psi)^{20^\circ}$ given its 2° conformation ($H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ}) =$ 3.279 nats) than it is to determine the 2° assignment itself ($H(\chi_{2^\circ}) = 1.069$ nats), indicating that even if near-perfect accuracy could be achieved by a 'fantasy' 2° structure prediction algorithm, the protein 3D conformation would be difficult to predict. We note that the quantity $H(\chi_{(\phi,\psi)^{20^\circ}})$ is magnitudes higher than $H(\chi_{2^\circ})$, pointing to the greater challenge in finding the actual phi–psi conformation of the backbone, compared to identifying its correct three-state secondary structure.

3.1.3. Including some sequence information in generating probability distributions helps lower the residual entropy associated with the backbone conformation (i.e. $H(\chi_{(\phi,\psi)}|\chi_{2^\circ}) > H(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{aa})$).

Generating amino acid-specific probability distributions lowers the residual entropy of the correct (ϕ, ψ) given full knowledge of the three-state 2° class. This procedure generates 20 different kinds of probability distributions, one for each amino acid z, $P(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{aa} = z)$. Because we would like to measure the effect of 2° assignment on the (ϕ, ψ) propensity, we take as the background distribution in this instance $P(\chi_{(\phi,\psi)}|Y_{aa} = z)$ for a specific amino acid z. The latter distribution is in turn derived using a similar hybrid method, with the universe of structures $P(\chi_{(\phi,\psi)})$ as background.

The average value per residue of $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ}, Y_{aa})$ is 3.118 nats, which is significantly lower than $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ}) = 3.279$ nats. Including some sequence knowledge reduces the uncertainty of the backbone conformation given its three-state 2° assignment. The entropy is still substantial, however, indicating that the problem is far from solved. Nonetheless, translating 2° state information of a residue into (ϕ, ψ) propensities should involve its amino acid identity because the resulting distribution is narrower.

3.1.4. The level of detail in the structural description affects the amount of information that can be extracted from the database.

It should come as no surprise that searching for the correct (ϕ, ψ) bin becomes more difficult as the number of unique bins increases. This is reflected in the increasing values for $H(\chi_{(\phi,\psi)})$ as the resolution is increased. Results using various levels of partition are summarized in Table 1. The same pattern can be seen for $H(\chi_{(\phi,\psi)}|\chi_{2^\circ})$ and $H(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{aa})$.

What we are more concerned with is the amount of information latent in knowledge of the three-state secondary structure. The quantity of interest is the information gain. Table 2 contains the relevant information gain

$$I_g(\chi_{2^\circ}, Y_{aa}) = H(\chi_{(\phi,\psi)}) - H(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{aa}) \quad (8)$$

Information gain measures the information extracted as a result of the introduction of one or more factors, calculated by subtracting the entropy of the distribution conditioned on the new factor(s) from the entropy without considering the same factors. In the analyses that follow, we use several forms of information gain. Our goal is to gauge the success of specific factors in extracting information from the protein data set.

One can see from Table 2 that the information gain peaks at 20° resolution. These calculations illustrate the fact, which we investigated in recent work [1], that there exists an optimal level of structural resolution given a finite data set. The amount of extractable information initially increases as the structural partition becomes finer, but diminishes as one reaches a resolution that cannot be supported by the data set size.

3.1.5. The difficulty in determining backbone conformation depends strongly on amino acid identity. The information gain due to knowledge of 2° state is the same for all amino acid residue types.

Residue identity plays an important role in determining backbone conformation. Calculations of the entropy associated with the (ϕ, ψ) distribution for each amino acid, summarized in the second column of Table 3, show the range of backbone conformation allowed by the various side chains. The values cover the range [2.659, 3.991 nats] for isoleucine and glycine, respectively. The residual entropy is decreased upon introduction of information relating to correct 2° state, $H(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{aa} = z)$, as shown in the third column of Table 3. Each amino acid benefits from this information, since specifying a 2° state restricts the range of dihedral angles it may take. The actual effect of 2° information is measured by the information gain

$$I_g(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{aa}) = H(\chi_{(\phi,\psi)}|Y_{aa}) - H(\chi_{(\phi,\psi)}|\chi_{2^\circ}, Y_{aa}) \quad (9)$$

for each amino acid, found in the fourth column of Table 3. The effect of knowing the 2° state is fairly uniform across all amino acids, with a range of [0.401, 0.596 nats] for aspartic

Table 2
Information gain from knowledge of correct secondary structure

Resolution[a] (°)	Information gain[b], (nats) $I_g(\chi_{2^\circ}, Y_{aa})$	Structural information stored in secondary structure[c], (%) $I_g(\chi_{2^\circ}, Y_{aa})/H(\chi_{(\phi,\psi)})$
90	0.646	37.9
45	0.686	26.7
20	0.741	19.2
10	0.721	14.2
5	0.647	10.1

[a] Resolution refers to the size of square bins used to evenly divide the (ϕ, ψ) plane.
[b] Information gain for (ϕ, ψ) structure from knowledge of true three-state 2° conformation, and using amino acid identity information of the residue.
[c] The proportion of uncertainty resolved by knowledge of true three-state 2° conformation, and using amino acid identity of information of the residue. This is effectively the fraction of (ϕ, ψ) structural information, under various resolutions, resolved by secondary structure.

acid and isoleucine, respectively, around an average of 0.504 nats, with the exception of proline at 0.277 nats. The ability to specify the backbone (ϕ, ψ) of every amino acid residue even with perfect knowledge of its 2° state is hampered by the diversity of possible backbone conformations within each 2° state.

3.2. Entropy resolved by 2° prediction algorithms of limited accuracy, and the systematic uncertainty caused by prediction errors.

Two parameters are used to describe the extent of these mispredictions: Q_3, or three-state accuracy, is the proportion of correct predictions; f_{bad} is the fraction of all predictions that mistake an H for an E and vice versa. We summarize the results of our Monte Carlo simulations as follows.

3.2.1. There is an optimum probability distribution that minimizes the residual entropy when 2° prediction is imprecise.

The search for the (ϕ, ψ) probability distribution which contains the lowest residual entropy $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ} = \chi 2_{^\circ P}, Y_{aa})$ is a compromise between believing the prediction $\chi 2^\circ P$ and hedging against errors. In a later section, we give an example of such a compromise, using the construction of optimal distributions of backbone conformation of alanine, given that it is predicted to be in the helical state. The pressure of these competing factors makes it necessary to reoptimize the value of γ in the hybrid distribution.

Examples of the optimization of the hybrid coefficient can be seen in Fig. 3. The procedure was applied to a range of accuracies, Q_3, while the fraction of bad predictions, f_{bad}, was set at 0.04. The residual entropy reaches clear minima

Table 3
Sequence-dependent information gain of the $(\phi, \psi)^{20^\circ}$ conformation due to knowledge of correct secondary structure

Amino acid z	Entropy[a], (nats) $H(\chi_{(\phi,\psi)^{20^\circ}}\|Y_{aa} = z)$	Residual entropy[b], (nats) $H(\chi_{(\phi,\psi)^{20^\circ}}\|\chi_{2^\circ}, Y_{aa} = z)$	Information gain[c], (nats) $I(\chi_{(\phi,\psi)^{20^\circ}}\|\chi_{2^\circ}, Y_{aa} = z)$
A	3.360	2.797	0.563
C	3.815	3.352	0.463
D	3.942	3.541	0.401
E	3.425	2.930	0.495
F	3.709	3.141	0.568
G	4.449	3.991	0.458
H	3.938	3.275	0.463
I	3.255	2.659	0.596
K	3.630	3.134	0.496
L	3.370	2.777	0.593
M	3.459	2.892	0.567
N	4.079	3.662	0.416
P	2.948	2.671	0.277
Q	3.521	3.005	0.516
R	3.618	3.091	0.527
S	3.814	3.375	0.439
T	3.748	3.252	0.496
V	3.346	2.774	0.572
W	3.553	3.040	0.513
Y	3.719	3.154	0.565

[a] Entropy of (ϕ, ψ) structure, at 20° resolution, with amino acid information only.
[b] Residual entropy after knowledge of correct 2° structure and with amino acid information.
[c] Information gain due to knowledge of correct 2° structure and with amino acid information.

