PROTEINS: Structure, Function, and Genetics 13:258–271 (1992)

Detection of Native-Like Models for Amino Acid Sequences of Unknown Three-Dimensional Structure in a Data Base of Known Protein Conformations

Manfred J. Sippl and Sabine Weitckus
Institute for General Biology, Biochemistry and Biophysics, Department of Biochemistry, University of Salzburg, A-5020 Salzburg, Austria

ABSTRACT **We present an approach which can be used to identify native-like folds in a data base of protein conformations in the absence of any sequence homology to proteins in the data base. The method is based on a knowledge-based force field derived from a set of known protein conformations. A given sequence is mounted on all conformations in the data base and the associated energies are calculated. Using several conformations and sequences from the globin family we show that the native conformation is identified correctly. In fact the resolution of the force field is high enough to discriminate between a native fold and several closely related conformations. We then apply the procedure to several globins of known sequence but unknown three dimensional structure. The homology of these sequences to globins of known structures in the data base ranges from 49 to 17%. With one exception we find that for all globin sequences one of the known globin folds is identified as the most favorable conformation. These results are obtained using a force field derived from a data base devoid of globins of known structure. We briefly discuss useful applications in protein structural research and future development of our approach.** © 1992 Wiley-Liss, Inc.

Key words: protein folding, protein modeling, knowledge-based prediction, molecular force field, statistical mechanics, globins

INTRODUCTION

The native conformation of proteins in general solely depends on the amino acid sequence and the surrounding solvent.[1] Therefore, it should be possible to model protein solvent systems by suitable force fields describing the inter- and intramolecular interactions which stabilize the native state. The basic requirement for such force fields is that the native conformation of a protein matches the global minimum on the corresponding energy surface or, more precisely, that the native state corresponds to the lowest local minimum accessible in the approach to equilibrium. The successful construction of such force fields would allow the prediction of native protein folds from amino acid sequences, which in turn would have an enormous impact in biochemistry, molecular biology, and related fields.

During the last two decades many research activities devoted to the protein folding problem have focused on the development of force fields based on electrostatic and van der Waals interactions.[2–6] Using such force fields it is still impossible to distinguish native folds from grossly misfolded models.[7,8] Recently we reported an approach for the construction of knowledge-based force fields consisting of potentials of mean force describing the intramolecular interactions within protein molecules.[9] The potentials are derived from a data base of known protein structures. We demonstrated that using this force field native folds can be identified among a large number of incorrect models indicating that our force field is indeed a reasonable model for protein solvent systems.[10] This was confirmed by the construction of model conformations for thymosin β-4, a polypeptide consisting of 43 amino acids, from the amino acid sequence alone which are in good agreement with conformations obtained from nuclear magnetic resonance studies.[11]

The number of possible conformational states of typical proteins (> 50 residues) is astronomically large. Using conventional optimization procedures the search for the global minimum is computationally prohibitive. It is therefore necessary to develop efficient algorithms and strategies to reduce the complexity of the search problem. At the present state of knowledge it seems likely that proteins adopt a limited number of spatial architectures and

Received March 14, 1991; revision accepted September 6, 1991.

Address reprint requests to Dr. Manfred J. Sippl, Department of Biochemistry, Institute for General Biology, Biochemistry, and Biophysics, the University of Salzburg, Hellbrunnerstrasse 34, A-5020 Salzburg, Austria.

that larger proteins are in many cases composed of modules which are found in a number of seemingly unrelated proteins (e.g., ref. 12). A most striking example is provided by mandelate racemase and muconate lactonizing enzyme. The enzymes are mechanistically distinct having nonhomologous sequences but identical three-dimensional folds.[13] Therefore, even in the absence of sequence homologies to proteins solved by X-ray analysis or nuclear magnetic resonance chances are that the unknown fold of a given sequence resembles the architecture of some protein of known structure. If such structures or modules can be identified in the data base they will provide excellent models for the unknown fold and subsequent refinement could yield the correct native conformation.

Here we present a method based on our knowledge based mean field which is able to detect native-like folds in a data base of known structures. Using sequences from the well-characterized globin family as a test case we show that the globin fold can be identified even for the most distantly related globin sequences (17% homology to globins of known structure). This is even possible when the data base used to compile the force field does not contain globins of known structure. In this case the force field is obtained from a recombination of measurements on a completely unrelated set of proteins containing no information on specific features of the globin fold.

The presentation of our approach will proceed along the following lines. We start with a brief review of the compilation of potentials of mean force and the parameters required in the calculations. Using amino acid sequences and conformations taken from the globin family we set up an experiment designed to demonstrate that unknown folds can be identified if related conformations exist in the data base. We finally discuss the implications of our method and we indicate future developments.

METHODS

In summary the procedure involves the following steps: (1) potentials of mean force are compiled from a data base of known protein structures, (2) a pool of conformations is obtained from a data base of known protein structures by generating all possible fragments of length L, the length of the amino acid sequence of interest, and (3) the amino acid sequence is combined with all fragments in the pool, the associated total energy is calculated, and the conformations are ranked with respect to their energy. Fragments which are likely candidates for the unknown fold must have low energy and will appear on top of the energy sorted list.

The compilation of potentials of mean force as well as the construction of the pool of conformations require a data base of known protein structures. In many cases these data bases are not identical for reasons which will be discussed below. We therefore, refer to the data base used to compile the potentials of mean force by FDB (force field data base) and we call the data base used to generate the pool PDB (pool data base).

Calculation of Potentials of Mean Force

We briefly review the calculation of potentials of mean force and define the parameters used in the present study. A full account of the concepts involved is found in refs. 9 and 10.

The net potentials of mean force $\Delta E_k^{ac,bd}(s)$ are calculated from a data base of known protein structures according to

$$\Delta E_k^{ac,bd}(s) = -kT \ln\left[\frac{1}{1 + \sigma m_{ab}} + \frac{\sigma m_{ab}}{1 + \sigma m_{ab}} \frac{g_k^{ac,bd}(s)}{g_k^{c,d}(s)}\right]$$

where the discrete variable s represents the distance between atoms c and d of amino acids a and b, respectively, and k is the separation of a and b along the amino acid sequence. On the right hand side the function $g_k^{ac,bd}(s)$ represents the relative frequency of c and d in the distance interval corresponding to s. The reference state $g_k^{c,d}(s)$ represents the relative frequency of c and d as a function of s averaged over all amino acid pairs. Via m_{ab} and σ the expression on the right-hand side takes care of the different frequencies of individual atomic pairs. $m_{a,b}$ is the total frequency of the amino acid pair a and b of sequential separation k in the data base and σ is the weight of a single observation of a particular distance. kT is Boltzmann's constant times absolute temperature ($T = 293$).

The net potential $\Delta E^{ac,bd}(s)$ can be considered as an average potential describing the interaction of atoms c and d of amino acids a and b with respect to a reference state which is defined as an average over all amino acid pairs. Such a potential contains electrostatic, van der Waals, as well as hydrophobic interactions and the form of the potential is a consequence of the relative strengths of the individual contributions.

Calculation of Total Conformational Energies

After compilation of the potentials of mean force from the data base (FDB) the potential surface for a given protein is obtained from a recombination of potentials as a function of the amino acid sequence. The total net energy $\Delta E(S,C)$ of an amino acid sequence S in a particular conformation C is computed as

$$\Delta E(S,C) = \sum_{ij} \Delta E_k^{ac,bd}(s_{ij})$$

where a, b, c, d, and k are functions of the atom indices i and j. Note that we consistently use the

TABLE I. Parameters Used to Compile the Potentials of Mean Force

Level*	Interaction	Intervals†
1	C_β–C_β	20
2	C_β–C_β	20
3	C_β–C_β	20
4	C_β–C_β	20
5	C_β–C_β	20
6	C_β–C_β	20
7	C_β–C_β	20
8	C_β–C_β	20
9	C_β–C_β	20
10	C_β–C_β	20
11–30	C_β–C_β	30
31–60	C_β–C_β	30
61–100	C_β–C_β	30
101–150	C_β–C_β	30

*Structural levels (sequential separation k) used to compile the potentials of mean force. Levels 1–10 are compiled independently. Higher levels are condensed to a single interaction type.
†Number of intervals used to compile the potentials on the respective levels.

