Journal of Computer-Aided Molecular Design, 7 (1993) 473–501
ESCOM

J-CAMD 210

Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures

Manfred J. Sippl

Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Jakob Haringer Straße 1, A-5020 Salzburg, Austria

Received 14 October 1992
Accepted 26 February 1993

Key words: Protein folding; Protein modelling; Knowledge-based prediction; Molecular force field; Statistical mechanics

SUMMARY

The data base of known protein structures contains a tremendous amount of information on protein-solvent systems. Boltzmann's principle enables the extraction of this information in the form of potentials of mean force. The resulting force field constitutes an energetic model for protein-solvent systems. We outline the basic physical principles of this approach to protein folding and summarize several techniques which are useful in the development of knowledge-based force fields. Among the applications presented are the validation of experimentally determined protein structures, data base searches which aim at the identification of native-like sequence structure pairs, sequence structure alignments and the calculation of protein conformations from amino acid sequences.

INTRODUCTION

The protein folding problem belongs to the most fascinating and important problems in contemporary biology. A satisfying solution would pave the way for a vast number of scientific and technological applications and it would ultimately yield a deep understanding of the protein folding process itself. Over the past decades the folding problem has resisted the attacks of intense theoretical research and it is only recently that the problem seems to give way to the invention of new strategies [1]. These strategies are based on the analysis of known three-dimensional (3D) structures of proteins using statistical procedures whose roots are in statistical physics. The results obtained so far give fair promise that in the near future the unknown structures of a substantial number of sequences will be discovered by computational methods.

In this review we summarize basic concepts and techniques developed in our laboratory, discuss the most important results obtained so far, assess the current state of the art and glance at the applications which are straight ahead of us and within sight. The field has gained momen-

tum. Therefore, the material presented will not only refer to published papers but to a substantial extent covers work in progress. In our presentation we tried to suit the general reader and we hope that we reach a large audience.

GENERAL OUTLINE

The 3D structures of proteins solved by X-ray analysis and nuclear magnetic resonance (NMR) contain a tremendous amount of information on the protein-solvent system. The Brookhaven protein data bank [2] currently holds in the order of 300 structures of unrelated or only distantly related proteins. The number of solved structures exceeds this number by far, since many researchers do not submit their structures to the data base. The public domain structures contain all the information on the relationship between amino acid sequences and associated native folds which we can trust with some confidence.

A long-standing goal in protein structure theory is the development of force fields and energy functions for the protein-solvent system, which could be used to calculate native folds solely from the information contained in amino acid sequences. As outlined in Fig. 1, the goal of the procedures summarized in this review is the extraction of a force field from a data base of known 3D structures, which reasonably models the protein-solvent system. If we succeed in this endeavour then the force field obtained can be employed to determine protein structures by computational means.

Our first goal is to present the basic techniques required for the compilation of knowledge-based force fields from a set of experimental data. The next section summarizes the basic physical principles, comments on the differences between the traditional *semi-empirical* and the new *knowledge-based* approach, and reviews the formal tools required. Subsequently, we focus on the application of these concepts to intramolecular pair interactions and protein-solvent interactions, we present techniques which are useful for the assessment of the predictive power of force fields, and we comment on the current predictive power of the force field developed in our laboratory.

Another section is devoted to applications in structural biology. We present techniques for analysis of conformational energies which are useful in the validation of experimentally determined structures. We outline the basic strategies used for sequence structure alignments required for data base searches. The task is the identification of native-like sequence structure pairs by combining known sequences with known structures. We then address the problem of calculating proteins structures from scratch, a most difficult endeavour due to the astronomical dimensions of conformation space. In the closing section we glance at related work in other laboratories and take an outlook on the developments straight ahead. The material presented in the following section is quite general. Readers not fond of formulas may skip this section in a first reading.

Before we proceed we need to clarify several basic terms. The term *energy* refers to the conformational energy of an individual polypeptide chain and its interaction energy with the surroundings. The energy is a function of the conformational variables of the system (e.g. Cartesian coordinates, distances between atoms, etc.). Taking the derivative of the energy with respect to the conformational variables we obtain the *force field* of the molecule. Energy functions and force fields therefore, are closely related physical quantities which constitute an energetic model of a real physical system. We interchangably use the terms energy function and force field when we

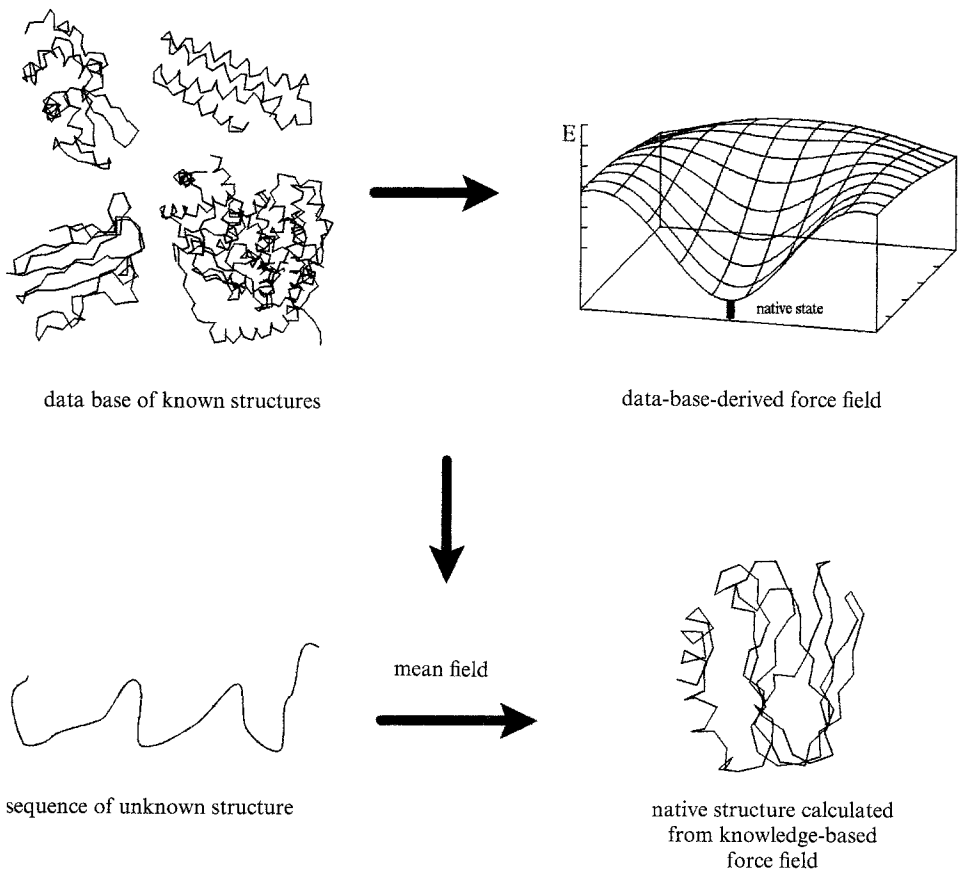

Fig. 1. Outline of the mean field approach to protein folding. The set of available 3D structures of proteins is used to extract a data-base-derived force field. If this attempt is successful the force field can be employed in the computational determination of protein structures.

refer to the energetic model of the system. The term *potential* or *potential function* is often confined to pair interaction energies but here we use this term as a synonym for energy functions.

Energy and force are *microscopic* quantities of the system. In contrast *free energy* is a *macroscopic* quantity referring to an ensemble of a large number of individual molecules. The free energy of a system depends on the energetic features of the individual molecules (enthalpic contributions) as well as on the distribution of the molecules among the various possible microscopic states (entropic contributions).

BASIC PHYSICAL PRINCIPLES IN MOLECULAR FORCE-FIELD DESIGN

The design of macromolecular force fields can be approached at least from two different directions. One strategy, the inductive approach, uses the results obtained from quantum-mechanical calculations on small molecules and thermodynamic and spectroscopic data derived from simple systems. These data are extrapolated to the macromolecular level using the hypothe-

sis that the complex phenomena of macromolecular systems result from the combination of a large number of the same type of interactions as found in the most basic molecular systems. The force fields obtained in this approach are called *semi-empirical* force fields [3–9].

The second strategy, the deductive or knowledge-based approach, departs from the opposite point of view. The forces encountered in large molecular systems are very complicated. To take full account of their complexity the known macromolecular structures are taken as the only reliable source of information. The goal is to extract the forces and potentials stabilizing native folds in solution from the set of known structures. If this approach is successful then the potentials derived from the data base of known structures can be recombined yielding a model for the force field of an amino sequence of a yet unknown structure. The resulting force field would then escort the computational determination of the protein's native fold.

The inductive approach has been extensively explored during the last two decades resulting in the design of several semi-empirical force fields. However, only recently the deductive or knowledge-based approach became a field of intense study due to the growing data base of experimentally determined 3D structures of proteins on the one hand, and the application of powerful concepts of statistical physics on the other.

Two important physical principles are common to both approaches. Both rely on the principle that, in equilibrium, thermodynamic systems attain the global minimum of free energy. The applicability of this principle to protein-solvent systems was first demonstrated by Anfinsen [10] on ribonuclease and was corroborated by unfolding and refolding studies on many proteins [11]. The *folding postulate* is a form of the minimum-energy principle adapted to protein-solvent systems: *In equilibrium the native state of the protein-solvent system corresponds to the global minimum of free energy.*

A macroscopic state of a molecular system contains a large number of individual molecules. At the global minimum of free energy the individual molecules may adopt one particular or many different microscopic states or conformations. In equilibrium, under physiological conditions, soluble globular proteins usually adopt one or several closely related conformations. In the case of short peptides however, the individual molecules are often distributed over a range of dissimilar conformations [12,13].

The distribution of molecules among the microscopic states is governed by *Boltzmann's principle*, the second principle common to the inductive and deductive approaches. This law connects the energy E of a system to its *probability density function* p. Using discrete variables Boltzmann's law can be written in the form

$$p_{ijk} = Z^{-1} \exp(-E_{ijk}/kT) \tag{1}$$

where k and T are Boltzmann's constant and the absolute temperature, respectively, and the subscripts i, j, k,... correspond to the variables of the system. The quantity Z

$$Z = \sum_{ijk} \exp\,(-E_{ijk}/kT) \tag{2}$$

is called Boltzmann's sum or partition function.

The general goal in statistical mechanics is the calculation of the partition function Z and probabilities p_{ijk} from a given function E_{ijk}. Then the macroscopic thermodynamic quantities of

the system can be derived from Z and p_{ijk}. This approach in statistical mechanics faces two very difficult problems which have to be solved successively: (1) It requires the design of an energy function which reasonably models the system and (2) It demands the calculation of the partition function by analytical or numerical techniques. In general the calculations succeed only for very simple systems or for systems which have only a small number of possible states.

