Novel Knowledge-based Mean Force Potential at Atomic Level

Francisco Melo and Ernest Feytmans*

Facultés Universitaires Notre-Dame de la Paix, Structural Molecular Biology, Rue de Bruxelles 61, 5000 Namur Belgium

We present a new approach at the atomic level for the development of knowledge-based mean force potentials (MFPs) that can be used in fold recognition, ab initio structure prediction, comparative modelling and molecular recognition. Our method is based on atom-type definitions, raising the total frequency of the pairwise distributions and leading to very accurate and specific distance-dependent energy functions.

Forty different heavy atom types were defined depending on their bond connectivity, chemical nature and location level (side-chain or backbone). Using this approach it has been possible to obtain average frequencies of pairwise contacts about 15 times higher than the ones obtained using the classic way of one heavy atom definition for each amino acid (i.e. \( \alpha \)-carbon, \( \beta \)-carbon, virtual centroid or virtual \( \beta \)-carbon co-ordinates).

In this paper we use this approach to develop a MFP that can be used in fold recognition and we compare it with a classic MFP at the amino acid level compiled from the \( \alpha \)-carbon distances between the different amino acid pairs. Both potentials involve all the pairwise contacts extracted from a non-redundant folds database of 180 protein chains with a sequence identity threshold of 25%.

The pairwise energy functions of the MFP at the atomic level have a deep and very well defined minimum for each pairwise interaction, in contrast to the same curves obtained from the MFP developed at the amino acid level, which generally have multiple minima with similar depth.

Our results also show that this MFP is able to produce very similar energy profiles for couples of proteins that share a very low sequence identity but are closely related at the structural level. When these profiles are plotted considering the structure-structure alignment, they are mostly superimposed, showing a correlation with the structure-structure similarity. In the same test, the MFP at the amino acid level fails to produce similar profiles.

We suggest that using this MFP at the atomic level in the last stages of fold recognition or threading, when some candidates are available, can improve the sequence-structure alignments and, therefore, the final models. We also discuss the possibility of using this approach in the development of new MFPs to be used in ab initio structure prediction, comparative modelling and molecular recognition procedures.

© 1997 Academic Press Limited

Keywords: knowledge-based mean force potentials; fold recognition; threading; protein structure prediction; protein folding

*Corresponding author

Introduction

The prediction of the three-dimensional structure of a protein from its amino acid sequence is today one of the most ambitious goals in biology. Currently there are around 5000 different known pro-
tein structures deposited at the Protein Data Bank (PDB: Bernstein et al., 1977). On the other hand, we know around hundreds of thousands of protein sequences. Although information in both databases is highly redundant, many sequences have still unknown structure and function. It has been estimated that the rate of sequence determination is at least 50-fold higher than the rate of structure determination by NMR techniques or X-ray crystallography (Bowie et al., 1991). This large existing gap between known protein sequences and structures will continue to increase due to the current high rate of DNA sequencing, especially in many genome determination projects where thousands of new genes will be discovered.

**Comparative modeling**

It is known that proteins with homologous sequences have similar structures and generally often related functions. Many proteins in the databases share high similarities at the sequence level with some proteins of known structure. In these cases, it is possible to build some successful models of the structure of a protein using comparative modelling techniques (Browne et al., 1969; Greer, 1981, 1990). However, in many cases where there is a very low or not significant similarity at the sequence level between a protein and all the proteins of known structure, it is not possible to use this approach. In these cases, as we will see below, it is possible to use less accurate techniques such as fold recognition or “threading” to obtain a tentative model or a general fold for these proteins.

**Fold recognition**

As a result of the development of multiple sequence alignment methods (Fitch & Smith, 1983; Devereux et al., 1984; Corpet, 1988; Higgins & Sharp, 1989; Vingron & Argos, 1989; Argos et al., 1991; Depiereux & Feytmans, 1991), structure-structure alignment techniques (Nishikawa & Ooi, 1974; Rossmann & Argos, 1976; Sippl, 1982; Mitchell et al., 1989; Taylor & Orengo, 1989; Sali & Blundell, 1990; Barakat & Dean, 1991; Vriend & Sander, 1991; Fischer et al., 1992; Holm et al., 1992; Holm & Sander, 1993c, 1994b; Murzin et al., 1995), molecular biology tools in general, and X-ray crystallography works carried out by many different groups, we currently know that two unrelated proteins at the sequence, functional or evolutionary level can sometimes share a surprising similarity in some folding motifs or in their overall fold or architecture (Chothia & Lesk, 1986, 1987; Holm & Sander, 1993a, 1993b, 1994a, 1995).