TABLE II. Data Base Used to Compile the Potentials of Mean Force*

155C	156B	1ABP	1ACX	1ALC
1BDS	1BP2	1CC5	1CMS	1CPV
1CRN	1CSE-E	1CSE-I	1CTF	1CTX
1CY3	1ETU-1	1ETU-2	1FC2-C	1FC2-D
1FDX	1FX1	1FXB	1GCN	1GCR
1GD1-Q	1GOX	1GOX	1GP11A	1GP12A
1HIP	1HMG-A	1HMG-B	1HMQ-A	1HNE-E
1HOE	1I1B	1IGE-A	1LDB-A	1LDB-C
1LDB-D	1LZ1	1LZT	1MCP-H	1MCP-L
1MLP-A	1MLT-A	1NXB	1PAZ	1PCY
1PFC	1PHH	1PP2-L	1PPT	1PRC-C
1PRC-H	1PRC-L	1PRC-M	1PYP	1R69
1REI-A	1RHD	1RN3	1SGT	1SN3
1TGS-I	1TIM-A	1TNF-A	1TON-A	1TON-B
1TPP	1UBQ	1UTG	1WSYAA	1WSYBA
1WSY-A	1WSY-B	2AAT	2ABX-A	2ACT
2ALP	2AZA-A	2B5C	2CA2	2CAB
2CCY-A	2CDV	2CI2-I	2CNA	2CPP
2CRO	2CTS	2CYP	2EST-E	2FB4-H
2FB4-L	2GLS-A	2GN5	2HLA-A	2HLA-B
2INS-A	2INS-B	2LBP	2LZM	2MEV-1
2MEV-2	2MEV-3	2MEV-4	2MHR	2MT2
2OVO	2PAB-A	2PKA-A	2PKA-B	2PLVB1
2PLVB4	2PLV-3	2PRK	2RNT	2RSPAA
2RSPBA	2SBT-1	2SGA	2SNS	2SOD-O
2SSI	2STV	2TAA-A	2TBV1A	2TBV2A
2TMV-P	2TS1-A	2TS1-B	2WRP-R	351C
3ADK	3APR-E	3C2C	3CLN	3DFR
3FXC	3GAP-A	3GRS	3ICB	3PFK
3PGK	3PGM	3RP2-A	3RXN	3TLN
3WGA-A	4ATC-A	4ATC-B	4CPA-I	4DFR-A
4FD1	4FXN	4HVPAA	4HVPBA	4MDH-A
4PEP	4RHV-1	4RHV-2	4RHV-3	4RHV-4
4SBV-A	5API-A	5API-B	5CHA1A	5CHA2A
5CPA	5CYT-R	5HIR	5PTI	5TNC
6LDH	7CAT-A	8ADH	8DFR	9PAP
1ECD	1LH1	1MBA	1MBD	2HHB-A
2HHB-B	2LHB			

*Codes are identical to the codes used in the Brookhaven Protein Data Bank. If the entry contains multiple chains the chain identifier is appended. The upper part corresponds to the FDB data base (no globins) used to compile the potentials of mean force. In addition PDB contains seven globins (lower two lines).

symbol $\Delta E(S,C)$ to distinguish the total net energy, which is defined with respect to a particular reference state from the total energy which is not.

In the present study we consider C_β–C_β iteractions only. All interactions among the remaining atoms are neglected. The potentials for sequential separations $k < 11$ were compiled independently. For larger separations we combined several k values.[10] Table I lists the parameters used to calculate the potentials on the various structural levels k.

Generation and Ranking of Model Conformations

Since the total net energy is evaluated with respect to C_β–C_β interactions we can directly use the conformations from a data base of known structures as possible candidates for the unknown conformation since with the exception of glycine all amino acid residues carry C_β atoms. For glycine residues we generated virtual C_β atoms so that the conformations in PDB resemble polyalanine chains. No additional manipulation of the conformations like optimization of side chain side chain contacts[8] is necessary to obtain consistent model folds.

To identify probable models for the unknown fold of a given amino acid sequence S we combine S with all possible conformations in the data base (PDB). We obtain a pool of conformations C_p by taking all possible fragments of length L from PDB, where L is the length of the amino acid sequence of interest. For moderate sizes of L ($L \approx 150$) we obtain in the order of 10,000 fragments from our current data base (Table II).

The sequence S is mounted on all fragments C_p, the total net energy is calculated and the conformations are sorted with respect to their energy. In terms of our knowledge-based mean field low total net energy is a necessary condition for the native conformation. The fragments of low net energy obtained from the pool are, therefore, possible candidates for the unknown conformation of S. It is clear, however, that this will be the case only when the architecture of the unknown fold is represented in the data base. In all other cases the conformations of lowest net energy cannot be suitable models of the unknown fold.

The total net energy $\Delta E(S,C)$ is comparable

TABLE III. Globin Sequences and Maximum Homologies*

Code†	Code‡	Ref.§	Globin	Homology**	
HHB-A	2HHB-A	18	Human hemoglobin α-chain	41	2HHB-B
HHB-B	2HHB-B	18	Human hemoglobin β-chain	41	2HHB-A
LHB	2LHB	19	Sea lamprey hemoglobin	30	2HHB-A
MBA	1MBA	20	Sea hare myoglobin	29	2LHB
MBD	1MBD	21	Sperm whale myoglobin	24	2HHB-A
ECD	1ECD	22	Midge larve hemoglobin	23	1MBA
LH1	1LH1	23	Yellow lupin leghemoglobin	22	1MBA
BBL	lgb1xvicfa	24	Broad bean leghemoglobin I	49	1LH1
MLG	glbxxchith	25	Midge larva globin CTT-X	39	1ECD
PNH	hbplxparad	26	Parasponia nonlegume hemoglobin I	39	1LH1
WSM	glbxcerrh	27	Water snail myoglobin	25	1MBA
BCG	glb1xanabr	28	Blood clam globin I	21	1MBA
MWE	glbxtylhe	29	Marine worm erythrocruorin	20	1MBD
MWG	glbaxtylhe	30	Marine worm globin IIA	19	2HHB-B
CEI	glb1xlumte	31	Common earthworm globin I	18	2HHB-A
UCH	hbf1xureca	32	Urechis caupo hemoglobin F-I	17	2LHB
CEC	glbcxlumte	33	Common earthworm globin C	17	2LHB

*The upper part summarizes the globins of known three dimensional structure used in this study (PDB). The lower part contains globin sequences whose native conformations are unknown.
†Codes used in this study.
‡Codes used in the Brookhaven Protein Data Bank and Swissprot Data Bank, respectively.
§Reference.
**Percentage of amino acid identities after optimal alignment.

among different conformations of a particular amino acid sequence but it is not comparable among different amino acid sequences.[9] Therefore, $\Delta E(S_a,C_i) - \Delta E(S_a,C_j)$ is related to the relative stabilities of C_i and C_j with respect to S_a, but $\Delta E(S_a,C_i) - \Delta E(S_b,C_i)$ cannot be used to estimate the relative stabilities of S_a and S_b in conformation C_i. Consequently the $\Delta E(S,C_1)$ values obtained for the conformation C_1 of minimum net energy in general cannot be used to judge the quality of this conformation as a model for the unknown fold.

In our previous study[10] we observed that the confidence in the identification of a native state is related to

$$\Delta E_{1,2} = \Delta E(S,C_1) - \Delta E(S,C_2)$$

where C_1 and C_2 are the conformations of lowest and second lowest energy, respectively. If the pool contains a group of closely related structures (e.g., several homologous proteins) the energy differences will be small among these structures. Such structures will be closely spaced in the energy sorted list. If such a group of fragments appears on top of the list $\Delta E_{1,2}$ is replaced by $\Delta E_{1,n}$ the energy difference between C_1 and C_n where C_n corresponds to the fragment of minimum energy which does not belong to the group containing C_1.