In this form (Eqs. 1 and 2) Boltzmann's law is used to calculate observable quantities from first principles. The important role of experimental data in the system of interest is to check the calculations and it is generally agreed that we have gained a deep physical understanding of a real physical system if the calculations match the experimental data. Here again we encounter the inductive approach.

The deductive approach starts from the inverse Boltzmann law [14], i.e.

$$E_{ijk} = -kT\ln(f_{ijk}) + kT\ln Z \tag{3}$$

where the energy function E_{ijk} is called *potential of mean force*. The energy of the state labelled by i, j, k,... is derived from the *relative frequencies* f_{ijk} obtained from measurements on this state. The relative frequencies f_{ijk} are equivalent to the probability densities p_{ijk} in the sense that, in the limit of infinitely many observations, relative frequencies converge to the probability densities, $\lim_{n\to\infty} f_{ijk} \equiv P_{ijk}$. Note that $\sum_{ijk} f_{ijk} = \sum_{ijk} p_{ijk} = 1$, i.e. relative frequencies and probability densities are *normalized.*

The partition function Z cannot be obtained directly from experimental measurements. However, at constant temperature, Z is a constant and does not affect the energy difference between particular states. The mean force energies are determined up to this constant term. Choosing Z = 1 so that

$$E_{ijk} = -kT\ln(f_{ijk}) \tag{4}$$

we obtain

$$Z = \sum_{ijk} \exp(-E_{ijk}/kT) = \sum_{ijk} \exp(kT\ln(f_{ijk})/kT) = \sum_{ijk} f_{ijk} = 1 \tag{5}$$

which shows that Z = 1 is consistent with the definition of the partition function in Eq. 2.

The inverted Boltzmann principle is indeed remarkable. The forces which act in solute-solvent systems are overwhelmingly complex. Interactions between atoms may be simple combinations of basic physical forces in vacuo, but in dense liquid systems the direct interactions are heavily disturbed and distorted by the surrounding molecules. Nevertheless, Boltzmann's inverse principle enables us to derive the combined action of all these forces in a single strike. The only requirement is that we sample a sufficient amount of experimental data of the system of interest. In the case of protein structures, of course, sampling of a sufficient amount of experimental data is an enormous task. The collection of protein structures solved over the last three decades is a rich substrate for Boltzmann's inverse principle.

However, the successful application of Boltzmann's inverse principle requires additional ingredients [14]. We have to define an appropriate *reference system* which can serve as a reference frame for the energies derived from Eq. 3. To facilitate the following discussion (but without loss

of generality) we assume that the system of interest can be described by a set of four variables i, j, k, l. Often we want to be able to consider some aspect of the system associated with a certain subset of variables only, e.g. k and l, keeping i and j at particular values. In this case we have to normalize the subsystem defined by particular values of i and j using

$$p_{kl}^{ij} = p_{ijkl}/p_{ij} \tag{6}$$

Here we used the Einstein convention of summation $p_{ij} = \sum_{kl} p_{ijkl}$, i.e. a missing subscript indicates summation over that subscript. Then

$$\sum_{kl} p_{kl}^{ij} = \sum_{kl} p_{ijkl}/p_{ij} = p_{ij}/p_{ij} = 1 \tag{7}$$

i.e. the subsystem is properly normalized. In this way the total system is decomposed into subsystems which are defined by the particular values of the superscripts i and j and whose independent variables are indicated by the subscripts k and l. Then the mean force energies of the subsystems are

$$E_{kl}^{ij} = -kT\ln(p_{kl}^{ij}) \tag{8}$$

In order to successfully apply the mean force concept we have to extract the general energetic characteristics of the system from the data base. These general characteristics are again mean force energies which serve as an energetic frame of reference [14]. Generally a useful reference system can be obtained by averaging over a particular set of subsystems. Continuing with the four-variable case we choose the subsystem

$$p_{kl} = \sum_{ij} p_{ijkl} \tag{9a}$$

where

$$\sum_{kl} p_{kl} = \sum_{ijkl} p_{ijkl} = 1 \tag{9b}$$

which is an average over all subsystems p_{kl}^{ij} with respect to i and j and we obtain the associated mean force energy

$$E_{kl} = -kT\ln(p_{kl}) \tag{10}$$

We are now in a position to compare the mean force energies of the subsystems to the reference system. We obtain the *net potential of mean force* ΔE_{kl}^{ij} from

$$\Delta E_{kl}^{ij} = E_{kl}^{ij} - E_{kl} = -kT\ln(p_{kl}^{ij}/p_{kl}) \tag{11}$$

The net mean force energy [14] is the difference between mean force energy and the mean force of reference. In subtracting the reference from the mean force we remove all forces which are

common to all subsystems. The net mean force energy contains only those components which are particular to the subsystem labelled by i and j.

The forces are obtained from the energies by taking the derivatives with respect to the variables in the subsystems

$$F_i^{ij}(k) = \frac{\delta}{\delta k}\Delta E_{kl}^{ij}, \quad F_k^{ij}(l) = \frac{\delta}{\delta l}\Delta E_{kl}^{ij} \tag{12}$$

Since in our formulation the mean force energies are functions of discrete variables, the derivatives have to be evaluated numerically. This by no means restricts the generality of the approach. The difference between discrete and continuous functions, when interpreted as models of physical systems, is largely a technical issue. Discrete functions may be represented by series of continuous functions which in turn can be differentiated analytically.

Up to this point our discussion has been very general and not restricted to protein-solvent systems. We are now ready to apply these concepts to proteins, which requires the identification of useful state variables, subsystems and reference frames.

KNOWLEDGE-BASED MOLECULAR FORCE FIELDS FOR PROTEIN-SOLVENT SYSTEMS

Mean force potentials for pair interactions

Native structures of proteins are stabilized by the mutual *intramolecular* interactions among the various protein atoms and the *intermolecular* interactions of these atoms with the surrounding solvent molecules. A particular intramolecular interaction depends on several variables as shown schematically in Fig. 2. The interaction is a function of the two participating amino acids a and b, the atom types c and d, the separation k of a and b along the amino acid sequence [15] and the spatial distance r between atoms c and d.

Atomic interactions are generally thought to be symmetric. However, interactions are symmetric only in the case of free particles. Amino acid sequences are linear chains directed from the N- to the C-terminus. Hence the inverse of a sequence grossly differs from the original. This asymmetry of the polypeptide backbone is passed on to the intramolecular interactions. Hence, the amino acid pairs (a,b) and atom pairs (c,d) are ordered quantities, so that $(a,b) \neq (b,a)$ and $(c,d) \neq (d,c)$ [14].

With the exception of the spatial distance r all variables are discrete. Since our goal is the compilation of relative frequencies, the continuous variable r has to be sampled in intervals, i.e. r is treated as a discrete variable. The set of variables (a, b, c, d, k, r) thus defined constitutes a model for pair interactions in proteins [14]. All the interactions we encounter in the structure of a particular protein can be described by this set of variables. The set a = Val, b = Glu, c = C^{α}, d = C^{β}, k = 3 and r = 10 (e.g. corresponding to the distance interval 5.0–5.5 Å) is an example for a particular state.

Using this model the compilation of relative frequencies f_{abcdkr} from a data base of known protein structures is straightforward. Our next task is the choice of useful subsystems and the definition of an appropriate reference frame by a proper partitioning of variables. The variables a, b, c, d and k define the nature of the interaction. They are constant if we observe a particular interaction in different folds but they change if the amino acid sequence is changed. On the other

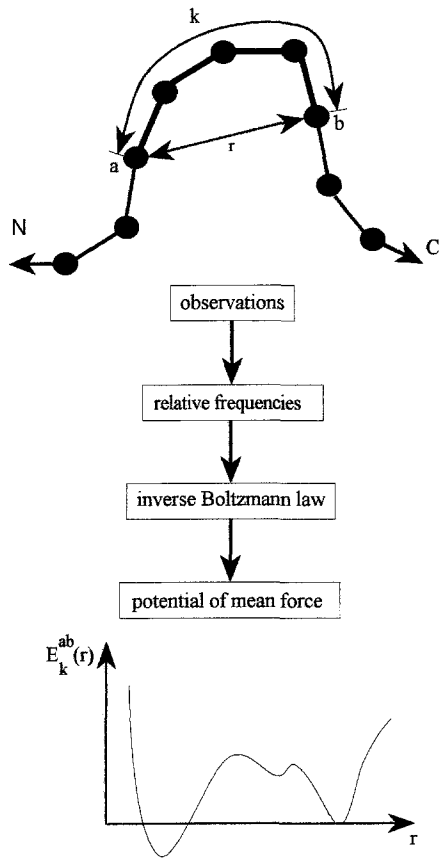

Fig. 2. Important variables of intramolecular pair interactions of proteins and the extraction of mean force energies. A particular intramolecular pair interaction depends on the amino acids a and b, atom types c and d (not shown), the separation of a and b along the sequence, and the spatial distance r between atoms c and d. In contrast to pair interactions of free particles the intramolecular potentials are asymmetric with respect to residues a and b and atoms c and d due to the asymmetry of amino acid sequences. The relative frequencies of particular pair interactions are functions of r. They are sampled in discrete intervals of r by scanning the proteins in the data base for particular values of a, b, c, d and k. When data acquisition is completed the inverse Boltzmann principle is used to transform relative frequencies to mean force potentials.

hand the variable r is conformation dependent but independent of the variation of the amino acid sequence.

The subsystems associated with these variables are the pair interaction potentials for amino acid pairs (a,b) and atom types (c,d) of sequential separation k. These potentials are functions of r. The energetic frame of reference is obtained by averaging over the amino acid pairs (a, b) [14]. The net mean force potential for a particular pair interaction is obtained from

$$\Delta E_r^{abcdk} = E_r^{abcdk} - E_r^{cdk} = -kT\ln(f_r^{abcdk}/f_r^{cdk}) \tag{13}$$

Figures 3 and 4 show several examples of mean force potentials compiled from a data base

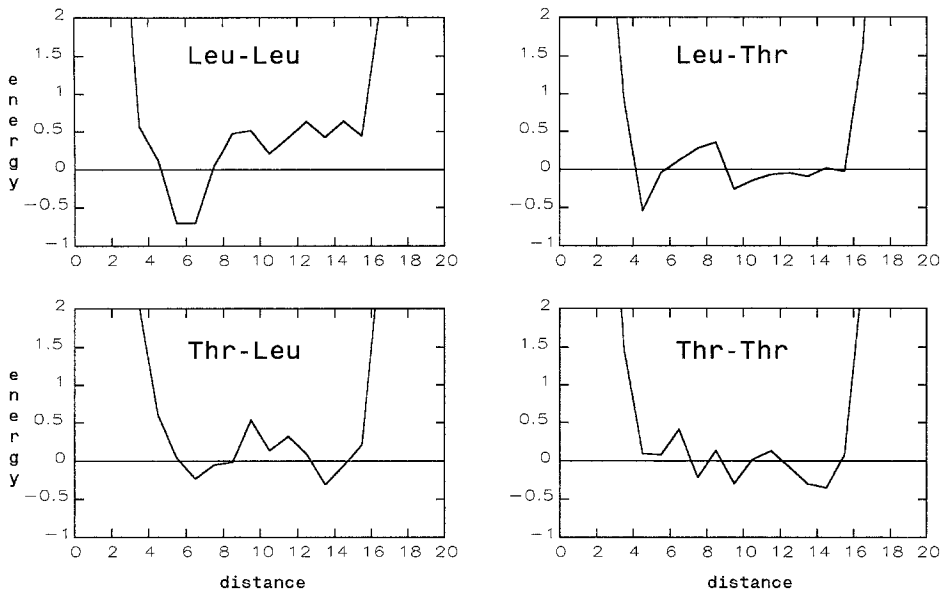

Fig. 3. Examples of C^{β}-C^{β} mean force potentials for separation k = 4 along the amino acid sequence. Energies are scaled in the form E/kT. For small values of k particular values of r correlate strongly with local structures. The deep minimum of Leu-Leu at r ≈ 6 Å reflects the strong preference for α-helical structures. In contrast, α-helical conformations are energetically unfavourable for Thr-Thr. The mixed pairs are intermediate. Thr-Leu, for example, has two minima of comparable depth at α-helical and extended conformations.

containing 160 individual protein chains corresponding to a total of ≈ 40 000 amino acid residues.

