J. Mol. Biol. (1990) 213, 859–883

Calculation of Conformational Ensembles from Potentials of Mean Force

An Approach to the Knowledge-based Prediction of Local Structures in Globular Proteins

Manfred J. Sippl

Institute for General Biology, Biochemistry and Biophysics
Department of Biochemistry, University of Salzburg
Hellbrunnerstraße 34, A-5020 Salzburg, Austria

(Received 16 August 1989; accepted 29 November 1989)

We present a prototype of a new approach to the folding problem of polypeptide chains. This approach is based on the analysis of known protein structures. It derives the energy potentials for the atomic interactions of all amino acid residue pairs as a function of the distance between the involved atoms. These potentials are then used to calculate the energies of all conformations that exist in the data base with respect to a given sequence. Then, by using only the most stable conformations, clusters of the most probable conformations for the given sequence are obtained. To discuss the results properly we introduce a new classification of segments based on their conformational stability. Special care is taken to allow for sparse data sets.

The use of the method is demonstrated in the discussion of the identical oligopeptide sequences found in different conformations in unrelated proteins. VNTFV, for example, adopts a β-strand in ribonuclease but it is found in an α-helical conformation in erythrocruorin. In the case of VNTFV the ensemble obtained consists of a single cluster of β-strand conformations, indicating that this may be the preferred conformation for the pentapeptide. When the flanking residues are included in the calculation the hepapeptide P-VNTFV-H (ribonuclease) again yields an ensemble of β-strands. However, in the ensemble of D-VNTFV-A (erythrocruorin) the major cluster is of α-helical type.

In the present study we concentrate on the local aspects of protein conformations. However, the theory presented is quite general and not restricted to oligopeptides. We indicate extentions of the approach to the calculation of global conformations of proteins as well as conceivable applications to a number of molecular systems.

1. Introduction

During the last two decades, great efforts have been made to solve the protein folding problem (Anfinsen, 1973) by a variety of different strategies including secondary structure prediction (Chou & Fasman, 1978; Schulz, 1988; Holley & Karplus, 1989), energy minimization (Momany *et al.*, 1975; Tanaka & Scheraga, 1975; Vasquez & Scheraga, 1985; Gibson & Scheraga, 1986; Hagler & Honig, 1978; Levitt, 1976), molecular dynamics (Karplus & Weaver, 1976; Karplus & McCammon, 1983; van Gunsteren & Karplus, 1981; Levitt, 1982; Levitt & Meirovitch, 1983; Skolnick *et al.*, 1989) and, more recently, by pattern recognition techniques (Lathrop *et al.*, 1987; Rooman & Wodak, 1988) and knowledge-based methods (Blundell *et al.*, 1987, 1988). Quite a large number of rules of protein folding have been discovered but there is still no method available to calculate the three-dimensional structure from the amino acid sequence when there is no related structure at hand.

A large data base of three-dimensional structures of globular proteins has accumulated through the continuous work of X-ray crystallographers. The detailed analysis of these structures revealed an overwhelming number of rules for the folding of specific sequence patterns and, *vice versa*, a number of constraints that certain structural elements impose on their sequence (e.g. see Chothia *et al.*, 1977; Eisenberg *et al.*, 1982; Richardson & Richardson, 1988; Sibanda *et al.*, 1989). In spite of the knowledge acquired, the recognition of a native fold amongst hypothetical decoys still remains a

0022-2836/90/120859-25 $03.00/0

difficult task (Novotny *et al.*, 1984; Baumann *et al.*, 1989).

In the present study we will concentrate on the local aspects of protein conformations and the forces that operate on the short range of a polypeptide backbone. Nuclear magnetic resonance (n.m.r.†) studies on the conformations of peptides in solution indicate that oligopeptides frequently adopt an ensemble of different conformational states rather than existing in one well-defined conformer (Dyson *et al.*, 1988*a*,*b*). Similar observations have been made for identical pentapeptides found in several globular proteins (Kabsch & Sander, 1984). As an example the pentapeptide VNTFV (the amino acid one-letter code can be found in the legend to Table 3) is found in an α-helical conformation in erythrocruorin but adopts a β-strand conformation in ribonuclease, indicating that interactions from outside of the pentapeptide stabilize different conformations of VNTFV in the two proteins.

More than half of the observed pentapeptide identities have dissimilar conformations. Hence, we have to conclude that a rather large fraction of the local backbone conformation in proteins is unstable or forced to a non-native conformation by long-range interactions. This fraction of unstable and forced segments puts an upper limit to the possible success of local structure calculations.

We present an approach, called the Boltzmann device, which is well adapted to these experimental findings. The approach is based on a physical model that has a long tradition in protein folding studies: the basic assumption is that the preferred conformations of an amino acid sequence are those of low energy. The Boltzmann device may be classified as a knowledge-based approach to the folding of polypeptide chains. However, instead of searching for specific rules, we try to deduce the specific interactions between atomic groups from the data base.

The energy of interaction of atomic groups depends on their separation. The shapes of the potentials are obtained by calculating the distribution of distances between interacting atoms from a data base of protein structures available from the Brookhaven protein data bank (Bernstein *et al.*, 1977). The observed frequencies are transformed with the help of Boltzmann's law to yield the potential of mean force of the interaction as a function of distance. These potentials contain the knowledge acquired from the data base.

The energy of an amino acid sequence of length L in a particular conformation is computed as the sum over the interactions, modelled by the potentials of mean force. In order to find low energy conformations the conformational space of a polypeptide has to be searched. This is a problem of non-linear optimization (e.g. see Gill *et al.*, 1981). To avoid the severe problem of local minima encountered in the optimization of molecular energy surfaces we adopt the following strategy. We take the conformations of all segments of length L of the data base as the set of admissible conformations. A particular sequence is mounted over each of these segments and the associated conformational energies are computed. Then the segments are sorted with respect to their energy, the low-energy species being the best candidates for the native conformation of that sequence. This corresponds to the information retrieval component of the Boltzmann device.

In a final step the ensemble of low-energy conformations is clustered in terms of conformational similarity. This identifies conformational states of high occupancy. Depending on the amino acid sequence the Boltzmann device yields four main types of conformational ensembles.

(1) Ensembles of stable segments consist of one cluster of very similar conformations.

(2) Ensembles of flip-flop segments consist of two or more large clusters of dissimilar conformations (for example, a cluster of α-helical conformations and a cluster of β-strand conformations).

(3) Ensembles of metastable segments consist of one predominant cluster and a number of small clusters.

(4) Ensembles of unstable segments have many clusters of dissimilar conformations with only one or few conformations in each cluster.

Our main concern is to provide an outline of the Boltzmann device and to discuss some results obtained for the peptide identities found in a number of proteins (Kabsch & Sander, 1984). A more detailed account of several aspects of the Boltzmann device will be presented elsewhere. All computations apply to short segments of polypeptide chains in globular proteins ranging from five to seven residues. Therefore, before we start to present the approach in some detail, a few comments on the problems associated with the computation of local structures in globular proteins may be useful.

2. Problems Associated with the Computation of Local Structures in Globular Proteins

One of the first approaches towards the prediction of the three-dimensional structure of proteins from the amino acid sequence has been an attempt to predict the local conformation of the residues along the polypeptide chain (Chou & Fasman, 1974). It is conceivable that a sufficiently accurate identification of secondary structure elements, i.e. α-helices, β-sheets and turns, and a proper assembly of the predicted parts could yield a complete model of the tertiary structure of a protein. Many variants of secondary structure prediction schemes have been devised and a large fraction of them is based on statistical analysis of X-ray structures. Schulz (1988) has reviewed the various techniques and their results.

The approach presented here yields the backbone conformations of peptides in full atomic detail. The results obtained cannot be adequately described in

† Abbreviations used: n.m.r., nuclear magnetic resonance; r.m.s., root-mean-square.

terms of a small number of conformational states (i.e. helix, sheet, random coil). To describe local structures in more detail we introduce a new classification of the conformational states of short peptide segments that differentiates the whole range of conformations in structural and energetical terms. The classification is based on the following experimental results and hypotheses on the folding of proteins.

Protein structures are metastable systems that are subject to considerable fluctuations when folded up in the native conformation (Gurd & Rothgeb, 1979; Petsko & Ringe, 1984; Kossiakoff, 1985; Frauenfelder *et al.*, 1988). Folding of the polypeptide chain from the the random coil to the native state seems to follow a loosly defined pathway. Folding starts with fluctuating secondary structural elements, where some segments may be stabilized by short-range interactions. Then the chain enters the molten globule state, where some secondary and supersecondary structures are stabilized by medium-range interactions. Finally the chain collapses to the native state, which is induced by long-range interactions between more or less stable folding units (Goldberg, 1985).

Results obtained from n.m.r. studies (Dyson *et al.*, 1988*a*,*b*) indicate that the conformations of oligopeptides in solution may be characterized as an ensemble of conformations. The nature of the ensembles depends on the amino acid sequence. Some peptides exist in a rather well-defined conformation, others seem to prefer the random coil state.

By investigating a data base of 62 protein structures Kabsch & Sander (1984) found 25 identical pentapeptides in unrelated proteins. More than half of the observed identities adopt different conformations. Therefore, this seems to be a quite frequent phenomenon. Since protein structures are presumably stabilized by interactions within the polypeptide chain this may indicate that (1) a considerable fraction of oligopeptides is unstable, i.e. does not have a preferred conformation, and/or (2) that a considerable number of oligopeptides is distorted from the preferred conformation by long-range interactions within the protein.

From this result we have to expect that, on average, it will be impossible to predict more that 50% of the local structures in globular proteins from interactions within short segments. This statement seems to contradict the success rate of secondary structure prediction schemes, which is estimated to be of the order of 60% (e.g. see Schulz, 1988). Note that these methods count the correct prediction of a random coil state as a predictive success. However, any two conformations in the random coil state usually have a large root-mean-square error of superposition, so that they cannot be considered as similar. Hence, the assignment of a segment to the random coil state is not a strong prediction since it does not yield the conformation of a segment, in contrast to the helix, sheet or turn assignments.

In view of these results it is appropriate to distinguish four major types of segments in globular proteins.

(1) *Stable segments* are sequences whose local interactions favour a particular conformation.

(2) *Flip-flop segments* are sequences whose local interactions favour two or a few different conformations.

(3) *Metastable segments* are sequences that favour one particular conformation as well as a range of other conformations.

(4) *Unstable segments* are sequences that do not prefer any particular conformation.

In addition, in the context of the whole tertiary structure of a protein we find *forced segments*. These are stable, flip-flop or metastable segments that are forced to a non-native conformation by medium and long-range interactions within the protein.

Regarding the design and success of local prediction schemes, several conclusions may be drawn.

(1) Only stable segments can be predicted correctly.

(2) The preferred conformations of flip-flop and metastable segments as well as unstable segments can be identified only if the calculation yields an ensemble of conformations.

(3) The conformation of forced segments cannot be predicted from the local interactions within the segment.

(4) The structures of stable, flip-flop and metastable segments are not necessarily of a regular secondary structure type like α-helix or β-strand. To calculate irregular local conformations prediction schemes have to take into account the whole range of possible conformations.

3. Topological Concepts and Physical Background of the Boltzmann Device

To facilitate the introduction of the basic concepts of the Bolzmann device we represent the conformation of a protein by its C^α atoms. The extension of the model to a more detailed representation will become obvious later.

The position $\mathbf{x}_i$ of the C^α atom of amino acid residue i, in a chain of N residues is specified by three co-ordinates in a Cartesian frame. The symmetric $N \times N$ matrix $\mathbf{D}$ contains all distances $d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|$ between C^α atoms i and j. $\mathbf{D}$ is a complete representation of the conformation of the C^α backbone. With the aid of embedding procedures, it is possible to obtain co-ordinates from $\mathbf{D}$ (Sippl & Scheraga, 1985, 1986). If $\mathbf{D}$ corresponds to a three-dimensional object, the d_{ij} values are redundant for $N > 4$. Then the co-ordinates can be recovered from a proper subset of $\mathbf{D}$. Therefore, the C^α distances in combination with the geometrical constraints for embeddability in three-dimensional space, may be used to represent the state space of conformations of the C^α backbone of protein chains.

The organization of the matrix $\mathbf{D}$ of C^α distances is closely related to the hierarchical organization and topology of protein conformations (Sippl, 1982). A diagonal k of $\mathbf{D}$ contains all distances $d_{i,i+k}$, $i =$

Table 1

Ninety-eight protein chains† from the Brookhaven protein data bank used to compile the net potentials $\Delta E_k^{ab}(s)$ *and to construct the pool of conformations*

1ABP	1INS-B	1RN3	2LHB	3FXC	5RXN
1ACX	1LH4	1SN3	2LZM	3FXN	8CAT-A
1BP2	1LZ1	1TGN	2MDH-A	3GPD-R	9PAP
1CC5	1MBD	1TIM-A	2MHR	3ICB	
1CPV	1MCP-H	2ABX-A	2MT2	3PGK	
1CRN	1MCP-L	2ALP	2PAB	3RP2-A	
1CSE-E	1MLT-A	2AZA-A	2PKA-A	3WGA-A	
1CTF	1OVO-A	2B5C	2PKA-B	4ADH	
1CTS	1PAZ	2C2C	2RHV-1	4APE	
1EST	1PCY	2CAB	2RHV-2	4CYT-R	
1FB4-L	1PFK-A	2CCY-A	2RHV-3	4LDH	
1FDX	1PP2-R	2CDV	2RHV-4	4PTI	
1GCN	1PPT	2CNA	2SBT	4SBV-A	
1GCR	1PRC-C	2CYP	2SGA	4TLN	
1GP1-A	1PRC-L	2EBX	2SOD-O	51C3	
1HIP	1PRC-M	2GN5	2SSI	55C1	
1HMG-A	1PRC-H	2GRS	2STV	5API-A	
1HMG-B	1PYP	2HHB-A	2TAA-A	5API-B	
1INS-A	1RDH	2HHB-B	2TBV-A	5CPA	

† The abbreviations used are identical with the codes used in the Brookhaven protein data bank. When there are multiple chains in a protein file the chain identifier is appended to the protein code.

$1, \ldots, N-k$ between two C^α atoms separated by k peptide units along the amino acid chain. The distances $d_{i,i+k}$ are confined to the interval $[l_k, u_k]$. l_k is the smallest value for $d_{i,i+k}$ which corresponds to the closest contact of two C^α atoms. Therefore, l_k is practically independent of k. u_k is the largest possible distance between two C^α atoms and is equal to the distance of the two terminal C^α atoms in a fully extended chain of length $k+1$.

A particular diagonal k carries information on certain topological aspects of the conformation of the C^α backbone. Small k values represent the local structure of the chain, medium and large values of k correspond to supersecondary structures and overall topology, respectively. Moreover, specific local conformations such as α-helices and β-strands have characteristic $d_{i,i+k}$ values. For example, 5·5 Å $< d_{i,i+4} <$ 6·5 Å in the case of an α-helix and 11·0 Å $< d_{i,i+4} <$ 14·0 Å for extended strands (1 Å = 0·1 nm). Therefore, the diagonal k of **D** is a suitable representation of the conformation of a protein on topological level k.