Experimental Setup and the FDB and PDB Data Bases

Our goal is to demonstrate that in the absence of significant sequence homology native-like folds of proteins can be identified in a data base of known structures. We set up a computer experiment using sequences and structures taken from the globin family as a test case. The Brookhaven Protein Data Bank[14] currently holds several globin entries from a variety of species. On the other hand, several hundred amino acid sequences are known for this protein family. A substantial fraction of these sequences does not have significant sequence homology to globins of known structure.

For the compilation of potentials of mean force we use a data base (FDB) which does not contain any globins (Table II). Hence the potentials do not contain any specific information on the relationship between globin sequences and globin folds. We prepare a second data base (PDB) used to generate a pool of fragments by adding seven globin conformations (Table II). A globin sequence is then mounted on all fragments in the pool constructed from PDB. The question is whether at least one of the globin folds yields lowest net energy so that this fold is identified as a suitable model for the given sequence.

We emphasize that by removing all information on the relationship between globin folds and globin sequences from the potentials of mean force (FDB) the problem of correct retrieval of globin folds is more difficult as compared to the usual applications of our approach. Usually the potentials of mean force will incorporate information on the relationship of the architecture of a protein family and at least one of its sequences. In fact, the experimental setup mimics the situation where several models for

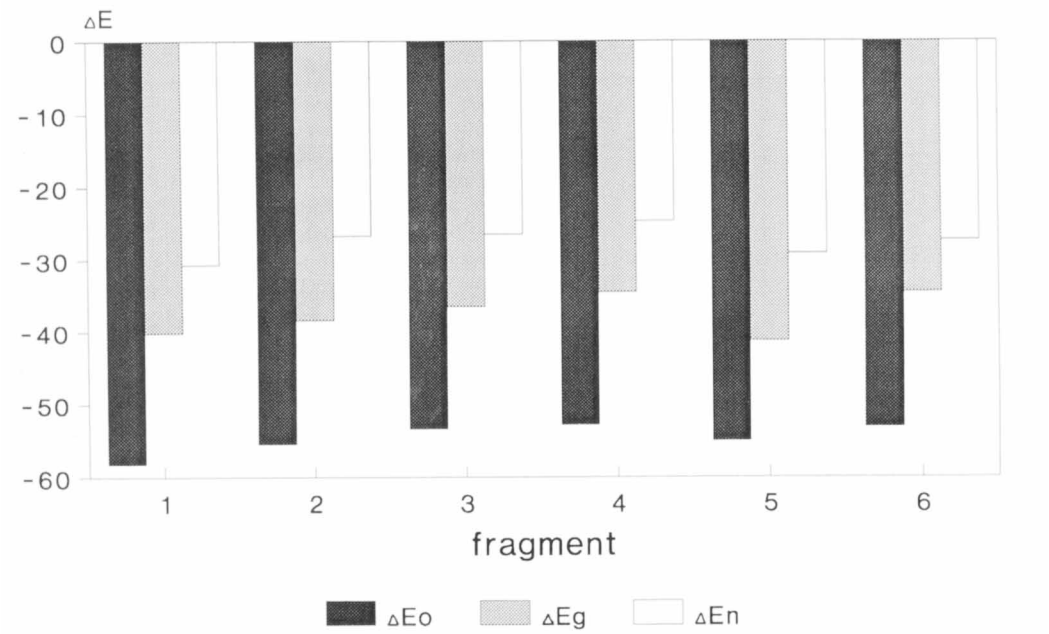

Fig. 1. Total net energies of the 6 fragments of the HHB-A sequence when combined with the conformations in the pool. ΔE_o is the total net energy of the respective fragment in its native conformation, ΔE_g is the total net energy of the most favorable nonnative globin conformation, and ΔE_n is the total net energy of the most favorable nonglobin conformation. For all fragments $\Delta E_o < \Delta E_g < \Delta E_n$, i.e., the native fold is identified as the most favorable conformation and a non-native globin fold is superior to any nonglobin fold. Note that in all calculations the data base used to compile the potentials of mean force did not contain any globin.

an unknown fold have been constructed and the problem is to rank these models in terms of their probable correctness.

The combination of sequences and conformations requires that both have the same number of residues. In our data base the shortest globin chain of known structure is ECD (the codes used in this study are defined in Table III) having $L = 136$ amino acid residues. We therefore, used a fragment size of 136 residues to generate a pool of conformations from PDB. Using this fragment size longer proteins contribute several fragments to the pool. HHB-A for example contains 141 residues. Using a fragment size of 136 we obtain $141 - 136 + 1 = 6$ fragments from this conformation (the number of fragments is $L - R + 1$ where L is the sequence length and R is fragment size). Using this constant fragment size the pool constructed from PDB contains 11,124 individual conformations.

In a similar way fragments of length R are obtained for a given sequence S. The first fragment (residues $1, \ldots, R$) is combined with all conformations in the pool, the net energies are computed, and the energy sorted list is constructed. The sequence is then scanned from the N- to the C-terminus for all $L - R + 1$ fragments.

Globin Sequences Used as Test Cases

In Table III we summarize the globins used in this study. The upper part lists the globins of known structure in PDB and the maximum homology to other globins in the PDB data base. The largest homology (41%) is found between HHB-A and HHB-B. The most distantly related globin is LH1 from the root nodules of yellow lupin with a maximum homology of 22% to MBA a sea hare myoglobin.

The lower part of Table III lists globin sequences of unknown structure and their maximum homology to globins in the PDB data base. The closest related pair is BBL, a broad bean nonlegume hemoglobin and the yellow lupin leghemoglobin LH1 (49%). Sequences of least homology (17%) to globins in PDB are GEC (hemoglobin C, annelida) and UCH (hemoglobin F-I, echiura).

RESULTS

Applications to Globins of Known Structure

Our first goal is to apply the procedure to globins whose conformations are contained in PDB. As a first example we choose HHB-A the α-chain of human hemoglobin. We are interested in the following points: (1) Is the native fold of HHB-A identified correctly among all conformations in the pool? and (2) If the native conformation is removed from the pool is the native fold of the HHB-A sequence identified as a globin fold?

There are three different types of conformations whose relative energy differences are of interest in investigating these points: (1) ΔE_0 the total net energy of the native state, (2) ΔE_g the minimum total

TABLE IV. Hemoglobin α-Chain HHB-A

F*	ΔE_0†	ΔE_1‡	Fragment§		$\Delta E_{0,g}$**	$\Delta E_{0,n}$††	$\Delta E_{g,n}$‡‡	$rms_{0,1}$§§
1	−58.20	−40.16	HHB-B	3	−18.04	−27.53	−9.49	5.48
2	−55.47	−38.39	HHB-B	4	−17.08	−28.78	−11.70	5.48
3	−53.43	−36.52	HHB-B	5	−16.91	−26.94	−10.03	5.47
4	−52.84	−34.50	HHB-B	6	−18.34	−28.21	−9.87	5.49
5	−54.99	−41.24	ECD	1	−13.75	−25.87	−12.12	5.41
6	−53.12	−34.51	HHB-B	8	−18.61	−25.83	−7.22	5.52

*Fragment number. Fragment 1 corresponds to residues 1–136, for example.
†Total net energy of the native state.
‡Total net energy of the most favorable nonnative structure obtained from the pool of conformations.
§Code and fragment number of the most favorable nonnative conformation.
**$\Delta E_{0,g} = \Delta E_0 - \Delta E_g$. Energy difference of the native and most favorable nonnative globin fold.
††$\Delta E_{0,n} = \Delta E_g - \Delta E_n$. Energy difference of the native and most favorable nonglobin fold.
‡‡$\Delta E_{g,n} = \Delta E_g - \Delta E_n$. Energy difference of the most favorable nonnative globin and nonglobin folds.
§§Root mean square error (rms) of optimal superposition between native (0) and most favorable nonnative conformation (1).

net energy among all nonnative globin fragments, and (3) ΔE_n the minimum total net energy among all nonglobin conformations. The difference $\Delta E_{0,g} = \Delta E_0 - \Delta E_g$ shows whether the native state can be identified among other globin folds and $\Delta E_{0,n} = \Delta E_0 - \Delta E_n$ shows whether the native fold can be distinguished from all non globin conformations. Negative values of $\Delta E_{g,n} = \Delta E_g - \Delta E_n$ indicate whether in the absence of the native fold a globin conformation is identified as the most favorable model. Note that $\Delta E_{0,n} = \Delta E_{0,g} + \Delta E_{g,n}$.