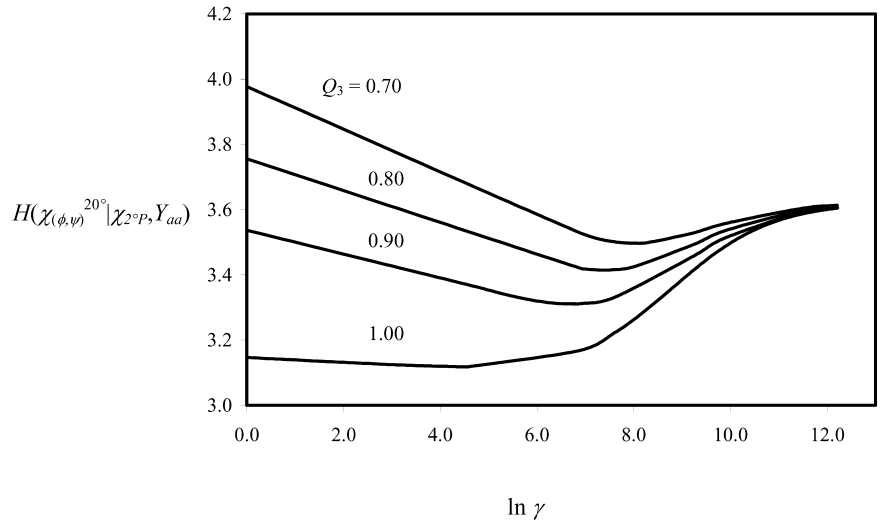


Fig. 3. Optimizing the residual entropy $H(\chi_{(\phi,\psi)^{20^\circ}}|\chi_{2^\circ P}, Y_{aa})$ at different accuracy levels Q_3 by varying the hybrid coefficient γ. The fraction of bad predictions, f_{bad}, was set at 0.04. The residual entropy reaches clear minima at 3.499, 3.415, and 3.312 nats with $\gamma = 3043$, 1592, and 810 for 70, 80, and 90%, respectively. In contrast, the curve for the optimization at perfect accuracy ($Q_3 = 100\%$) has a minimum of 3.118 nats at $\gamma = 98$.

at 3.499, 3.415, and 3.312 nats with $\gamma = 3043$, 1592, and 810 for $Q_3 = 0.7$, 0.8, and 0.9, respectively. From these examples, the residual entropy can be seen to decrease as the prediction accuracy increases. The curve for the optimization at perfect accuracy, also in Fig. 3, shows a shallow minimum of 3.118 nats at $\gamma = 98$.

The pattern of decreasing hybrid coefficients γ illustrates the correct behavior of Eq. (5d) in balancing the two opposing forces. Low hybrid coefficients favor the belief that the prediction is correct, while high hybrid coefficients favor the opposite possibility. The optimum hybrid coefficient falls somewhere in the middle. As the prediction accuracy rises, the belief that the prediction is correct grows stronger, favoring lower hybrid coefficients.

3.2.2. The detrimental effect of low accuracy and a high fraction of bad predictions on the latent information in the output of 2° prediction algorithms is significant

The residual entropy increases as prediction accuracies decrease. Another way to understand the influence of the quality of prediction on the amount of extractable information is to measure the information gain directly. The quantity of interest here is the information gain attributed to the knowledge of the prediction,

$$I_g(\chi_{2^\circ P}) = H(\chi_{(\phi,\psi)}|Y_{aa}) - H(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{aa}) \quad (10)$$

Fig. 4 shows the effect of varying accuracy and fraction of bad predictions on the amount of information latent in predictions. *The increase in information gain is exponential with accuracy* Q_3. At low accuracies, the information gain is negligible; only at accuracy levels above 0.55 does the gain become significant. The exponential behavior also points to the fact that increases in accuracy beyond the current level will be rewarded by higher increases in information extraction.

The fraction f_{bad} should also affect the amount of residual entropy; i.e. the greater the fraction of catastrophic misprediction, the higher the entropy. This pattern is observable in Figs. 4 and 5. While the effect of f_{bad} on information gain is not as dramatic as that of accuracy, the decrease in information gain becomes severe at large values of f_{bad}. Limiting catastrophic mispredictions leads to an improvement in prediction quality even without actually strengthening the accuracy.

3.2.3. The effect of structure discretization is significant.

As previously shown, the amount of information from predicted secondary structure depends on the resolution of the backbone structural descriptor. Figs. 5 and 6 show this clearly. In this work, the 20° resolution seems to be the most efficient. Describing the conformation at lower resolution misses details that can decrease the residual entropy, while higher resolution is counterproductive because of limited data availability.

3.2.4. Generating optimum conditional probability distributions.

Analysis of the set of probability distributions corresponding to the alanine residue predicted as H, $p(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = \mathrm{H}, Y_{aa} = \mathrm{A})$, in Fig. 7, reveals the effect of prediction errors. The first figure of the series of contour maps is the (ϕ, ψ) distribution of all helical alanine residues found in the data set, $P(\chi_{(\phi,\psi)}|\chi_{2^\circ} = \mathrm{H}, Y_{aa} = \mathrm{Ala})$, and the last is the distribution of all alanines, irrespective of 2° state. The first and last figures in the series depict the extreme cases of information carried by a prediction of H. If a prediction scheme carries a 100% accuracy rate, then the former figure represents the (ϕ, ψ) distribution of those alanine residues assigned as helical. If the accuracy rate is extremely low (approaching randomized assignment), the latter (ϕ, ψ) distribution is the most conservative choice to

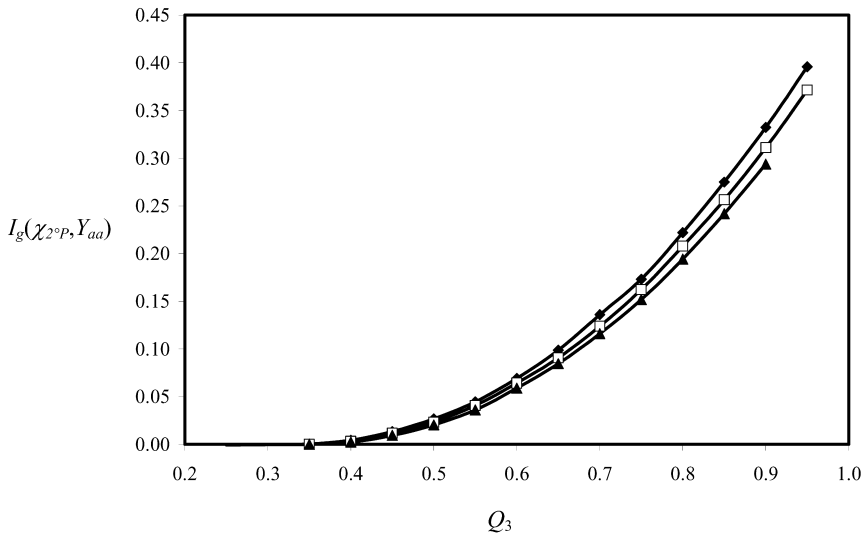


Fig. 4. The effect of varying accuracy (Q_3) and fraction of bad predictions (f_{bad}) on the amount of information latent in predictions, $I_g(\chi_{2^\circ P})$. The values of f_{bad} used are 0.02 (filled diamonds), 0.04 (empty squares), and 0.06 (filled triangles). At low accuracies, the information gain is negligible; only at accuracy levels above 0.55 does the gain become significant.

describe the backbone conformation of alanine residues predicted as helical. At such low accuracy rates, the 2° prediction is irrelevant, and our ignorance of the secondary structure demands that we describe the (ϕ, ψ) of any alanine residue (whether predicted as H, E, or C) as simply $p(\chi_{(\phi,\psi)}|Y_{\text{aa}} = \text{A})$.