The average probability density of protein conformations on topological level k can be sampled by compiling the distributions of the distances $d_{i,i+k}$ from a data base. In the present study the data base consists of a subset of protein chains available from the Brookhaven protein data bank (Table 1). Sampling is carried out by counting the number of distances whose values lie in the interval $[h_s, h_{s+1}]$, where $h_s = l_k + s(u_k - l_k)/m$, $s = 0, \ldots, m-1$ and m is the number of intervals. The distributions obtained for several topological levels are shown in Figure 1 and Table 2. For example, the distribution for $k = 4$ has a large peak in the interval [5·9, 6·5]

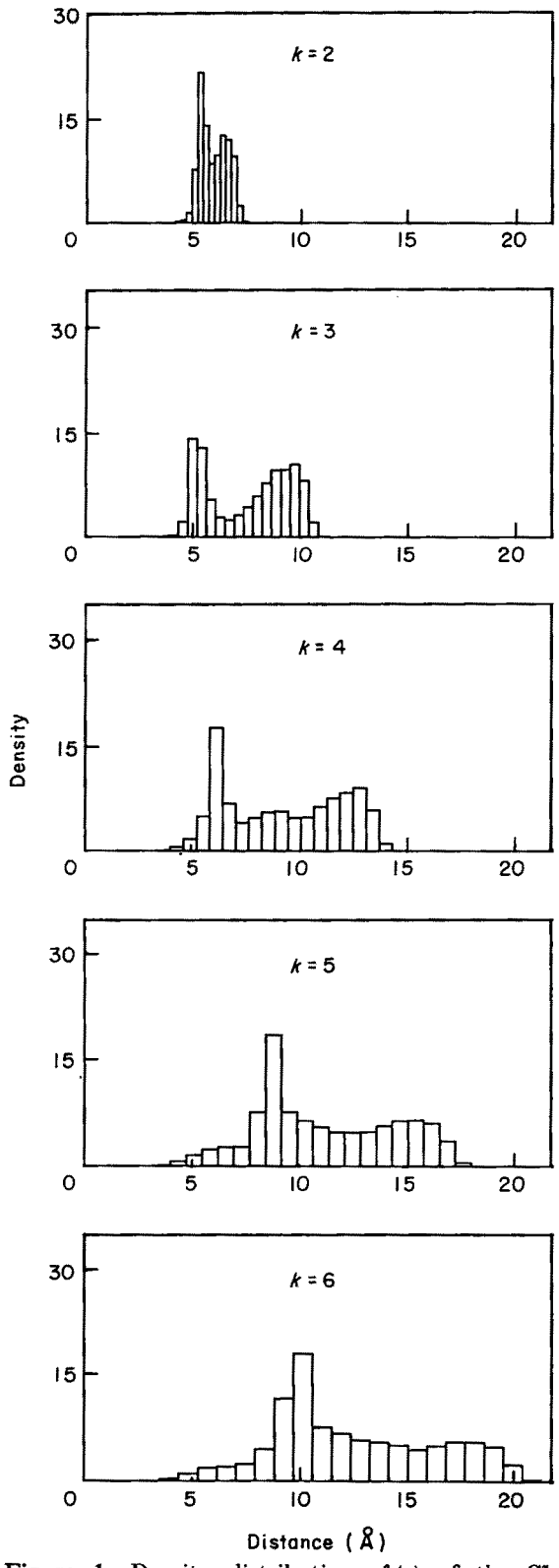

Figure 1. Density distributions $f_k(s)$ of the C^α–C^α distances $d_{i,i+k}$ on the topological levels $k = 2$, 3, 4, 5 and 6. The distributions are obtained from interval sampling with 20 intervals on each level. The bound l_k and u_k and the lengths of the intervals are assembled in Table 2. The distributions are drawn on the same distance scale to emphasize the importance of separating different topological levels. Note that the large helix peak moves towards larger distances with increasing topological level.

Table 2

Lower bounds l_k, *upper bounds* u_k *and lengths of intervals for* C^α–C^α *distances*

Level	Lower bound l_k (Å)	Upper bound u_k (Å)	Interval length (Å)†
2	2·6	8·0	0·27
3	2·6	11·5	0·45
4	2·8	15·2	0·62
5	3·2	18·1	0·75
6	3·5	21·7	0·91

† Numbers refer to $m = 20$ intervals.

and a broad shoulder in [11·4, 13·3]. This reflects the high frequency of α-helical and β-strand conformations.

The observed densities of distances are closely related to the potentials of mean force of interaction between two C^α atoms separated by k peptide units along the chain. According to the law of Boltzmann a particular state **x** of a physical system in equilibrium is occupied with probability $f(\mathbf{x})$, which is proportional to the Bolzmann factor of state **x**:

$$f(\mathbf{x}) = \frac{1}{Z}\exp\left[-\frac{E(\mathbf{x})}{kT}\right]. \tag{1}$$

The state integral Z is defined as:

$$Z = \int \ldots \int \exp\left[-\frac{E(\mathbf{x})}{kT}\right] d\mathbf{x}, \tag{2}$$

k and T are Boltzmann's constant and absolute temperature, respectively. In the case of a discrete state space consisting of n states s with associated energy $E(s)$ the corresponding equations are:

$$f(s) = \frac{1}{Z}\exp\left[-\frac{E(s)}{kT}\right], \tag{3}$$

where Z is the Boltzmann sum:

$$Z = \sum_{s=1}^{n} \exp\left[-\frac{E(s)}{kT}\right]. \tag{4}$$

Therefore, if the energies of all states **x** or s are known, the probability densities can be computed.

Conversely, if we are able to measure the probability density functions $f(\mathbf{x})$ or $f(s)$ of a system we can get the energy from:

$$E(\mathbf{x}) = -kT \ln [f(\mathbf{x})] - kT \ln [Z], \tag{5}$$

or in the case of discrete systems:

$$E(s) = -kT \ln [f(s)] - kT \ln [Z]. \tag{6}$$

However, by measuring the density of states we will not be able to evaluate the state integral Z and, therefore, the energy can be determined up to the additive constant $-kT \ln [Z]$ only.

Applying equation (6) to the density $f_k(s)$ of distances obtained from the data base we get $E_k(s)$, the potential of mean force of the interaction of two

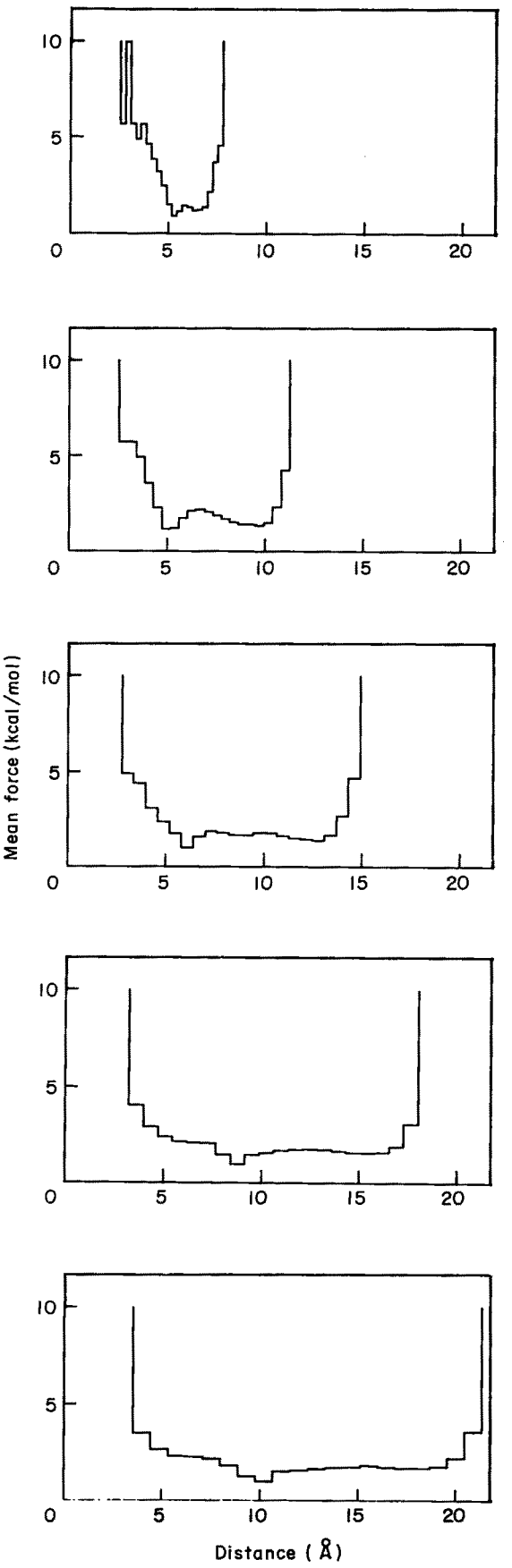

Figure 2. Potentials of mean force between C^α atoms obtained from the distributions shown in Fig. 1 by transforming $f_k(s)$ by eqn (6). The temperature was set to 293K so that RT = 0·582 kcal/mol (1 cal = 4·184 J).

C^α atoms on topological level k. This potential includes contributions from the different interactions within a polypeptide chain of length $k+1$. It contains the interactions of the segment with the rest of the protein, its interactions with the solvent molecules within the protein crystals, all geometrical constraints imposed by covalent interactions, and it is an average over all amino acid pairs a and b at position i and $i+k$, respectively. Moreover, since the potential is derived from interval sampling it is an average over all conformations (and all their interactions) whose distances $d_{i,i+k}$ map into the interval $[h_s, h_{s+1}]$.

The potentials obtained from the distributions for several k values are shown in Figure 2. For $d_{i,i+k} \rightarrow l_k$ the energy approaches infinity due to the collision of the C^α atoms and their substituents. At u_k the energy increase is due to the distortion of chemical bonds. Models of inter- and intramolecular forces approach zero when the distance between two interacting particles becomes large. However, models of pair interactions such as the Lennard–Jones potential or Coulomb's law apply to free particles. In the case of a polypeptide chain, separating two C^α atoms beyond u_k requires the distortion of valence angles and the disruption of covalent bonds. The associated high energies contribute to the averaged potential.

The potentials shown in Figure 2 represent an average over all amino acid pairs. They do not contain any specific information on the relationship between amino acid sequences and conformational states. However, provided that the folding of proteins is stabilized by atomic interactions, we should be able to relate conformational states and amino acid sequences by compiling the densities of C^α distances for individual amino acid pairs such as $(\text{Ala, Val})_k$.

On every topological level there are 400 distributions. Symmetric pairs such as $(\text{Ala, Val})_k$ and $(\text{Val, Ala})_k$ are not equivalent since polypeptide chains are asymmetric due to their N and C termini. We denote these relative frequencies by $f_k^{ab}(s)$, where a and b correspond to one of the 20 amino acids, k stands for the diagonal and s for the interval $[h_s, h_{s+1}]$.

From the pair distributions $f_k^{ab}(s)$ we get the potentials $E_k^{ab}(s)$ by:

$$E_k^{ab}(s) = -kT \ln [f_k^{ab}(s)] - kT \ln [Z_k^{ab}]. \quad (7)$$

For the *net potential* of the pair (a, b) we write:

$$\Delta E_k^{ab}(s) = E_k^{ab}(s) - E_k(s), \quad (8)$$

which yields the energy contribution of the amino acid pair (a,b) to the averaged energy $E_k(s)$. $\Delta E_k^{ab}(s)$ is tied to the averaged energy $E_k(s)$ which, by the above definition, is the *reference state* of the net potential. The net potentials $\Delta E_k^{ab}(s)$ arise from the specific interactions of the pairs (a,b). They can be computed from the probability densities $f_k(s)$ and $f_k^{ab}(s)$ by:

$$\Delta E_k^{ab}(s) = -kT(\ln [f_k^{ab}(s)] - \ln [f_k(s)] + \ln [Z_k^{ab}] - \ln [Z_k]) \quad (9)$$

or:

$$\Delta E_k^{ab}(s) = -kT \ln \left[\frac{f_k^{ab}(s)}{f_k(s)}\right] - kT \ln [Z_k^{ab}/Z_k]. \quad (10)$$

The Boltzmann sums Z_k^{ab} and Z_k cannot be obtained from the probability density functions. Note, however, that Z_k, Z_k^{ab} and hence $-kT \ln (Z_k^{ab}/Z_k)$ are constants. They do not depend on the state variable s. To simplify the following discussion we assume that $Z_k \approx Z_k^{ab}$ so that $-kT \ln (Z_k^{ab}/Z_k) \approx 0$ and:

$$\Delta E_k^{ab}(s) = -kT \ln \left[\frac{f_k^{ab}(s)}{f_k(s)}\right]. \quad (11)$$

We will return to the term $-kT \ln [Z_k^{ab}/Z_k]$ of equation (10) in a later section and discuss it in some detail.

In the present study the temperature is not a variable quantity. T is tied to the temperature used in the X-ray investigation of the respective proteins. Not all structures of the data base were solved at exactly the same temperature. Therefore, the potentials obtained are an average over the temperature range of the X-ray determinations. This average temperature is part of the definition of the reference state. For the average temperature we used $T = 293\text{K}$. Note that the choice of the average temperature is not very critical since T is a multiplicative factor applied to all potentials.

4. The Problem of Small Data Sets

It is necessary to distinguish the genuine probability densities $f_k(s)$ and $f_k^{ab}(s)$ from the density distributions $g_k(s)$ and $g_k^{ab}(s)$ obtained from the data base. In principle, the densities $f_k(s)$ and $f_k^{ab}(s)$ can be approximated by the relative frequencies $g_k(s)$ and $g_k^{ab}(s)$. The more data we have for evaluating $g_k(s)$ and $g_k^{ab}(s)$ the better will be the approximation to $f_k(s)$ and $f_k^{ab}(s)$, respectively.

When compiling the pair distributions on a particular diagonal k we immediately run into the problem of small data sets. Currently the data base holds roughly 100 proteins that do not have strong sequence homologies amongst each other. These proteins contain approximately 18,000 residues. The number of distances for small k values is of the same order of magnitude. This figure is large enough to evaluate the total density distribution $f_k(s)$ so that $f_k(s) \approx g_k(s)$. However, for a particular pair (a,b) we may expect only $18{,}000/400 = 45$ cases.

This number is far too small for conventional statistical procedures. Amino acids occur with quite different frequencies. The actual number on $k = 3$ is 17,804 distances. The pair distribution with the largest number of cases is Ala-Ala (161 distances) and the pair with the smallest number is Met-Trp (1 distance). Distributions of more than 50 cases contain a certain amount of information but the significance breaks down gradually for the rare amino acid pairs. Therefore, many of the $g_k^{ab}(s)$ obtained from the data base are weak approxima-

tions to the probability densities $f_k^{ab}(s)$ and they cannot be used to calculate the net potentials $\Delta E_k^{ab}(s)$.

On the other hand, every single observation of the conformational state of an amino acid pair (a,b) yields information on the potential $\Delta E_k^{ab}(s)$ and, of course, we are interested in gathering as much information on the conformational preferences as possible. To use all the information contained in the data base, we need a transformation that maps the relative frequencies $g_k^{ab}(s)$ to suitable models of the net interactions, in a manner that is independent of the number of measurements on the individual pairs.

In order to derive a suitable transformation we will consider the compilation of the distribution for a particular pair (a,b) as a dynamic process. A single measurement may be described by:

$$\begin{aligned} &\delta(s) = 0 \quad \text{for } s \neq t \\ &\delta(s) = 1 \quad \text{for } s = t \\ &\sum_{s=1}^{n} \delta(s) = 1. \end{aligned} \tag{12}$$

Note that $\delta(s)$ is just the distribution of relative frequencies obtained from a single measurement and as such is the smallest quantum of information that we obtained from the system. Therefore, if we observe the pair (a,b) m times we get m quanta $\delta_i(s)$ and by summing up and normalization we get the relative frequencies:

$$g_k^{ab}(s) = \frac{1}{m} \sum_{i=1}^{m} \delta_i(s). \tag{13}$$

When we start to evaluate the distribution for (a,b) no measurement has been made on the system and we do not have any information on the specific interactions $\Delta E_k^{ab}(s)$. However, we do know the total densities $f_k(s)$ and for the lack of more detailed information $f_k(s)$ may serve as a useful first approximation to the probability density $f_k^{ab}(s)$. Therefore, before any measurements are carried out we set $f_k^{ab}(s) = f_k(s)$ and from equation (11) we get $\Delta E_k^{ab}(s) \equiv 0$.