Figure 1 and Table IV show the results obtained for the HHB-A sequence. The sequence yields 6 fragments of 136 residues, the size of the fragments in the pool of conformations. Fragment 1, for example corresponds to residues 1–136 of HHB-A. Table IV contains ΔE_0, the total net energy of the native state of the individual fragments and ΔE_1, the minimum total net energy of all nonnative conformations in the pool (note that $\Delta E_1 = \Delta E_g$ if $\Delta E_g \leq \Delta E_n$ or else $\Delta E_1 = \Delta E_n$). Fragment 1 of the HHB-A sequence yields $\Delta E_1 = \Delta E_g = -40.16$ kcal/mol when mounted on fragment 3 of HHB-B. Since $\Delta E_0 < \Delta E_1$ the native fold of HHB-A-1 (we append the fragment number to the protein code) is identified correctly.

The difference $\Delta E_{0,g}$ between native fold and HHB-B-3 indicates that there are structural differences between the HHB-A-1 and HHB-B-3 conformations. Indeed the optimal alignment of HHB-A and HHB-B requires two gaps the largest spanning six amino acids. Hence the sequence of HHB-A-1 does not fit exactly on the HHB-B-3 fold which is reflected by the rather large value of $\Delta E_{0,g} = -18.04$ kcal/mol. The root mean square error (rms) of optimal superposition[15] of HHB-A-1 and HHB-B-3 is 5.48 Å (Table IV). The energy difference between the most favorable globin fold and the most favorable nonglobin fold of $\Delta E_{g,n} = -9.49$ kcal/mol shows that if the native fold of HHB-A-1 is unknown, with considerable confidence HHB-B-3 is identified as the best available model conformation.

As shown in Figure 1 and Table IV for all six fragments of HHB-A the native fold is reliably identified. In all cases $\Delta E_{0,g}$ as well as $\Delta E_{0,n}$ have large negative values. In the absence of the native fold the most suitable models are fragments from HHB-B and ECD. Fragment 5 of HHB-A has lowest net energy when combined with ECD-1 (which is just the complete conformation of ECD) instead of HHB-B-7 as might be expected from Table IV. The energy difference for the native conformation of HHB-A-5 and ECD-1 being $\Delta E_{0,g} = -13.75$ kcal/mol is lower as compared to the $\Delta E_{0,g}$ values of the remaining fragments.

In terms of the total net energy the ECD-1 conformation is more favorable than the HHB-B-7 fragment. These two structures are indistinguishible in terms of the structural similarity to the HHB-A-5 conformation. The rms deviations between HHB-A-5 and ECD-1 (5.50 Å) and between HHB-A-5 and HHB-A-7 (5.54 Å) are identical and the rms value for ECD-1 and HHB-B-7 (4.74 Å) is in the same range. Hence the structural similarity among HHB-A-5, HHB-B-7, and ECD-1 is of the same order of magnitude but the total net energy with respect to the HHB-A-5 sequence clearly distinguishes between these three conformations.

Figure 2 and Table V summarize the results obtained for HHB-B. The large negative $\Delta E_{0,g}$ and $\Delta E_{0,n}$ values show that for all 11 fragments the native fold is correctly identified. $\Delta E_{g,n}$ reveals that among all nonnative conformations a globin fold is found to be the most favorable model. The minimum total net energies for nonnative folds are obtained for fragments of sperm whale myoglobin MBD. Note that the fragments obtained from MBD are in line with the fragments of HHB-B, i.e., fragment i of HHB-B yields the minimum energy on fragment $i + 1$ of MBD. This indicates that the complete sequence of HHB-B (residues 1–146) can be fitted on the MBD conformation from residues 2–147. In fact these conformations are superimposable with an rms devia-

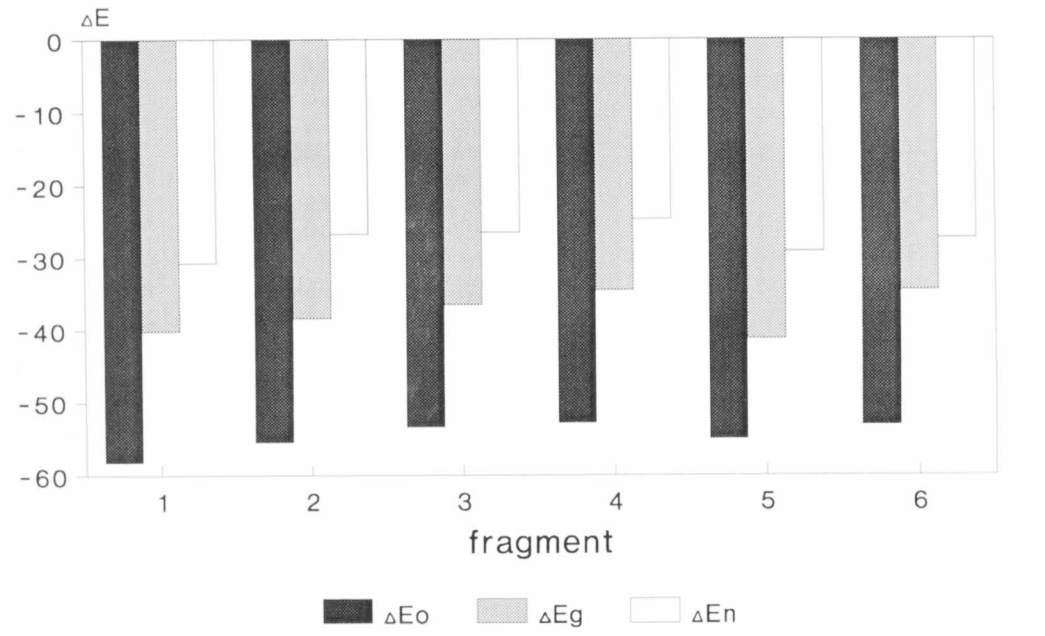

Fig. 2. Total net energies of the 11 fragments of the HHB-B sequence (see legend to Fig. 1).

TABLE V. Hemoglobin β-Chain HHB-B*

F	ΔE_0	ΔE_1	Fragment		$\Delta E_{0,g}$	$\Delta E_{0,n}$	$\Delta E_{g,n}$	$rms_{0,1}$
1	−65.34	−53.57	MBD	2	−11.77	−27.61	−15.84	2.38
2	−67.58	−56.18	MBD	3	−11.40	−29.34	−17.94	2.36
3	−68.42	−57.68	MBD	4	−10.74	−28.15	−17.41	2.33
4	−66.31	−56.85	MBD	5	−9.46	−29.63	−20.17	2.30
5	−67.42	−57.67	MBD	6	−9.75	−28.80	−19.05	2.30
6	−71.54	−60.88	MBD	7	−10.66	−28.45	−17.79	2.26
7	−72.08	−63.29	MBD	8	−8.79	−24.68	−15.89	2.21
8	−73.54	−64.84	MBD	9	−8.70	−22.24	−13.54	2.19
9	−72.10	−63.26	MBD	10	−8.84	−20.04	−11.20	2.15
10	−73.99	−65.92	MBD	11	−8.07	−20.05	−11.98	2.08
11	−73.42	−65.99	MBD	12	−7.43	−19.50	−12.07	2.07

*See legend to Table IV.

tion of 2.2 Å. The structural similarity of these conformations results in smaller $\Delta E_{0,g}$ values of ≈ 9.0 kcal/mol as compared to the values obtained for the HHB-A sequence (Table IV).

