Recognition of Errors in Three-Dimensional Structures of Proteins

Manfred J. Sippl
Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, A-5020 Salzburg, Austria

ABSTRACT A major problem in the determination of the three-dimensional structure of proteins concerns the quality of the structural models obtained from the interpretation of experimental data. New developments in X-ray crystallography and nuclear magnetic resonance spectroscopy have accelerated the process of structure determination and the biological community is confronted with a steadily increasing number of experimentally determined protein folds. However, in the recent past several experimentally determined protein structures have been proven to contain major errors, indicating that in some cases the interpretation of experimental data is difficult and may yield incorrect models. Such problems can be avoided when computational methods are employed which complement experimental structure determinations. A prerequisite of such computational tools is that they are independent of the parameters obtained from a particular experiment. In addition such techniques are able to support and accelerate experimental structure determinations. Here we present techniques based on knowledge based mean fields which can be used to judge the quality of protein folds. The methods can be used to identify misfolded structures as well as faulty parts of structural models. The techniques are even applicable in cases where only the C_α trace of a protein conformation is available. The capabilities of the technique are demonstrated using correct and incorrect protein folds.
© 1993 Wiley-Liss, Inc.

INTRODUCTION

Knowledge of the three-dimensional structure of proteins is essential for understanding their function and it is a prerequisite for engineering their properties or for the design of drugs interfering with their biological function. Two powerful, complementary methods, X-ray analysis and nuclear magnetic resonance spectroscopy, are available to obtain the structures of proteins. However, in several cases experimentally determined structures have been proved to contain major errors and the recognition of errors in three-dimensional structures has recently become a subject for open discussion. In their recent commentary Bränden and Jones remark[1]: "The biological community has regarded X-ray structures as gospels of truth because X-ray diffraction data in principle contain sufficient information to extract the correct structure. This faith has been shaken by studies showing that there are serious errors in several recently published X-ray structures."

Prominent examples are ferredoxin from *Azotobacter vinelandii*, the small subunit of Rubisco, the HIV-proteinase and the Ha-*ras* oncogene product p21 (reviewed in refs. 1,2). Errors in these structures have been revealed by independent parallel or subsequent structure determinations on the same molecule. Erroneous interpretation of experimental data is not confined to X-ray analysis. An early nuclear magnetic resonance structure of a plasminogen kringle fragment was corrected only after a related X-ray structure was published.[2]

Even a reader sophisticated enough to interpret data of published structure determinations is usually unable to judge the quality of reported folds. Hence, the biological community is confronted with a fast growing number of experimentally determined protein folds, but it remains uncertain, whether or not the experimental data have been correctly interpreted. Exceptions are folds obtained by repeated structure determination or high resolution structures. But even in cases of moderate resolution and low *R*-factors structural models may contain errors.

A satisfying method for the quality assessment of experimentally determined protein folds should be complementary to experimental data, like resolution and *R*-factors obtained in a single structure determination. An ideal technique relies only on the

final coordinates obtained from the interpretation of experimental data. A major step toward the development of such techniques has been achieved by Eisenberg and co-workers.[3] They employ statistically derived amino acid properties like secondary structure preferences and solvent accessibilities. Using these parameters the quality of fit of a given amino acid sequence and an all atom model of its three-dimensional structure can be evaluated. The incompatibility of a sequence with a given fold is indicated by unfavorable profiles and scores obtained from the preference parameters.

In the present study methods derived from knowledge-based mean fields are used to analyze the energy distribution in protein folds. The technique is able to recognize misfolded structures and reveals incorrect parts of a given fold. A major advantage is that the technique is applicable to low resolution structures so that the incompatibility of a sequence and a given fold can be detected even in cases where only the C_α trace is available. A survey of a substantial number of public domain structures shows that there are only a few structures in the current Brookhaven data base[4] which appear to be problematic.

METHODS

Our approach to the development of force fields for protein solvent systems is based on Boltzmann's principle.[5] The set of three-dimensional structures solved by experimental means contains a tremendous amount of information on the forces which stabilize native folds in solution. Using Boltzmann's principle these forces are extracted from a data base of known structures in the form of potentials of mean force. The force field for a particular protein of known or unknown structure is then obtained by a recombination of these potentials as a function of the amino acid sequence. The applicability of this approach to protein folding has been explored recently and the mean field has been successfully applied to a number of problems in structural biology.[6-10] A most important feature is the ability of the force field to distinguish native folds from misfolded decoys.[6,7] This feature is the key which enables the recognition of incorrect structure determinations. To be specific, we are not interested in problems which arise from close contacts or other violations of basic steric principles but rather in the correct arrangement of the protein chain in three dimensions.