A single measurement on (a,b) modifies the density $f_k(s)$ by the information quantum $\delta(s)$. We get the modified density $f_k'(s)$ from:

$$f_k'(s) = \frac{1}{z} [f_k(s) + \sigma\delta(s)], \tag{14}$$

where $\sigma > 0$ represents the weight of an information quantum $\delta(s)$ with respect to $f_k(s)$. For z, which is required for normalization, we obtain:

$$z = \sum_{s=1}^{n} [f_k(s) + \sigma\delta(s)] = 1 + \sigma. \tag{15}$$

If we take m measurements we may write:

$$f_k^{ab}(s) \approx \frac{1}{1+m\sigma} \left[f_k(s) + \sigma \sum_{i=1}^{m} \delta_i(s) \right], \tag{16}$$

or:

$$f_k^{ab}(s) \approx \frac{1}{1+m\sigma} f_k(s) + \frac{m\sigma}{1+m\sigma} g_k^{ab}(s). \tag{17}$$

In the limit $m \to \infty$ of a large data base, the right-hand sides of equations (16) and (17) converge to $f_k^{ab}(s)$, where the rate of convergence depends on σ. By transforming all distributions $g_k^{ab}(s)$ obtained from the data base by equation (17) using a suitable value for σ we obtain probability densities that are close to $f_k(s)$ for small m values but are good approximations to the genuine densities $f_k^{ab}(s)$ for large values of m.

We emphasize how an information quantum $\delta(s)$ enters the net interaction potential $\Delta E_k'(s)$. Using equations (14) and (15) we obtain:

$$f_k'(s) = \frac{1}{1+\sigma} f_k(s) + \frac{\sigma}{1+\sigma} \delta(s) \tag{18}$$

and:

$$\frac{f_k'(s)}{f_k(s)} = \frac{1}{1+\sigma} \left[1 + \sigma \frac{\delta(s)}{f_k(s)} \right], \tag{19}$$

and therefore using equation (11) we get for the modified net potential:

$$\Delta E_k'(s) = -kT \ln \left[1 + \sigma \frac{\delta(s)}{f_k(s)} \right] + kT \ln [1+\sigma]. \tag{20}$$

Since $\delta(s) = 1$ for $s = t$ and zero elsewhere:

$$\begin{aligned} \Delta E_k'(s) &= kT \ln [1+\sigma] - kT \ln [1+\sigma/f_k(s)] \quad \text{for } s = t \\ &= kT \ln [1+\sigma] \quad \text{for } s \neq t. \end{aligned} \tag{21}$$

Thus, the effect of a single measurement on $\Delta E_k'(s)$ is twofold. The energy $\Delta E_k'(s)$ of all states is shifted to higher energies by an amount of $E_c = kT \ln [1+\sigma]$. In addition the energy of state t is lowered by $E_t = -kT \ln [1+\sigma/f_k(t)]$. Since $f_k(t) < 1$ the net effect on state t is $\Delta E_k'(t) = E_c + E_t < 0$, i.e. in any case $\Delta E_k'(t)$ is negative.

When $f_k(t)$ is small E_t is a large negative quantity. In a sense E_t is a measure of the information content of the observation $\delta(s)$ relative to what is known about the system. If $\delta(s)$ maps into the α-helix region (Fig. 1) the effect on $\Delta E_k'(t)$ is small since the result is expected with high probability. On the other hand if t hits a region of lower density the result is rather unusual, and $\Delta E_k'(t)$ is lowered by a much larger amount. It follows that the total energy take up E_T induced by $\delta(s)$ is not a constant quantity but depends on $f_k(t)$:

$$\begin{aligned} E_T &= \sum_{s=1}^{n} \Delta E_k'(s) = nE_c + E_t \\ &= nkT \ln [1+\sigma] - kT \ln [1+\sigma/f_k(t)]. \end{aligned} \tag{22}$$

Along the same line we get $\Delta E_k^{ab}(s)$ for the case of m_{ab} observations of the states of (a,b) from

equations (11) and (17):

$$\Delta E_k^{ab}(s) = -kT \ln \left[\frac{f_k^{ab}(s)}{f(s)}\right] = kT \ln [1+m_{ab}\sigma] - kT \ln \left[1+m_{ab}\sigma \frac{g_k^{ab}(s)}{f_k(s)}\right]. \quad (23)$$

The energies associated with these m_{ab} measurements are:

$$E_{k,(c)}^{ab} = nkT \ln [1+m_{ab}\sigma]$$
$$E_{k,(t)}^{ab} = -kT \sum_{s=1}^{n} \ln [1+m_{ab}\sigma g_k^{ab}(s)/f_k(s)] \quad (24)$$
$$E_{k,(T)}^{ab} = E_{k,(c)}^{ab} + E_{k,(t)}^{ab}.$$

Note that if $g_k^{ab}(s) = f_k(s)$ for all states s we get $E_T^{ab} = 0$ and $\Delta E^{ab}(s) \equiv 0$. In this case the measurements do not provide any additional information on the system.

In equations (23) and (24) we have combined the weighting scheme equation (17) for the different numbers of occurrences m_{ab} of the pairs (a,b) with the net potential of interaction of the C^α atoms of these residues. By this device we are able to use any single bit of information on the energetics of a certain conformational state. Moreover, since the models $\Delta E_k^{ab}(s)$ of the net interactions of all pairs (a,b) converge to the genuine potentials in the limit $m_{ab} \to \infty$, the models for the pair interactions compiled from the data base will improve with the increasing number of X-ray structures available from the Brookhaven protein data bank.

5. Properties of Individual Distributions

If the distributions $g_k^{ab}(s)$ were very similar to the total density $f_k(s)$ then $\Delta E_k^{ab}(s) \approx 0$. This would imply that the individual potentials do not carry any information on the conformational energies of amino acid sequences. In this case they would be of no use in calculating conformational states of amino acid chains.

The distributions and potentials obtained from the protein data base (Table 1) seem to be sufficiently different from the total density, so that they should be useful for the calculation of conformations. In Figure 3 we present a few examples of individual distributions $g_k^{ab}(s)$ and the related potentials $\Delta E_k^{ab}(s)$ calculated from equation (23) for $k = 3$. In the following discussion we are aware of the small data set, so that we do not stress the statistical significance of the individual distributions and potentials and we will not try to interpret the origin of certain features of the potentials. This would be rather hazardous in view of the small data base as discussed above. The main point is to show that there are interesting differences amongst the individual potentials, which may reflect the physical basis of the stability and folding of amino acid chains.

To begin with, the density in the α-helix region of $f_3^{A\text{-}P}(s)$ is very low in accordance with the well-known α-helix breaking property of proline residues. There is also very limited tendency for extended structures. The two deep minima around 6·5 Å and 8 Å do not correspond to regular secondary structures so that the Ala-Pro pair on $k = 3$ seems to favour rather irregular conformational states.

At a first glance the distributions $f_3^{A\text{-}T}(s)$ and $f_3^{A\text{-}V}(s)$ seem to be very similar. However, the potentials $\Delta E_3^{A\text{-}T}(s)$ and $\Delta E_3^{A\text{-}V}(s)$ exhibit some important differences. $\Delta E_3^{A\text{-}V}(s)$ has a minimum at very short distances which is absent in $\Delta E_3^{A\text{-}T}(s)$. $\Delta E_3^{A\text{-}T}(s)$ is negative in the α-helix region, whereas $\Delta E_3^{A\text{-}V}(s)$ is zero, indicating a slightly higher α-helix preference of Ala-Thr over Ala-Val. Moving towards larger distances both potentials have a positive and then a negative peak, the former being much larger in Ala-Thr than in Ala-Val. Around 9 Å $\Delta E_3^{A\text{-}T}(s)$ is negative and $\Delta E_3^{A\text{-}V}(s)$ is positive. Both pairs have negative values for large distances, indicating a preference for extended structures.

The symmetric pairs Ala-Val and Val-Ala show interesting differences. There are 127 cases for the Ala-Val distributions and 128 observations for the Val-Ala pair so that the observed differences may have some significance. The tendency to adopt α-helical conformations is higher in Val-Ala than in Ala-Val. In addition, Val-Ala has a strong tendency for extended conformations which is lower in Val. The example shows that it is important to separate symmetrical pairs. The same difference occurs at the next topological level. The α-helix peak for $k = 4$ is around 6 Å (see Fig. 1). Again Val-Ala has a higher density in this region as compared to Ala-Val (Fig. 4).

There are 400 different pairs on every topological level k. The question arises whether there are groups of pairs with similar potentials and conformational preferences and it may be interesting to investigate the physical and chemical properties of such groups. However, the preliminary results obtained so far indicate that the individual pairs do not have a strong tendency to form related groups. In view of the large number of distributions and the problems associated with the small number of occurrences of individual pairs an appropriate analysis is quite demanding.

6. Locating Low-energy Conformations and Calculation of Ensembles

Once the potentials $\Delta E_k^{ab}(s)$ are compiled from the data base, the energy $\Delta E(S_q,C_p)$ of a given sequence S_q of length L with respect to a particular conformation C_p can be computed:

$$\Delta E(S_q,C_p) = \Delta E(\mathbf{x}) = \sum_{i=1}^{L} \sum_{j=i+1}^{L} \Delta E_k^{ab}(d_{i,j}(\mathbf{x})) = \sum_{k=1}^{L-1} \sum_{i=1}^{L-k} \Delta E_k^{ab}(d_{i,i+k}(\mathbf{x})). \quad (25)$$

a and b refer to the amino acids at sequential

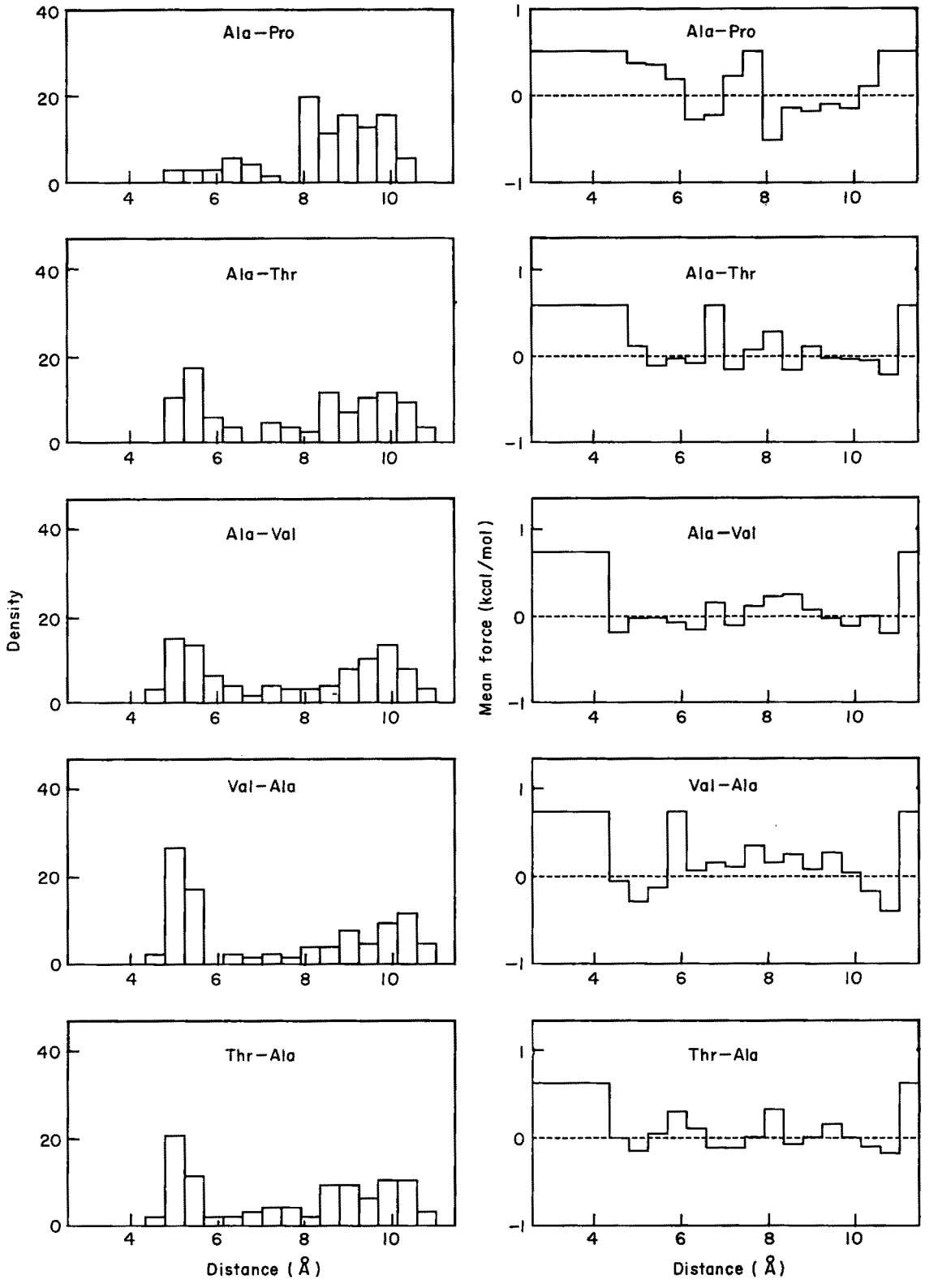

Figure 3. (a) Distance distributions $g_3^{ab}(s)$ of Ala-Pro, Ala-Thr, Ala-Val, Val-Ala and Thr-Ala on topological level $k = 3$. (b) Net potentials $\Delta E_3^{ab}(s)$ obtained from $g_3^{ab}(s)$ through eqn (23). Again $RT = 0{\cdot}582$ kcal/mol. The weight of one information quantum σ was set to 1/50, so that on 50 observations of the pair (a,b) $g_k^{ab}(s)$ and $f_k(s)$ have the same weight. The numbers of individual pairs in the data base are (Ala-Pro; 71), (Ala-Thr; 87), (Ala-Val; 127), (Val-Ala; 128) and (Thr-Ala; 97).