The advantage of using the hybrid method to approximate probability distributions becomes apparent when dealing with algorithms of intermediate accuracy. To construct such probability distributions, we use Fig. 7(A) as one component of the hybrid distribution and Fig. 7(E) as the other, weighted by a hybrid coefficient γ selected to return the lowest residual entropy. For instance, Fig. 7(B)–(D) shows the backbone distributions, generated automatically by the optimization procedure for cases where the accuracy rates are 90, 70, and 55% respectively, and the f_{bad} are 0.02, 0.04, and 0.10, respectively. In these cases, we observe that the helical (ϕ, ψ) distribution, concentrated around (−60°, −40°), is supplemented by a significant occupancy in the extended region (positive values of ψ). Moreover, the extent of mixing depends on the accuracy of prediction and the proportion of catastrophic predictions.

In these compromise distributions, any true E that is wrongly predicted as H will be correctly found more frequently using $P(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = \text{H}, Y_{\text{aa}} = \text{A})$ rather than $P(\chi_{(\phi,\psi)}|\chi_{2^\circ} = \text{H}, Y_{\text{aa}} = \text{A})$. The unavoidable trade-off is that the backbone conformation of an alanine correctly

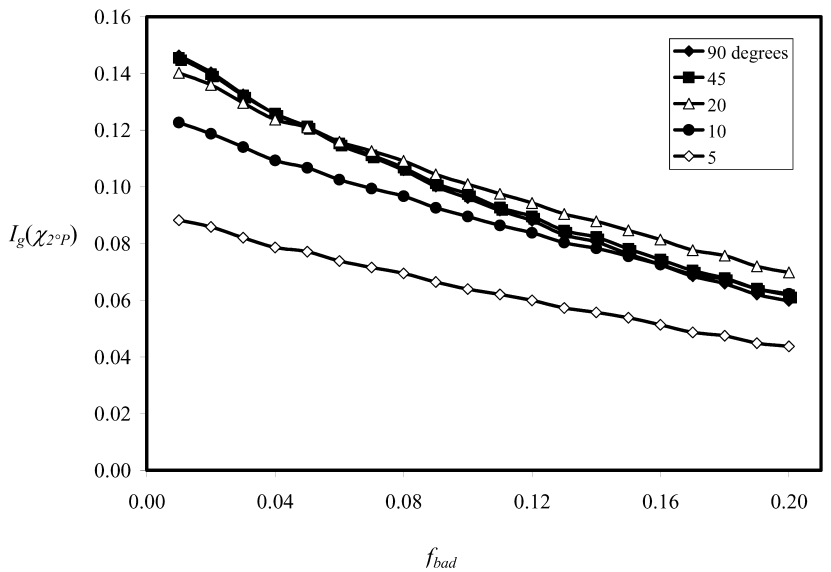


Fig. 5. The influence of the (ϕ, ψ) resolution on information gain $I_g(\chi_{2^\circ P})$ at different levels of fraction f_{bad}. The accuracy level Q_3 is kept at 70%. At most levels of f_{bad}, the resolution which yields the highest information gain is 20°.

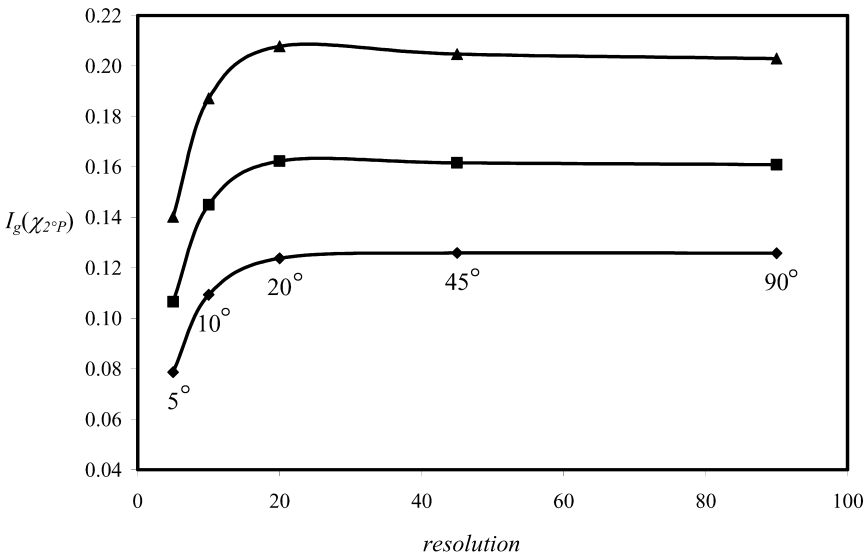


Fig. 6. The influence of the (ϕ, ψ) resolution on information gain $I_g(\chi_{2^\circ P})$ at different levels of accuracy Q_3. The fraction f_{bad} is kept at 0.04. The drop in information gain at higher resolutions (5–10°) is caused by the limited size of the structural data set.

predicted as H is easier to locate using $p(\chi_{(\phi,\psi)}|\chi_{2^\circ} = \mathrm{H})$ than $p(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = \mathrm{H})$. These competing situations are balanced by the hybrid procedure. While H and E residues with canonical backbone conformations favor lower γ, the significant fraction of catastrophic predictions necessitate that peripheral regions are represented with non-zero probabilities, and this is facilitated by a high γ.

Given a particular Q_3 and f_{bad}, our procedure generates contour maps $P(\chi_{(\phi,\psi)}|\chi_{2^\circ P}, Y_{aa})$ for each of the 20 amino acids Y_{aa}, in the three kinds of 2° prediction $\chi_{2^\circ P}$, for a total of 60 maps. Each pair of unique values for the parameters Q_3 and f_{bad} generate 60 characteristic contour maps, which incorporate the optimal correction for the accuracy of prediction.

3.2.5. The possibility of prediction errors results in substantial loss in information gain.

The net information gain brought about by a 2° prediction algorithm with a given accuracy (and its associated f_{bad}), can be measured by Eq. (10). This can be compared to Eq. (9), the information gain due to knowledge of the correct 2° state, to yield the fraction of information retained despite errors in prediction:

$$\theta_{ret}(\chi_{2^\circ P}) = I_g(\chi_{2^\circ P})/I_g(\chi_{2^\circ}) \tag{11}$$

We are interested in the dependence of this fraction on the prediction accuracy Q_3 and the fraction of bad predictions f_{bad}. A value of θ_{ret} near zero means that, because of the extensive prediction errors, no information can be gathered from a faulty 2° prediction output. Conversely, a value of θ_{ret} near unity characterizes successful prediction schemes. Fig. 8 shows our calculations for a range of Q_3 [0.25–0.95] and a corresponding range of f_{bad}, beginning at 0.01 and ranging up to the highest level permitted by Q_3.