Finally, the mutual mean force acting on atoms c and d of amino acids a and b at sequential separation k and at spatial distance r is

$$F_r^{abcdk} = \frac{\delta}{\delta r}\Delta E_r^{abcdk} \tag{14}$$

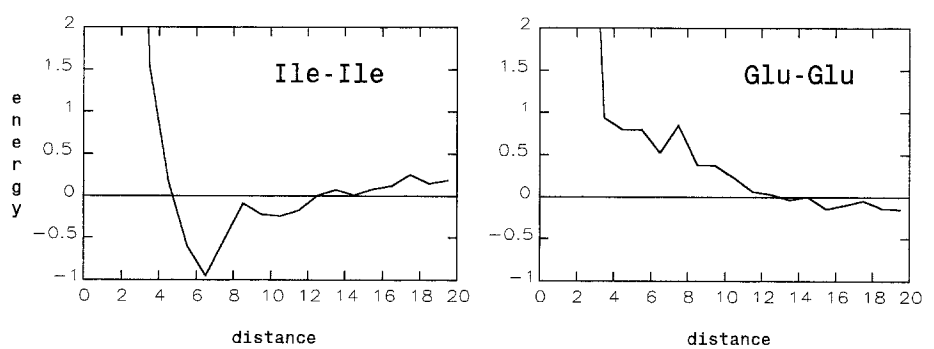

Fig. 4. Examples of C^{β}-C^{β} interactions for large separations (k > 10) along the sequence. Ile-Ile has a minimum at close contacts which is characteristic of hydrophobic pairs. Close contact energies of Glu-Glu are unfavourable. The potential has a negative energy for large distances only.

Complexity of pair interactions

The set of variables used to model the pair interactions produces a description of protein structures which is of considerable complexity. There are 400 amino acid pairs and, restricting the model to the backbone atoms N, C^{α}, C′, O and C^{β}, we have 25 different atom pairs. Thus there are 10 000 individual potentials for a particular sequential separation k.

Since we have to derive the relative frequencies from a data base, we run into the problem of sparse data. For small k values the number of observations we may expect for a particular interaction is in the order of N/400, where N is the total number of amino acids in the protein structure data base. For N = 40 000 the average number of measurements is 100, but for rare amino acid pairs the actual number is much smaller and the potentials derived from the raw data will be quite unreliable. Procedures have been developed which overcome this problem, allowing the extraction of useful potentials even in the case of extremely sparse data [14].

Surface energy

Proteins strongly interact with the surrounding solvent. David Eisenberg and co-workers [16,17] and Bowie et al. [18] demonstrated that solvent exposure of amino acids is a sensitive parameter which can be used to model the energetic features on the protein-solvent boundary. These parameters as well as the solvent preference of amino acids derived independently by Sander and co-workers [19,20] capture an important feature of the protein-solvent system.

Mean force potentials for the interactions of the protein atoms with solvent molecules can be obtained in the same way as we have demonstrated for the intramolecular interactions. However, solvent molecules are mobile and only a small fraction of the solvent molecules can be located in X-ray determinations. Hence, the required experimental information is missing and we have to resort to an indirect approach.

The neighbourhood of atoms buried in the protein interior is fully occupied by protein atoms. On the other hand, for surface atoms only a fraction of the surrounding volume is filled by protein atoms. The complement of this volume is occupied by solvent molecules. The variable s, corresponding to the number of protein atoms in a sphere of radius R, serves as a quantitative measure of the solvent exposure of an atom located at the centre of this sphere. In addition, solvent exposure depends on the atom type c and amino acid a. Again, using Boltzmann's inverse principle the mean force of solvent exposure is obtained from the relative frequencies f_{acs} by

$$\Delta E_s^{ac} = -kT\ln(f_s^{ac}/f_s^c) \tag{15}$$

Calculation of conformational energies for sequence structure pairs

Once the net potentials of mean force are compiled from a data base, they can be used to calculate the conformational energy of a protein. In Fig. 5 we outline the computations involved. The total intramolecular pair interaction energy of an amino acid sequence S in some conformation C is

$$\Delta P(S,C) = \sum_{ij} \sum_{cd} \Delta E_r^{a(i),b(j),c,d,k} \tag{16}$$

The summation is over all positions i and j in the amino acid sequence and over all atom pairs (c,d). a(i) and b(j) are the amino acids at sequence positions i and j, respectively and $k = |i - j|$.

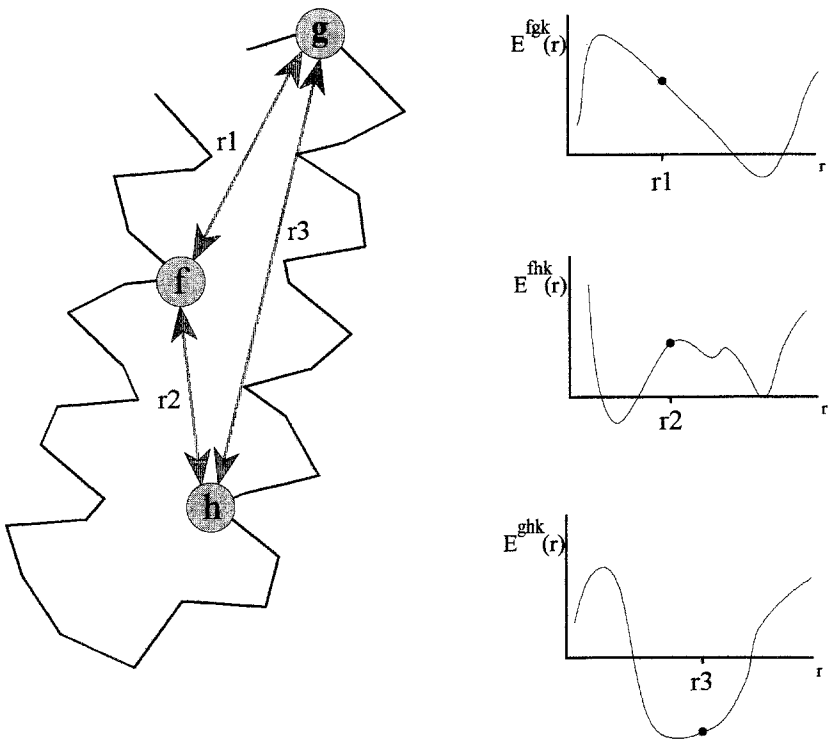

Fig. 5. Outline of the computation of the total pair interaction energy of proteins. The distances between atoms are calculated. The residue types a and b, atom types c and d, the separation k along the sequence determine the type of potential used to evaluate the energy at distance r. The total pair interaction energy is obtained by summing over all atom pairs in the molecule.

The distance interval r is derived from the Cartesian coordinates of atoms c and d of amino acids a(i) and b(j), respectively.

The total surface energy is

$$\Delta S(S,C) = \sum_{i} \Delta E_s^{a(i),c} \tag{17}$$

and finally the total combined pair and surface energy is

$$\Delta E(S,C) = \omega_p \Delta P(S,C) + \omega_s \Delta S(S,C) \tag{18}$$

where ω_p and ω_s are the relative weights of the individual contributions. The molecular force field F(S,C) is obtained by differentiating $\Delta E(S,C)$ with respect to the conformational variables of C.

The most important question is, of course, whether $\Delta E(S,C)$ is a reasonable energetic model for protein-solvent systems. To address this question we need reliable tools for the quality assessment of macromolecular force fields which is the subject of the next section.

ASSESSMENT OF THE PREDICTIVE POWER OF FORCE FIELDS

The successful development of reasonable energetic models of protein-solvent systems strongly depends on techniques which can be used to judge the quality of the model at each stage of development. The folding postulate requires that the native fold has lowest energy among all other alternatives. This is a necessary and sufficient criterion which must hold for all sequences, including the degenerate case of sequences which have energies similar to the global minimum for a range of dissimilar conformations.

A rigorous proof for a particular sequence S_p requires the computation of $E(S_p,C_q)$ for all possible conformations q = 1,...,n which is computationally prohibitive, since n is astronomically large. However, if we look at a subset of conformations which includes the native fold C_N of S_p, then the folding postulate requires $E(S_p,C_N) < E(S_p,C_q)$ where q now labels the structures in the subset and $q \neq N$. If this condition is violated then it will be impossible to identify the native fold by minimizing the energy $E(S_p,C_q)$. The condition is necessary but not sufficient, since there may be structures outside the test set which have lower energy than the native fold.

Novotny et al. [21] were the first to apply this principle in their pioneering study on the predictive power of semi-empirical force fields. Using two proteins of different architecture they prepared two native sequence structure pairs (S_1,C_1) and (S_2,C_2) and two misfolded pairs (S_1,C_2) and (S_2,C_1) by exchanging the sequences. They demonstrated that the force field applied was unable to distinguish misfolded from native pairs. In contrast, an early prototype of a knowledge-based force field as defined in Eq. 16 has no difficulties in picking the correct native pairs, i.e. $\Delta E(S_1,C_1) << \Delta E(S_1,C_2)$ and $\Delta E(S_2,C_2) << \Delta E(S_2,C_1)$ [22].

Hide-and-seek on a polyprotein

Identification of the native sequence structure pair is, of course, harder if the native fold is hidden among a large number of nonnative decoys provided the set of alternative structures contains genuine protein folds similar to folds determined by experimental methods. The set of all possible fragments C_q of length l derived from the structures in the data base is a convenient source of alternative conformations. Here l is the length of the test sequence S_p whose native fold C_N is hidden in the pool of fragments. The guiding principle in our search is the energy $\Delta E(S_p,C_q)$. The native sequence structure pair is successfully identified if $\Delta E(S_p,C_N) < \Delta E(S_p,C_p)$ for all $q \neq N$.