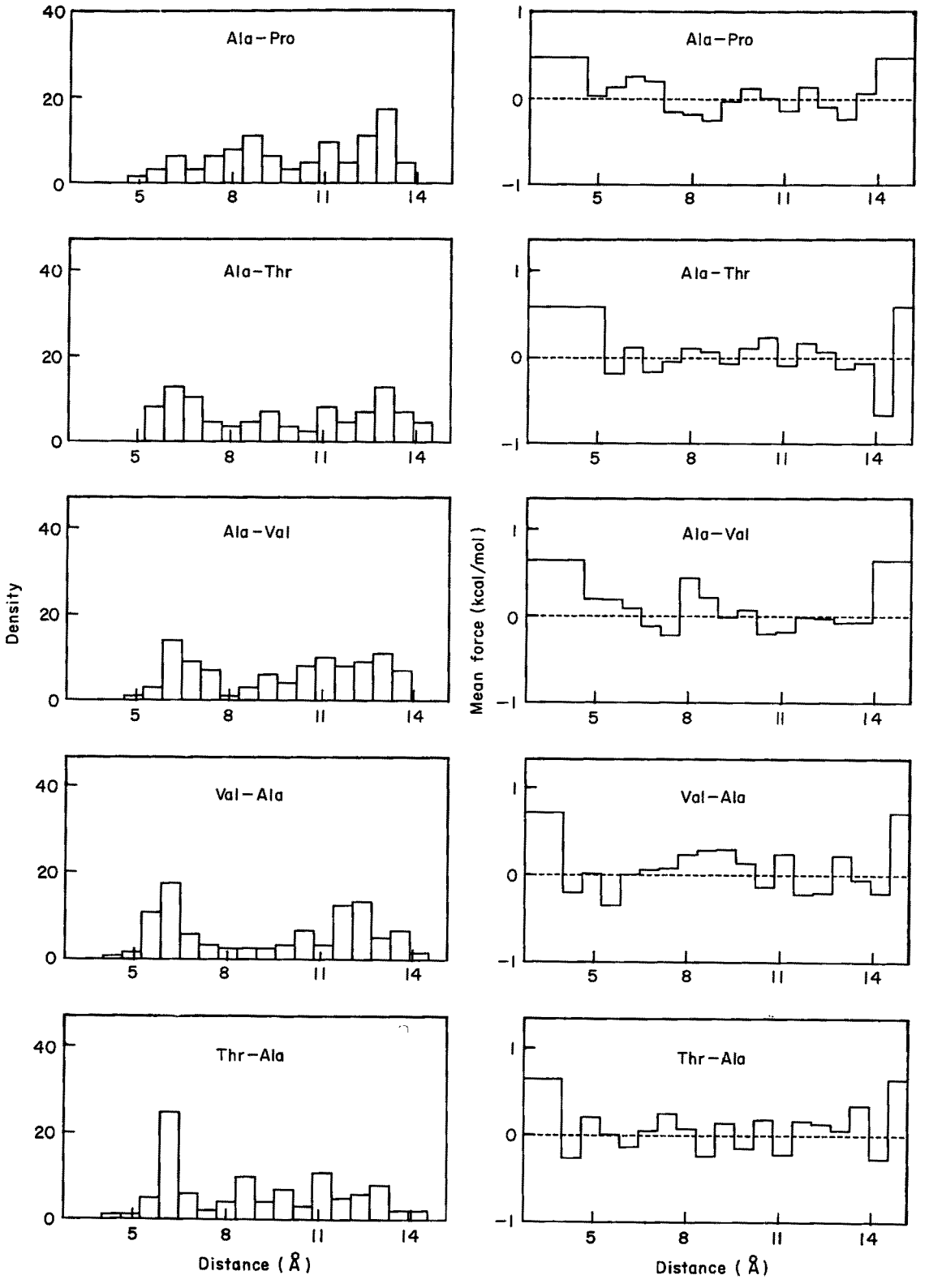

Figure 4. Same as Fig. 3 for topological level $k = 4$. The numbers of individual pairs in the data base are (Ala-Pro; 64), (Ala-Thr; 87), (Ala-Val; 101), (Val-Ala; 120) and (Thr-Ala; 102).

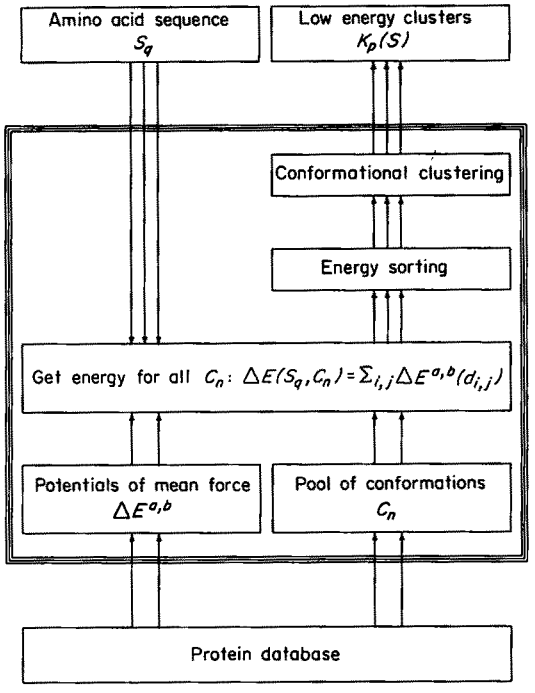

Figure 5. Main components of the Boltzmann device and the flow of information through its components. The Boltzmann device is initialized by the protein data base. The potentials of mean force are compiled from the distances in the data set and the pool of conformations is prepared. To calculate an ensemble for a particular sequence, the Boltzmann device is loaded with the sequence and the energies of the conformations in the pool are calculated. The segments in the pool are sorted with respect to their energies and the representative ensemble is sampled from the conformations of lowest energy. In the last step probable conformations of the input sequence are obtained by conformational clustering of the ensemble.

positions i and j, respectively, **x** represents the Cartesian co-ordinates of the conformation C_p, and the distances d_{ij} are functions of these co-ordinates. Since the net potentials are defined with respect to the reference states $E_k(s)$ (eqn (8)), the appropriate reference state of $\Delta E(S_q,C_p)$ is the data base of protein structures used to compile the net potentials.

To identify conformations of low net energy, one has to solve the problem of minimizing equation (25). This is a non-linear problem that cannot be approached efficiently by methods of optimization, due to the large number of local minima of the energy surface of ΔE (e.g. see Gill *et al.*, 1981; Hall & Lyons, 1980; Purisima & Scheraga, 1986).

To avoid the problems of minimization we again resort to the proteins of the data base. By taking all conformations C_p of segments of length L from the data base we obtain a set of possible conformational states. For $L < 10$ residues the current data base of 100 proteins yields $r \approx 18{,}000$ segments. This set is not complete, since we cannot assume that all possible conformations of segments of length L are realized in the proteins of the data base. For small L values, however, the available sample is a useful representation of the possible conformational states. Since the number of possible conformations explodes with Y^L, where Y is the number of different conformations of one residue, the pool dramatically loses its representative character with increasing L.

By mounting the sequence S_q over each of the conformations C_p, we calculate the associated energies $\Delta E(S_q,C_p)$ for $p = 1, \ldots, r$. Then the conformations C_p are sorted with respect to their net energies and we get the low-energy conformations on top of the list. Taking the first n conformations of lowest energy we obtain a representative ensemble of highly probable conformations.

The representative ensemble can be investigated further by searching for clusters of very similar conformations. The clusters are obtained by calculating the root-mean-square error (r.m.s.) of optimal superposition of any two conformations C_i and C_j, $i,j = 1, \ldots, n$; $i < j$ in the representative ensemble. Using Kruskal's (1956) algorithm the minimal spanning tree is constructed so that all conformations are connected by their minimal r.m.s. values. All branches larger than a suitable cutoff c are then removed from the minimal spanning tree (see page 206 of Kohonen (1984)). The conformations in the resulting subtrees constitute the clusters of the ensemble. The pattern of clusters obtained from the ensemble is appropriately classified in terms of the segment types presented above.

We have now defined the major components of the Boltzmann device. In Figure 5 we summarize the design of the device and the information flow through its components.

7. The Significance of the Reference States

We have to justify the use of the net potentials $\Delta E_k^{ab}(s)$ instead of the potentials $E_k^{ab}(s)$ in equation (25). In analogy to equation (25) we write:

$$E(S_q,C_p) = \sum E_k^{ab}(s). \qquad (26)$$

From equation (8) we get $E_k^{ab}(s) = \Delta E_k^{ab}(s) + E_k(s)$, and inserting this in equation (26) we obtain:

$$E(S_q,C_p) = \sum \Delta E_k^{ab}(s) = \Delta E(S_q,C_p) + \sum E_k(s). \qquad (27)$$

Hence, $E(S_q,C_p)$ is dominated by the sum over the reference states $E_k(s)$, which are averages over the pair potentials $E_k^{ab}(s)$, and do not contain any sequence-specific information.

To obtain a closer view of the difference between $\Delta E(S_q,C_p)$ and $E(S_q,C_p)$ we define the averaged amino acid X by its interactions $E_k^{xx}(s) \equiv E_k(s)$ so that the associated net potentials are $\Delta E_k^{xx}(s) \equiv 0$. We now calculate the ensembles for homopolypeptides of X.

By calculating $E(S_q,C_p)$ from equation (26) the output from the Boltzmann device is an ensemble of α-helical conformations due to the energy minimum of $E_k^{xx}(s)$ in the α-helix region on all levels (Fig. 2). This is not the desired result. We are not interested

in the conformation of lowest energy but in the appropriate representative ensemble of conformations. Using $E_k^{xx}(s)$ the representative ensemble consists of 100% helix in contrast to about 30% helix in the pool.

Using $\Delta E(S_q,C_p)$ (eqn (25)) we obtain the desired result. All conformations in the pool have $\Delta E(S_q,C_p) = 0$. The effect of energy sorting is equivalent to a random redistribution of the conformations. The sample of the first n structures on top of the list is a small representative sample of the conformations in the pool and we will find 30% helix.

Fortunately the net potentials $\Delta E_k^{ab}(s)$ obtained from the data base are different from zero. The recombination of potentials $\Delta E_k^{ab}(s)$ is unique for a given sequence and induces a characteristic redistribution of weights in the pool of conformations. Small variations of the net potentials $\Delta E_k^{ab}(s)$ may have a dramatic effect on the representative ensemble sampled from the pool.

We emphasize that the pool of conformations is not an essential component of the Boltzmann device. It is merely a useful tool to circumvent the problems associated with the minimization of the non-linear function $\Delta E(S_q,C_p)$. In addition the pool is particularly well adapted to the calculation of local conformations in proteins since the conformations in the pool are already weighted with respect to their energies.

8. The Significance of the Boltzman Sums Z_k^{ab}

Previously we dropped the term $-kT \ln (Z_k^{ab}/Z_k)$ from equation (10). This is equivalent to the assumption that $Z_k^{ab} \approx Z_k$ for all a and b. We want to investigate how this term affects $\Delta E(S_q,C_p)$. Using equation (10) for $\Delta E_k^{ab}(s)$ we obtain from equation (25):

$$\Delta E(S_q,C_p) = \sum \Delta E_k^{ab}(s) = -kT \sum \ln \left[\frac{f_k^{ab}(s)}{f_k(s)}\right] - kT \sum \ln \left[\frac{Z_k^{ab}}{Z_k}\right], \quad (28)$$

where the intervals s depend on the co-ordinates of C_p. The sum involving Z_k^{ab} and Z_k is independent of the conformational variables s. Hence, the term is constant for a particular sequence and the net energies $\Delta E(S_q,C_p)$ of all conformations are shifted by a constant term as compared to the assumption $\Delta Z_k^{ab} \approx Z_k$. This does not affect energy sorting nor conformational clustering. However, the term is sequence-dependent. The relative stabilities $\Delta E(S_i,C_p)$ and $\Delta E(S_j,C_p)$ of two different sequences S_i and S_j in the same conformation C_p are not comparable unless the Z_k^{ab}/Z_k ratios are known.

The question remains whether or not $Z_k^{ab} \approx Z_k$ is a useful approximation. Z_k is the Boltzmann sum over the reference state $E_k(s)$. This reference state is an average over the potentials $E_k^{ab}(s)$. Therefore, Z_k has the quality of an average over Z_k^{ab}. We may argue that many of the Z_k^{ab} values are indeed close to Z_k, provided that the variance of Z_k^{ab} around Z_k is small. On the other hand, it is conceivable that the Boltzmann sums of interactions involving amino acids of rather unusual conformational properties like Pro and Gly may be quite different from Z_k. At present, none of the Z_k^{ab} values is known, so this issue remains speculative.

It may be possible to estimate Z_k^{ab} experimentally, for example, by using a series of different but similar oligopeptides that differ in only a few of their interactions. If it is possible to observe the relative conformational stabilities of these peptides, e.g. by n.m.r. measurements (Dyson *et al.*, 1988*a*), one could possibly get an estimate of the ratios Z_k^{ab}/Z_k^{cd}.

9. Recombining Information on the Conformational Properties of Sequences

When we started to present the Boltzmann device we confined our model to the C^α atoms of a polypeptide chain and the distance distributions between the C^α atoms. It is straightforward to generalize the model to any set of distances (e.g. C^β–C^β, C^α–C^β, N–C^β, etc.). However, more types of distance increase the data and computational load. A more detailed discussion of useful distance sets and their impact on computing resources will be presented elsewhere. In the present study the computations included the net potentials for the atomic pairs C^α–C^α, C^β–C^β, N–N, O–O, O–N and N–O.

In previous sections we pointed out that the size of the current data base is small with respect to the distributions $g_k^{ab}(s)$ of amino acid pairs. We want to emphasize that equation (25) represents a recombination of several distributions $g_k^{ab}(s)$ as a function of the sequence S_q. By calculating the energy $\Delta E(S_q,C_p)$ we put together small pieces of information on the conformation of S_q that sums up to a considerable amount. For example, if $L = 7$ we have $L(L-1)/2 = 21$ amino acid pair distributions $g_k^{ab}(s)$ used in the energy calculation. It depends on the sequence S_q which of the distributions are used, but in general they will be different, except for special sequences such as homopolypeptides or co-polypeptides. In the mean there are 45 observations for one distribution. For $L = 7$ and the six different atomic interactions used there are $6 \times 21 \times 45 = 5670$ observations of the conformational preferences of S_q that are used to discriminate between the conformations C_p.

10. Model-specific Parameters

The Boltzmann device has a few important parameters that have to be specified in the calculations. First there is the number of intervals used to compile the distance distributions. It is desirable to choose the intervals as small as possible in order to highlight the fine details. In addition the size of the intervals is important in distinguishing different conformations. Since spacing is inherited to the

potentials of mean force, two conformations whose distances map into the same intervals will have identical energies. The sparse data set restricts spacing to rather large intervals but we find that 20 intervals on each k value is a good compromise. The lengths of the resulting intervals for different k values are assembled in Table 2.

A second parameter is σ, the weight of one information quantum with respect to the total frequency distribution $f_k(s)$. We use $\sigma = 1/50$ which means that after 50 observations of (a, b) $g_k^{ab}(s)$ and $f_k(s)$ have the same weight (eqn (17)).

The third parameter is the segment L. Segments should not be longer than ten residues. For longer segments the current pool cannot be regarded as a valid representation of the conformational space. In the examples discussed below we calculated ensembles for penta-, hexa- and heptapeptides. The ensemble of highly probable structures consists of the $n = 20$ structures of lowest energy.

A suitable cutoff parameter c_L used to obtain the clusters from the minimal spanning tree for different segment lengths L was estimated in the following way. For a given segment length L we randomly draw conformations from the pool and calculate the mean r.m.s. value between all conformations. c_L is defined as one-third of the respective average. The numerical values of c_L in the present study are $c_5 = 0{\cdot}67$, $c_6 = 0{\cdot}83$ and $c_7 = 0{\cdot}97$ Å.

For the set of proteins used in the calculations we extracted a set of protein chains from the Brookhaven protein data bank (assembled in Table 1), which for the present purpose are sufficiently unrelated in their sequences. We checked the sequences in the data base for sequence homologies using the algorithm of Needleman & Wunsch (1970). The two proteins that show the largest homology in our data base are 55C1 and 2C2C with 40·5% amino acid identities. The second strongest homology is found between 1EST and 3RP2-A with 32·4% identity.