Figure 3 and Table VI display the results for LHB (sea lamprey hemoglobin V). The molecule has an N-terminal extension of 9 residues which has no counterpart in other globins of known structure (Fig. 4). Again for all fragments the native fold is identified correctly (large negative values of $\Delta E_{0,g}$ and $\Delta E_{0,n}$). However, $\Delta E_{g,n}$ is positive for the N-terminal fragments. The pool does not contain globin conformations which are suitable models for fragments of LHB containing the N-terminal extension. The best model available for these fragments is identified as fragment 153–296 of the 1PRC-C conformation.

Fragments 10 to 14 of LHB are in line with MBD $(i,i-9)$ so that the MBD conformation (residues 1–143) is identified as the most favorable model for residues 10–152 of the LHB sequence. The ΔE_1 values for the N-terminal fragments are much higher as compared to the C-terminal fragments and they are correlated with the rms deviation. Between fragments 9 and 10 of LHB ΔE_1 changes from −38.20 to −63.36 kcal/mol indicating a significant change in the quality of the available models for the respective fragments. Hence, although in general ΔE_1 is in-

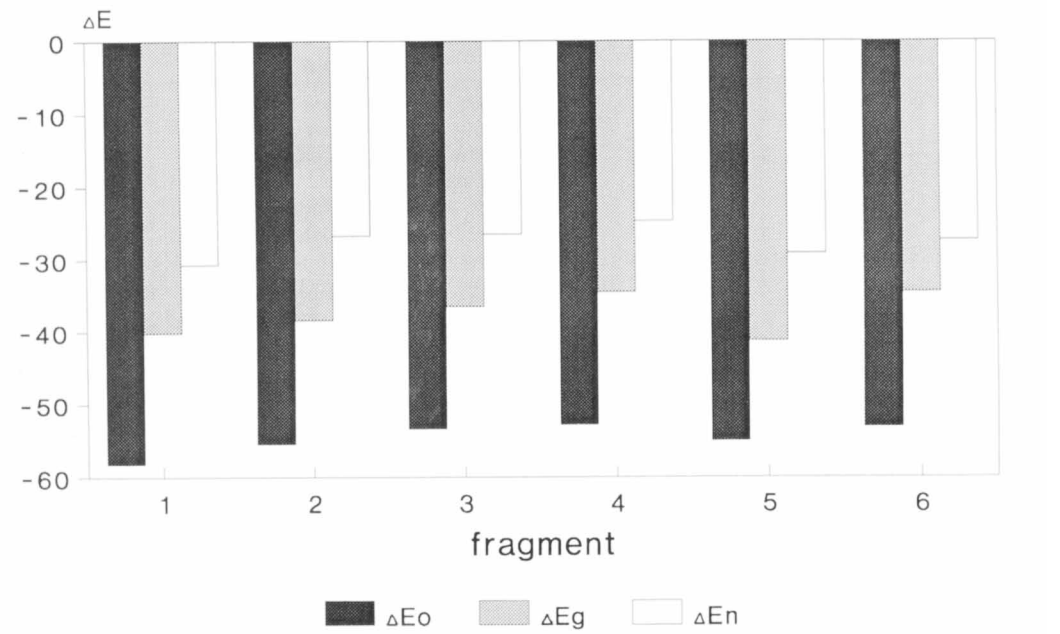

Fig. 3. Total net energies of the 14 fragments of the LHB sequence (see legend to Fig. 1). The native fold is identified correctly for all fragments. The N-terminal extension of LHB has no counterpart in the remaining globins of known structure, so that for the 6 N-terminal fragments no nonnative globin conformation is identified as a suitable model. However, a strong globin signal is obtained for the C-terminal fragments 10–14.

TABLE VI. Sea Lamprey Hemoglobin LHB*

F	ΔE_0	ΔE_1	Fragment		$\Delta E_{0,g}$	$\Delta E_{0,n}$	$\Delta E_{g,n}$	$rms_{0,1}$
1	−62.48	−35.20	1PRC-C	153	>−52.00	−27.28	>25.00	15.74
2	−65.76	−36.39	1PRC-C	154	−54.01	−29.37	24.64	15.84
3	−67.47	−31.63	1PRC-C	155	>−56.00	−35.84	>20.00	15.96
4	−66.07	−31.13	1PRC-C	156	>−54.00	−34.94	>20.00	16.07
5	−71.03	−29.85	1PRC-C	157	−54.06	−41.18	12.88	16.12
6	−75.97	−28.83	1TIM-A	74	−54.84	−47.14	7.70	15.90
7	−77.94	−32.49	MBD	1	−45.45	−49.03	−3.58	7.05
8	−80.24	−35.30	MBA	1	−44.94	−47.24	−2.30	5.60
9	−82.08	−38.20	MBA	2	−43.88	−50.72	−6.84	5.56
10	−85.08	−63.36	MBD	1	−21.72	−50.92	−29.20	4.28
11	−84.32	−63.12	MBD	2	−21.20	−51.02	−29.82	4.34
12	−81.63	−60.05	MBD	3	−21.58	−47.68	−26.10	4.38
13	−81.50	−60.42	MBD	4	−21.08	−47.70	−26.62	4.42
14	−82.71	−61.42	MBD	5	−21.29	−47.36	−26.07	4.46

*See legend to Table IV.

comparable among different sequences these dramatic changes clearly indicate that for the C-terminal fragments of LHB the conformations obtained from the pool are suitable models whereas no satisfying conformations are found for the N-terminal fragments.

Compared to the results obtained for HHB-A, HHB-B, and LHB the behavior of LH1 (Fig. 5 and Table VII) is quite unusual. For fragments 1–5 and fragments 9–18 the native conformation of LH1 is identified correctly. However, for fragments 6–8 the HHB-A conformation (fragments 1–3) yields lower energies than the native fold. Comparison of the ΔE_0, ΔE_1, and $\Delta E_{0,g}$ values reveals that with respect to the LH1 sequence the HHB-A fold and the native LH1 conformation are indistinguishable in terms of the total net energy. This is surprising since the rms deviation of LH1 and HHB-A is rather large (≈ 6.6 Å).

Fragments containing the 5 N-terminal residues

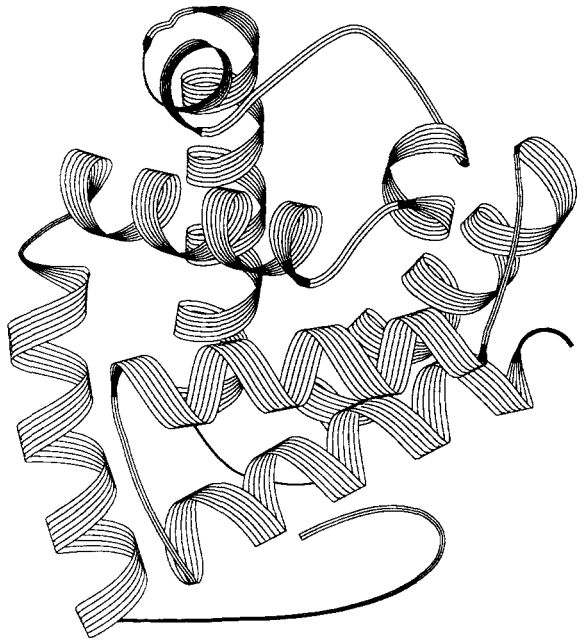

Fig. 4. Priestle cartoon of the lamprey hemoglobin fold. The N-terminal extension which is not found in other globins of known structures is shown at the bottom.

of LH1 do not yield globin folds as the most favorable models (positive values of $\Delta E_{g,n}$). Starting at fragment 6 the LH1 sequence is clearly identified as a globin fold. Similar to the N-terminal extension of LHB the N-terminal fold of LH1 is unique. This motif is not found in other globins of known structure.