The mean field employed consists of potentials of mean force which model the energetic features of intramolecular pair interactions as a function of the spatial separation of two atoms associated with particular amino acids.[5] In a given structure the interaction energy e_{ij} between amino acid residues at positions i and j along the chain is the sum of the interaction energies between the atoms of the respective residues. In the calculations presented below e_{ij} corresponds to C_α–C_α or C_β–C_β interactions. The total pair interaction energy E is obtained from the sum over all pair interactions, $E = \frac{1}{2} \Sigma_{ij} e_{ij}$. The total energy of a protein is a function of its sequence S and conformation C, as expressed by the symbol $E_{S,C}$. The native fold corresponding to sequence S will be denoted by N. A model fold derived from experiment or modeling studies will be denoted by a different symbol X, since native and observed folds are not necessarily identical.

If X corresponds to (or is at least closely related to) the native fold N of sequence S then the energy $E_{S,X}$ should have its lowest value as compared to a large set of alternative conformations. This feature can be tested in a straightforward computer experiment. The backbone of the experimentally determined fold X is hidden among a large number of nonnative decoys C. In using only the backbone atoms (including C_β) the amino acid sequence cannot be derived from the remaining scaffold. Now the task is to seek X where the energy calculated from the force field is used as the guiding principle. The task is solved successfully if the sequence S has lowest energy when combined with the observed fold X as compared to all alternatives.

Hide and seek is performed on a polyprotein constructed from a set of known three-dimensional structures.[11] The structures are joined using short fragments from the proteins in the data base with the requirement that there are no close contacts between modules along the polyprotein and that the local geometry in the linker regions does not violate basic steric principles. This ensures that any fragment taken from the polyprotein corresponds to a reasonable conformation. The polyprotein used in this study consists of 230 proteins of known structure with a total length in the order of $L \approx 50,000$ residues.

In the hide and seek experiment the amino acid sequence S of length l of a protein is shifted along the polyprotein and the mean force energy $E(S,C)$ is evaluated at each position C. The set of conformations $C = 1, \ldots, L - l + 1$ encountered represents the conformation space of sequence S accessible in the experiment. Since $l << L$ the number of available conformations is practically independent of l and close to $L \approx 50,000$. The energies $E_{S,C}$ express the fitness of the sequence structure pair (S,C) in terms of the mean field. These energies are transformed to z-scores by $z_{S,C} = (E_{S,C} - \bar{E}_S)/\sigma_S$, where $\bar{E}_S = \Sigma_C E_{S,C}/(L - l + 1)$ and σ_s is the associated standard deviation. In particular $z_{S,X}$ is the z-score of the target fold X. The target fold X is successfully identified if $E_{S,X} < E_{S,C}$ and equivalently $z_{S,X} < z_{S,C}$, for all $C \neq X$. $z_{S,X}$ can be interpreted as a measure of the predictive power of the force field with respect to the protein of sequence S whose observed structure is X.

In this hide and seek experiment we use the hypothesis that the target conformation X corresponds to the native fold N of S and the goal is to verify this assumption. Deliberately misfolded proteins are obtained by assigning sequences S to arbitrary folds X. Here again we use the hypothesis that such folds are the most favorable for the associated sequence. In the case of misfolded proteins the hypothesis is, of course, false and we expect high energies and insignificant scores in these cases.

RESULTS

Figure 1 summarizes the results obtained in a survey of a set of 163 public domain structures using a force field consisting of C_α (left) and C_β interactions only (right). The number of decoys used for each test is in the order of 50,000 conformations. The ranges of energies encountered by individual sequences depend on sequence length and are not comparable, but the transformation of energies to z-scores enables a comparison of the results for all proteins. For most experimentally determined proteins the combination of their sequence S with the associated observed fold X yields the most favorable energy and z-score. On the other hand deliberately misfolded proteins are easily recognized by their high energies/z-scores. The result demonstrates that the force field provides a reasonable energy model for virtually all proteins in the public domain data base. The results obtained for the two different sets of potentials yield comparable results, but the scores obtained from the C_β interactions are generally larger (in absolute value). This may indicate that the C_β atoms carry a larger amount of information on the energetic features of protein folds as compared to C_α atoms.

It is noteworthy that the force field employed is derived from a set of soluble globular proteins. Hence, it is unreasonable to assume that the force field appropriately models the interactions encountered in hydrophobic environments. It is, therefore, indeed surprising that the observed folds of membrane associated and integral proteins (photosynthetic reaction center) as well as those of virus coat proteins are successfully identified as the most favorable structures among more than 50,000 alternatives. In general the z-scores of these proteins are, however, less significant from those of soluble globular proteins of comparable size.