In our early studies the knowledge-based force field was able to identify the native fold for a large number of proteins [22], indicating the applicability of the mean force approach to protein folding. The results obtained have been used recently as a bench-mark to judge the quality of alternative force fields or sets of preference parameters [23]. It should be noted however, that the force field used by Hendlich et al. [22] was an early prototype.

The number of decoys obtained from the data base depends on the length of the test sequence. This number is large for small proteins. On the other hand only few fragments of the size of large proteins can be obtained. Hence, a successful identification is hard for the smaller proteins in the test set but for the largest proteins the test is insignificant. Moreover, the results obtained for proteins of different length are not comparable. This is a severe drawback prohibiting the calculation of a meaningful performance measure over the whole test set.

To circumvent this problem we constructed a polyprotein from the structures in the data base

(manuscript in preparation). Starting at the N-terminus the test sequence S_p glides along the polyprotein chain, one amino acid at a time, and the energy $\Delta E(S_p,C_q)$ is recorded at each position q. Finally the native pair energy $\Delta E(S_p,C_N)$ is calculated. Again, if the native pair energy is less than $\Delta E(S_p,C_q)$ for all q then the native fold is successfully identified. In constructing the polyprotein, particular care has been taken in designing the linker regions between protein modules. This ensures that fragments containing linker regions do not violate general characteristics of protein folds making the identification of the native fold as hard as possible.

Our current version of the polyprotein is composed of 160 experimentally determined folds with a total length in the order of $n \approx 40\,000$ residues. The number of folds encountered by each test sequence is $n - 1$ and since $1 << n$ these numbers are close to the total length n of the polyprotein for all test sequences. Thus the problem of identifying the correct fold using hide-and-seek on a polyprotein is comparably hard for all sequences in the test set.

The energies sampled along the polyprotein can be transformed to z-scores

$$z_{pq} = (\Delta E(S_p,C_q) - \overline{E}_p)/\sigma_p \tag{19}$$

where $\overline{E}_p = \sum_q \Delta E(S_p,C_q)$ is the average energy of all conformations C_q with respect to sequence S_p and σ_p is the corresponding standard deviation. The z-score $z_{p,N}$ of the native fold is a quantitative measure of the predictive power of a force field with respect to test protein p. The average

$$\overline{z} = \frac{1}{n}\sum_{p}^{n} z_{p,N} \tag{20}$$

obtained from the test set of n sequences is a measure of the overall performance of the force field.

Predictive power of mean force potentials

Table 1 summarizes the predictive power of the mean force potentials. The test set consists of 157 individual protein chains of known structure. The force field used in this study was compiled for the C^β atoms only. For all proteins in the test set hide-and-seek was performed on a polyprotein of $\approx 40\,000$ residues, and the number of decoys is of the same order of magnitude. The pair energy successfully identifies 148 native conformations. The success of the surface energy is lower (134 successful trials) but it is noteworthy that the average native z-score for the surface energy is higher although a smaller fraction of native folds is successfully identified.

The combined force field identifies all native folds. The average native z-score almost doubles when the pair and surface energies are combined. This indicates that the information contained in the two terms is complementary. The results also indicate the usefulness of the average native

TABLE 1
PREDICTIVE POWER OF MEAN FIELDS

Force field	Total correct[a]	% correct	Average native z-score
Pair	148	94	−5.40
Surface	134	85	−6.33
Combined	157	100	−8.02

[a] Number of proteins whose native fold is successfully identified. The number of chains in the test set is 157.

z-score obtained from the polyprotein as a performance measure of force fields. Even if all native folds are identified correctly improvement or deterioration of the force field is reliably monitored by this parameter.

COMPUTATIONAL DETERMINATION OF PROTEIN STRUCTURES

From the results obtained on the test set we may assume with some confidence that the mean-force approach is a powerful tool for the development of force fields for the protein-solvent system. At the same time we should be aware of the fact that our present force field is neither complete nor optimized. The pair interactions are calculated for the backbone atoms only and the surface term is a crude approximation of the actual protein-solvent interactions. In addition the quality of the force field depends on several parameters. Such parameters are the grid size used to sample intramolecular distances, cut-off distances for the calculation of energies, and approximations to the probability density functions from relative frequencies, to mention only a few. Optimization of these parameters is laborious but with the help of hide-and-seek the task is manageable.

Theoretical work reaches a state of maturity when the concepts developed can be applied to verify experimental data and when the calculations correctly predict the state of a physical system ahead of experimental observation. The following sections are devoted to applications of the current mean field in protein structural problems. Before we start we introduce the notion of *sequence structure space* which will be useful in the discussion of some peculiarities of protein folding and its twin the *inverse folding* (e.g. Ref. 16).

Folding and inverse folding in sequence structure space

At a particular instant of time a protein molecule is identified by two basic features, its amino acid sequence and its 3D fold. In Fig. 6 we plot the space of all amino acid sequences versus the space of all conformations. Each point in this 2D representation corresponds to a sequence structure pair, i.e. a particular sequence S_p folded in a particular conformation C_q, where the subscripts p and q run over all possible sequences and conformations, respectively.

By traversing sequence structure space across vertical lines we explore the conformational space of a particular sequence S_p. Folding of proteins proceeds along such lines of constant sequence. Travelling along horizontal lines we encounter all sequences folded in conformation C_q. Movements along such lines require the instantaneous replacement of one or more amino acids whereby the conformation is not allowed to relax. Inverse folding is confined to such lines of constant conformation.

Each pair (S_p,C_q) has an associated energy $E(S_p,C_q)$ which in a sense reflects the fitness of the pair. Some points in sequence structure space correspond to native pairs. According to the folding postulate, native sequence structure pairs have the lowest energy along lines of constant sequence. Along such lines we may find only one native pair corresponding to a distinct global energy minimum. In equilibrium almost all molecules of an ensemble of this sequence will be found at this point in sequence structure space. This corresponds to the native state of a globular protein with the free energy of the system at its global minimum.

However, along lines of constant sequence we may also encounter several pairs whose energies are close to the global minimum. Such sequences do not fold to a unique structure. In equilibrium

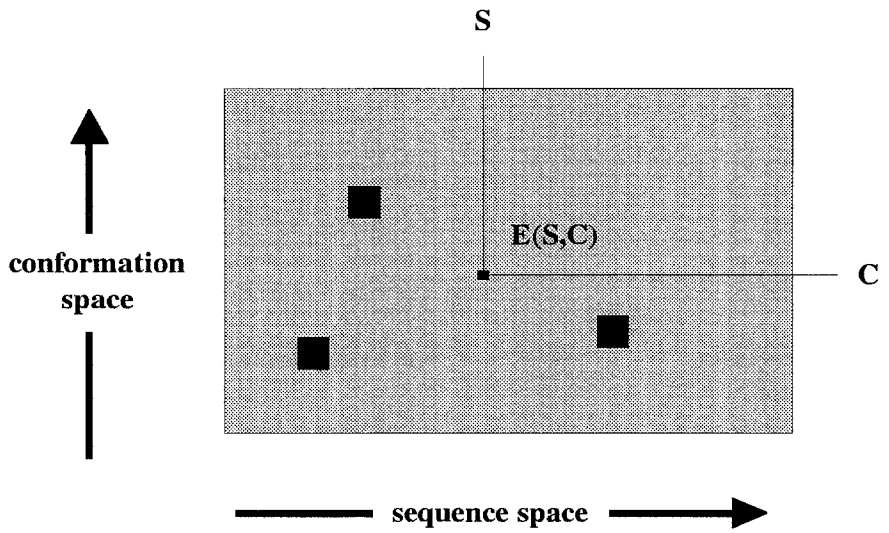

Fig. 6. Sequence structure space. In this representation the universe of all amino acid sequences is plotted against the universe of all conformations. Each point in the resulting plane corresponds to a particular sequence S_p folded into a particular conformation C_q. For each point (S_p,C_q) there is an associated energy $E(S_p,C_q)$, which in a sense reflects the fitness of the sequence structure pair. When plotted in three dimensions the energy forms a landscape over the sequence structure plane. Some points in the plane correspond to native sequence structure pairs (represented by solid squares). According to the folding postulate these points correspond to global minima along the vertical lines (i.e., constant sequence) through these points.

the individual molecules will be distributed over these low-energy pairs. This situation often applies to short polypeptides where, in equilibrium, individual chains travel along paths of constant sequence resting at conformations of comparable energy.

The protein folding problem requires the determination of the native structure from the information contained in the amino acid sequence alone by computational or theoretical methods. In terms of sequence structure space the problem can be approached in two successive steps. The first step requires the construction or design of an energy function whose global minima along lines of constant sequence correspond to the native state(s) of amino acid sequences. If we succeed in this endeavour the second step requires that we locate the global minima along lines of constant sequence.

Inverse folding pursues a different goal. Here the problem is to identify sequences which fit into a given conformation. Hence, in the inverse folding problem sequence structure space is explored along lines of constant conformation. This does not correspond to a physical process. Sequences change in the course of evolution but these changes take place on the DNA level and not in a protein folded into a rigid structure.

Proteins frequently adopt similar 3D folds even if they are completely unrelated on the sequence level. The current data bases hold a large number of known sequences but only a relatively small number of known structures. The native folds of a substantial number of these sequences will be similar to a known structure. The goal of inverse folding is to identify the sequences in the data base which fit into a known fold. As in the original folding problem, the guiding principle is the energy or fitness of sequence structure pairs.

There is, however, a subtle difference between the original folding problem and its inverse. In the original folding problem the native sequence structure pairs correspond to global minima along lines of constant sequence. However, there is no physical principle which guarantees that global minima along lines of constant conformation correspond to native sequence structure pairs.

This puts some constraints on the applicability of inverse folding. In general a particular conformation does not correspond to the native fold of an amino acid sequence. Noncompact and random-coil chains belong to this class. However, there are less trivial cases. To illustrate this point let us start at a native pair (S_1,C_1) in sequence structure space, see Fig. 7. The folding postulate guarantees that along constant S_1 the energy $E(S_1,C_1)$ is a global minimum. However, along constant C_1 we may find sequences which have lower energy, where S_2 is the sequence of most favourable energy. The pair (S_2,C_1) is not necessarily native. Minimizing along constant S_2 we may find some conformation C_2 as the native fold of S_2. The conclusion is that a search along lines of constant conformation does not necessarily locate a native sequence structure pair. This is only the case if at the same time the pair is at a global minimum along the corresponding line of constant sequence.