In summmary in all four examples presented the native structures are identified as the most favorable folds and in the absence of the native conformation folds from the globin family are identified as the most favorable models.

Applications to Globins of Unknown Structure

In Tables VIII to XI we summarize the results obtained for 10 globin sequences of unknown structure. Since the native conformations of these globins are not available all quantities involving ΔE_0 as well as $rms_{0,g}$ are omitted. $\Delta E_{g,n} < 0$ indicates that a particular fragment is identified as a globin fold and its magnitude is a qualitative measure for the reliability.

We emphasize that the magnitude of positive $\Delta E_{g,n}$ values is not correlated with the reliability of the identification of the respective conformation. For example, in the case of fragment 1 of PNH (Table VIII) the conformation yielding lowest net energy ($\Delta E_1 = -30.83$ kcal/mol) is PRC-M-79. The conformation of second lowest energy is 7CAT-A-360 with $\Delta E_2 = -29.65$ kcal/mol. Hence $\Delta E_1 - \Delta E_2 = -1.18$ kcal/mol indicating that none of the conformations is particularily favored. This is also corroborated by the large rms deviation of 22.6 Å of these conformations.

We first discuss the results obtained for PNH a nonlegume hemoglobin whose sequence homology of 39% to LH1 (leghemoglobin), a globin of known structure, is significant. Table VIII shows the results obtained for the 26 fragments of PNH. Fragments 7 to 26 of PNH are all identified as globin conformations. The $\Delta E_{g,n}$ values range from -5.00 to -14.00 kcal/mol indicating that the identification is quite reliable. None of the globin fragments in the data base is a favorable model for the six N-terminal fragments of PNH.

Table IX summarizes the results obtained for the three most distantly related sequences CEI, UCH, and CEC. The maximum homology to globins of known structure being 17% is quite low. All three proteins are identified as globins. The smallest $\Delta E_{g,n}$ values are obtained for CEI but the most significant value of -4.6 kcal/mol of fragment 4 indicates a reliable detection of the globin fold.

Table XI summarizes the results obtained for all ten globin sequences. Only the best score for each sequence is shown. With the exception of BCG all sequences are identified as globin folds, although the confidence for MWE (most significant $\Delta E_{g,n}$ of -2.99 kcal/mol) is small. None of the fragments of BCG is identified as a globin fold (Table X). The conformation of minimum net energy with respect to this sequence is 1GOX (glycolate oxidase).

Effect of Gaps in Structure Sequence Alignments

Our approach to the identification of native-like models can be interpreted as the alignment of amino acid sequences with protein conformations. Favorable sequence structure alignments are obtained when sequence structure combinations yield low conformational energies.

We emphasize that in the present study we do not allow for gaps in the sequence structure alignment. This is a severe restriction since in most cases optimal alignment is only possible if gaps are introduced. For example, the optimal sequence alignment of HHB-A and HHB-B requires two gaps of two (HHB-A) and six (HHB-B) residues (Fig. 6). On the other hand the optimal alignment of hemoglobin β and myoglobin (MBD) requires only one gap of two residues in the HHB-B sequence (Fig. 6).

If these gaps are neglected the superposition of the respective conformations yields rather large rms errors. Between the HHB-A and HHB-B fragments these errors are in the order of 5.5 Å (Table IV). The rms error between HHB-B and MBD where only one small gap is required for optimal alignment yields much lower rms values of $\approx$ 2.2 Å (Table V).

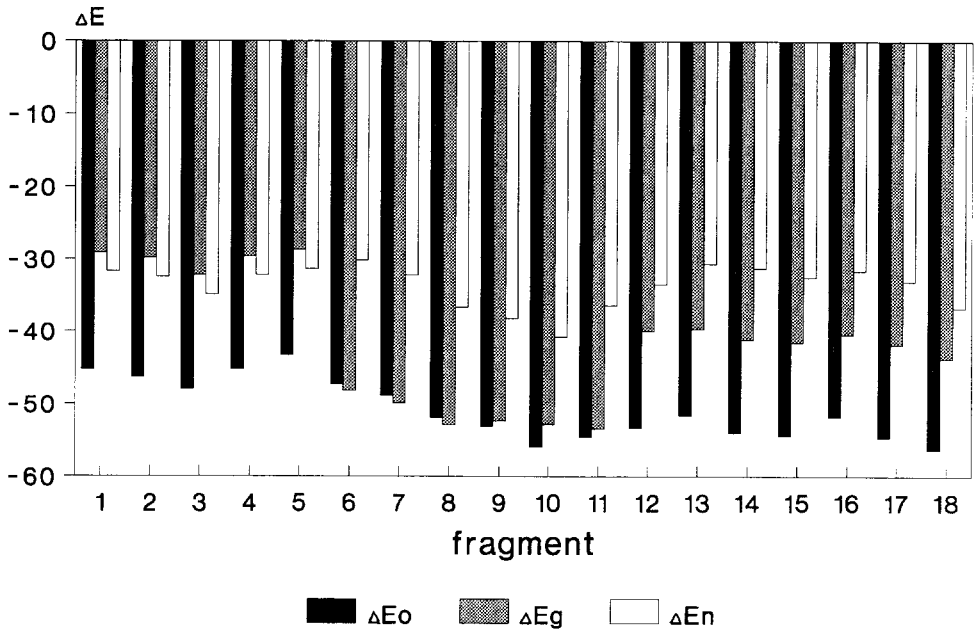

Fig. 5. Total net energies of the 18 fragments of the LH1 sequence (see legend to Fig. 1). Note that for fragments 6–11 the total net energy of the native fold is indistinguishable from the most favorable nonnative globin fold.

TABLE VII. Leghemoglobin LH1*

F	ΔE_0	ΔE_1	Fragment		$\Delta E_{0,g}$	$\Delta E_{0,n}$	$\Delta E_{g,n}$	$rms_{0,1}$
1	−45.27	−31.70	1ABP	107	−16.10	−13.57	2.53	16.57
2	−46.29	−32.45	8ADH	170	−16.44	−13.84	2.60	16.11
3	−47.92	−34.89	8ADH	171	−15.69	−13.03	2.66	15.90
4	−45.19	−32.24	8ADH	172	−15.52	−12.95	2.57	15.71
5	−43.27	−31.37	8ADH	173	−14.59	−11.90	2.69	15.51
6	−47.24	−48.12	HHB-A	1	0.88	−17.10	−17.98	6.71
7	−48.80	−49.86	HHB-A	2	1.06	−16.58	−17.64	6.64
8	−51.86	−52.93	HHB-A	3	1.07	−15.18	−16.25	6.62
9	−53.05	−52.35	HHB-A	4	−0.70	−14.84	−14.14	6.61
10	−55.93	−52.82	HHB-A	5	−3.11	−15.14	−12.03	6.61
11	−54.53	−53.43	HHB-A	6	−1.10	−18.07	−16.97	6.63
12	−53.33	−39.95	MBD	8	−13.38	−19.80	−6.42	6.85
13	−51.58	−39.69	MBD	9	−11.89	−20.87	−8.98	6.85
14	−53.97	−41.18	MBD	10	−12.79	−22.69	−9.90	6.82
15	−54.39	−41.52	MBD	11	−12.87	−21.83	−8.96	6.79
16	−51.79	−40.45	MBD	12	−11.34	−20.15	−8.81	6.78
17	−54.64	−41.83	MBD	13	−12.81	−21.52	−8.71	6.77
18	−56.31	−43.75	MBD	14	−12.56	−19.54	−6.98	6.74

*See legend to Table IV.