The results summarized in Figure 1 contain four exceptions. The observed fold X of gene 5 DNA binding protein[12] (2GN5; the protein codes used are identical to the codes used in the Brookhaven protein data bank[4]) is not recognized as the most favorable fold. The rank of the 2GN5 fold with respect to C_β interactions is 3,326, i.e., there are 3,325 conformations which are more favorable for the 2GN5 sequence as compared to the observed fold of 2GN5. The result obtained for 2GN5 is most similar to
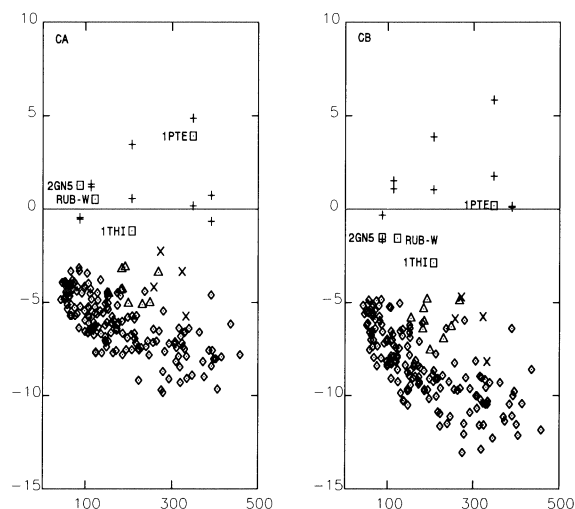

Fig. 1. z-scores obtained from hide and seek on a polyprotein plotted as a function of sequence length. z-scores calculated from C_α–C_α interactions are shown on the left and those calculated from C_β–C_β interactions on the right. Each symbol corresponds to the z-score $z_{S,X}$ of an individual protein. The different symbols denote soluble globular proteins ($\diamond$), chains of the membrane associated photosynthetic reaction center ($\times$), viral coat proteins ($\triangle$), incorrect and problematic folds ($\square$), and deliberately misfolded proteins ($+$). With the exception of RUB-W, 2GN5, 1THI, 1PTE, and deliberately misfolded proteins all observed sequence structure pairs have the most favorable energy and z-score as compared to the 50,000 alternatives derived from the polyprotein. The proteins used to compile the mean force potentials (marked by $\diamond$) are 1abp, 1acx, 1ak3-a, 1alc, 1bb-d, 1bp2, 1c5a, 1ca2, 1cbp, 1ccr, 1cd8, 1cho-i, 1cla, 1cms, 1col-b, 1crn, 1cro, 1ctf, 1eca, 1f3g, 1fb4-h, 1fb4-1, 1fc1-a, 1fc2-c, 1fdx, 1fkf, 1fx1, 1fxb, 1fxd, 1gcr, 1gd1-p, 1gky, 1gmf-b, 1hip, 1hne-e, 1hoe, 1il8-a, 1l01, 1ldm, 1lh1, 1lrd-4, 1lz1, 1mba, 1mbd, 1mle, 1ntp, 1p07-a, 1paz, 1pcy, 1pom, 1pp2-r, 1pyp, 1rhd, 1rn3, 1rnb, 1rsm, 1rve-a, 1sdh-a, 1sgt, 1sn3, 1snc, 1snv, 1tec-e, 1tgs-i, 1tnf-c, 1tpk-a, 1trb, 1ubq, 1utg, 1wsy-b, 1ypi-a, 256b-a, 2aat, 2act, 2apr, 2aza-a, 2c2c, 2ccy-a, 2cd4, 2chy, 2cna, 2cpp, 2cro, 2cyp, 2enl, 2er7-e, 2fbj-i, 2gbp, 2hhb-a, 2hhb-b, 2hip-a, 2hmg-a, 2kai-a, 2kai-b, 2lbp, 2lhb, 2ltn-a, 2mhr, 2pab-a, 2phh, 2prk, 2scp-a, 2sec-i, 2sni-e, 2sni-i, 2sod-b, 2ssi, 2trx-a, 2tsc-a, 2wrp-r, 351c, 3adk, 3b5c, 3blm, 3cpa, 3dfr, 3ebx, 3est, 3fxc, 3gap-a, 3hla-a, 3hla-b, 3icb, 3mt2, 3rp2-b, 3sgb-e, 4dfr-a, 4fd1, 4fxn, 4i1b, 4icd, mdh-a, 4pfk, 4sgb-i, 4tmn-e, 4xia-a, 5cpv, 5hvp-b, 5pti, 5rxn, 5tnc, 7at1-c, 7at1-c, 8adh, 8dfr, 9api-a, 9rub-b, 9wga-b, act1-a, hcrtn-b, hcys-i, hgpf-a, hhir-h, hmar, hmbm-i, hphy-k, hphy-l, hppst-i, nitro-a, ras1, rop2-a, sblp. Proteins whose codes start with "h" were obtained from R. Huber. nitro-a and sblp are courtesy of D. Rees and P. Sigler, respectively. ras1, rop2 and act1 were obtained from the EMBL file server. In all calculations presented the respective protein is subtracted from the data base and the potentials are recompiled. This ensures that the force field of a protein does not contain any specific information derived from experimental data of this molecule. For 1THI and 1PTE only C_α coordinates are available from the Brookhaven protein data bank[4]. C_β coordinates were obtained from Liisa Holm and Chris Sander. Misfolded pairs ($+$) are constructed by exchanging the sequences of experimentally determined folds. Pairs are formed using conformations and sequences of proteins of comparable length. This ensures that misfolded pairs have a globular fold. The sequence structure pairs shown in the figure are 1HMQ-2MCP, 2MCP-1HMQ, 1TPK-2GN5, 2GN5-1TPK, 1THI-3GAP, 3GAP-1THI, 1PTE-1MLE, 1MLE-1PTE, 1NSB-2PHH, 2PHH-1NSB.