We emphasize that in all calculations the protein chain under investigation was excluded from the data base. Before any calculations were carried out the protein chain was removed from the list (Table 1) and the distributions as well as the net potentials were recompiled.

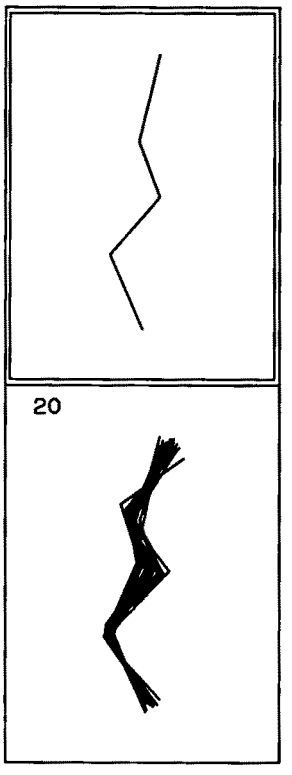

Figure 6. Ensemble obtained from the Boltzmann device for the pentapeptide VNTFV of ribonuclease. In Figs 6 to 18 the conformation of the peptide in the respective protein is shown on top. The clusters are drawn from top to bottom in order of decreasing size. Only the C^α backbone is shown. The number of segments in each cluster is found in the respective boxes. Clusters were oriented by optimal superposition on the conformation in the protein as follows. First the conformation of the cluster with the lowest r.m.s. deviation to the conformation in the protein was identified. Then all other conformations in the cluster were superimposed on that conformation. Finally the whole cluster was superimposed on the conformation of the protein segment by the matrix of the best superposition of the conformation of lowest r.m.s. deviation. Average r.m.s. values and the number of conformations in the clusters are assembled in Tables 3 to 8.

11. Applications to Identical Peptides Found in Different Proteins

For any computational scheme that aims at the prediction of conformations from the amino acid sequence it is most important to estimate the range of validity of the approach. This is usually done by calculating the conformations of proteins of known structure. The success and failure of the method are then investigated by comparing the results of the calculations to the native structure.

However, comparing computational results with native conformations may not be the method of choice if the success of local predictions has to be assessed. We have to expect that a number of local structures are distorted from their preferred conformations by interactions with the remainder of the protein. The fraction of such forced segments in proteins is unknown, but in view of the variable conformations of identical pentapeptides found in globular proteins (Kabsch & Sander, 1984) we have to expect that forced and unstable segments occur quite frequently. The pentapeptide VNTFV for example adopts an α-helical conformation in erythrocruorin but is in a β-strand conformation in ribonuclease. A calculation that predicts an α-helical conformation is a 100% success in the case of erythrocruorin, but it is a 100% failure in ribonuclease yielding a combined predictive success of 50%.

Obviously, the results of computations should be compared to the ensemble of conformations of the isolated peptide which, however, is not available. Therefore, unless the conformational properties of

Table 3

Ensembles of several peptides of ribonuclease 1RN3 obtained from the Boltzmann device

Peptide[a]	Start[b]	Type[c]	Clusters[d]	C1[e]	R1[f]	C2	R2	C3	R3	C4	R4
A. *Pentapeptides*											
KP-VNT	42	Stable	3	18	1·69	1	1·73	1	2·17		
P-VNTF	43	Stable	1	20	0·65						
VNTFV	44	Stable	1	20	0·58						
NTFV-H	45	Metastable	9	10	3·37	3	2·92	1	0·75	1	0·97
TFV-HE	46	Metastable	6	12	2·63	4	3·16	1	0·72	1	1·02
B. *Hexapeptides*											
KP-VNTF	42	Stable	1	20	1·94						
P-VNTFV	43	Stable	3	18	0·90	1	1·37	1	1·44		
VNTFV-H	44	Stable	3	18	0·75	1	1·88	1	2·29		
NTFV-HE	45	Metastable	8	13	3·49	1	2·40	1	2·88	1	3·36
C. *Heptapeptides*											
KP-VNTFV	42	Stable	1	20	1·99						
P-VNTFV-H	43	Stable	3	18	1·08	1	1·34	1	1·37		
VNTFV-HE	44	Metastable	8	12	2·88	2	2·23	1	0·84	1	1·09

[a] Sequence of the peptide in amino acid 1- letter code. A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr. (—) The start or end of a pentapeptide identity.

[b] Sequential number of the peptides first residue relative to the protein sequence.

[c] Segment type obtained from the Boltzmann device. The representative ensemble consists of the 20 conformations of lowest energy. Segments are classified as stable if the ensemble has a cluster of more than 16 members, as metastable if there is a cluster at least 10 members and a number of small clusters, as flip-flop if there are at least 2 clusters with at least 4 elements and a small number of additional clusters, and as unstable if there are many small clusters.

[d] Number of clusters in the ensemble.

[e] C1, the number of conformations in the largest cluster, C2 in the second largest cluster and so on. Only the 4 largest clusters are shown.

[f] R1, R2 etc. refer to the average r.m.s. error of the conformations in the clusters C1, C2 etc. to the conformations of the peptide in the respective proteins. The r.m.s. values were calculated for the C^{α} atoms only.

the isolated pentapeptide are determined experimentally, there is no obvious way to judge the results of a computation even if several protein structures with this peptide are available from X-ray analysis. But even then there is the complication of the free amino and carboxyl groups, which are absent when the peptide is part of a protein.

Confronted with the lack of immediate experimental evidence, pentapeptide identities are useful test cases for local structure calculations. We concentrate on the following questions.

If the peptide assumes two different conformations in the proteins, does the method calculate one of the alternatives?

Table 4

Ensembles of several peptides of erythrocruorin 1ECD obtained from the Boltzmann device

Peptide	Start	Type	Clusters	C1	R1	C2	R2	C3	R3	C4	R4
A. *Pentapeptides*											
AD-VNT	78	Unstable	10	6	3·12	3	1·31	2	1·74	2	2·52
D-VNTF	79	Unstable	8	8	0·67	3	3·20	3	2·69	1	0·94
VNTFV	80	Stable	1	20	3·30						
NTFV-A	81	Stable	3	17	3·36	2	0·65	1	2·44		
TFV-AS	82	Stable	4	17	3·32	1	2·65	1	2·72	1	2·86
B. *Hexapeptides*											
AD-VNTF	78	Stable	1	20	3·90						
D-VNTFV	79	Flip-flop	2	16	3·71	4	3·07				
VNTFV-A	80	Stable	1	20	3·76						
NTFV-AS	81	Metastable	6	14	3·68	2	3·26	1	1·97	1	3·15
C. *Heptapeptides*											
AD-VNTFV	78	Stable	2	19	4·14	1	3·32				
D-VNTFV-A	79	Flip-flop	4	10	0·43	6	4·12	3	3·63	1	3·77
VNTFV-AS	80	Unstable	5	9	4·18	7	3·95	2	3·54	1	3·62

See the legend to Table 3 for details.

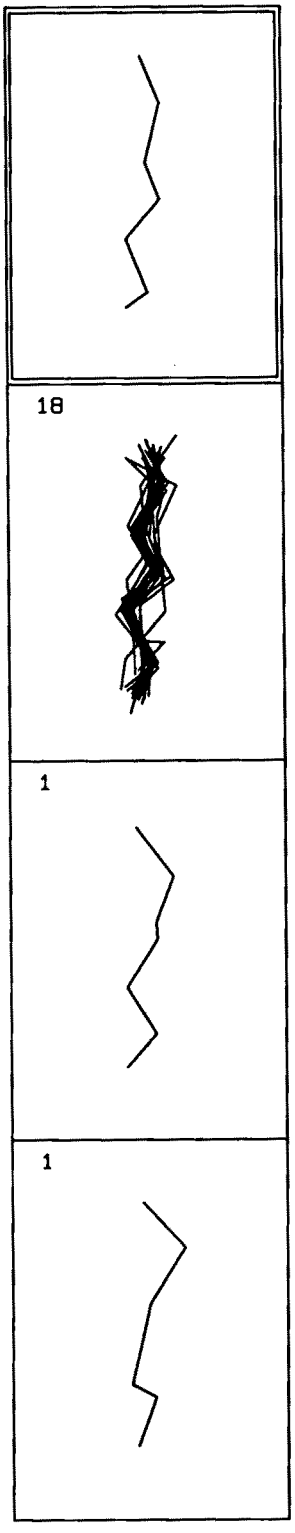

Figure 7. Ensemble obtained for the heptapeptide P-VNTFV-H of ribonuclease.

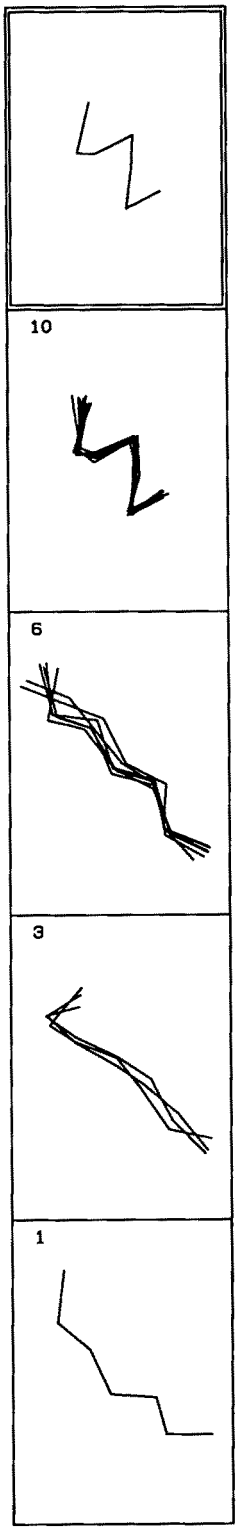

Figure 8. Ensemble obtained for the heptapeptide D-VNTFV-A of erythrocruorin.

If the peptide assumes identical conformations in two different proteins, does the method calculate this conformation?

What can be inferred from the results obtained for the flanking regions of the peptide identities?

The ensembles calculated from the Boltzmann device for several peptide identities are summarized in Figures 6 to 18 and Tables 3 to 8.

We start with VNTFV found in ribonuclease and erythrocruorin. The ensemble of VNTFV obtained

Table 5

Ensembles of several peptides of carbonic anhydrase 2CAB obtained from the Boltzmann device

Peptide	Start	Type	Clusters	C1	R1	C2	R2	C3	R3	C4	R4
A. *Pentapeptides*											
LQ-KVL	153	Stable	1	20	0·25						
Q-KVLD	154	Flip-flop	3	11	0·39	8	1·47	1	2·16		
KVLDA	155	Stable	1	20	0·13						
VLDA-L	156	Stable	1	20	0·17						
LDA-LQ	157	Stable	1	20	0·66						
B. *Hexapeptides*											
LQ-KVLD	153	Stable	1	20	0·27						
Q-KVLDA	154	Stable	1	20	0·17						
KVLDA-L	155	Stable	1	20	0·21						
VLDA-LQ	156	Stable	1	20	0·67						
C. *Heptapeptides*											
LQ-KVLDA	153	Stable	1	20	0·23						
Q-KVLDA-L	144	Stable	1	20	0·20						
KVLDA-LQ	155	Stable	1	20	0·65						

See the legend to Table 3 for details.

from the Boltzmann device consists of one cluster of β-sheet conformations (Fig. 6). Since there is only one preferred conformation VNTFV is a stable segment in our terminology. Moreover, in terms of the Boltzmann device the segment adopts its native conformation in ribonuclease but it is a forced segment in erythrocruorin.

To see whether the ensemble is affected by the flanking residues of VNTFV we calculated the conformations of the two heptapeptides D-VNTFV-A (erythrocruorin) and P-VNTFV-H (ribonuclease). Again for the ribonuclease segment (Fig. 7) the ensemble consists of a predominant cluster of β-strands and a few irregular strands, but the erythrocruorin segment yields an ensemble where the predominant cluster is α-helical and the rest are strand-like conformations (Fig. 8 and Tables 3 and 4).

A second example is the pentapeptide KVLDA that adopts an α-helix in carbonic anhydrase and a β-strand in prealbumin. KVLDA is a stable α-helical segment that adopts the native conformation in carbonic anhydrase (Fig. 9). It is forced to an extended conformation in prealbumin (Fig. 12). The segment Q-KVLDA-L of carbonic anhydrase is again in a stable α-helical conformation (Fig. 11).

Table 6

Ensembles of several peptides of prealbumin 2PAB obtained from the Boltzmann device

Peptide	Start	Type	Clusters	C1	R1	C2	R2	C3	R3	C4	R4
A. *Pentapeptides*											
LMV-KV	3	Stable	2	19	0·42	1	2·74				
MV-KVL	4	Stable	2	19	0·29	1	0·98				
V-KVLD	5	Stable	3	18	0·90	1	2·01	1	1·51		
KVLDA	6	Stable	1	20	3·17						
VLDA-V	7	Stable	1	20	2·95						
LDA-VR	8	Stable	1	20	2·12						
B. *Hexapeptides*											
LMV-KVL	3	Stable	3	18	0·42	1	1·18	1	1·73		
MV-KVLD	4	Stable	1	20	0·55						
V-KVLDA	5	Stable	3	16	3·73	3	4·21	1	1·21		
KVLDA-V	6	Stable	1	20	3·34						
VLDA-VR	7	Stable	1	20	2·74						
C. *Heptapeptides*											
LMV-KVLD	3	Stable	4	16	0·77	2	3·92	1	1·91	1	4·36
MV-KVLDA	4	Stable	1	20	4·19						
V-KVLDA-V	5	Flip-flop	5	14	3·91	3	4·32	1	0·48	1	1·44
KVLDA-VR	6	Stable	1	20	3·19						

See the legend to Table 3 for details.

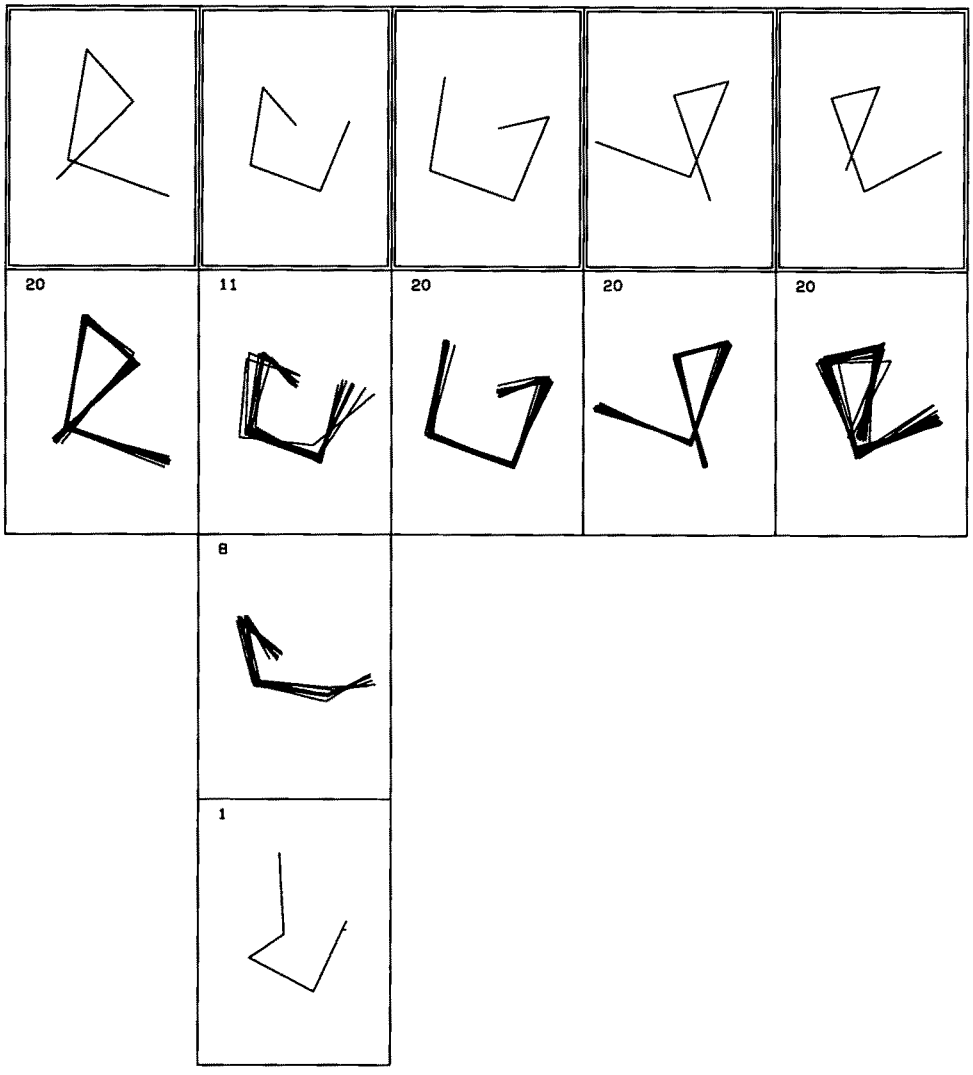

Figure 9. Ensembles obtained for the pentapeptides LQ-KVL, Q-KVLD, KVLDA, VLDA-L and LDA-LQ of carbonic anhydrase (from left to right).