The loss of information due to inaccurate prediction is severe. For any level of f_{bad}, predictions with accuracies of 55% or below retrieve less than 10% of the information available to specify backbone conformation from secondary structure. Typical accuracies of 70–75%, with $f_{bad} =$ 0.03–0.05, lose as much as 70% of the information due to mispredictions. Moreover, even an algorithm giving 95% accuracy returns less than 75% of the information encoded in secondary structure, although this level of success in predicting secondary structure would be exceptional.

3.2.6. What is a bad prediction?

The so-called bad predictions are those that confuse helical with sheet segments, and vice versa. We can measure the information loss specific to such catastrophic mispredictions. For every wrong 2° prediction, we can measure the amount of entropy it took to locate the observed $(\phi, \psi)^{20^\circ}$ in the wrong probability distribution $p(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = w)$ and compare it to the amount of entropy it would have taken to find the observed $(\phi, \psi)^{20^\circ}$ state in the correct distribution $p(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = c)$. The entropy cost for specific instances for all six possible misprediction pairs $(c, w) = (\mathrm{H}, \mathrm{C})$ (an H residue mispredicted as C), (H,E), (E,C), (E,H), (C,H), and (C,E) could be averaged to measure the amount of information lost in improperly predicting the 2° state:

$$\langle I_{loss}(c,w)\rangle = [1/n(c,w)] \sum_{i}^{n(c,w)} \{\ln[-p(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = w)_i] - \ln[-p(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = c)_i]\} \tag{12}$$

where $n(c, w)$ is the number of mispredictions in the Monte Carlo procedure, and the index i goes through every instance. The average information losses, computed for all

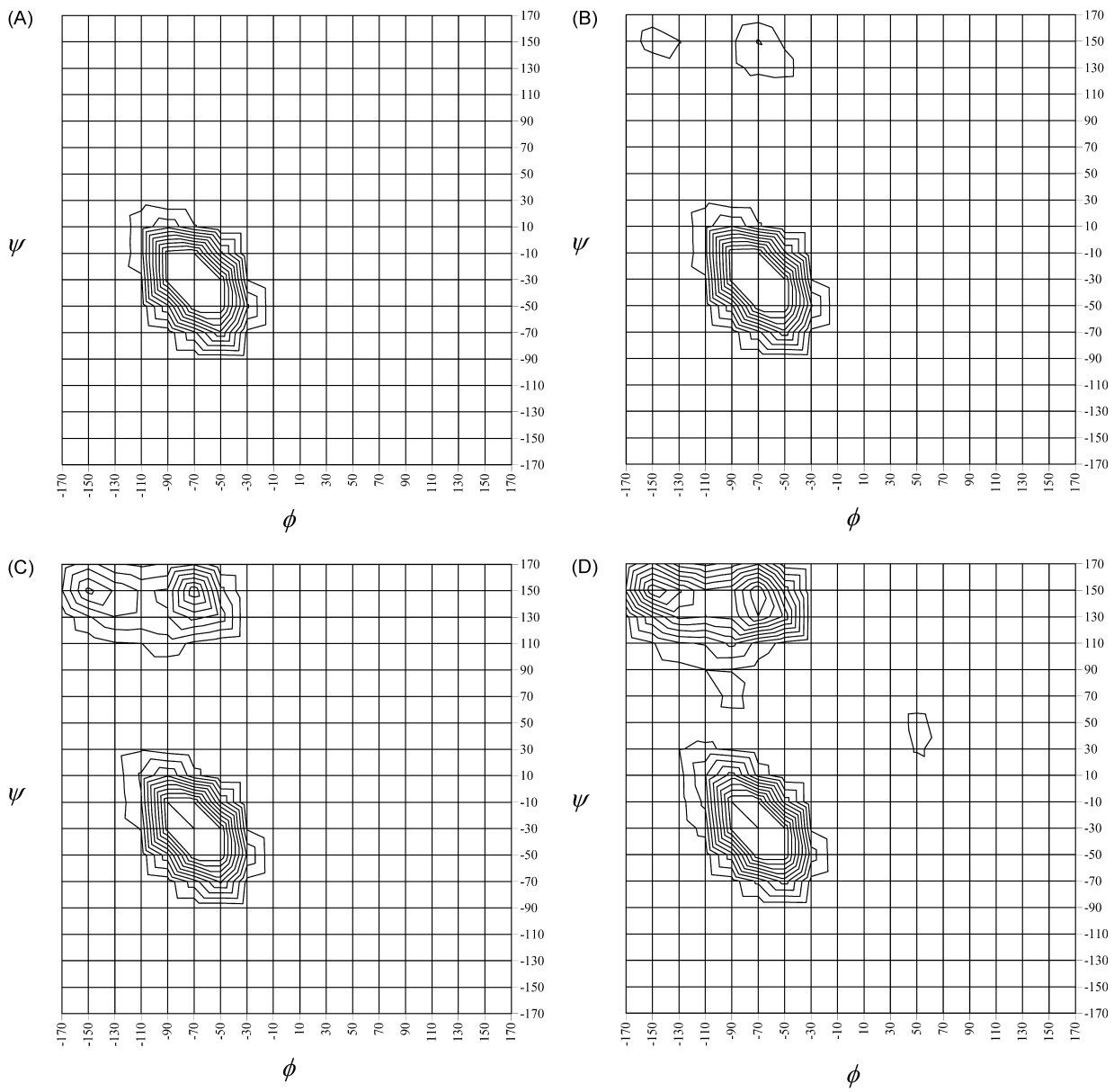


Fig. 7. Propensity distributions in the $(\phi, \psi)^{20^\circ}$ dihedral angle space of alanine predicted as helical, or $p(\chi_{(\phi,\psi)}|\chi_{2^\circ P} = \mathrm{H}, Y_{\mathrm{aa}} = \mathrm{A})$. The figures represent contour maps at different levels of accuracy Q_3 and f_{bad} : (A) 100%, (B) 90% ($f_{\mathrm{bad}} = 0.02$), (C) 70% (0.04), (D) 55% (0.10). (E) The (ϕ, ψ) distribution of *all* alanine residues in the data set. The intersections of the grid lines represent the midpoint of the 20° bin. For example, the intersection at $(-70, 150^\circ)$ represents the bin bounded by the ϕ and ψ ranges of $[-80^\circ, -60^\circ]$ and $[140, 160^\circ]$, respectively. Contour lines represent grades in probability in increments of 0.0015, starting from zero probability. Only the first 10 contour lines are included in this figure for the purpose of illustration, with the tenth contour line representing probabilities greater than 0.015. The peaks in probability of the helical phi–psi angle pair occur in the vicinity of 0.25. This and other details (pertaining to $p > 0.015$) are preserved in the actual probability distributions computed, but are not included in these figures for simplicity.

six misprediction pairs, are listed in Table 4. Not surprisingly, the two pairs that have been called bad, (E,H) and (H,E), have the two highest information losses for all levels of prediction accuracy. Another pair, (H,C), shows a comparable information loss. This is the case when a residue in a helical segment is incorrectly predicted to be in the coil state. The search for a helical dihedral angle pair in the probability distribution of the coil state, Fig. 1(C), is not as difficult, since the helical region is amply represented in the distribution. However, helical dihedral angles are the easiest to find if the correct probability distribution is used (Fig. 1(A)), and the loss in the efficiency is reflected in the high $\langle I_{\mathrm{loss}}(\mathrm{H}, \mathrm{C})\rangle$. The least serious mispredictions are the (C,E) and (E,C) pairs, as the probability distribution of coil and extended dihedral angles have similar coverage.