RUB-W, an incorrect variant of the small subunit of ribulose biphosphat carboxylase (rubisco) producing a score of -1.57 (rank 2,211). The revised rubisco model ranks on first position and the score of -8.16
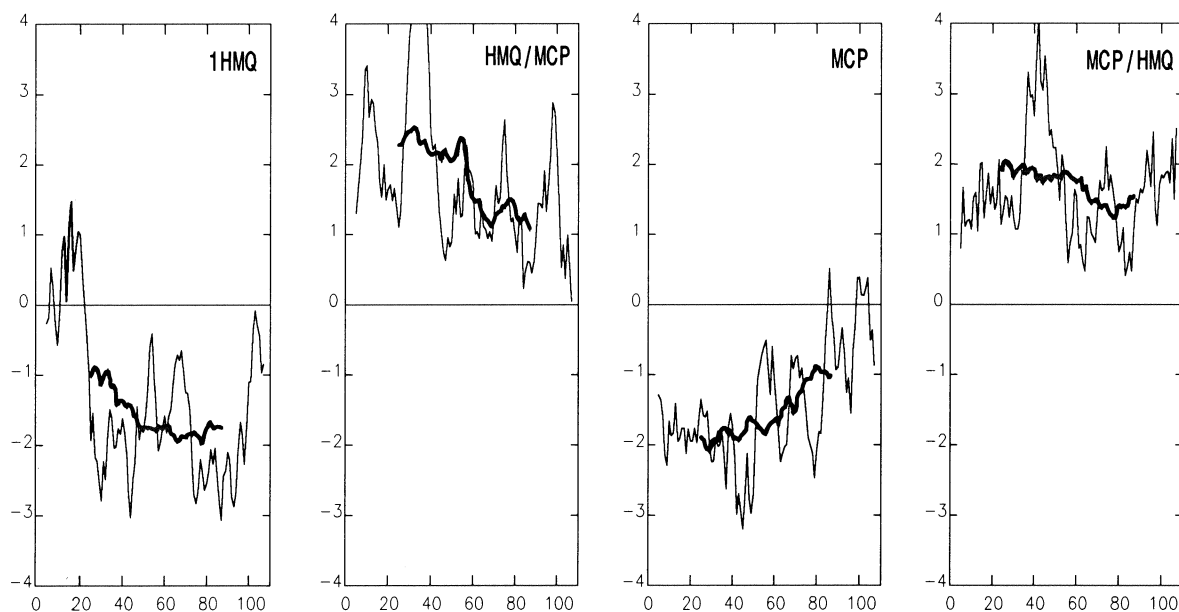
Fig. 2. Energy graphs of several observed and deliberately misfolded proteins calculated from C_β–C_β interactions. The graphs are smoothed by a window size of 10 (thin lines) and 50 (bold lines) residues. Energies are represented in units of E/kT. The figure compares the observed hemerythrin (1HMQ) and variable domain of a mouse myeloma immunoglobulin (2MCP) structures and two deliberately misfolded pairs. The misfolded pairs are obtained by exchanging the sequences of 2MCP and 1HMQ. 1HMQ/2MCP corresponds to the 1HMQ sequence folded in the 2MCP conformation. 2MCP/1HMQ corresponds to the 2MCP sequence folded in the hemerythrin conformation. Graphs of misfolded proteins (high energies) are in marked contrast with graphs obtained for observed sequence structure pairs (low energies).
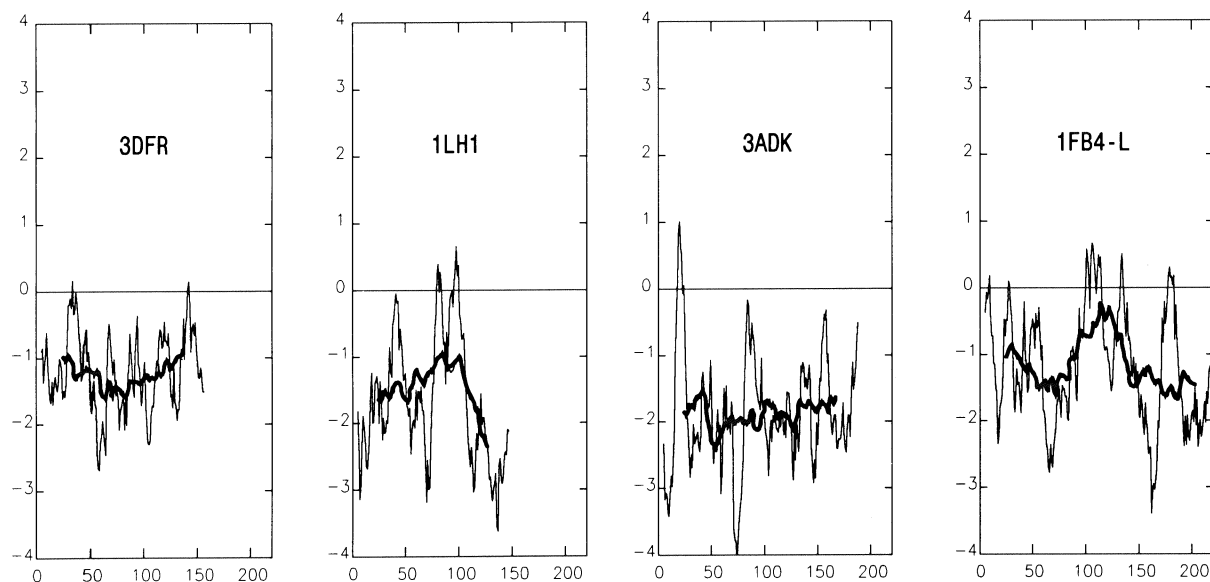