The effect of the neglect of these gaps on the conformational energies can be inferred from Tables IV and V. The $\Delta E_{0,g}$ values for the HHB-B sequence mounted on the MBD conformation are in the order of −9.00 kcal/mol. The values for HHB-A mounted on HHB-B being approximately −17.0 kcal/mol are much larger. Similarily, $\Delta E_{g,n} \approx -17.0$ kcal/mol for the HHB-B sequence is considerably larger as compared to $\Delta E_{g,n} \approx -10.0$ kcal/mol for the HHB-A sequence. This indicates that the conformation of

TABLE VIII. Parasponia Nonlegume Hemoglobin I PNH

F	ΔE_1	Fragment		$\Delta E_{g,n}$
1	−30.83	1PRC-M	79	7.82
2	−29.93	1TIM-A	108	4.15
3	−33.24	1TIM-A	109	9.71
4	−33.91	1TIM-A	110	13.33
5	−34.14	1TIM-A	111	9.72
6	−34.04	1TIM-A	112	6.96
7	−34.84	MBD	1	−6.44
8	−34.94	MBD	2	−7.48
9	−36.00	MBD	3	−7.52
10	−39.08	MBD	4	−5.74
11	−38.14	MBD	5	−8.27
12	−38.24	MBD	6	−8.68
13	−41.06	MBD	7	−9.81
14	−43.47	MBD	4	−8.74
15	−46.05	LH1	8	−6.22
16	−46.57	LH1	9	−5.65
17	−50.20	LH1	10	−7.98
18	−49.27	MBD	8	−11.40
19	−50.59	HHB-B	8	−13.16
20	−49.09	MBD	10	−13.07
21	−49.32	LH1	14	−10.35
22	−48.76	MBD	12	−9.76
23	−47.10	MBD	13	−11.41
24	−46.26	MBD	14	−12.60
25	−45.58	MBD	15	−13.57
26	−45.18	MBD	16	−13.00

TABLE IX. Sequences of Least Homology to Globins of Known Structure: CEI, UCH, CEC

Sequence	F	ΔE_1	Fragment		$\Delta E_{g,n}$
CEI	1	−31.89	2PRK	42	1.15
	2	−34.15	2PRK	43	2.57
	3	−31.73	LHB	13	−1.47
	4	−35.62	HHB-B	3	−4.60
	5	−34.46	MBD	5	−4.23
	6	−33.00	MBD	6	−1.09
	7	−37.15	HHB-A	1	−3.77
UCH	1	−36.60	MBD	1	−4.43
	2	−38.05	MBD	2	−2.28
	3	−39.16	MBD	3	−4.87
	4	−40.66	MBD	4	−5.81
	5	−42.83	MBD	5	−5.85
	6	−40.72	MBD	6	−4.40
CEC	1	−47.14	MBA	3	−6.61
	2	−45.19	MBA	4	−7.79
	3	−49.63	HHB-A	1	−10.41
	4	−48.24	HHB-A	2	−8.58
	5	−48.05	HHB-A	3	−4.86
	6	−47.93	HHB-A	4	−7.51
	7	−48.07	HHB-A	5	−9.53
	8	−50.02	HHB-A	6	−11.51
	9	−44.78	MBA	11	−7.31

TABLE X. Blood Clam Globin I BCG

F	ΔE_1	Fragment		$\Delta E_{g,n}$
1	−32.46	3PGK	246	>16.00
2	−35.48	4MDH-A	36	>16.00
3	−34.77	4MDH-A	37	>16.00
4	−34.33	3PGK	245	>15.00
5	−35.36	3PGK	246	>16.00
6	−37.06	1TIM-A	94	16.00
7	−38.22	1GOX	2	11.25
8	−40.75	1GOX	3	15.52
9	−42.05	1GOX	4	14.71
10	−42.33	1GOX	5	13.45
11	−46.01	1GOX	6	14.66

MBD is a much better model for the HHB-B sequence than the HHB-B conformation for the HHB-A sequence.

It is clear that depending on the number and size of gaps required sequence structure misalignments dramatically affect the conformational energies. If gaps are neglected it will be impossible to detect a native-like fold if the effect on the conformational energies is too severe. However, sequence structure misalignments affect energy contributions from individual structural levels in a different manner. For small sequential separations k only few distances in the vicinity of a gap are misplaced. Hence the effect of a gap on the short range energy contributions (small k values) is small. This is completely different for medium and long-range interactions since in this case the introduction of a gap affects a large fraction of the interactions.

To investigate whether the failure to identify BCG as a globin fold is due to the neglect of gaps we calculated the net energies in the short range $k = 1, \ldots, 10$, since the short-range energy should be less affected by misalignments. Table XII shows the results obtained for BCG in the short range. Indeed for fragments 9 to 11 of BCG the most favorable conformations are fragments from a globin fold (LH1).

DISCUSSION

We have demonstrated that the approach presented is able to retrieve native-like protein folds from a data base of known protein conformations. With the exception of BCG all globin sequences investigated in this study yield most favorable net energies when combined with globin conformations. This is achieved by the calculation of conformational energies using potentials of mean force derived from a data base devoid of globins.

In the present study we used examples from the well-characterized globin family mainly because several structures are available for this class of pro-

TABLE XI. Fragments of Maximum $\Delta E_{g,n}$ of 10 Globin Sequences

Sequence	F	ΔE_1	Fragment		$\Delta E_{g,n}$	Homology
BBL	6	−37.69	HHB-A	6	−5.97	49
MLG	15	−34.92	MBA	11	−9.20	39
PNH	25	−45.58	MBD	15	−13.57	39
WSM	7	−52.79	MBA	4	−23.73	25
BCG	7	−38.22	1GOX	2	+11.25	21
MWE	4	−37.33	MBA	4	−2.99	20
MWG	10	−42.75	MBA	8	−9.73	19
CEI	4	−35.62	HHB-B	3	−4.60	18
UCH	5	−42.83	MBD	5	−5.85	17
CEC	8	−50.02	HHB-A	6	−11.51	17

teins. However, the approach is not restricted to globins. We have shown previously[10] that correct native folds can be identified for a large number of proteins. Examples where the method fails to identify the native conformation as the most favorable fold are generally restricted to proteins containing large prosthetic groups or ions (e.g., Fe-S clusters) required to stabilize the native conformation. In addition we find that in several cases small proteins containing a large number of disulfide bridges are not identified correctly. Such examples are crambin (1CRN) and α-bungarotoxin (2ABX) although the native fold is usually among the most favorable conformations (i.e., small positive values for $\Delta E_1 - \Delta E_0$).

The globin folds are identified correctly although the native fold contains a heme group. This may indicate that the globin fold is stabilized mainly by interactions within the protein moiety. This is also supported by experimental results on myoglobin indicating that in the absence of the heme group the protein retains approximately 60% of its α-helix content.

The resolution of our knowledge-based force field is high enough to discriminate a native globin fold from several closely related conformations. The only exceptions found in the present study are three fragments from LH1. These fragments (Table VII) yield more favorable energies in the HHB-A conformation but the energy differences to the native conformation are quite small. We emphasize, however, that the force field used in this study is quite incomplete since we calculated the C_β–C_β interactions only. The failure to resolve the differences between the LH1 and HHB-A conformations with respect to the LH1 sequence may be due to this approximation. On the other hand for the HHB-A sequence the native fold is reliably identified and not confused with the LH1 conformation. This may point to unusual features of the LH1 conformation. However, at the present stage of development the interpretation of such effects is rather hazardous. It is likely that the resolution of the force field can be improved by adding interactions of backbone and additional side chain atoms.[9] This will allow us to investigate such problems in more detail.

Optimal sequence structure alignments generally require the introduction of gaps in the amino acid sequence and/or conformation. It is remarkable that in spite of the neglect of gaps the procedure correctly identifies native-like folds in most cases. We anticipate that the performance of our approach can be improved by the introduction of gaps in the structure sequence alignment. We are currently proceeding along these lines.

The method presented is not restricted to model conformations derived from X-ray analysis or nuclear magnetic resonance studies. It is possible to include model conformations derived from any experimental or theoretical studies in the pool of conformations in order to check whether these conformations can be considered as useful models for the unknown native fold. In addition the technique is a useful tool in judging early models obtained from the interpretation of electron densities especially in cases of low resolution and it is helpful in discriminating among various models consistent with distance constraints obtained from nuclear magnetic resonance.