This general trend—increasing cost of misprediction as the accuracy level rises—highlights yet again the competing factors in building a compromise distribution. As the canonical (ϕ, ψ) distributions are favored at high accuracy

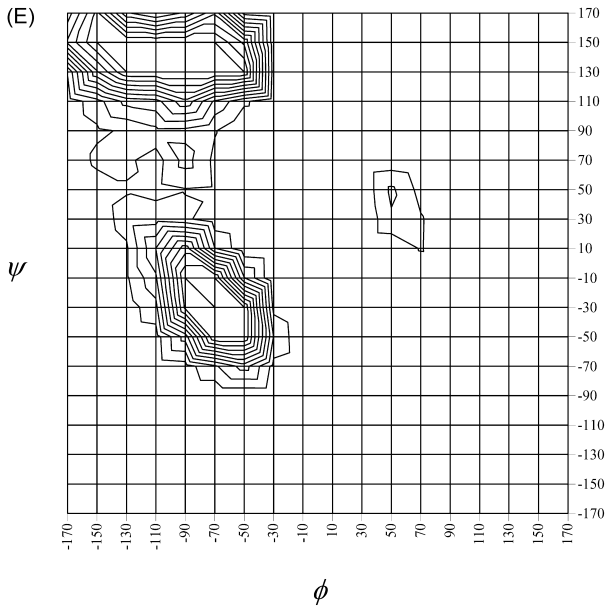


Fig. 7 (*continued*)

levels, the regions where mispredictions occur will not have as high a probability value as one would like. On the other hand, the actual number of mispredictions is low, while correct predictions see higher probability values, causing a general increase in information gain. However, the problem is magnified because of the sizeable penalty for locating the correct (ϕ, ψ) values in the unfavorable contour map arising from an erroneous prediction.

3.2.7. Single-state prediction outputs accompanied by an accuracy index show a modest increase in average information gain when one calibrates hybrid distributions specific to the accuracy index.

Ideally, we would like 2° prediction algorithms to output a confidence index in order to assess the likelihood that the prediction is true. This additional output can be used to tailor the structural distribution to a particular prediction conveniently, if one knows the precise relationship between

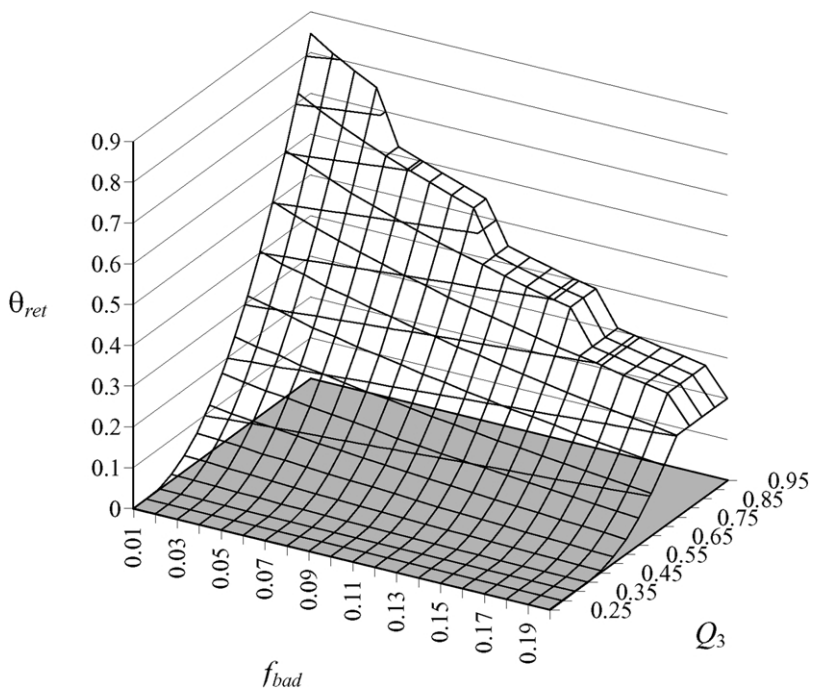


Fig. 8. The dependence of the fraction of information retention θ_{ret} on the accuracy level Q_3 and fraction of bad predictions f_{bad}. The figure summarizes calculations made for a range of Q_3 [0.25–0.95] and a corresponding range of f_{bad}, beginning at 0.01 and ranging up to the highest level permitted by Q_3.

Table 4
Average cost of misprediction

Q_3 and f_{bad}[a]	Secondary structure[b]		Cost of misprediction[c], $< I_{loss}(c, w) >$
	True, c	Predicted, w	
$Q_3 = 60\%, f_{bad} = 0.05$	H	E	0.717
	H	C	0.678
	E	H	0.881
	E	C	0.394
	C	H	0.471
	C	E	0.207
$Q_3 = 70\%, f_{bad} = 0.04$	H	E	1.083
	H	C	0.967
	E	H	1.348
	E	C	0.552
	C	H	0.719
	C	E	0.313
$Q_3 = 80\%, f_{bad} = 0.03$	H	E	1.535
	H	C	1.258
	E	H	1.922
	E	C	0.695
	C	H	1.050
	C	E	0.455
$Q_3 = 90\%, f_{bad} = 0.02$	H	E	2.141
	H	C	1.563
	E	H	2.661
	E	C	0.819
	C	H	1.475
	C	E	0.726

[a] Q_3, three-state accuracy; f_{bad}, fraction of bad predictions (H ↔ E).
[b] Three-state secondary structure: H = helix; E = extended; C = coil.
[c] Calculated using Eq. (12).

the confidence index and the actual fraction of correct predictions. While it is reasonable to assume that there is a linear relationship between these two variables, it is still necessary to translate the index range (usually within 0–9) into a form more directly related to Q_3 (i.e. % accuracy).

The accuracy rate of a particular prediction instance, which we shall denote as q_3, is operationally similar to the global accuracy rate Q_3, for the purpose of generating structural distributions. For instance, if presented with an alanine predicted as helical with a confidence index which corresponds to 55% accuracy ($q_3 = 0.55$), we generate the propensity shown in Fig. 7(D). On the other hand, if at another location, we encounter an alanine predicted as helical but with an index corresponding to 90% accuracy, the distribution shown in Fig. 7(B) more appropriately describes its structural propensity. The important element in utilizing the additional information from such confidence indices is a description of the relationship between the confidence index and the actual fractional success of prediction (e.g. [32]).

Tailoring the hybrid distribution to the confidence index gives a modest increase in the average information gain. As established previously, information gain increases exponentially with accuracy. That is, $E(f(x)) \geq f(E(x))$ for any convex function f, implying $E(I_g(q_3)) > I_g(E(q_3))$ (since the equality does not occur within the region of interest). The latter quantity is just $I_g(Q_3)$, the information gain arising from a prediction scheme with a known global accuracy rate Q_3. As an illustration, let us take information gain measurements from three different Q_3 (all with $f_{bad} = 0.04$): 0.0641, 0.1237, and 0.2077 nats for $Q_3 = 0.60$, 0.70, and 0.80, respectively (Fig. 8). If we are given two predictions of $q_3 = 0.60$ and 0.80, the mean information gain is $(0.0641 + 0.2077)/2 = 0.1359$ nats. However, the mean $q_3 = (0.60 + 0.80)/2 = 0.70$, which gives an information gain of 0.1237 nats. Therefore, indicating the accuracy index for each prediction should result in a modest increase in the information gain.

3.3. Entropy resolved by secondary structure predictions outputting probability distributions, and the systematic uncertainty caused by prediction errors.

We summarize results of our Monte Carlo simulation of outputs involving three-state probability outputs below.

3.3.1. Three-state probability outputs not consistent with the underlying distribution, $P_Q(\chi_{2^\circ})P_{true}(\chi_{2^\circ})$, lead to a significant reduction in the ability to extract structural information from 2° prediction.