Fig. 3. Energy graphs of dihydrofolate reductase (3DRF), leghemoglobin (1LH1), adenylate kinase (3ADK), and immunoglobulin light chain (1FB4-L). The graphs are typical for native sequence structure pairs. Graphs obtained from a small window size ($\approx$ 10 residues) have only few small positive peaks. Graphs obtained from large window sizes ($\approx$ 50 residues) stay below zero. The examples cover a wide range of protein architectures. 1LH1 belongs to the all α class and 1FB4-L is an all β protein.

is in the range typical for native folds. Data obtained from recent nuclear magnetic resonance studies on 2GN5[13] are in conflict with the model obtained from the X-ray determination, a result which is in line with the result obtained from the hide and seek experiment.

A second case where the native fold is not identified as the most favorable structure is D-alanyl car-
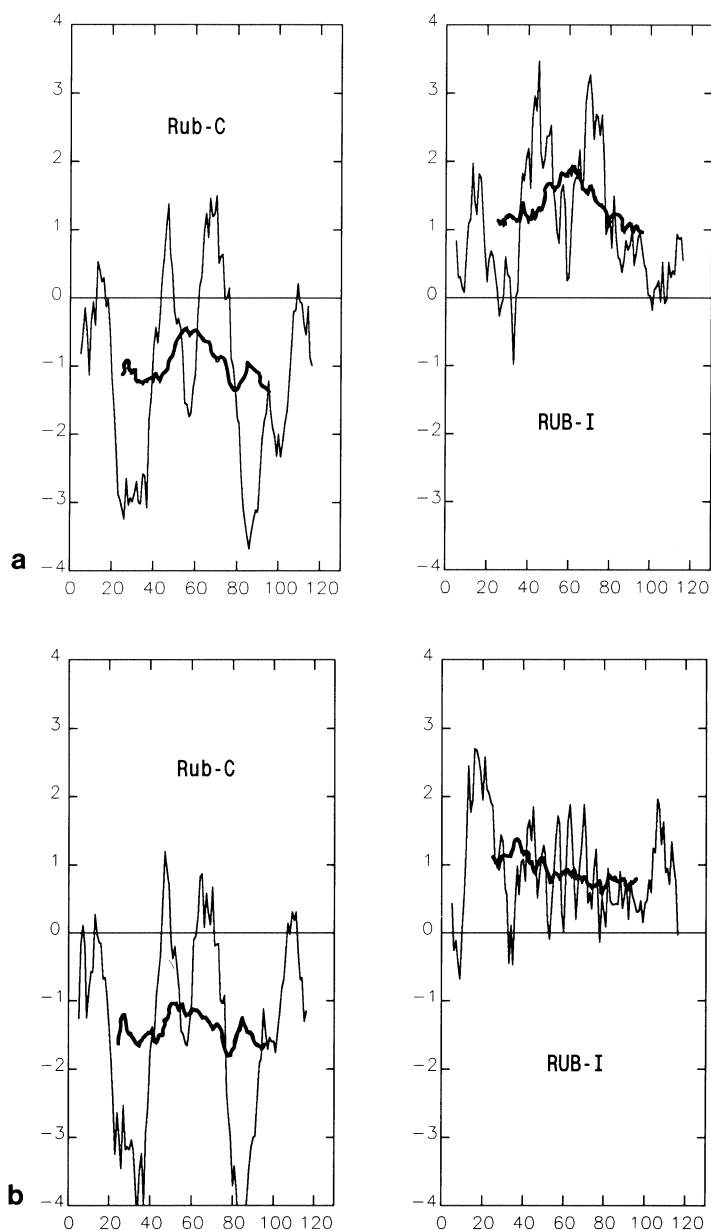
Fig. 4. Energy graphs of the correct (left) and incorrect (right) model of rubisco, based on C_α (a) and C_β interactions (b).