In several cases structures derived from X-ray analysis have been proved to contain major errors.[16,17] The difficulty to correct gross misinterpretions of electron densities is partly due to the lack of efficient criteria which indicate whether a proposed conformation is consistent with the general features of a native fold. The method presented here is a powerful tool which in many cases will detect such errors. This can be achieved for example by combining the sequence of the protein under investigation with all conformations in the data base. If the current model is not found among the most favorable structures chances are that it is incorrect.

Finally we want to emphasize that our force field depends on the set of available protein conformations. It will certainly improve with the growing number of known protein folds. It is also obvious that the chance to identify a native-like fold in the data base increases with the number of known pro-

```
HHB-A .VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQVKGHGKKVADALTN
HHB-B VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD

HHB-A AVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
HHB-B GLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

HHB-B VHLTPEEKSAVTALWGKVNVDEVGGEA--LGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD
MBD  .VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGA

HHB-B GLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH.....
MBD  ILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
```

Fig. 6. Sequence alignments of HHB-A versus HHB-B and HHB-B versus MBD.

TABLE XII. Blood Clam Globin I (BCG): $\Delta E(S,C)$ Evaluated for the Short-Range Levels $k = 1, \ldots, 10$ Only

F	ΔE_1	Fragment		$\Delta E_{g,n}$
1	−10.73	1PRC-L	32	2.98
2	−10.89	3TLN	160	2.58
3	−11.07	3TLN	161	1.58
4	−10.97	1PRC-L	35	1.15
5	−11.16	2CPP	117	0.78
6	−11.33	2CPP	118	0.71
7	−11.24	2CPP	119	0.50
8	−11.18	2CPP	120	0.08
9	−11.85	LH1	5	−0.79
10	−12.20	LH1	6	−1.02
11	−13.11	LH1	7	−1.65

tein structures. Moreover, if we succeed to combine our force field with powerful energy minimizers or molecular dynamic procedures it seems possible that we will be able to calculate native folds from amino acid sequences in the near future.

ACKNOWLEDGMENTS

We thank Manfred Hendlich and Peter Lackner for assistance with their source code. We are indebted to all X-ray crystallographers who submitted coordinates to the Brookhaven Protein Data Bank. This work was supported by the Fonds zur Förderung der wissenschaftlichen Forschung (Austria) under project number 7262-BIO.

REFERENCES

1. Anfinsen, C.B. Principles that govern the folding of protein chains. Science 181:223–230, 1973.
2. Weiner, P.K., Kollman, P.A. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. In: J. Comp. Chem. 2:287–299, 1981.
3. Burkert, U., Allinger, N.L. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. "Molecular Mechanics." Washington DC: American Chemical Society, 1982.
4. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D. States, D.J., Swaminathan, S., Karplus, M. J. Comp. Chem. 4:187, 1983.
5. van Gunsteren, W.F., Berendsen, H.J.C., Hermans, J., Hol, W.G.J., Postma, J.P.M. Computer simulation of the dynamics of hydrated protein crystals and its comparison with X-ray data. Proc. Natl. Acad. Sci. U.S.A. 80:4315–4319, 1983.
6. Carson, M., Hermans, J. In "Molecular Dynamics and Protein Structure," Hermans, J. (ed). Chapel Hill: University of North Carolina, 1985: 156–166.
7. Novotny, J., Brucceroli, R.E., Karplus, M. An analysis of incorrectly folded protein models. Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
8. Novotny, J., Rashin, A.A., Brucceroli, R.E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4:19–30, 1988.
9. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins. J. Mol. Biol. 213:859–883, 1990.
10. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M.J. Identification of native protein folds amongst a large number of incorrect conformations. The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216:167–180, 1990.
11. Sippl, M.J., Hendlich, M., Lackner, P. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments. Development of strategies and construction of models for myoglobin, lysozyme and thymosin β_4. Protein Science, in press, 1991.
12. Neidhart, D.J., Kenyon, G.L., Gerit, J.A., Petsko, G.A. Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. Nature (London) 347:692–694, 1990.
13. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tsumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
14. Sippl, M.J., Stegbuchner, H. Superposition of three dimensional objects: A fast and numerically stable algorithm for the calculation of the matrix of optimal rotation. Comput. Chem. 15:73–78, 1991.
15. Bränden, C.I., Jones, T.A. Between objectivity and subjectivity. Nature (London) 343:687–689, 1990.
16. Janin, J. Errors in three dimensions. Biochimie 72:705–709, 1990.
17. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. The crystal structure of human deoxyhemoglobin at 1.74 A resolution. J. Mol. Biol. 175:159–174, 1984.
18. Honzatko, R.B., Hendrickson, W.A., Love, W.E. Refinement of a molecular model for lamprey hemoglobin from Petromyzon marinus. J. Mol. Biol. 184:147–164, 1985.
19. Bolognesi, M., Onesti, S., Gatti, G., Coda, A., Ascenzi, P. Aplysia limacina myoglobin. Crystallographic analysis at 1.6 A resolution. J. Mol. Biol. 205:529–544, 1989.
20. Phillips, S.E.V. Structure and refinement of oxymyoglobin at 1.6 A resolution. J. Mol. Biol. 142:531–554, 1980.
21. Steigemann, W., Weber, E. Structure of Erythrocruorin in different ligand states refined at 1.4 A resolution. J. Mol. Biol. 127:309–316, 1979.
22. Arutyunyan, E.G., Kuranova, I.P., Vainshtein, B.K., Steigemann, W. X-ray structural investigation of leghemoglobin. Structure of acetate-ferrileghemoglobin at a resolution of 2.0 A. Kristallografiya 25:80–92, 1980.
23. Richardson, M., Dilworth, M.J., Scawen, M.D. The amino acid sequence of leghemoglobin I from root nodules of broad bean (Vicia faba 1.). FEBS Lett. 51:33–37, 1975.
24. Lalthantluanga, R., Braunitzer, G. Primary structure of one of the dimeric hemoglobin (Erythrocruorin) components, CTT-X, of Chironomus-thummi-thummi (Diptera). *Hoppe-Seyler's Z. Physiol. Chem.* 360:99–101, 1979.
25. Kortt, A.A., Trinick, M.J., Appleby, C.A. Amino acid sequences of hemoglobins I and II from root nodules of the nonleguminous Parasponia rigida-rhizobium symbiosis, and a correction of the sequence of hemoglobin I from Parasponia andersonii. Eur. J. Biochem. 175:141–149, 1988.
26. Takagi, T., Tobita, M., Shikama, K. Amino acid sequence

of dimeric myoglobin from Cerithidea rhizophorarum. Biochim. Biophys. Acta 745:32–36, 1983.
27. Furuta, H., Kajita, A. Dimeric hemoglobin of the bivalve mollusc Anadara broughtonii: Complete amino acid sequence of the globin chain. Biochemistry 22:917–922, 1983.
28. Suzuki, T., Takagi, T., Gotoh, T. Amino acid sequence of the smallest polypeptide-chain containing heme of extracellular hemoglobin from the Tylorrhynchus heterochaetus. Biochim. Biophys. Acta 708:253–258, 1982.
29. Suzuki, T., Gotoh, T. The complete amino acid sequence of giant multisubunit hemoglobin from the polychaete Tylorrhynchus heterochaetus. J. Biol. Chem. 261:9257–9267, 1986.
30. Shishikura, F., Snow, J.W., Gotoh, T., Vinogradov, S.N., Walz, D.A. Amino acid sequence of the monomer subunit of the extracellular hemoglobin of Lumbricus terrestris. J. Biol. Chem. 262:3123–3131, 1987.
31. Garey, J.R., Riggs, A.F. The hemoglobin of Urechis caupo. J. Biol. Chem. 261:16446–16450, 1986.
32. Jhiang, S.M., Riggs, A.F. The structure of the gene encoding chain c of the hemoglobin of the earthworm, Lumbricus terrestris. J. Biol. Chem. 264:19003–19008, 1989.