It seems that, during evolution, only a limited number of building blocks of protein structures have been generated. Consistent with this, around 50% of the newly solved experimental structures appear to be related to known folding motifs (Blundell & Doolittle, 1992). It is currently accepted that the total number of possible folds is limited, although there is discussion about their number (Chothia, 1992; Orengo et al., 1994). One of these estimations is that the maximum number of protein families would be around 1000 (Chothia, 1992) and the total number of different folds would be approximately half this number (Blundell & Johnson, 1993). We currently know around 200 different folds, and assuming that the above figures are correct, it means that we know about 50% of the total possible different folds. This knowledge gives us the possibility to find a known fold on which a particular protein sequence with unknown structure can fit in a good way. When a protein does not exhibit a high similarity with any known protein structure at the sequence level, the above procedure could be carried out. This approach is called fold recognition or “threading” and consists in forcing the fold of a sequence on a library of different known folds. The sequence-structure alignments are evaluated depending on the energies obtained using MFPs derived from a database of non-redundant folds.

Fold recognition procedure has been conceived by David Eisenberg’s group (Bowie et al., 1991) and several groups have followed or independently pursued the idea, bringing important improvements and contributions (Sippl & Weitckus, 1992; Jones et al., 1992; Godzik et al., 1992; Ouzounis et al., 1993; Maiorov & Crippen, 1992; Goldstein et al., 1992; Wilmans & Eisenberg, 1993; Bryant & Lawrence, 1993; Johnson et al., 1993; Kocher et al., 1994). There are three major components in fold recognition or threading procedures: the mean force potential, the sequence-structure alignment method and the criteria used to discriminate between wrong and correct alignments. Each of these components are directly or indirectly related to each other. The different groups working in fold recognition have developed threading methods that present variations in each component. The description of many existent knowledge-based MFPs and sequence-structure screening methods has been reviewed (Wodak & Roostan, 1993). Bryant & Altschul (1995) proposed a criterion for evaluating the results of threading, which is based in Z-scores relative to the composition-corrected score distribution.

**How accurate are the currently used fold recognition techniques?**

A large scale experiment was performed in order to determine the current state of the different methods in protein structure prediction (Moult et al., 1995). This experiment was a blind test, where the structures to be predicted were not known by the participants. In December 1994, a meeting was held at the Asilomar conference centre in California to examine the predictions. There, it was discussed why some predictions were right and others were wrong. The results of this experiment and the meeting were reported in a special issue (Proteins: Structure, Function, and
Genetics 23, 1995). Based on this experiment, an evaluation of the current threading methods was performed (Lemer et al., 1995). The conclusion of this evaluation was that, in general, all the threading methods were able to detect the correct fold, when it was present in the database of known structures. However, the quality of the sequence-structure alignments, in the case of correctly recognized folds, is not good when compared with the structure-structure alignments. This means that even when the methods are able to detect the fold, the placement of the residues on the structure is not correct. It seems that current fold recognition is achieved through a maximization of hydrophobic interactions in the protein core. This is consistent with the recent finding of Levitt’s group (Huang et al., 1995), who have shown that a simple arrangement of polar and hydrophobic residues, i.e. a binary pattern, is very effective in recognizing native folds in general. However, they used the most simple conceivable alignment method, of course unable to produce even a tentative model of the protein in most of the cases.

Mean force potentials currently used in fold recognition

All the current threading methods use mean force potentials developed at the amino acid level, i.e. they involve distance-dependent amino acid interactions between two residues, contacts between two or three amino acids, angles between residue pairs, etc. (Lemer et al., 1995). They do not consider all the side-chain atoms of the different amino acids. Although this approximation works relatively well in some cases as a secondary structure prediction, just considering the whole amino acid as a central body of one interaction is perhaps a rough simplification of the energetic of the real system. All the atoms of an amino acid are considered as one unique set under the same name, even if these atoms belong to different chemical functional groups.

The current MFPs used in threading are statistical potentials developed from a non-redundant database of known structures (Lemer et al., 1995; Sippl, 1995). It is possible to relate probability and energy of a pairwise contact using the inverse Boltzmann’s law (Sippl, 1990, 1993a). This approach is statistical in nature and the connections to the real physical interactions are not yet clear. The energies obtained using this approach do not represent real energies but its usefulness and consistency have been proved in many cases (Jones et al., 1992; Sippl, 1993a,b).

The non-redundant databases currently used for the derivation of these MFPs do not contain enough occurrences for some pairwise residue contacts. Although some MFPs include a correction for sparse data or a weighting scheme depending on the number of observations, in some cases, for the less frequent amino acids, the very low number of observations gives inaccurate relative frequencies and poorly estimated probabilities. This can be solved defining that different atoms sharing similar physico-chemical characteristics can be grouped under a same name, thus reducing the total number of different atoms and considering as the same atom the ones that are different in a very strict physico-chemical point of view. This approximation enables us to derive an accurate MFP at the atomic level, which contains a very high number of observations for almost all the different atom pairwise interactions. Also, because of its atomic level definition, this MFP contains more specific pairwise interactions than the existent MFPs at the amino acid level.