boxypeptidase[14] (1PTE). The fold determined by X-ray analysis to a resolution of 2.8 Å is defeated by 28,857 decoys. In the third case, the very sweet protein thaumatin[15] (1THI), the native fold ranks at position 82. In the hide and seek test these examples, like 2GN5 and RUB-W, behave more or less like deliberately misfolded proteins. The structure of thaumatin has been recently refined to a resolution of 1.65 Å. During refinement several frameshift errors were corrected and several loops were remodeled.[16] The coordinates of the refined model are not yet available so that we are presently unable to report the score for the revised structure.

The z-scores obtained from hide and seek can be interpreted as an overall quality index of a particular fold. A more detailed view of the energy distributions in protein folds is obtained from the residue interaction energies e_{ij}. The interaction energies e_{ij}, $i, j = 1, \ldots, l$ form the energy matrix $\mathbf{E}$ of a conformation, where the sequence length l corresponds to the dimension of the matrix. From $\mathbf{E}$ the interaction energy $e_i = \Sigma_j e_{ij}$ of a particular residue i with respect to all other residues in the molecule is derived. When e_i is plotted as a function of i we obtain an energy graph displaying the energy distribution of a sequence structure pair in terms of sequence posi-
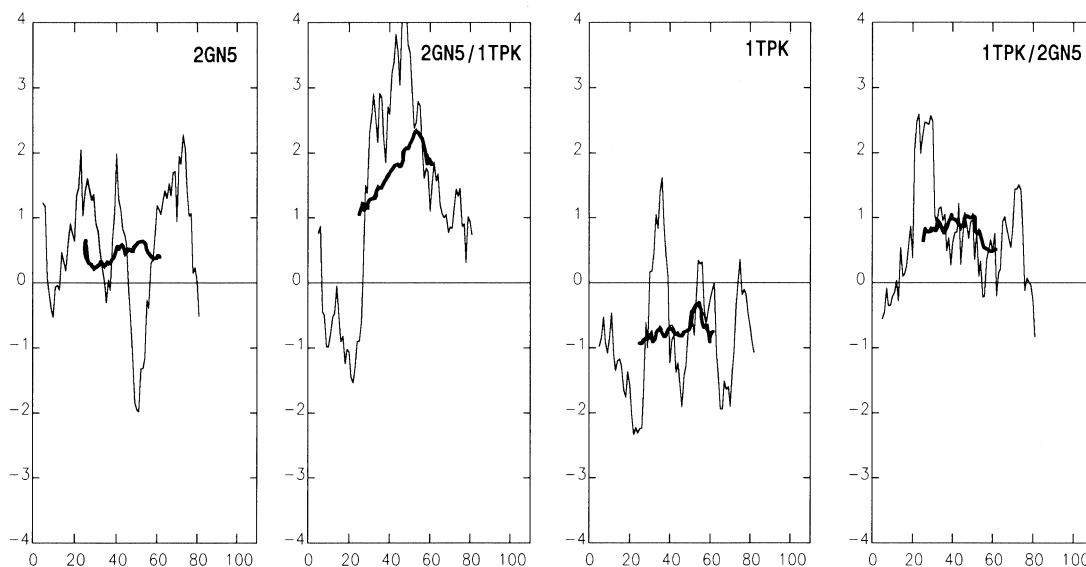
Fig. 5. Comparison of the observed 2GN5 and 1TPK folds with deliberately misfolded pairs. 2GN5 on 1TPK corresponds to the 2GN5 sequence folded in the 1TPK conformation and 1TPK on 2GN5 corresponds to the 1TPK sequence folded in the 2GN5 conformation. The graph of 1TPK is typical for native like sequence structure pairs. The 2GN5 graph resembles misfolded energy graphs.

tion. In energy graphs positive values point to strained sections of the chain whereas negative values correspond to stable parts of the molecule.