It is important that $P_Q(\chi_{2^\circ})$ approximate $P_{true}(\chi_{2^\circ})$ as close as possible. The difference in residual entropy

between output probabilities $P_Q(\chi_{2^\circ})$ that are equivalent to $P_{\text{true}}(\chi_{2^\circ})$ and those that are uncorrelated, is large (shown in columns 4 and 5 of Table 5). The degradation in information caused by outputs barely resembling the underlying probabilities is highlighted by the fact that the magnitude of $H(\chi_{(\phi,\psi)}|Y_{\text{aa}})$, the structural entropy prior to any 2° prediction, is comparable to that of $H(\chi_{(\phi,\psi)}|\chi_{2^\circ P3}, Y_{\text{aa}})$, the entropy after introduction of a three-state probability prediction, but with $P_Q(\chi_{2^\circ}) \neq P_{\text{true}}(\chi_{2^\circ})$. Whereas there is a modest information gain when one uses a prediction algorithm of moderate quality ($Q_3 = 0.70$ and $f_{\text{bad}} = 0.04$) and which outputs $P_Q(\chi_{2^\circ}) = P_{\text{true}}(\chi_{2^\circ})$ (i.e. $3.62 - 3.44 = 0.18$ nats), there is actually a *loss* in information when one uses another algorithm with the same global prediction accuracy but with $P_Q(\chi_{2^\circ})$ unrelated to $P_{\text{true}}(\chi_{2^\circ})$ (i.e. $3.62 - 3.66 = -0.04$ nats). When faced with such algorithms (high global Q_3, but $P_Q(\chi_{2^\circ}) \neq P_{\text{true}}(\chi_{2^\circ})$), it is actually advisable to force the $P_Q(\chi_{2^\circ})$ into a majority-rules decision, and then proceed with methods designed for single-state prediction outputs to assemble a more informative ensemble of (ϕ, ψ) structure distributions.

3.3.2. The improvement in structural information extraction in going from the single-state 2° output to the three-state probability output is marginal.

Comparison between values in columns 3 and 4 of Table 5 reveals that the improvement in going from single-state to three-state probability outputs is below 0.10 nats. Moreover, the amount of structural information generated by well-performing single-state prediction algorithms may eclipse three-state prediction algorithms of lower accuracy. For instance, a single-state prediction with $Q_3 = 0.8$ and $f_{\text{bad}} = 0.04$ yields more information than a three-state probability prediction with $Q_3 = 0.7$ and $f_{\text{bad}} = 0.04$ (3.41 and 3.44 nats, respectively).

It should be emphasized that the ability to extract information from single-state 2° predictions relies on implementation of the methodology, described in this work, to generate optimal hybrid distributions. The added burden of approximating $P_{\text{true}}(\chi_{2^\circ})$ well, in order to reap this small benefit, may not be worth the risk of mis-information.

3.3.3. Information from generated hybrid distributions is relatively stable with respect to slight perturbations of the three-state probability outputs.

The perturbation of probabilities was accomplished by altering the underlying $p_{\text{true}}(\chi_{2^\circ} = \text{correct})$ by a small fraction δ, and then adjusting the probabilities of the other two structural states by $\delta/2$ each in the other direction. If the adjustment causes any probability value to exceed the allowable range of probability ($0 \leq p \leq 1$), then the perturbation is aborted, and the original underlying probability $P_{\text{true}}(\chi_{2^\circ})$ is kept as $P_Q(\chi_{2^\circ})$. The intention is not to measure the residual entropy exactly at the particular perturbation, but to gauge the stability of the residual entropy when perturbations are made to the underlying probability distribution $P_{\text{true}}(\chi_{2^\circ})$.

From results summarized in Table 6, it seems that at small levels of δ examined ($\pm 0.05, \pm 0.10$), the residual entropy does not increase drastically (≤ 0.02 nats). While it is important that the $P_Q(\chi_{2^\circ})$ approximate $P_{\text{true}}(\chi_{2^\circ})$, the two distributions do not have to be identical: it is acceptable that $P_Q(\chi_{2^\circ}) \approx P_{\text{true}}(\chi_{2^\circ})$. However, at slightly larger perturbations (± 0.20, in Table 6), the increase in residual entropy becomes significant (≤ 0.07 nats). The degradation in information should become severe as the deviation between $PQ(\chi 2^\circ)$ and $P_{\text{true}}(\chi_{2^\circ})$ increases.

3.3.4. How does one determine if $P_Q(\chi_{2^\circ}) \approx P_{\text{true}}(\chi_{2^\circ})$?

The effectiveness of the three-state probability outputs in generating informative structural distributions rests on the condition that $P_Q(\chi_{2^\circ}) \approx P_{\text{true}}(\chi_{2^\circ})$. It is difficult to establish that this condition holds, for $P_{\text{true}}(\chi_{2^\circ})$ is essentially unknowable. However, one may more confidently assume that the resulting $P_Q(\chi_{2^\circ})$ is good when

Table 5
Residual entropy from three-state probability outputs

Global prediction conditions[a]		Residual entropy[b], (nats) $H(\chi_{(\phi,\psi)^{20^\circ}}\|\chi_{2^\circ P}, Y_{\text{aa}})$	Residual entropy[c], (nats) $H(\chi_{(\phi,\psi)^{20^\circ}}\|\chi_{2^\circ P3}, Y_{\text{aa}})$ and $P_Q(\chi_{2^\circ}) = P_{\text{true}}(\chi_{2^\circ})$	Residual entropy[d], (nats) $H(\chi_{(\phi,\psi)^{20^\circ}}\|\chi_{2^\circ P3}, Y_{\text{aa}})$ and $P_Q(\chi_{2^\circ}) \neq P_{\text{true}}(\chi_{2^\circ})$
Q_3	f_{bad}			
0.65	0.05	3.53	3.49	3.68
0.70	0.04	3.50	3.44	3.66
0.70	0.07	3.51	3.44	3.70
0.75	0.04	3.46	3.39	3.63
0.80	0.04	3.41	3.33	3.58
0.85	0.03	3.37	3.28	3.48
0.90	0.02	3.29	3.20	3.35
0.95	0.01	3.21	3.16	3.21

[a] Q_3, three-state accuracy; f_{bad}, fraction of bad predictions (H ↔ E).
[b] Residual entropy from single-state output.
[c] Residual entropy from three-state probability schemes which have as outputs the true underlying probabilities.
[d] Residual entropy from three-state probability schemes which have randomly generated outputs, without regard to the true underlying probabilities.

Table 6
Residual entropy from three-state probability outputs with perturbation from underlying probabilities

Global prediction conditions[a]		Residual entropy[b], (nats) $H(\chi_{(\phi,\psi)^{2^\circ}}\|\chi_{2^\circ P3}, Y_{aa})$, $P_Q(\chi_{2^\circ}) = P_{true}(\chi_{2^\circ})$	Perturbation[c] δ	Residual entropy[d], (nats) $H(\chi_{(\phi,\psi)^{2^\circ}}\|\chi_{2^\circ P3}, Y_{aa})$, $P_Q(\chi_{2^\circ}) = P_{true}(\chi_{2^\circ}) \pm \delta$
Q_3	f_{bad}			
0.65	0.05	3.49	−0.20	3.55
			−0.10	3.51
			−0.05	3.49
			+0.05	3.49
			+0.10	3.50
			+0.20	3.51
0.70	0.04	3.44	−0.20	3.51
			−0.10	3.46
			−0.05	3.45
			+0.05	3.44
			+0.10	3.44
			+0.20	3.45
0.75	0.04	3.39	−0.05	3.39
			+0.05	3.39
0.80	0.04	3.33	−0.05	3.34
			+0.05	3.34
0.85	0.03	3.28	−0.05	3.29
			+0.05	3.28

[a] Q_3, three-state accuracy; f_{bad}, fraction of bad predictions (H ↔ E).
[b] Residual entropy from three-state probability schemes which have as outputs the true underlying probabilities.
[c] Perturbation made to the highest of the three probability values comprising $P_Q(\chi_{2^\circ})$, when possible (i.e. perturbed values must fall between 0 and 1).
[d] Residual entropy from three-state probability schemes which have as outputs slightly perturbed values from the true underlying probabilities.

the algorithms arise from direct probabilistic methodologies, like the GOR algorithm [35]. Computational methods utilizing black boxes, such as neural nets, while they may be implicitly based on probabilistic models, are trained to reproduce only on–off data. They may make predictions with high accuracy, but the resulting decision scores may be meaningless beyond their function as votes in a majority-rules decision. Simply normalizing them to resemble probabilities does not turn them into viable estimates for $P_{true}(\chi_{2^\circ})$.