This has some consequences for the applicability of energy calculations to proteins. In general it will not be possible to design proteins by proposing a desired 3D fold followed by the construction of a sequence compatible with this fold by minimization in sequence space. The optimal sequence may form a native pair with a different conformation, or the sequence may be unstable

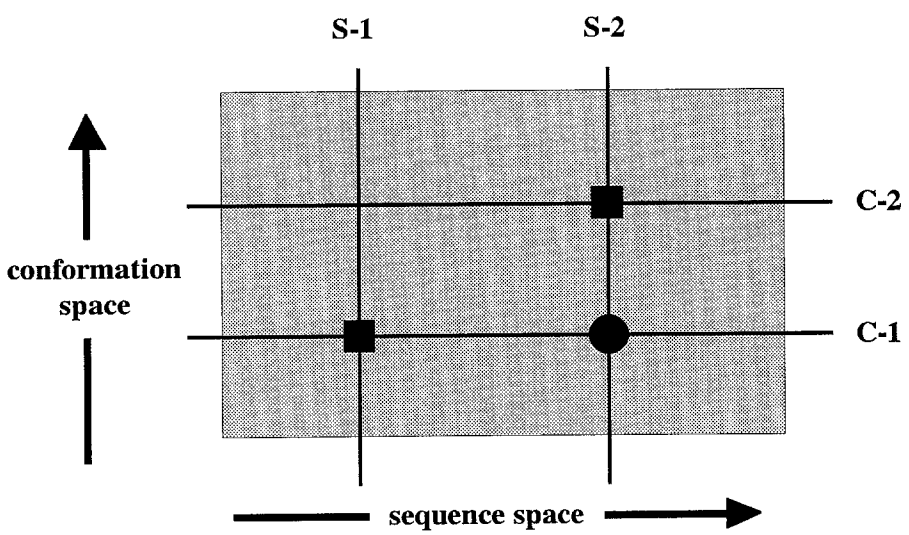

Fig. 7. Folding and inverse folding in sequence structure space. If point (S_1,C_1) corresponds to a native sequence structure pair, then the folding postulate guarantees that $E(S_1,C_1)$ is a global minimum along the straight line S_1 = constant. However, we have no physical principle at hand which prevents that along C_1 = constant we find a sequence S_2 and hence pair (S_2,C_1) (filled circle) of lower energy, i.e. $E(S_2,C_1) < E(S_1,C_1)$. Even if (S_2,C_1) is a global minimum along C_1 = constant, the pair is not necessarily native. The chances are that S_2 forms a native pair with a different conformation C_2. In summary, if the energy relation between the pairs is of the form $E(S_1,C_1) < E(S_2,C_1)$, minimization along lines of constant conformations in general will not yield physically meaningful results.

and may not fold to a unique structure. Hence, a subsequent search in conformation space is necessary to keep the designed sequence constant. If this search reveals structures of lower energy the designed sequence structure pair is likely to be unstable. We encounter a similar problem in sequence structure alignments, which we will discuss below. We now turn to applications of knowledge-based mean fields.

Validation of experimentally determined structures

Validation of experimentally determined structures is a difficult problem [24,25]. Several mistraced structures have been discovered by repeated structure determination on the same molecule. Validation of experimentally determined folds is one of the most important duties of computational techniques in structural biology. It is only recently that the development of computational tools has reached a level of sophistication where this task can be approached with some success. The profile method invented by Eisenberg and co-workers [17] belongs to the most advanced methods in this field.

At the present stage of development the knowledge-based mean field can be used to analyse the distribution of energies in experimentally determined structures. The resulting profiles display native-like or nonnative-like features of protein folds. Conformational analysis is performed on the pair energy matrix (Fig. 8). Each element e_{ij} of this matrix corresponds to the pair interaction energy of amino acids i and j in the amino acid sequence. The matrix can be analysed in several

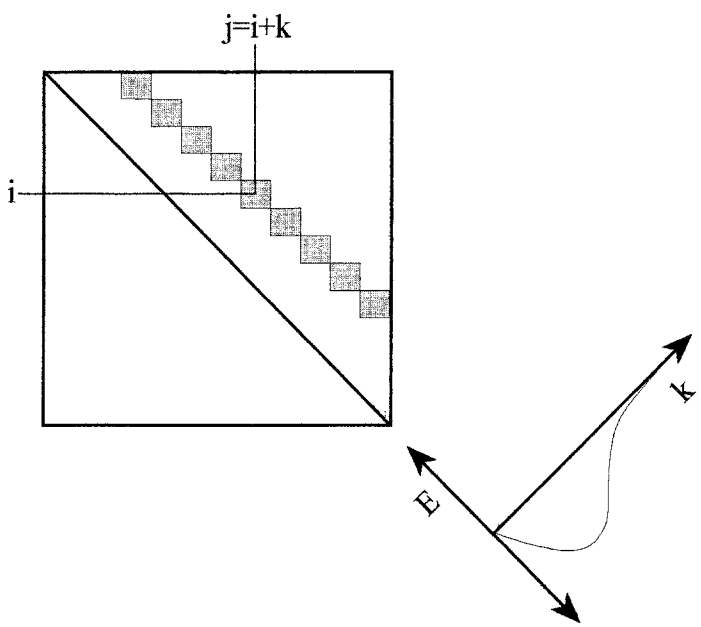

Fig. 8. k-profile calculated from the pair interaction matrix. Elements e_{ij} of the energy matrix correspond to the pair interaction energy of two residues at sequence positions i and j in the amino acid sequence. The sum over the subset of elements $\Sigma_i e_{i,i+k}$ collects all energies of constant sequence separation k. When plotted as a function of k these values constitute an energy profile for a particular sequence structure pair. As indicated k-plots of native-like sequence structure pairs have a pronounced minimum at short sequence separations.

ways. As shown in Fig. 8 the sum over diagonals yields the energy content of the matrix as a function of sequential separation k. The values are normalized by the number of interactions for a particular value of k to account for the variations in sequence length.

Native sequence structure pairs have a pronounced minimum at small sequential separations whose depth is in the order of −0.1. Maxima or positive values indicate a nonnative sequence structure pair. Figure 9 presents several k-plots calculated from structures obtained from X-ray analysis. Two of the plots (2GN5, 1PTE) have a nonnative appearance.

A more detailed view of the energy distribution within a conformation is obtained from the total interaction energies of individual residues along the amino acid sequence (Fig. 10). The resulting plot serves as an energetic finger print. Figure 11 shows several examples. In the case of native folds the energy is below zero for most sequence positions and only occasionally we encounter small positive peaks. Large positive peaks indicate strained parts of the chain. 2GN5, for example, has positive peaks along the whole sequence. The electron-density interpretation of this protein appears to be problematic. We found several entries (e.g. 1PTE) in the Brookhaven data base whose total mean force energy and profiles are nonnative. A detailed analysis of these structures is in progress.

Energy profiles are powerful tools. They can be used to identify incorrect chain tracings and

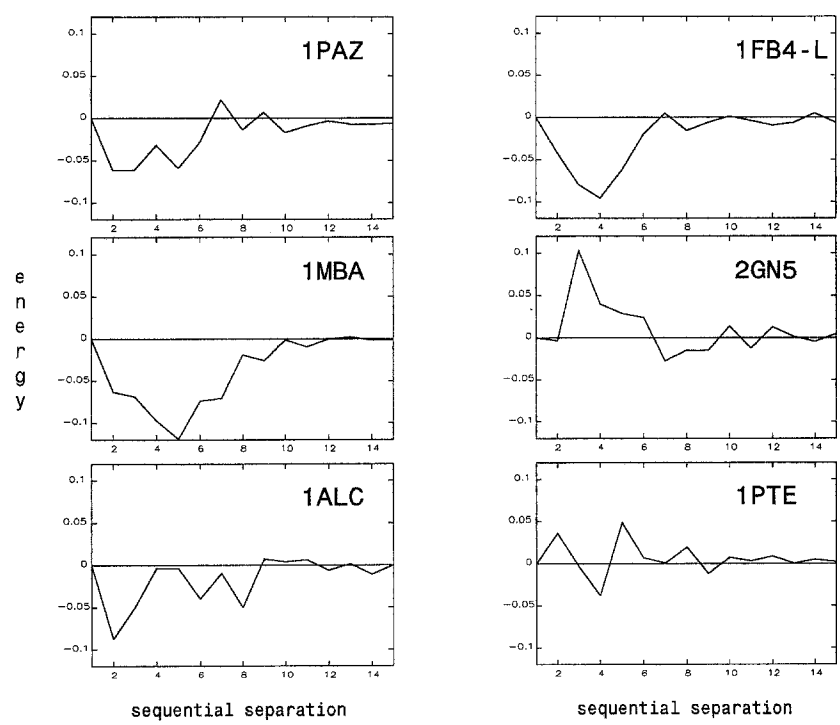

Fig. 9. k-profiles of several protein structures determined by X-ray analysis. The energies were calculated from C^{β} interactions only. Plastocyanin (1PAZ), myoglobin (1MBA), α-lactalbumin (1ALC), and immunoglobulin light chain (1FB4-L) have native-like profiles. The profiles of bacteriophage M13 gene 5 DNA binding protein (2GN5) and streptomyces D-alanyl-*D-alanine carboxypeptidase (1PTE) do not resemble native-like profiles. The energies are normalized by the number of terms in each diagonal to account for variations in protein size.

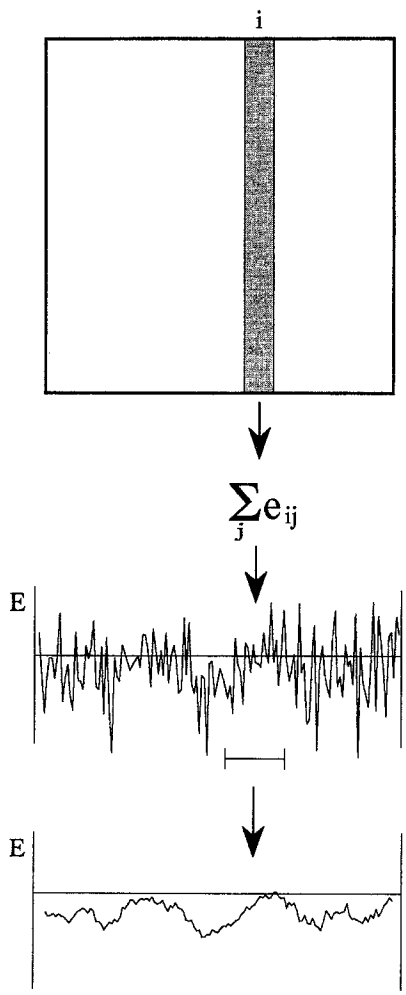

Fig. 10. Residue profile. The sum $\Sigma_j e_{ij}$ corresponds to the total interaction energy of a residue at sequence position i with respect to all other residues. The resulting profile shows the energy distribution in the molecule as a function of sequence position. Generally the residue profiles fluctuate strongly. Fluctuations can be damped using gliding averages along the chain.

problematic structures obtained from NMR studies. They can be employed to guide early interpretations of electron densities and they can be used to supplement structure calculations from NMR studies especially in cases where distance information is sparse. The power of these energetic tools in experimental structural biology is just beginning to unfold.