Figure 2 shows several energy graphs calculated from observed (hemerythrin and variable domain of an immunoglobulin) and misfolded sequence structure pairs.[17] Misfolded pairs have positive energies e_i for most positions i which is particularly apparent when a large window is used for sliding averages. Such structures are highly strained. In general, the energy graphs of observed sequence structure pairs have negative values and only occasionaly we observe small positive peaks as shown in Figure 3. Figure 4 compares the energy graphs calculated from the incorrect and correct structure determination of the small subunit of rubisco. The figure shows that the general features of the C_α graphs (Fig. 4a) are very similar to the C_β graphs (Fig. 4b), demonstrating that it is possible to identify faulty structures even if there is only a C_α trace available.

Figure 5 compares the energy graph of the observed structure of 2GN5 to the graphs of the observed 1TPK, the misfolded 1TPK-structure-2GN5-sequence and the misfolded 2GN5-structure-1TPK-sequence pairs. As expected from the high total energy the 2GN5 graph is similar to those of misfolded sequence structure pairs.

The graph of 1THI has some interesting features (Fig. 6). The N-terminal part is highly strained but the energy drops toward the C-terminus, reminescent of a partially misfolded chain. The high energy regions agree with those parts of the 1THI model, which were subsequently revised in the high resolution study.[16] The 1PTE structure again produces a

graph of high energy resembling energy graphs of misfolded chains (Fig. 7).

DISCUSSION

The high energies observed in the energy graphs of RUB-W, 2GN5, 1THI, and 1PTE are not a consequence of violations of basic steric requirements. In all calculations presented the energies e_{ij} are calculated for distances r in the range $4 \text{ Å} \le r \le 15 \text{ Å}$. Hence, possible close contacts do not contribute to the residue interaction energies e_i or total energies E. On the other hand mean force energies at large distances r are mainly determined by large proteins in the data base used to compile the mean force potentials. Inclusion of these energies may introduce side effects which derive from specific features of a particular protein. Consequently, to avoid such possible effects, the potentials were cut at 15 Å.

An important feature of mean field calculations is that they do not depend on minute structural details of protein folds. They can be applied in cases where only the C_α trace of a fold is available. Such techniques are particularly useful in the interpretation of electron density maps of low resolution. In a similar way mean field calculations can support the computation of structures from distance constraints derived from nuclear magnetic resonance studies, especially in cases where distance information is sparse or insufficient. In the context of the present work, the most important point is, however, that the computational tools presented complement experimental structure determinations allowing the objective judgment of the quality of a structure. This is
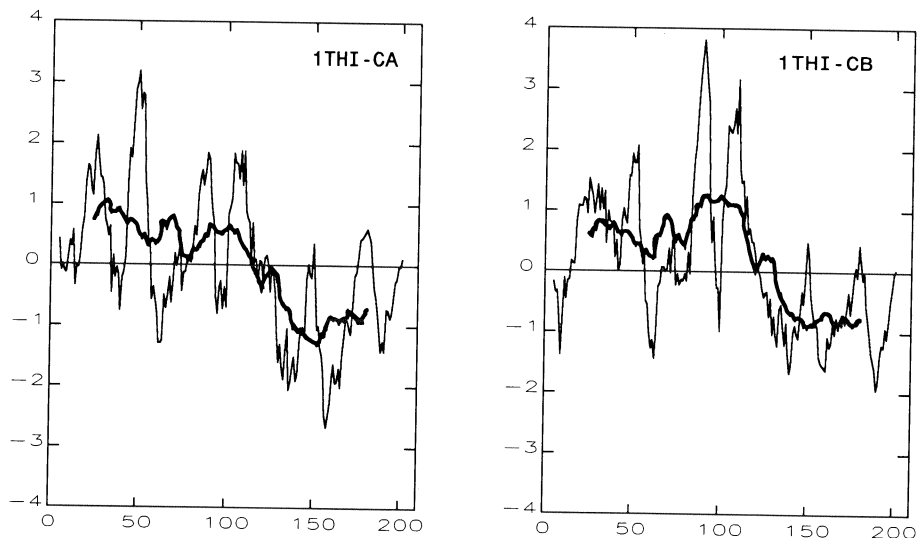
Fig. 6. Energy graphs of thaumatin (1THI) calculated from C_α–C_α interactions **(left)** and C_β–C_β interactions **(right)**.
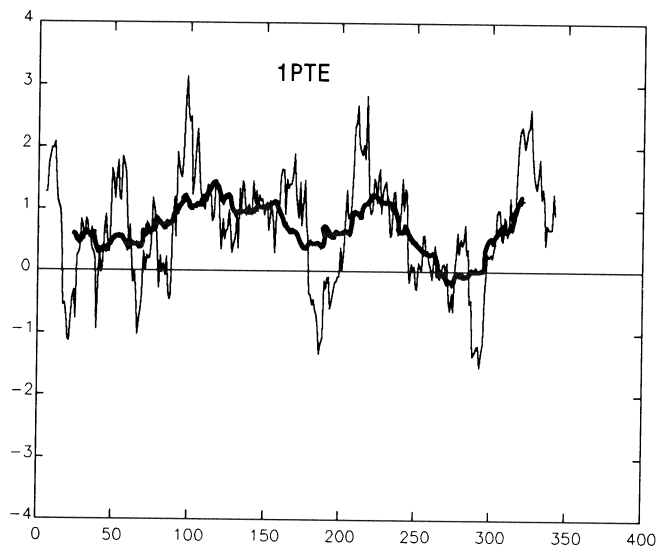