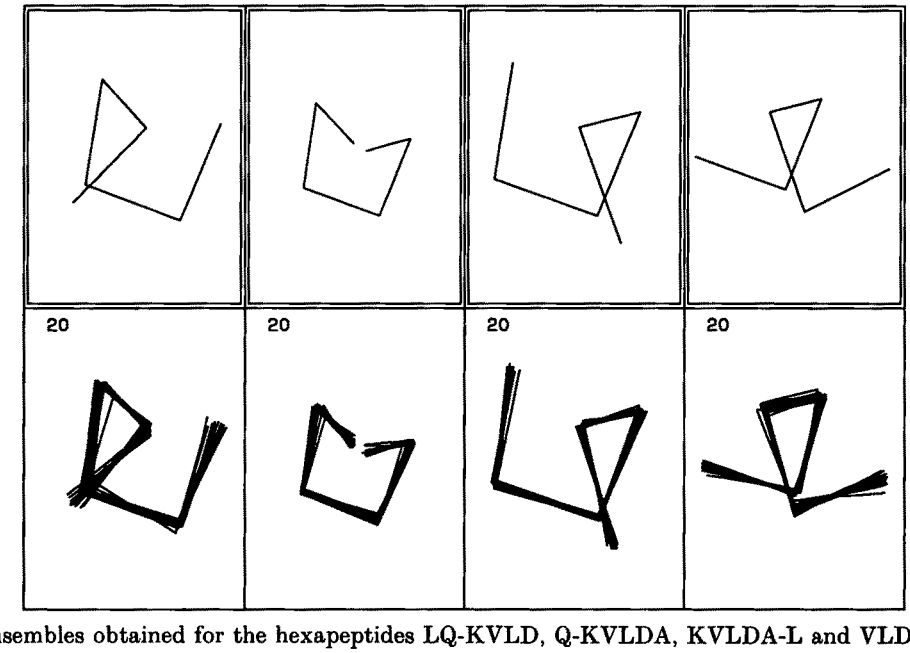

Figure 10. Ensembles obtained for the hexapeptides LQ-KVLD, Q-KVLDA, KVLDA-L and VLDA-LQ or carbonic anhydrase.

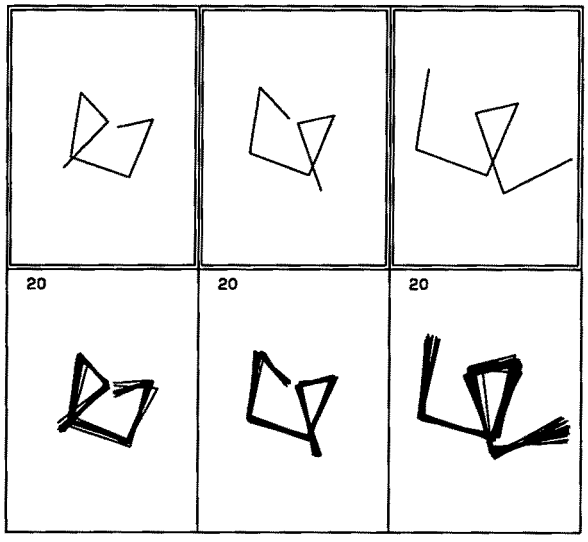

Figure 11. Ensembles obtained for the heptapeptides LQ-KVLDA, Q-KVLDA-L and KVLDA-LQ of carbonic anhydrase.

V-KVLDA-V, which is extended in prealbumin, produces an ensemble with a large α-helical cluster, a smaller cluster of helix termini and a few strand-like conformations (Fig. 14). In this case also the heptapeptide is in a forced conformation, but it has some affinity for extended conformations.

To investigate the region around KVLDA in carbonic anhydrase and prealbumin we calculated a series of penta-, hexa- and heptapeptides in the neighbourhood of KVLDA in both proteins. In Figures 9 to 14 and Tables 5 and 6 we summarize the results. All five pentapeptides LQ-KVL, Q-KVLD, KVLDA, VLDA-L and LDA-LQ of carbonic anhydrase produce ensembles consisting of helical clusters. All segments are stable with the exception of Q-KVLD which is a flip-flop segment consisting of an α-helix cluster, a cluster of C-terminal helix ends and a cluster of one element which is an N-terminal helix end. The heptapeptides LQ-KVLDA, Q-KVLDA-L and KVLDA-LQ are stable α-helix clusters. Therefore, the slight instability of Q-KVLD is ruled out when more topological levels are included in the computation.

Figure 12 and Table 6 show the ensembles for five pentapeptides of prealbumin. The ensemble of MV-KVL consists of one cluster of strands. V-KVLD is a stable segment, the major cluster consisting of strands with a starting turn on the C terminus. KVLDA, VLDA-V and LDA-VR are stable segments of α-helical type.

Similar results are obtained for the hexapeptides of prealbumin (Fig. 13). MV-KVLD is a stable strand and V-KVLDA is a stable segment whose ensemble consists of a large helix cluster, a smaller cluster of helix ends and a cluster of one conformation of strand type. KVLDA-A and VLDA-VR produce stable helices.

The heptapeptides LMV-KVLD, V-KVLDA-V and KVLDA-VR show the same conformational preferences (Fig. 14). Starting at the N terminus there is a strong tendency to form strands which suddenly switch over to strong α-helix potential. Note that a few strand-like conformations appear in the ensembles of MV-KVLDA and V-KVLDA-V, indicating that in spite of the predominating helix clusters there is still some tendency to form strands. The results indicate that within the pentapeptide KVLDA a high strand potential from the N-terminal side collides with a high helix potential from the C-terminal side, so that the energy in this part of the polypeptide chain may be delicately balanced.

As a third example we discuss the pentapeptide LPASQ which is found in acid proteinase and carboxypeptidase. The conformations of this peptide are very similar in both proteins yielding an r.m.s. deviation of 0·3 Å. Kabsch & Sander (1983) classified the conformations as helix end in acid proteinase and as 3_{10} helix in carboxypeptidase. In contrast to the examples presented above this peptide has a rather irregular conformation and it adopts quite similar conformations in the two proteins.

LPASQ is a stable segment whose conformational ensemble consists of a single cluster (Figs 15 and 17, Tables 7 and 8). The conformations in the cluster have an average r.m.s. deviation of 0·32 Å from the conformation in acid proteinase and 0·31 Å from the conformation in carboxypeptidase. This result indicates that the Boltzmann device is able to produce stable segments for conformations that are not of a regular secondary structure type.

We also examine the flanking regions of LPASQ in both proteins (Figs 15 to 18, Table 7 and 8). The pentapeptide E-LPAS of acid proteinase is a flip-flop segment consisting of two clusters of 15 and 5 members whose average r.m.s. value is 0·42 Å and 3·07 Å, respectively. PASQ-Q has one major cluster of 17 members (average r.m.s. = 0·31 Å) and two small clusters. The ensembles of the heptapeptides around LPASQ in acid proteinase are shown in Figure 16. TE-LPASQ is a stable segment with just one conformation in a second cluster. E-LPASQ-Q is a flip-flop segment and LPASQ-QS is a stable segment. Both have a tendency to form helical conformations. Note that the C terminus of LPASQ-QS seems to be very flexible.

The pentapeptide L-LPAS of carboxypeptidase (Fig. 17) is a metastable segment whose major cluster has an average r.m.s. value of 0·46 Å to the conformation in the protein. PASQ-I is a stable segment of 0·25 Å r.m.s. deviation to the X-ray structure. The two additional clusters have one element each, with an r.m.s. value of 0·83 Å and 1·54 Å and are also quite similar to the conformation in the protein.

The segment FL-LPASQ (Fig. 18) has a major cluster of α-helix (2·86 Å r.m.s.) and two small clusters similar to the conformation in the protein (1·24 Å and 0·87 Å r.m.s.). For L-LPASQ-I (flip-flop) and LPASQ-II (stable) the major clusters are again very similar to the conformation in the pro-

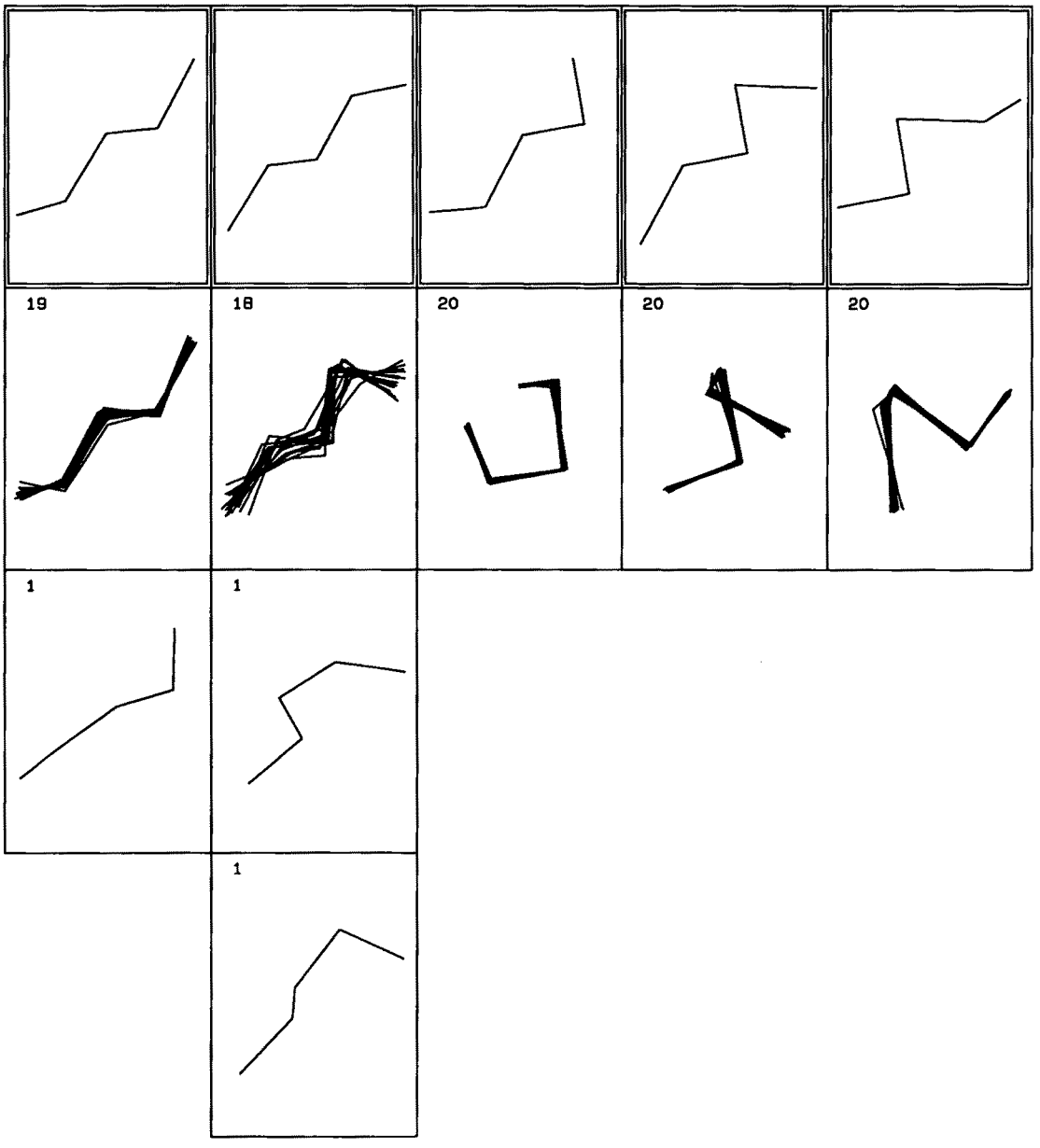

Figure 12. Ensembles obtained for the pentapeptides MV-KVL, V-KVLD, KVLDA, VLDA-V and LDA-VR of prealbumin.

Table 7

Ensembles of several peptides of acid proteinase 2APP obtained from the Boltzmann device

Peptide	Start	Type	Clusters	C1	R1	C2	R2	C3	R3	C4	R4
A. *Pentapeptides*											
TE-LPA	44	Metastable	8	12	0·82	2	1·95	1	1·77	1	1·82
E-LPAS	45	Flip-flop	2	15	0·42	5	3·07				
LPASQ	46	Stable	1	20	0·32						
PASQ-Q	47	Stable	3	17	0·31	2	1·38	1	2·22		
ASQ-QS	48	Flip-flop	8	10	2·33	3	1·32	2	2·47	1	0·72
B. *Hexapeptides*											
TE-LPAS	44	Metastable	6	14	1·48	2	2·44	1	1·82	1	2·10
E-LPASQ	45	Stable	2	18	0·52	2	2·93				
LPASQ-Q	46	Stable	1	20	0·53						
PASQ-QS	47	Flip-flop	5	13	1·18	4	2·34	1	2·12	1	2·34
C. *Heptapeptides*											
TE-LPASQ	44	Stable	2	19	1·38	1	1·78				
E-LPASQ-Q	45	Flip-flop	3	10	2·95	9	0·72	1	2·89		
LPASQ-QS	46	Stable	2	19	1·19	1	2·33				

See the legend to Table 3 for details.