A necessary but not sufficient condition on a probability-based $P_Q(\chi_{2^\circ})$ score is that the average of the highest element (or the best guess) $p_Q(\chi_{2^\circ})$ must tend to Q_3, or

$$E\{p_Q(\chi_{2^\circ} = \text{single-state guess})\} = Q_3 \tag{13}$$

It is easy to understand that the expected success of the prediction for each residue must approach the over-all accuracy of the algorithm. Thus, one can do two direct measurements from an application of the algorithm to a large sample (e.g. a non-redundant data set): first, a simple average of the confidence of each prediction, and second, the tally of prediction success to arrive at Q_3. One can imagine that many distributions can satisfy this condition without bearing any relation to the true underlying distribution. However, those that do not satisfy it are inappropriate as coefficients in the hybrid distribution (Eq. (6)).

4. Summary and concluding remarks

4.1. Building optimal probability distributions in the (ϕ, ψ) space for each amino acid given a 2° prediction of a known accuracy.

One can build propensities in backbone structure which incorporates the extent of belief for a particular prediction. The method we introduce in this work takes account of the following factors:

(a) the amino acid identity of the residue of interest;
(b) the global prediction accuracy Q_3;
(c) proportion of bad predictions (H ↔ E) f_{bad};
and any of the following output types:
(d) the 2° state predicted for a residue;
(e) the accuracy index at every prediction instance, q_3; or
(f) the 3-state probability output.

Propensities generated by the methods suggested here reflect states of knowledge – no biophysical relationships are implied. For instance, when the prediction scheme is highly confident in a particular prediction, the optimal propensity distribution resembles the (ϕ, ψ) distribution of backbones known to be in the predicted three-state 2° state. However, at low confidence levels, the optimal distribution is buffered by boosting probabilities outside the canonical regions of the predicted state's phi–psi space, to account for

the increased possibility that the prediction may be wrong. The proper balance between the two extremes is determined by an optimization scheme.

In this work, we introduced a procedure to generate the most informative phi–psi structural distribution given a single-state prediction, through a straightforward optimization of the hybrid coefficient γ. There is some improvement in information extraction when using algorithms which output a confidence index for every prediction, brought about by tailoring the structural distribution to each residue at its particular level of 2° prediction confidence. Finally, when given a three-state probability output, one must first ascertain its probabilistic nature before proceeding further. Because one does not have to carry out an optimization procedure, it is convenient to use the three-state probabilities to generate hybrid structural distributions. Specifically, one simply takes the state probabilities as coefficients in a weighted linear combination of the canonical distributions for H, E, and C. However, the threat of significant degradation of information if the state probabilitiies do not correspond with the underlying true prediction probabilities requires caution.

4.2. Information content of secondary structure predictions of limited accuracy.

We summarize our results relating to the information content of predicted secondary structures and to their potential use in the effort to build tertiary structural models.

a. If the prediction accuracy is below 50%, virtually no advantage is gained from using the 2° prediction to build backbone structural propensities. The high rate of mispredictions cancels any benefit from the correct assignments. The fraction of structural information from secondary structure retained, θ_{ret}, given prediction accuracy $Q3$ and fraction of bad predictions f_{bad} is shown in Fig. 9. The amount of information loss at $Q_3 = 60\%$ is at least

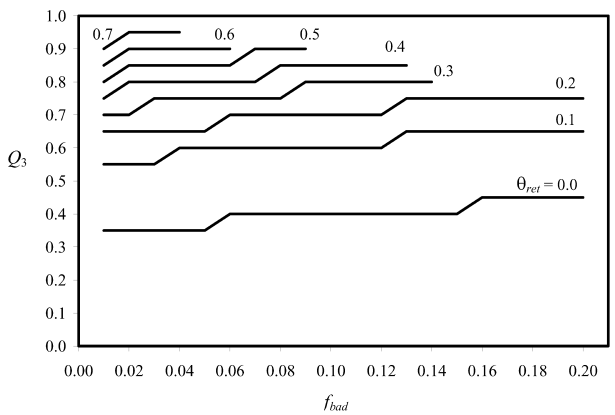


Fig. 9. Contour map showing the fraction of structural information from secondary structure retained, θ_{ret}, given prediction accuracy Q_3 and fraction of bad predictions f_{bad}. The contour lines represent the level of θ_{ret} at different values of Q_3 and f_{bad}.

90%, a sizeable casualty of the still substantial rate of mispredictions. Even at high levels of Q_3, the level of information loss remains high: for instance, at $Q_3 = 95\%$, $f_{\mathrm{bad}} = 0.03$, the amount of information retained is only 70%.

b. Even at perfect prediction accuracy ($Q_3 = 100\%$), the task of specifying the actual backbone structure given its 2° state remains difficult. The amount of information gained by knowing the secondary structure fully, $I_g(\chi_{2^\circ}, Y_{\mathrm{aa}})$ (Eq. (8)), is 0.504 nats, which is only 13.9% of the initial backbone structure uncertainty $H(\chi_{(\phi,\psi)^{20^\circ}}|Y_{\mathrm{aa}}) = 3.622$ nats. The remainder (3.118 nats) is a formidable amount of residual entropy facing any tertiary structure prediction strategy, even after the secondary structure state of every residue in the chain is known.

c. Small improvements in prediction accuracy have a significant effect on the amount of information extracted by the backbone propensity distributions. Fig. 10(A) plots the increase in the fraction of information salvaged by a single point increase in Q_3 vs. the particular Q_3 level. At levels around $Q_3 = 75\%$, the average increase in the fraction of information returned by improving Q_3 by 1% to 76% is around 1.5%. The increase in information increases at higher levels of Q_3, as shown by the positive slope (0.047) of the linear regression line. At Q_3 levels around 80%, the increase in information by a single point improvement in Q_3 is close to 1.8%. Minute increases in Q_3 give a non-trivial improvement in information extraction.

d. Decreasing the fraction of bad predictions causes a corresponding increase in the fraction of information returned by the backbone propensity distributions. Fig. 10(B) plots the increase in the fraction of information salvaged by a single point decrease in f_{bad} vs. the particular f_{bad} level. The increase in information is around 1% when the fraction f_{bad} is decreased by 1% at around $f_{\mathrm{bad}} = 0.04$ level. The improvement in information extraction by improving prediction quality, as gauged by f_{bad}, is not as large as the effect of increasing Q_3. Nonetheless, limiting the number of bad predictions generates more informative structural distributions.

e. Secondary structure prediction schemes which provide an index to quantify prediction confidence at each prediction site yield a better ensemble of structural distributions. Predictions with high confidence index values will have propensity distributions that resemble canonical distributions of H, E, or C (whichever the predicted state is); on the other hand, low confidence index values will have broader distributions, to reflect the uncertainty associated with the particular prediction. A modest increase in the information gain can be expected in tailoring the structural distribution to reflect prediction confidence.