Fig. 7. Energy graph of D-alanyl carboxypeptidase (1PTE) (based on C_β interactions).

even possible for low resolution studies, and in cases where the C_α trace is available only.

In addition the techniques presented are relevant for protein structure prediction. z-scores and energy graphs can be calculated from any model, derived from experimental data or theoretical calculations. In other words the technique can be used to judge the quality of calculated or predicted structures. This in turn implies a possible route to protein structure prediction: Change the structural model as long as z-scores and energy graphs indicate a misfolded structure. We are currently proceeding along these lines.

In summary we conclude that the current version of the knowledge based mean field provides a rea-

sonable energy model for virtually all protein folds examined and that most of these structures have energetic features expected for native protein folds. It is clear that no comments can be made in the case of structures which have been published, but whose coordinates have not been made available to the scientific community. We suggest that reports of structure determinations are supplemented by the z-scores and energy graphs obtained from the respective structural models, since these parameters indicate the quality of the model.

ACKNOWLEDGMENTS

X-ray crystallographers who submitted coordinates to the Brookhaven data bank, Robert Huber and Paul Sigler for several structures, and Liisa Holm and Chris Sander for the backbone coordinates of 1THI and 1PTE. This work was supported by the Fonds zur Förderung der Wissenschaftlichen Forschung (Austria), project number 8361-CHE. A program running on silicon graphics workstations is available (e.g., sippl@agnes.came.sbg.ac.at).

REFERENCES

1. Bränden, C.I., Jones, T.A. Between objectivity and subjectivity. Nature (London) 343:687–689, 1990.
2. Janin, J. Errors in three dimensions. Biochimie 72:705–709, 1990.
3. Lüthy, R., Bowie, J.U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature (London) 356:83–85, 1992.
4. 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. The protein data bank: A computer based archival file macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
5. 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.
6. 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 models. J. Mol. Biol. 216:167–180, 1990.
7. Sippl, M.J., Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271, 1992.
8. 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 Sci. 1:625–640, 1992.
9. Casari, G., Sippl, M. Structure derived hydrophobic potential. J. Mol. Biol. 224:725–732, 1992.
10. Sippl, M.J. Boltzmann's principle, knowledge based mean fields and protein folding. J. Comput. Aided Mol. Design 7:473–501, 1993.
11. Sippl, M.J., Jaritz, M. Predicitive power of mean force pair potentials. J. Mol. Biol. 1993, submitted.
12. Brayer, G.D., McPherson, A. Refined structure of the gene 5 DNA binding protein from bacteriophage fd. J. Mol. Biol. 169:565–596, 1983.
13. Folkers, P.J., vanDuynhoven, P.M., Jonker, A.J., Harmsen, B.J., Konings, R.N., Hilberts, C.W. Sequence-specific ^{1}H-NMR assignment and secondary structure of the Tyr41→His mutant of the single stranded DNA binding protein, gene V protein, encoded by the filamentous bacteriophage M13. Eur. J. Biochem. 202:349–360, 1991.
14. Kelly, J.A., Knox, J.R., Moews, P.C., Hite, G.J., Bartolone, J.B., Zhao, J. 2.8 Å structure of penicillin-sensitive D-alanyl carboxypeptidase-transpeptidase from Streptomyces R61 and complexes with β-lactams. J. Biol. Chem. 260:6449–6458, 1985.
15. DeVos, A.M., Hatada, M., Van der Wel, H., Krabbendam, H., Peerdeman, A.F., Kim, S.-H. Three dimensional structure of Thaumatin I, an intensely sweet protein. Proc. Natl. Acad. Sci. U.S.A. 82:1406–1409, 1985.
16. Ogata, C.M., Gordon, P.F., de Vos, A.M., Kim, S.-H. Crystal structure of a sweet tasting protein thaumatin I, at 1.64 Å resolution. J. Mol. Biol. 228:893–908, 1992.
17. Novotny, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded protein models. Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.