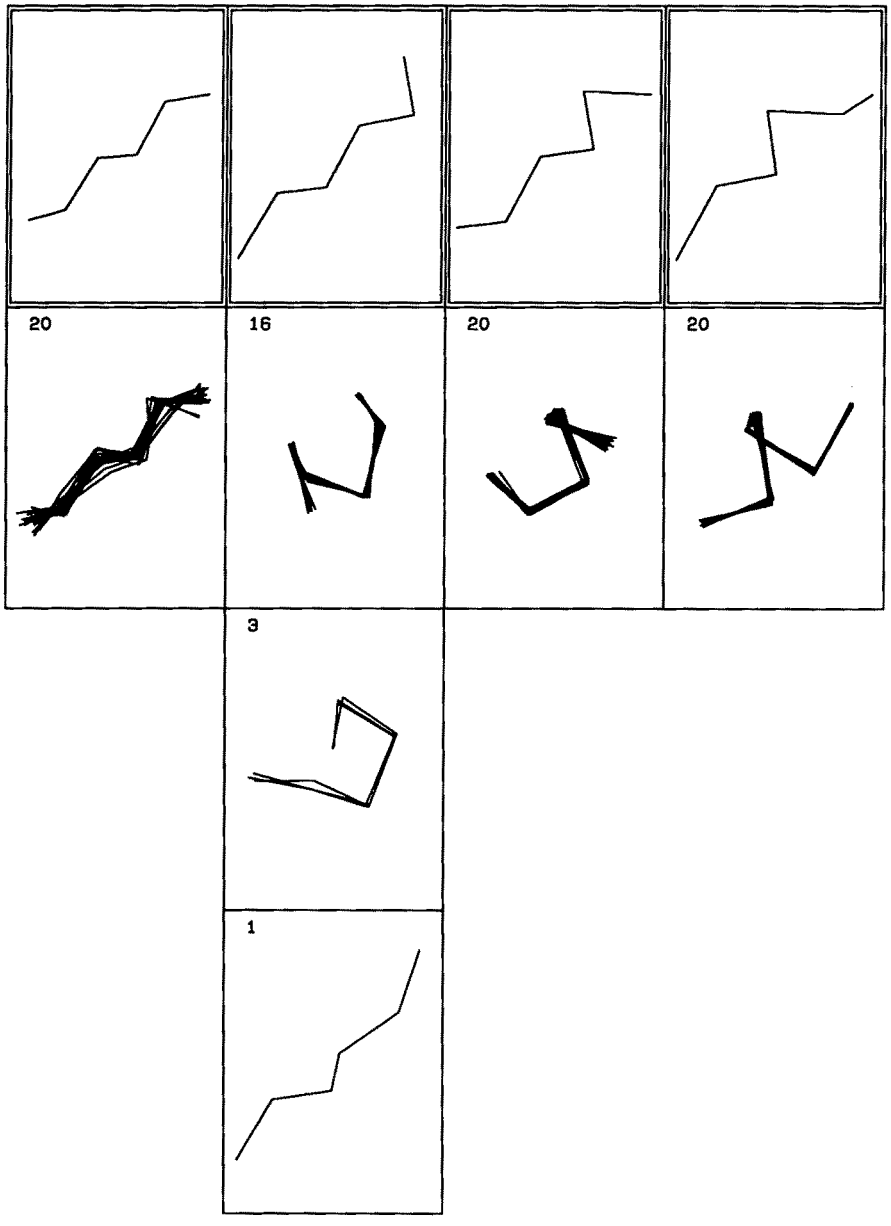

Figure 13. Ensembles obtained for the hexapeptides MV-KVLD, V-KVLDA, KVLDA-V and VLDA-VR of prealbumin.

Table 8

Ensembles of several peptides of carboxypeptidase 5CPA obtained from the Boltzmann device

Peptide	Start	Type	Clusters	C1	R1	C2	R2	C3	R3	C4	R4
A. *Pentapeptides*											
FL-LPA	278	Stable	4	17	2·37	1	1·15	1	1·89	1	2·11
L-LPAS	279	Unstable	9	9	0·46	4	2·83	2	1·50	2	2·87
LPASQ	280	Stable	1	20	0·31						
PASQ-I	281	Stable	3	18	0·25	1	0·83	1	1·54		
ASQ-II	282	Stable	2	19	0·92	1	1·93				
B. *Hexapeptides*											
FL-LPAS	278	Flip-flop	3	15	2·97	4	1·65	1	2·66		
L-LPASQ	279	Flip-flop	3	9	0·62	8	2·63	3	2·69		
LPASQ-I	280	Stable	1	20	0·51						
PASQ-II	281	Stable	3	17	0·93	2	0·76	1	1·68		
C. *Heptapeptides*											
FL-LPASQ	278	Metastable	3	16	2·86	3	1·24	1	0·87		
L-LPASQ-I	279	Flip-flop	2	16	0·80	4	2·69				
LPASQ-II	280	Stable	1	20	0·92						

See the legend to Table 3 for details.

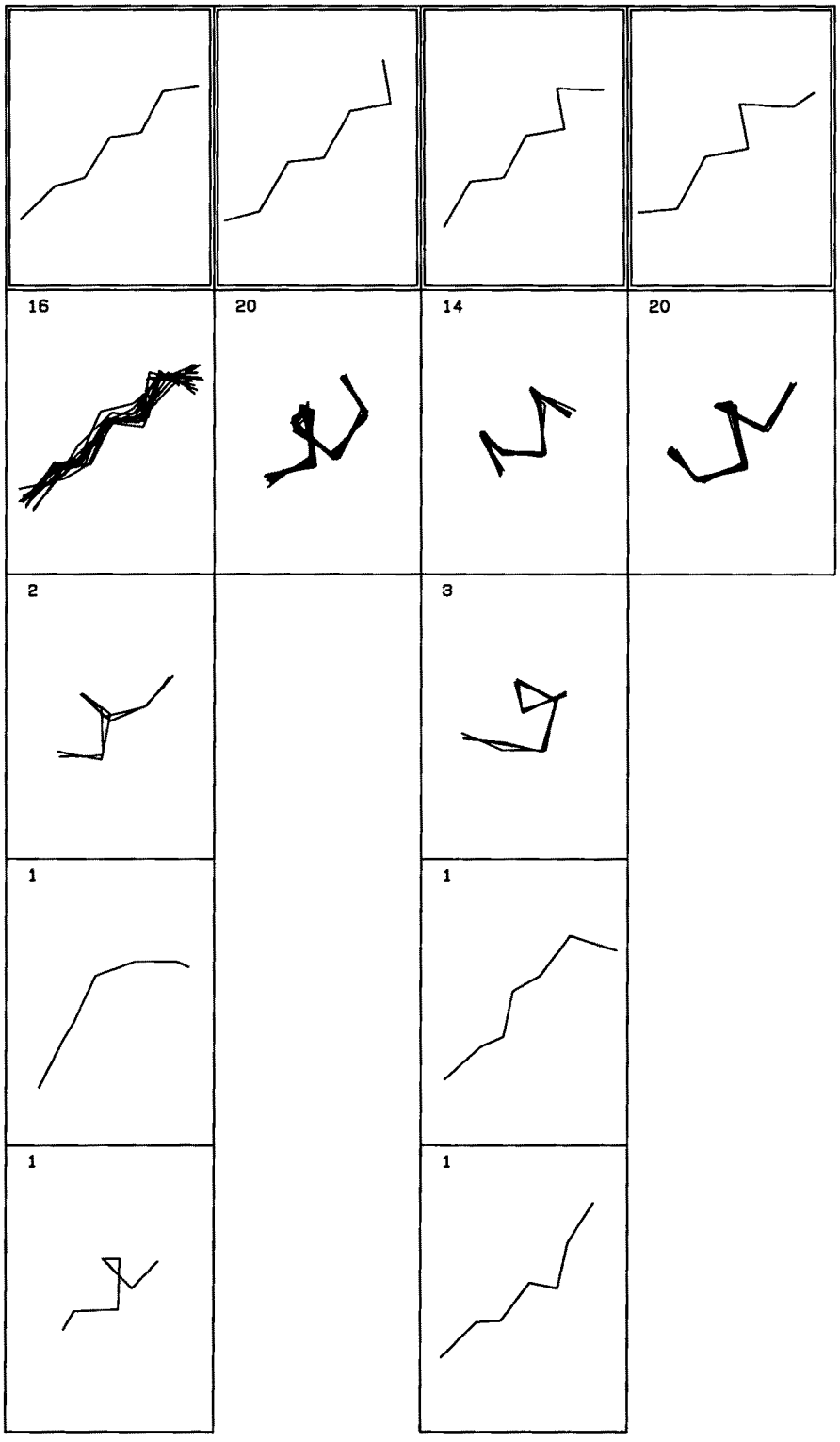

Figure 14. Ensembles obtained for the heptapeptides LMV-KVLD, MV-KVLDA, V-KVLDA-V and KVLDA-VR of prealbumin.

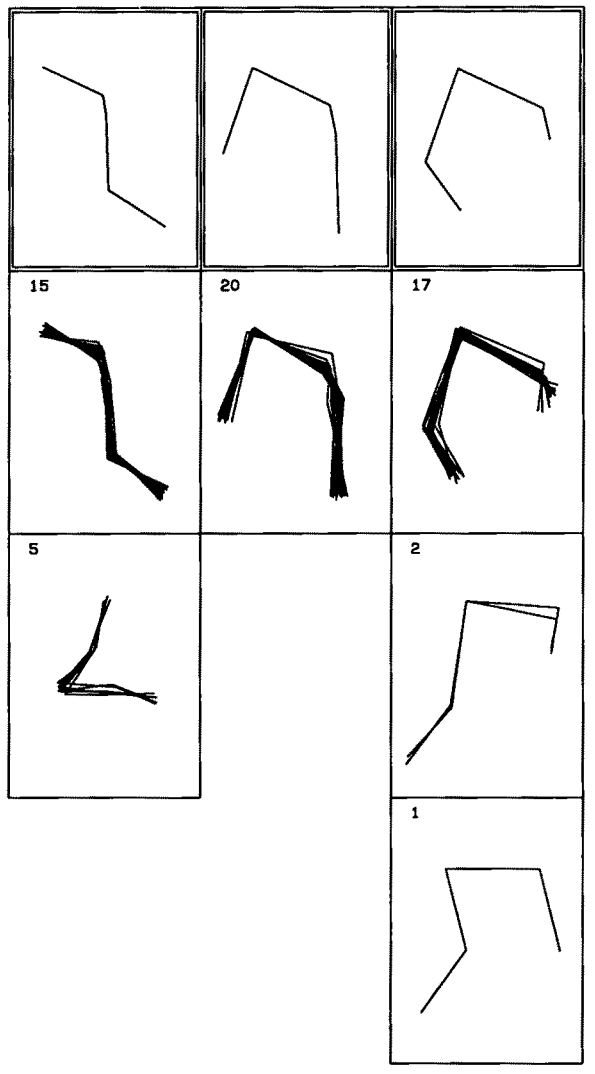

Figure 15. Ensembles obtained for the pentapeptides E-LPAS, LPASQ and PASQ-Q of acid proteinase.

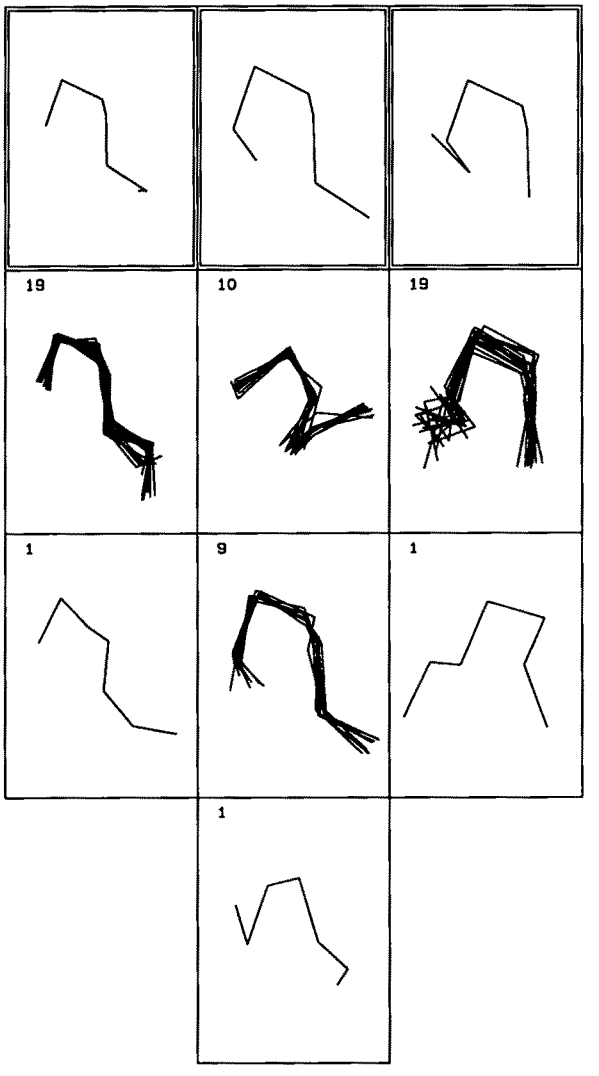

Figure 16. Ensembles obtained for the heptapeptides TE-LPASQ, E-LPASQ-Q and LPASQ-QS of acid proteinase.

tein with average r.m.s. values of 0·80 Å and 0·92 Å, respectively.

In summary the output of the Boltzmann device provides a consistent picture of the pentapeptide identities presented. VNTFV is a strand-forming peptide adopting its native conformation in ribonuclease. In erythrocruorin the peptide is forced to a helix by the flanking amino acids. KVLDA is an α-helix-forming peptide adopting its native conformation in carbonic anhydrase. In prealbumin a high helix potential from the C terminus collides with a strong strand-forming potential so that the energy in the KVLDA region is delicately balanced. The observed X-ray conformation of this region is probably determined by long-range interactions of KVLDA with other parts of prealbumin. Finally, LPASQ is found in similar conformations in acid proteinase and carboxypeptidase. The Boltzmann device calculates a stable segment whose structure matches the conformations found in the proteins. This may indicate that LPASQ is a segment of very stable conformation so that it may serve as a helix start or stop signal in protein conformations.

Kabsch & Sander (1984) found 25 pentapeptide identities. Over the last few years the data set of proteins in the Brookhaven protein data bank has grown considerably, so that we may expect a few more sequence identities in the enlarged set. A detailed computation on these identities will be published elsewhere.

12. Discussion

The results obtained indicate that the Boltzmann device might be a useful approach to the calculation of ensembles of local structures of peptides and proteins. They also show that the conformational behaviour of ensembles of short segments is quite complex and delicately balanced. Some sequences produce ensembles of one preferred type of confor-

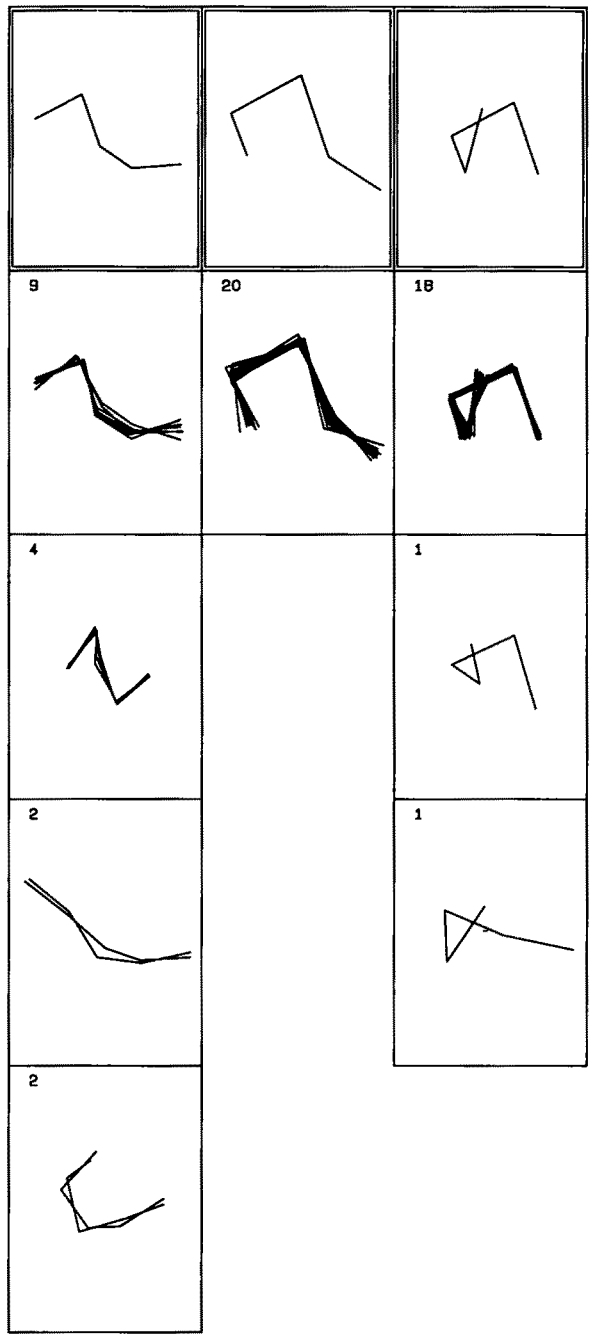

Figure 17. Ensembles obtained for the pentapeptides L-LPAS, LPASQ and PASQ-I of carboxypeptidase.