f. We reiterate the importance of the choice of structural partition in maximizing the information gain. Given a limited database, an optimum level of resolution exists for

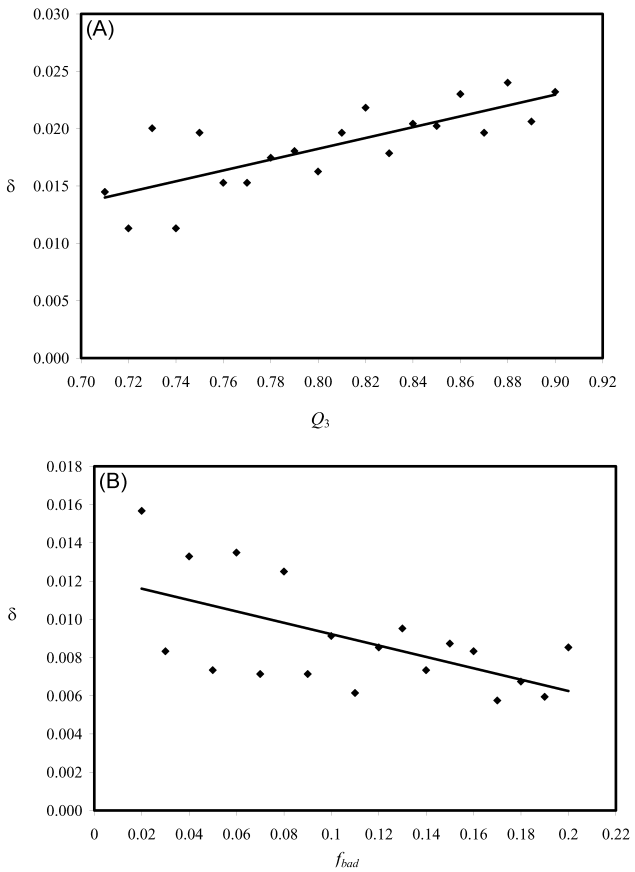


Fig. 10. Increase in information measured upon small changes in accuracy Q_3 and fraction of bad predictions f_{bad}. (A) The fraction of information salvaged (δ) by a single point increase in Q_3 (i.e. 1% increase) vs. the particular Q_3 level. (B) The fraction of information salvaged (δ) by a single point decrease in f_{bad} (i.e. 0.01 decrease) vs. the particular f_{bad} level.

every purpose. Our procedure incorporates a search for this optimum by direct measurement of the information gain as a function of structure discretization. Table 2 shows a typical set of results generated by varying the level of structural detail.

g. More sophisticated algorithms which output three-state probabilities instead of a single-state best guess have the potential to simplify the generation of structural distributions (by a simple linear combination of the canonical distributions of H, E, and C, weighted by the three-state probability output) and marginally increase information gain. However, to reap these benefits, the relation $P_Q(\chi_{2^\circ}) \approx P_{\text{true}}(\chi_{2^\circ})$ must be assured. Any latent information a prediction may have can be erased by unprobabilistic outputs which mislead the search for the native backbone conformation.

Acknowledgements

This work is funded by a grant from the National Library of Medicine of the National Institutes of Health (Grant Number R01 LM06789). The authors thank Dr Igor Kuznetsov for assistance in organizing data for the protein data set used in this work.

References

[1] Solis AD, Rackovsky S. Proteins: Struct Funct Genet 2002;48: 463–86.
[2] Solis AD, Rackovsky S. Proteins: Struct Funct Genet 2000;38:149–64.
[3] Kabsch W, Sander C. Biopolymers 1983;22:2577–637.
[4] Rost B, Sander C. In: Webster DM, editor. Protein structure prediction: methods and protocols. Totowa, New Jersey: Humana Press; 2000.
[5] Rost B. J Struct Biol 2001;134:204–18.
[6] Ramachandran GN, Sasisekharan V. Adv Protein Chem 1968;23: 283–437.
[7] Rackovsky S, Scheraga HA. Acc Chem Res 1984;17:209–14.
[8] Rackovsky S. Proteins: Struct Funct Genet 1990;7:378–402.
[9] DeWitte RS, Shakhnovich EI. Protein Sci 1994;3:1570–81.
[10] Vasquez M, Nemethy G, Scheraga HA. Chem Rev 1994;94:2183–239.
[11] Rost B, Schneider R, Sander C. J Mol Biol 1997;270:471–80.
[12] An Y, Friesner RA. Proteins: Struct Funct Genet 2002;48:352–66.
[13] Marsden RL, McGuffin LJ, Jones DJ. Protein Sci 2002;11:2814–24.
[14] Monge A, Friesner RA, Honig B. Proc Natl Acad Sci USA 1994;91: 5027–9.
[15] Eyrich VA, Standley DM, Felts AK, Friesner RA. Proteins: Struct Funct Genet 1999;35:41–57.
[16] Yue K, Dill KA. Protein Sci 2000;9:1935–46.
[17] Fain B, Levitt M. J Mol Biol 2001;305:191–201.
[18] Rost B, Sander C. J Mol Biol 1993;232:584–99.
[19] Frishman D, Argos P. Proteins: Struct Funct Genet 1997;27:329–35.
[20] Lesk AM, Lo Conte L, Hubbard TJP. Proteins: Struct Funct Genet 2001;Suppl. 5:98–118.
[21] Rost B, Sander C, Schneider R. J Mol Biol 1994;235:13–26.
[22] Emberly EG, Mukhopadhyay R, Wingreen NS, Tang C. J Mol Biol 2003;327:229–37.
[23] Chan AWE, Hutchinson EG, Harris D, Thornton JM. Protein Sci 1993;2:1574–90.
[24] Pedersen JT, Moult J. Proteins: Struct Funct Genet 1997;Suppl. 1: 179–84.
[25] Ortiz AR, Kolinski A, Skolnick J. J Mol Biol 1998;277:419–48.
[26] Ortiz AR, Kolinksi A, Rotkiewicz P, Ilkowski B, Skolnick J. Proteins: Struct Funct Genet 1999;Suppl. 3:177–85.
[27] Xia Y, Huang ES, Levitt M, Samudrala R. J Mol Biol 2000;300: 171–85.
[28] Kihara D, Lu H, Kolinski A, Skolnick J. Proc Natl Acad Sci USA 2001;98:10125–30.
[29] Salamov AA, Solovyev VV. J Mol Biol 1997;268:31–6.
[30] Rychlewski L, Godzik A. Protein Engng 1997;10:1143–53.
[31] Jones DT. J Mol Biol 1999;292:195–202.
[32] Hua S, Sun Z. J Mol Biol 2001;308:397–407.
[33] Yi T-M, Lander ES. J Mol Biol 1993;232:1117–29.
[34] Rost B, Sander C, Schneider R. Comput Appl Biosci 1994;10:53–60.
[35] Garnier J, Gibrat J-F, Robson B. Methods Enzymol 1996;266: 540–53.
[36] Shannon CE. Bell Syst Technol J 1948;27:379–423.
[37] Rooman MJ, Kocher J-PA, Wodak SJ. J Mol Biol 1991;221:961–79.
[38] Lambert MH, Scheraga HA. J Comput Chem 1989;10:798–816.
[39] Feldman HJ, Hogue CWV. Proteins: Struct Funct Genet 2002;46: 8–23.
[40] DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. Proteins: Struct Funct Genet 2003;51:41–55.
[41] Hobohm U, Sander C. Protein Sci 1994;3:522–4.
[42] Rost B. Proteins: Struct Funct Genet 1997;Suppl. 1:192–7.