Data base searches

Frequently unrelated sequences adopt similar folds. In fact it is becoming rather unusual to find a completely novel fold in a newly determined protein structure [26]. Prominent examples of proteins unrelated in sequence but similar in structure are hexokinase/actin and 44K heat shock-cognate protein, and mandelate racemase and muconate lactonizing enzyme [27].

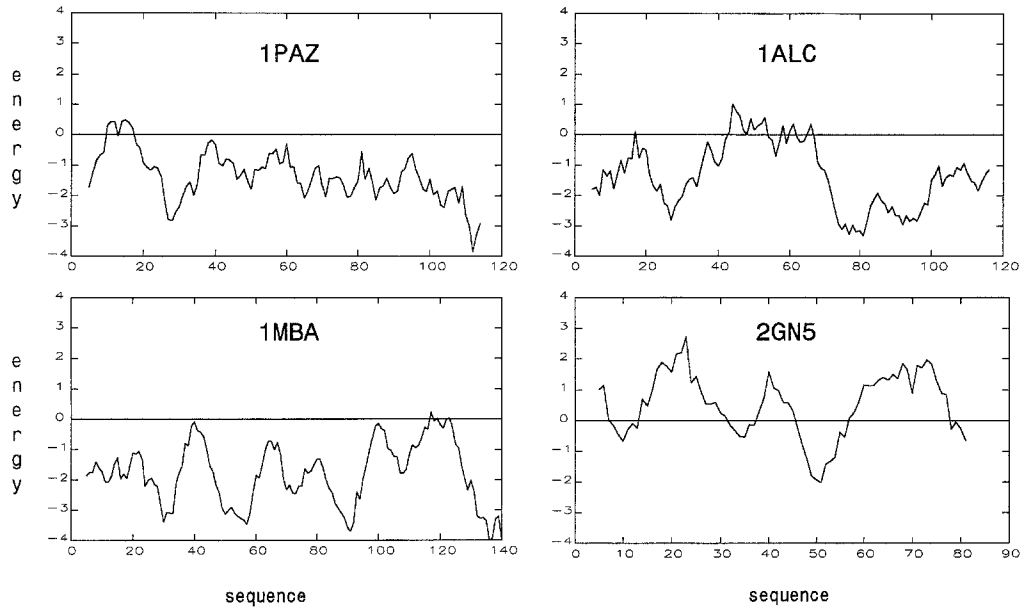

Fig. 11. Residue profiles for several protein structures determined by X-ray analysis. The energies were calculated from C^{β} interactions only. In the plastocyanin (1PAZ), myoglobin (1MBA), and α-lactalbumin (1ALC) profiles the energy remains mostly below zero. Only occasionally we encounter small positive peaks. In contrast, the residue profile of 2GN5 contains large positive peaks. The conformation appears to be extremely strained. It is noteworthy that this strain is not a consequence of steric overlap. The energies for all distances r less than 5 Å were excluded from the calculations. The window used for gliding averages amounts to 10 residues.

The current sequence data bases hold a large number of sequences whose structure and function is unknown. The native fold of a considerable number of these sequences will be similar to some fold in the data base of known structures. Using the mean force energy, proteins unrelated in sequence but similar in structure can be discovered [28].

Figure 12 outlines the data base search currently being performed in our laboratory. Sequences in the SWISSPROT data base (and other sources) are combined with all available conformations. The known structures are joined to a polyprotein and using hide-and-seek the conformational mean force energies of a particular sequence are sampled along the polyprotein and transformed to z-scores. If the lowest z-score is similar to scores expected for native sequence structure pairs, a detailed analysis of the conformational energy is performed.

Searching the current data bases the number of possible sequence structure pairs is in the order of $50\,000 \times 300 = 15 \times 10^6$, equivalent to 50 000 searches along a polyprotein assembled from 300 structures. A complete search needs considerable computing resources and, therefore, our initial search is confined to sequences of less than 200 amino acids. At the current stage our search has identified several sequence structure pairs of native-like z-scores. One example is the SWISSPROT entry IMMF$BPPH1, the IMMF control region protein (10 kD) of bacteriophage Φ-105, forming a native-like sequence structure pair with Brookhaven data base entry 1LRD, lambda repressor. Another example is COAB$BPFD, the major coat protein precursor of bacteriophages FD, F1 and M13, scoring native-like when combined with residues 245–316 of 2TS1 (a fragment

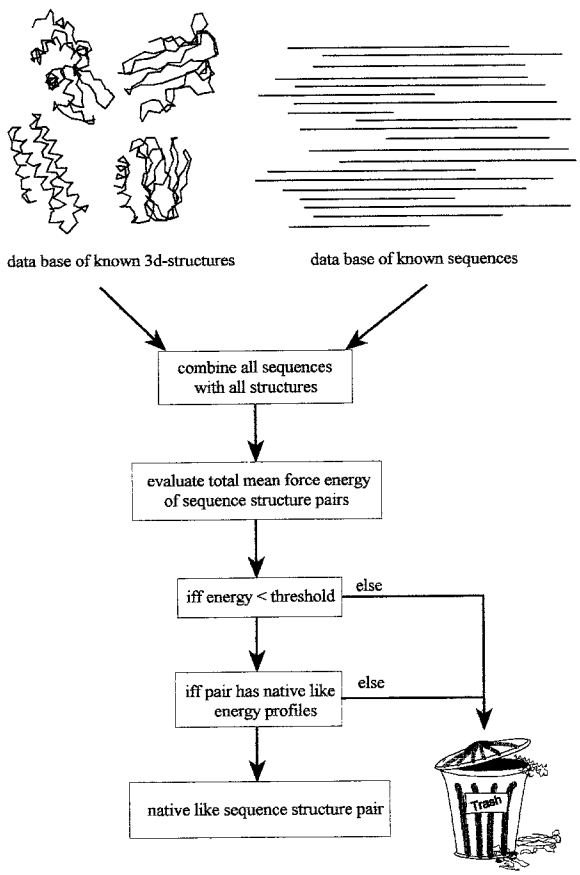

Fig. 12. Data base search for native-like sequence structure pairs. A polyprotein is constructed from the set of available protein structures. Sequences taken from the SWISSPROT and similar data bases are sent through the polyprotein and the energies and z-scores are recorded for each sequence. If the most favourable z-score of a particular sequence is native-like, the corresponding fragment is cut from the polyprotein and the profiles are investigated. If the profile shape is native-like we may assume with some confidence that we have discovered the native fold of the corresponding sequence. The sequences of many of the native-like sequence structure pairs encountered in our search have very low but still discernible homology to the sequence associated with the high scoring fold (20 to 30%). The most interesting native-like sequence structure pairs are those whose sequence identity is below 20%. Figure 13 shows two examples.

from tyrosyl-tRNA-synthase). Figure 13 shows a detailed energy analysis of these two examples. A detailed report on the native-like sequence structure pairs identified in this search is in preparation.

Sequence structure alignments

In sequence structure alignment the goal is to optimally thread a given sequence through a given fold. Reasonable alignments are generally possible only if gaps are allowed in the sequence and/or structure. Although hide-and-seek is a powerful tool for the development and quality assessment of force fields, its value as a tool to identify native-like sequence structure pairs is

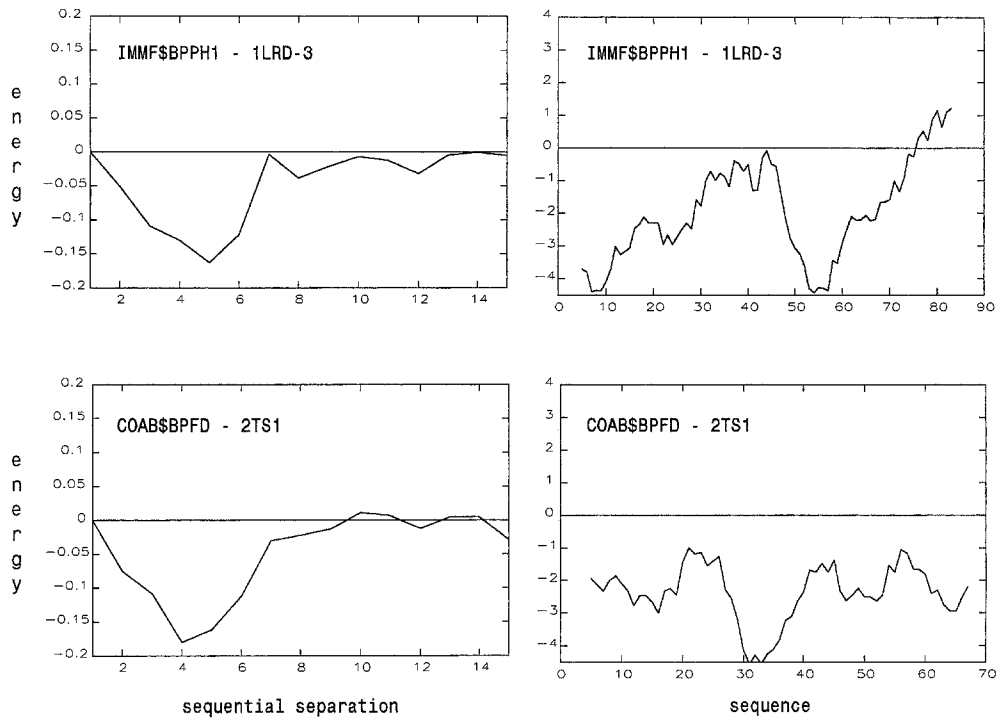

Fig. 13. Native-like sequence structure pairs found in a data base searches. The sequence of the IMMF control region 10-kD protein of bacteriophage Φ-105 (SWISSPROT entry IMMF$BPPH1) forms a native-like pair with the λ-repressor structure (Brookhaven entry 1LRD). The amino acid identity to the 1LRD sequence is 18%. At the C-terminus the residue profile points to a strained conformation. The native fold of IMMF$BPPH1 may deviate from the 1LRD structure in this region. COAB$BPFD, the coat protein B precursor of bacteriophages FD, F1 and M13 forms a native-like pair with fragment 245-316 of 2TS1, tyrosyl-tRNA-synthase. The amino acid identity of COAB$BPFD to the 2TS1 fragment is 7%.

restricted due to the neglect of gaps. Therefore, many native-like sequence structure pairs will be missed in the data base search discussed above.

More sophisticated techniques are necessary which account for the possibility of gaps. Such techniques can be derived from dynamic-programming algorithms similar to the techniques employed in conventional sequence alignment [29,30]. Sequence alignment techniques based on dynamic programming generally require two steps: (1) the calculation of a comparison matrix where each matrix element c_{ij} measures the similarity between amino acids a_i and b_j (where a_i is at position i in sequence A and b_j is at position j in sequence B, respectively) and (2) the calculation of an optimal path through the comparison matrix.