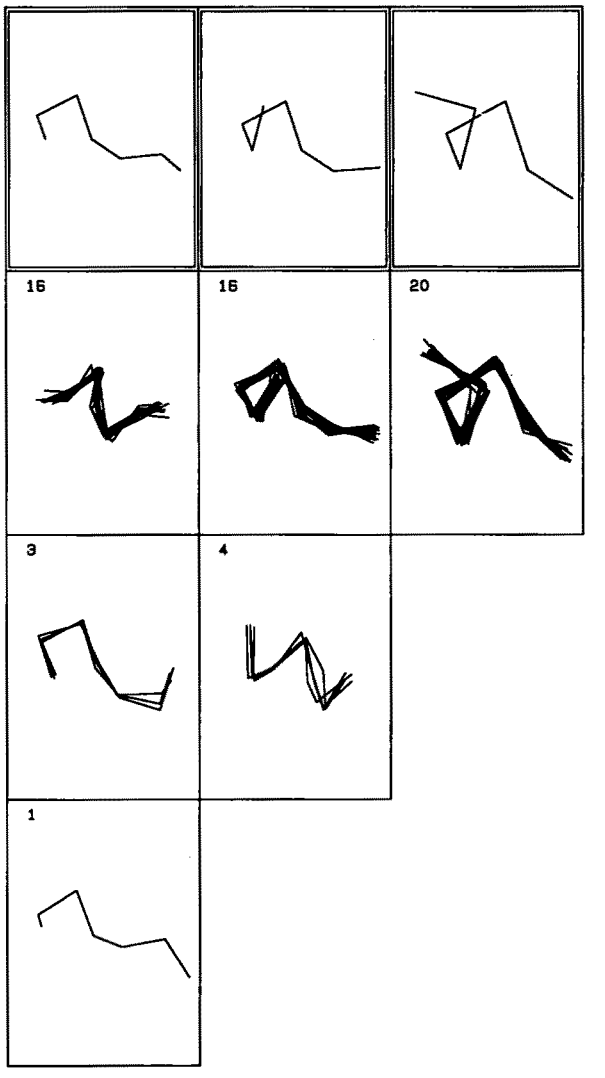

Figure 18. Ensembles obtained for the heptapeptides FL-LPASQ, L-LPASQ-I and LPASQ-II of carboxypeptidase.

mation, others yield ensembles with a spectrum of conformations. Conformational ensembles may change dramatically when sliding along the chain by a single residue. VKLDA of prealbumin, for example, prefers extended conformations whereas KVLDA prefers α-helix, although the pentapeptides have four residues in common.

We emphasize that the Boltzmann device as presented is a prototype. The few adjustable parameters of the device have not been tuned to optimize the results. With the exception of the weighting scheme equation (23), no transformations have been applied to the distributions $g_k^{ab}(s)$ and net potentials $\Delta E_k^{ab}(s)$.

We point out that there is a numerical problem associated with the distributions at the extreme points l_k and u_k. With the assumption $Z_k^{ab} \approx Z_k$ the net potentials $\Delta E_k^{ab}(s)$ should approach zero for $d_{ij} \to l_k$ and $d_{ij} \to u_k$. As shown in Figures 3 and 4 the potentials in most cases have rather large positive values near l_k and u_k. The boundaries correspond to high-energy regions and the densities $f_k(s)$ and $f_k^{ab}(s)$ are very small at these points. Hence, there are only a few observations of distances in this region. Therefore, at the boundaries the term $g_k^{ab}(s)/f_k(s)$ of equation (23) is numerically unstable since a small quantity is divided by another small quantity. It should be easy to stabilize this term by some useful weighting scheme. In the present study this is not a severe problem since, whenever $f_k(s)$ is small, there are only a few conformations in the pool whose distances match these extreme values.

Apart from the problems associated with sparse data sets there should be ample room for improvement of the Boltzmann device. We find that every additional type of distance refines the output considerably. There are two important reasons for the improvements. Firstly, additional distances define the conformations more precisely in geometrical terms, and secondly every type of distance (e.g. N–O) corresponds to an individual atomic interaction that helps to refine the results in energetical terms. Therefore, the informational load of the Boltzmann device is increased by every additional type of distance. So far we used the C^{α}–C^{α}, C^{β}–C^{β}, N–N, O–O, N–O and O–N distances. There are still many more distances in the polypeptide backbone that can be included in the computations (C^{α}–C^{β}, C^{α}–C′, C^{α}–N, etc.).

We might expect that the Boltzmann device will improve with the growing set of proteins in the Brookhaven protein data bank. Currently, in the mean, we have 45 measurements for each amino acid pair. It is likely that the data set will double over the next few years so that we may expect of the order of 100 measurements on the individual pairs. This will refine considerably the approximations to the net potentials $\Delta E_k^{ab}(s)$ and may lead to quite reliable calculations on conformational ensembles of short peptides.

Recently Rooman & Woodak (1988) concluded from their studies, that a data base of roughly 1500 unrelated protein structures is required in order to obtain residue patterns of reasonable predictive power. In view of the theory and the results presented here we estimate that 200 to 300 structures will suffice to obtain a set of quite reliable potentials.

We note that in the case of a larger data base some pairs like $(\text{Met, Trp})_k$ will still occur with very low frequency. Since this pair is rare there will be only a few sequences of the type $(\text{Met, Trp})_k$. Hence, the lower the frequency of a given pair, the lower the probability that we will find that pair in the sequence under consideration. If the sequence at hand contains $(\text{Met, Trp})_k$ the pair will usually be surrounded by amino acids (and pairs) of higher frequencies. Hence, the contribution of $(\text{Met, Trp})_k$ to $\Delta E(S_q,C_p)$ will be small and we should get a useful ensemble as long as the interactions of $(\text{Met, Trp})_k$ do not critically determine the conformational stability of the peptide.

The net potentials $\Delta E_k^{ab}(s)$ obtained from the data base are free energies since entropic effects are included in the crystal structures. This is an appealing feature of the Boltzmann device, since the evaluation of the entropy of states is cumbersome and often impossible on computational grounds (e.g. see Israelachvili, 1985). Similar problems prevail with respect to the nature and physical origin of electrostatic and solvent forces in proteins. Still their physical basis remains largely unknown and modelling of these forces is a difficult theoretical task and computational burden (Israelachvili, 1985; Mathew, 1985; Klapper *et al.*, 1986; Gilson & Honig, 1988*a*,*b*; Harvey, 1989; Nemethy *et al.*, 1981; Pratt & Chandler, 1977, 1980). As discussed above all types of interactions whether electrostatic, solvent or steric in nature, contribute to the net potentials. Therefore, using the Boltzmann device, we do not have to use models of electrostatic and solvent effects. They are included in the net potentials. The net potentials $\Delta E_k^{ab}(s)$ may even serve to investigate the physical basis of these forces once the data set is large enough so that the approximations are sufficiently close to the actual potentials.

In the present study we concentrated on the local aspects of protein conformations. However, the approach presented is not restricted to oligopeptides or short segments in proteins. In Figures 9 to 18 we present clusters for a series of overlapping segments. From Figures 9, 10 and 11 it is obvious that the overlapping segments can be joined to yield a single consistent conformation for the nine residues LQKVLDALQ. The same applies to the clusters shown in Figures 15 and 17. Again the largest clusters can be joined to yield a contiguous conformation for ELPASQQ and LLPASQI, respectively. In addition, the procedure to join overlapping segments can be used to choose an appropriate cluster from flip-flop, metastable and unstable segments. An example can be found in Figure 16. The second largest cluster of LLPASQI can be connected to the largest clusters of the adjacent segments to yield a consistent conformation for FLLPASQII. By joining overlapping segments we may be able to assemble quite extensive regions of protein chains. We are currently proceeding along these lines.

Using the concepts presented the calculation of medium and long-range forces in globular proteins is straightforward. The distinction between topological levels becomes less important with increasing k values. Therefore, we are able to compute the potentials of mean force from a combination of several topological levels, which considerably increases the statistical significance of the potentials. It is conceivable that these potentials will be useful in energy minimization, molecular dynamics, and modelling of globular proteins by homology. Applications in these fields are in progress.

Finally we note that the approach is not confined to distances. Any other physical state variable may be included in the Boltzmann device. Moreover, the Boltzmann device is a general device that is applicable to any molecular or microscopic system provided a sufficiently large number of measurements is available from experiment. It is conceivable that the theory presented can be used to calculate inter- and intramolecular net potentials of mean force for a number of molecular systems, e.g. nucleic acids, carbohydrates and inorganic compounds, for which large data bases have accumulated through the immense work of X-ray crystallographers over the last few decades. Once more we emphasize that valuable applications will critically depend on the appropriate choice of the reference states.

I am indebted to all X-ray crystallographers who submitted co-ordinates to the Brookhaven protein data bank. A major part of the summary was copied from a referees report. I acknowledge the immensely helpful comments made by the referees. The points raised yielded a more concise version of the manuscript. Hence, also the interested reader is indebted to the referees. Thanks to the confidential scientific editing system the referees will remain uncited.

Large portions of the computer programs used were written by Peter Lackner. Drawings of relative frequencies and potentials (Figs 1 to 4) were prepared by Karl Gottsbacher. Figures on conformational ensembles (Figs 6 to 18) were designed by Manfred Hendlich. Hans B. Strack and George Casari made useful suggestions and Sylvia Sippl wiped out useless phases. This work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung under project number P7262-BIO.

References

Anfinsen, C. B. (1973). *Science*, **181**, 223–230.

Baumann, G., Frömmel, C. & Sander, C. (1989). *Protein Eng.* **2**, 329–334.

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

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

Blundell, T. L., Carney, D., Gardner, S. Hayes, F., Howlin, B., Hubbard, T., Overington, J. Singh, D. A., Sibanda, B. L. & Sutcliff, M. (1988). *Eur. J. Biochem.* **172**, 513–520.

Chothia, C., Levitt, M. & Richardson, D. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 4130–4134.

Chou, P. Y. & Fasman, G. D. (1974). *Biochemistry*, **13**, 222–245.

Chou, P. Y. & Fasman, G. D. (1978). *Annu. Rev. Biochem.* **47**, 251–276.

Dyson, H. J., Rance, M., Houghten, R. A., Lerner, R. A. & Wright, P. E. (1988*a*). *J. Mol. Biol.* **201**, 161–200.

Dyson, H. J., Rance, M., Houghten, R. A., Wright, P. E. & Lerner, R. A.(1988*b*). *J. Mol. Biol.* **201**, 201–217.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982). *Nature (London)*, **299**, 371–374.

Frauenfelder, H., Parak, F. & Young, R. D. (1988). *Annu. Rev. Biophys. Biophys. Chem.* **17**, 451–479.

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

Gill, P. E., Murray, W. & Wright, M. H. (1981). *Practical Optimization*, Academic Press, New York.

Gilson, M. K. & Honig, B. (1988*a*). *Proteins*, **3**, 32–52.

Gilson, M. K. & Honig, B. (1988*b*). *Proteins*, **4**, 7–18.

Goldberg, M. E. (1985). *Trends Biochem. Sci.* **10**, 388–391.

Gurd, F. R. N. & Rothgeb, T. M. (1979). *Advan. Protein Chem.* **33**, 73–165.

Hagler, A. T. & Honig, B. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 554–558.

Hall, D. & Lyons, P. J. (1980). *Comput. Chem.* **4**, 69–72.

Harvey, S. C. (1989). *Proteins*, **5**, 78–92.

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

Israelachvili, J. N. (1985). *Intermolecular and Surface Forces*, Academic Press, London.

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

Kabsch, W. & Sander, C. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 1075–1078.

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

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

Klapper, I., Hagstrom, R., Fine, R., Sharp, K. & Honig, B. (1986). *Proteins*, **1**, 47–59.

Kohonen, T. (1984). *Self-Organization and Associative Memory*, Springer Verlag, Berlin.

Kossiakoff, A. A. (1985). *Annu. Rev. Biochem.* **54**, 1195–1227.

Kruskal, J. B. (1956). *Proc. Amer. Math. Soc.* **71**, 48–50.

Lathrop, R. H., Webster, T. A. & Smith, T. F. (1987). *Commun. ACM*. **30**, 909–921.

Levitt, M. (1976). *J. Mol. Biol.* **128**, 59–107.

Levitt, M. (1982). *Annu. Rev. Biophys. Bioeng.* **11**, 251–271.

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

Mathew, B. W. (1985). *Annu. Rev. Biophys. Bioeng.* **14**, 387–417.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A.. (1975). *J. Phys. Chem.* **79**, 2361–2381.

Needleman, S. & Wunsch, C. (1970). *J. Mol. Biol.* **48**, 443–453.

Nemethy, G., Peer, W. J. & Scheraga, H. A. (1981). *Annu. Rev. Biophys. Bioeng.* **10**, 459–497.

Novotny, J., Bruccoleri, R. E. & Karplus, M. (1984). *J. Mol. Biol.* **177**, 787–818.

Petsko, G. A. & Ringe, D. (1984). *Annu. Rev. Biophys. Bioeng.* **13**, 331–371.

Pratt, L. & Chandler, D. (1977). *J. Chem. Phys.* **67**, 3683–3704.

Pratt, L. & Chandler, D. (1980). *J. Chem. Phys.* **73**, 3434–3441.

Purisama, E. O. & Scheraga, H. A. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 2782–2786.

Richardson, J. S. & Richardson, D. C. (1988). *Science*, **240**, 1648–1652.

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

Schulz, G. E. (1988). *Annu. Rev. Biophys. Biophys. Chem.* **17**, 1–21.

Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). *J. Mol. Biol.* **206**, 759–777.

Sippl, M. J. (1982). *J. Mol. Biol.* **156**, 359–388.

Sippl, M. J. & Scheraga, H. A. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 2197–2201.

Sippl, M. J. & Scheraga, H. A. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 2283–2287.

Skolnick, J. S., Kolinski, A. & Yaris, R. (1989). *Proc. Nat. Acad. Sci., U.S.A.* **86**, 1229–1233.

Tanaka, S. & Scharaga, H. A. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 3802–3806.

van Gunsteren, W. F. & Karplus, M. (1981). *Nature (London)*, **293**, 677–678.

Vasquez, M. & Scheraga, H. A. (1985). *Biopolymers*, **24**, 1437–1447.

Edited by R. Huber