In the calculation of sequence structure alignments we can employ the same techniques. The major problem is the calculation of an appropriate comparison matrix whose elements measure the mean force energy or fitness of amino acid a_i of sequence A at location b_j in conformation B. To evaluate this energy the positions of all other amino acids in conformation B must be known. This is a severe complication as compared to sequence alignments where the matrix elements c_{ij} are strictly local quantities. One approach to overcome this problem was recently proposed by Jones et al. [31].

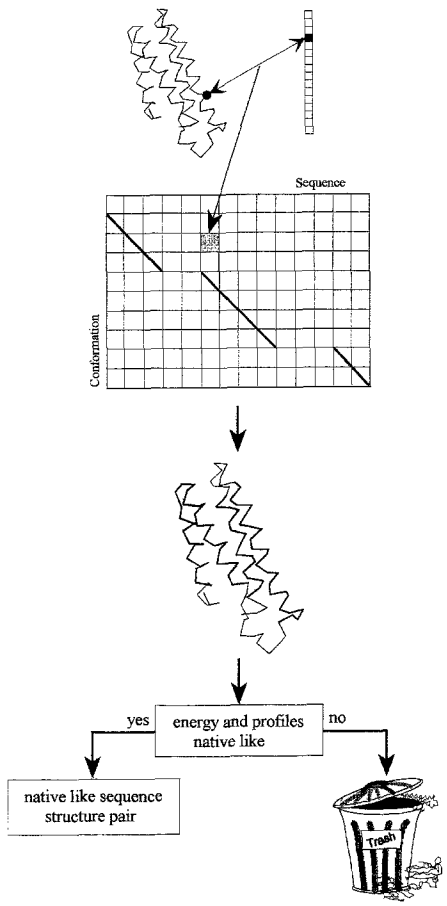

Fig. 14. Outline of sequence structure alignment. The amino acid at position i in the sequence is placed at position j in the conformation and the total interaction energy of this residue is calculated. The energy obtained constitutes element c_{ij} of the comparison matrix. Repeating the process for all sequence structure positions yields the complete comparison matrix. As indicated, optimal paths through the comparison matrix are obtained from dynamic programming techniques. Then all unpaired sequence structure positions are removed. For the aligned pairs the amino acids in the structure are replaced by the corresponding residues in the sequence (bold lines in the lower structure). Alignments can be calculated for arbitrary sequence structure pairs. Native-like energies and profiles are necessary conditions for reasonable alignments.

Our current approach, summarized in Fig. 14, uses the field produced by the native structure retaining the original amino acid sequence. A similar approach is currently being explored by David Eisenberg and co-workers (personal communication). The amino acid at position b_j in conformation B is replaced by a_i and the interaction and surface energy of this residue are calculated yielding the matrix element c_{ij}. In this way every amino acid in sequence A is placed at every position of structure B yielding the complete comparison matrix. The calculation of the optimal path through this matrix, using standard techniques borrowed from sequence-sequence alignment, is straightforward. Then, for all aligned pairs, the residue in conformation B is replaced by the corresponding amino acid in sequence A. The final aligned sequence structure pair is obtained by removing all unpaired residues from A and B.

Alignments can be calculated for arbitrary sequence structure pairs. Thus, a most important last step is the validation of the alignment. The quality of the alignment is judged by the mean force energy calculated from the aligned sequence structure pair. The energy of native sequence structure pairs is known and the energy of native-like sequence structure alignments must be relatively close to this energy value. If the energy and the profiles derived from the energy matrix are native-like we may assume with some confidence that the aligned regions comprise a reasonable model for the unknown fold.

In Fig. 15 we summarize the results obtained in the case of plastocyanin and azurin. When the plastocyanin sequence is aligned with all structures in the data base the native plastocyanin and the related azurin structures stand out as the energetically most favourable. An additional interesting question is whether the sequence structure alignment calculated from the mean field is

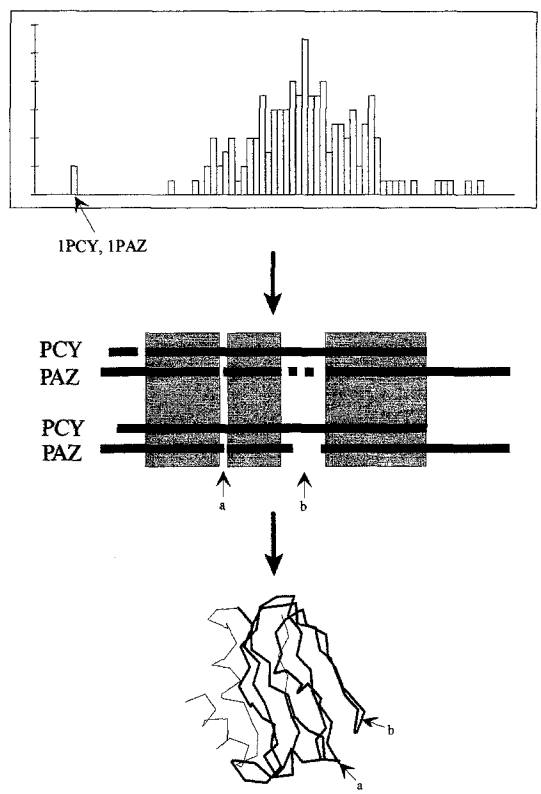

Fig. 15. Alignment of the plastocyanin sequence with the azurin structure. Plastocyanin (PCY) and azurin (PAZ) are unrelated at the sequence level but they have similar 3D folds. The figure summarizes the results obtained from the sequence structure alignment technique outlined in Fig. 14. The plastocyanin sequence is aligned with 160 conformations in our data base and the corresponding mean force energies are calculated. When aligned with the plastocyanin sequence, the native plastocyanin and the related azurin conformations stick out with the most favourable mean force energies. The structure structure alignment of PCY versus PAZ and the corresponding sequence structure alignment obtained from our algorithm are closely related (shaded areas) with only minor deviations in some details. The bold lines on the PAZ conformation (bottom) indicate the aligned sequence structure positions. a and b indicate the location of deletions in the PAZ conformation relative to the PCY sequence.

similar to the alignment obtained from the known 3D structures of these molecules. As sketched in Fig. 15, with the exception of some details the two alignments match satisfactorily.

Sequence structure alignments are currently used in our laboratory to identify unknown folds by aligning the sequence of interest with all structures in the data base. This may reveal additional native-like pairs which the hide-and-seek approach is unable to detect. It is likely, however, that sequence structure alignments may fail in some situations, even if two proteins have very similar structures. In terms of sequence structure space, gap opening is equivalent to changes in sequence. Therefore, the folding postulate is not applicable in sequence structure alignments and the chances are that the alignment procedure goes astray as indicated in Fig. 16. Hence, sequence structure alignments may fail even if the force field used to calculate the comparison matrix c_{ij} is of excellent quality.

Computation of structures from scratch

Similar to conventional sequence alignments, a twilight zone exists [32] for sequence structure alignments. Proteins related in structure may share a similar architecture but the details may differ. If these differences are large, search techniques will fail. In addition, for the unknown fold of many sequences there will be no related structure in the data base. In such cases we have to resort to techniques which enable the computation of structures from scratch.

A major strength of data base searches is their computational efficiency. All structures encountered in the data base are real and have all the characteristics of native folds. If we leave the firm ground of real structures we have to deal with steric overlap, excluded volume effects and similar phenomena. A large fraction of structures generated in energy minimization, Monte Carlo and molecular dynamics (MD) studies violate basic steric requirements. These structures will not

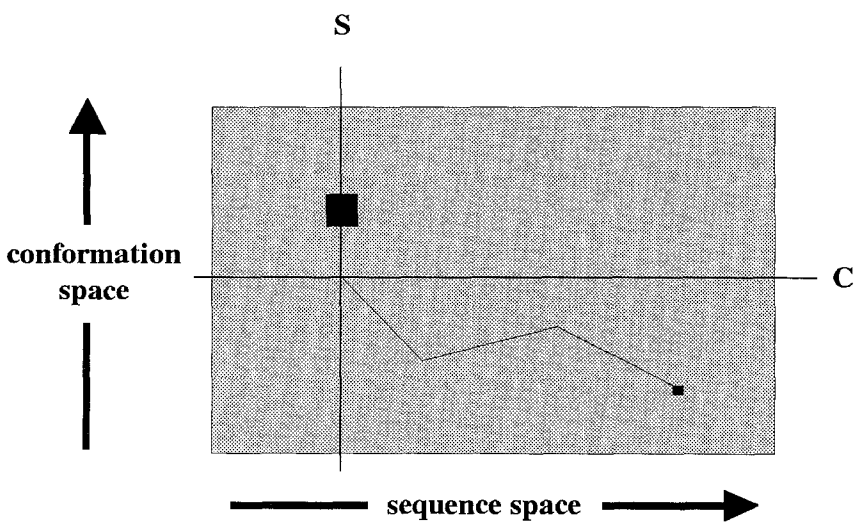

Fig. 16. Sequence structure alignment in terms of sequence structure space. In the alignment of sequence S with conformation C there is no guarantee that we find a native-like sequence structure pair. The introduction of gaps changes the sequence S, hence the folding postulate is not applicable, and the chances are that the alignment goes astray. Nonnative-like alignments are unmasked by their nonnative energies and profiles but it may be impossible to find the native-like pair (filled square) even if we start at a point (S,C) in its immediate neighbourhood.

survive, but production of such structures and evaluation of their energy is extremely time-consuming. It is therefore advantageous to stick to experimentally determined structures as long as possible.

For small fragments the data base is a rich source of conformations. The accessible conformational space of oligopeptides is reasonably modelled by the set of fragments obtained from the data base. They cover almost the entire range of possible conformations. Mean force potentials have been applied to study the conformational preferences of oligopeptides and sequence fragments of proteins. The collection of procedures employed in this approach is called Boltzmann Device [14]. The major components are: (1) calculation of conformational energies for all fragments derived from the data base; (2) energy ranking and extraction of low-energy structures from the pool of conformations; and (3) conformational analysis, which facilitates the visualization and processing of the preferred ranges of conformations.

Depending on the sequence, ensembles of short peptides may prefer only one, several or a range of conformations [14]. In addition, in the context of the whole protein, parts of the amino acid chain may be forced to adopt conformations which are unfavourable for the free oligopeptide. Kabsch and Sander [33] discovered pairs of identical pentapeptides which have different folds depending on the parent protein. In studies on the conformational ensembles of these sequences, we have been able to account for the observed differences [14].

We recently extended this approach to the calculation of complete models for protein backbone conformations [34]. Low-energy ensembles of overlapping fragments along the amino acid chain are assembled to complete conformations for entire proteins. The resulting conformations are optimized with respect to local interactions along the chain. A large fraction of the local structures assembled for lysozyme, myoglobin and thymosin are close to the structures obtained from X-ray and NMR studies. Such locally optimized folds provide reasonable starting points for subsequent energy minimization using the complete set of interaction and surface energies.

A large body of knowledge and skill has been acquired on energy minimization, Monte Carlo and MD calculations on proteins using semi-empirical force fields (for reviews see Refs. 7,9,35,36). These techniques combined with knowledge-based mean fields should be a powerful tool for the calculation of native protein folds from scratch. As in the past, the success or failure of these techniques will depend on the quality of the force field employed. At the present stage of development the knowledge-based force field is able to detect virtually all native folds hidden in the polypeptide. As pointed out in previous sections this does not necessarily mean that the global minimum of the force field corresponds to the native fold of a protein.

The current force field is compiled for the backbone atoms (including C^{β}) only and is therefore an incomplete model for the protein-solvent system. In addition there are still numerous ways to improve the backbone version using hide-and-seek on a polyprotein as a supporting tool. In spite of its incompleteness it is of course tempting to study the performance of the force field in combination with energy minimization, Monte Carlo and MD. The results obtained so far are encouraging. For example, Monte Carlo calculations on the antennapedia homeodomain in conjunction with mean force potentials yield the local structure and overall topology as determined by NMR [37] with only minor deviations in the structural details. We are currently exploring the performance of the mean field in Monte Carlo and minimization studies on a number of different proteins.

SUMMARY AND OUTLOOK

In this final section we touch on a few issues concerning the physical basis and range of applicability of knowledge-based mean fields and summarize several important features of the mean-field approach to protein folding. Mean force energies are derived from experimental data. Hence, they are only valid for the kind of system from which they have been compiled although they *may be applicable* for an extended range of systems.

The current force field is derived from a data base of soluble globular proteins. Since the resulting energetic model strongly depends on the surrounding solvent it cannot be applied to membrane proteins which maintain their structures in hydrophobic environments. Surprisingly however, the force field can be used to calculate the local structures of membrane proteins with some success. If the total pair energy is calculated for $k < 10$ hide-and-seek is able to identify the native fold of membrane proteins hidden in the polyprotein. Seemingly, the local mean forces along polypeptide backbones are less dependent on the surrounding solvent as compared to long-range ($k > 10$) and surface forces. A meaningful complete model of the energetic features of proteins in hydrophobic environments requires the compilation of mean force energies from a set of membrane proteins of known structure. This will be possible once a larger set of membrane protein structures is available from experiment.

A similar question arises in the case of short peptides. The mean force energies are derived from complete globular folds and, therefore, they are not necessarily applicable to oligopeptides in solution. However, calculations on oligopeptides reveal that the structural preferences derived from the mean field largely agree with experimental data obtained from NMR studies [12,13].

The intramolecular pair interactions depend on the separation along the sequence. Again, these interactions are not necessarily applicable to intermolecular interactions between individual protein chains. Amino acid residues within protein chains are forced to interact with each other due to the covalent linkage along the polypeptide chain. The intermolecular forces encountered in the association of two protein chains lack this constraint. Therefore, the intermolecular forces should be compiled from the intermolecular distances between oligomers and molecular associations as found in the crystal structures. On the other hand, the intramolecular potentials corresponding to large separations along the sequence may be useful models for these intermolecular forces.

The mean force concept is a statistical approach to protein folding but there is a strong connection to basic physical principles. The mean force potentials are combinations of all the basic forces which stabilize the native folds of proteins. For example, it is possible to extract hydrophobic forces from the mean force potentials [38]. Single amino acid hydrophobicities obtained in this study correlate with experimentally observed amino acid solubilities and transfer energies. In a recent study Casari and Beyer [39] were able to determine the electrostatic contribution to mean force potentials. Their results point to an exponential law of the e^{-r}/r type, consistent with the Debye-Hückel theory. It is noteworthy that the Debye-Hückel theory is a mean-field approach to electrostatic phenomena in solution. The correspondence of mean-force electrostatic energies with the Debye-Hückel theory indicates that mean force potentials are a productive source for the study of complex physical phenomena which are difficult to handle by approaches based on first principles.

Statistical analysis of protein structures has a long tradition in protein structure theory. The basic feature which distinguishes the approach based on the inverse Boltzmann principle from

other approaches is its connection to very general physical principles. In a sophisticated logical study Rooman and Wodak [40] have estimated that the extraction of parameters of sufficient predictive power requires a data base of at least 1500 unrelated protein structures. In contrast, employing the mean force approach [14], parameters of considerable predictive power could be derived from a data base as small as 100 proteins [41,42].

There have been independent attempts by other groups to apply Boltzmann's principle to protein folding. Miyazawa and Jernigan [43], for example, derived a contact potential from a data base of known structures. Others proposed contact potentials without explicit reference to Boltzmann's principle (e.g. Ref. 23). Wilson and Doniach [44] performed dynamic simulations on simplified representations of proteins using knowledge-based potentials. A detailed account of these studies is beyond the scope of this review.

Instead, we summarize several of the most important features which are required for a successful application of Boltzmann's inverse principle. These features are: (1) definition of a consistent energetic frame of reference; (2) use of a structural model which preserves the characteristics of genuine protein folds; (3) identification of important variables; and (4) a procedure which enables the assessment of the predictive power of the energetic model.

The most important item in the mean field approach to protein folding is the set of experimentally determined protein structures. Without them the approach would be meaningless. The data have been collected over the last three decades by X-ray crystallographers and NMR spectroscopists in an often painstaking and tiresome effort and we owe a large part of the current knowledge on protein folding to the people who determined the structures. Of course, this applies only to those who made the structures available. Solved and published structures whose coordinates remain undisclosed for years are of little use to the scientific community.

The prospects for the knowledge-based mean field approach to protein folding are exciting. The number of experimentally determined structures is increasing at an accelerating rate. The growing number of available structures enhances the quality of the potentials as well as the chances to find native-like sequence structure pairs in data base searches. At the present stage of development the force field supports the experimental determination of protein folds, aids in the verification process and reveals native-like sequence structure pairs in data base searches. The most ambitious goal is the calculation of native protein folds from scratch. The chances are that this will be achieved in the near future.

ACKNOWLEDGEMENTS

I am most grateful to Hannes Flöckner who prepared most of the illustrations, notably Figs. 12 and 14. For a large part, the results discussed in this work have been obtained in collaboration with Peter Lackner, Hannes Flöckner, Manfred Hendlich and Georg Casari. The work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (project number 8361-CHE).

REFERENCES

1 Lesk, A.M. and Boswell, D.R., Curr. Biol., 2 (1992) 491.
2 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.

3 Momany, F.A., McGuire, R.F., Burgess, A.W. and Scheraga, H.A., J. Phys. Chem., 7 (1975) 2361.
4 Weiner, P.K. and Kollman, P.A., J. Comput. Chem., 2 (1981) 287.
5 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
6 Van Gunsteren, W.F., Berendsen, H.J.C., Hermans, J., Hol, W.G.J. and Postma, J.P.M., Proc. Natl. Acad. Sci. U.S.A., 80 (1983) 4315.
7 McCammon, J.A. and Harveys, S.C., Dynamics of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1987.
8 Brünger, A.T., Clore, T., Gronenborn, A.M. and Karplus, M., Proc. Natl. Acad. Sci. U.S.A., 83 (1986) 3801.
9 Karplus, M. and Petsko, G.A., Nature, 347 (1990) 631.
10 Anfinsen, C.B., Science, 181 (1973) 223.
11 Jänicke, R., Prog. Biophys. Mol. Biol., 49 (1987) 117.
12 Dyson, H.J., Rance, M., Houghton, R.A., Lerner, R.A. and Wright, P.E., J. Mol. Biol., 201 (1988) 161.
13 Dyson, H.J., Rance, M., Houghton, R.A., Wright, P.E. and Lerner, R.A., J. Mol. Biol., 201 (1988) 201.
14 Sippl, M.J., J. Mol. Biol., 213 (1990) 859.
15 Sippl, M.J., J. Mol. Biol., 156 (1982) 359.
16 Bowie, J.U., Lüthy, R. and Eisenberg, D., Science, 253 (1991) 164.
17 Lüthy, R., Bowie, J.U. and Eisenberg, D., Nature, 356 (1992) 83.
18 Bowie, J.U., Clarke, N.D., Pabo, C.O. and Sauer, R.T., Proteins, 7 (1990) 257.
19 Ouzounis, C., Sander, C., Scharf, M. and Schneider, R., J. Mol. Biol. (1992) in press.
20 Holm, L. and Sander, C., J. Mol. Biol. (1992) in press.
21 Novotny, J., Bruccoleri, R.E. and Karplus, M., J. Mol. Biol., 177 (1984) 787.
22 Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. and Sippl, M.J., J. Mol. Biol., 216 (1990) 167.
23 Maiorov, V.N. and Crippen, G.M., J. Mol. Biol., 227 (1992) 876.
24 Bränden, C.I. and Jones, T.A., Nature, 343 (1990) 687.
25 Janin, J., Biochimie, 72 (1990) 705.
26 Overington, J.P., Curr. Opin. Struct. Biol., 2 (1992) 394.
27 Neidhart, D.J., Kenyon, G.L., Gerit, J.A. and Petsko, G.A., Nature, 347 (1990) 692.
28 Sippl, M.J. and Weitckus, S., Proteins, 13 (1992) 258.
29 Needleman, S.B. and Wunsch, C.D., J. Mol. Biol., 48 (1970) 443.
30 Smith, T.F. and Waterman, M.S., J. Mol. Biol., 147 (1981) 195.
31 Jones, D.T., Taylor, W.R. and Thornton, J.M., Nature, 358 (1992) 86.
32 Sander, C. and Schneider, R., Proteins, 9 (1991) 56.
33 Kapsch, W. and Sander, C., Proc. Natl. Acad. Sci. U.S.A., 81 (1984) 1075.
34 Sippl, M.J., Hendlich, M. and Lackner, P., Protein Sci., 1 (1992) 625.
35 Karplus, M. and McCammon, J.A., Annu. Rev. Biochem., 53 (1983) 263.
36 Van Gunsteren, W.F., Protein Eng., 2 (1988) 5.
37 Billeter, M., Qian, Y., Otting, G., Müller, M., Gehring, W.J. and Wüthrich, K., J. Mol. Biol., 214 (1990) 183.
38 Casari, G. and Sippl, M.J., J. Mol. Biol., 224 (1992) 725.
39 Casari, G. and Beyer, A., (1992) submitted.
40 Rooman, M.J. and Wodak, S.J., Nature, 335 (1988) 45.
41 Rooman, M.J., Kocher, J.-P. and Wodak, S.J., J. Mol. Biol., 221 (1991) 961.
42 Rooman, M.J., Kocher, J.-P. and Wodak, S.J., Biochemistry, 31 (1992) 10226.
43 Miyazawa, S. and Jernigan, R., Macromolecules, 18 (1985) 534.
44 Wilson, C. and Doniach, S., Proteins, 6 (1989) 193.