In this paper we describe a new MFP at atomic level, which involves 40 different atom type definitions, and we compare it with a MFP developed at the amino acid level from the same database of non-redundant folds. We also discuss the possibility of using this novel MFP in threading or fold recognition as a possible solution to obtain better sequence-structure alignments and therefore better structural models for unknown protein sequences.

Results

We have derived two MFPs: a new MFP at the atomic level (ATL) considering our heavy atom type definitions (Figure 1) which will be called ATL-MFP, and a MFP at the amino acid level (AAL) considering the a-carbon for each amino acid which will be called AAL-MFP. This last MFP has been derived in order to compare different aspects with the ATL-MFP.

Statistics of the MFPs

Our protein data set contains 180 different chains, which correspond to 47,197 amino acid residues and 367,140 heavy atoms. In this set, the lowest represented amino acids are cysteine and tryptophan with 666 (1.41%) and 687 (1.46%) occurrences, respectively. The highest represented amino acids are alanine (4022, 8.52%), glycine (3881, 8.22%) and leucine (3873, 8.21%).

After performing the calculations, 15,749,848 atomic pairwise contacts and 277,545 amino acid pairwise contacts were obtained. The total number of observations is strongly dependent on the sequence separation. Figure 2 shows the average total observations for all the different distributions in both cases at the different sequence separations. At the amino acid level, the average of total observations ranges between 120 and 12, while at the atomic level it fluctuates between 1600 and 150. Through the whole spectrum of sequence separation, the average atom pairwise observations are higher than the average amino acid pairwise observations, by a factor of 10 to 15-fold. The average number of observations, considering all the sequence separations, is about 205 for an atom pairwise contact and about 14 for an amino acid pairwise contact. This means that the ATL-MFP
Figure 1. Diagram showing the 40 atom types defined for all the heavy atoms of the 20 amino acids. The definition was made according to the atom location (side-chain or backbone), connectivity and chemical nature. All the backbone atoms contain the same definition among the different amino acids, with the exception of the $\alpha$-carbon of glycine and the nitrogen of proline. The nitrogen atom of the extreme amino-terminal and the oxygen atom of the extreme carboxy-terminal were defined as the types 20 and 28, respectively.
contains, on average, about 15 times more observations for the different pairwise contact distributions than the AAL-MFP.

Energy functions of the MFPs

In order to compare the energy functions of both MFPs, we choose as an example the amino acid pairs Ala-Ala and Cys-Cys as representatives of the most and less frequently observed pairs, respectively. Also, they can be easily compared with the atomic pairs 6-6 and 19-19. The atom type 6 describes a methyl group bounded to a carbon with hybridization sp$^3$ through a simple bond and the atom type 19 represents the sulphur atom of a cysteine (Figure 1). The atom type 6 is present in many amino acid side-chains and is one of the most observed in our data set. In contrast, the atom type 19 shows the same low frequency as Cys, because it is only present in this amino acid side-chain.

Initially, as a global picture of the MFPs, we compared the energy functions for the pairs Ala-Ala and 6-6 in the whole range of sequence separations (Figure 3). The atomic pairwise energy functions show a very good definition along all the sequence separations, with a deep minimum between 3.5 and 4.0 Å. This is consistent with the expected optimum interaction distance between two methyl groups. Due to the very high number of observations, it is possible to obtain an energy curve shape very similar to a classical electrostatic atom pairwise energy function in vacuum: repulsion at very short distances, one deep minimum with negative energy at a certain distance and zero energy at very long distances. Although the Ala-Ala pair is one of the most well represented in our data set, the amino acid pairwise energy functions show multiple and not very well defined minima, especially when the sequence separation is increased.

When individual pairwise energy functions at some specific sequence separations are examined, the differences are much more clear. Figure 4 shows the energy curves for these pairs at some particular sequence separations. The 6-6 pair shows a defined minimum that does not change its position for the different sequence separations. The Ala-Ala pair always shows more than one minimum, not very deep and whose positions vary for different sequence separations.
Interestingly, a similar result is observed when we compare the energy functions of two pairs that present a very low number of observations. We have considered the atomic pair 19-19 and the amino acid pair Cys-Cys. Figure 5 shows these energy curves for some particular sequence separations. The atomic pair exhibits a single and very deep minimum between 2.0 and 3.0 Å, which does not change with the sequence separation. The amino acid pair sometimes exhibits two not very deep minima. The first minimum between 5.0 and 5.5 Å is conserved in all the sequence separations considered.

Energy profiles of misfolded proteins

In order to roughly test the discriminative power of both MFPs, we calculated the energy profiles of some misfolded proteins. All these misfolded proteins contain a backbone conformation that belongs to an unrelated protein structure. This means that some residues are not correctly located on the structure and that many side-chain atoms have a wrong environment (see Methods). In all the cases, the ATL-MFP shows a very high discriminative power between the correct and the misfolded protein (Figure 6). In contrast, the AAL-MFP is not able to discriminate clearly between them (see Discussion).

Structure-structure alignments over the energy profiles

The FSSP database contains all the structure-structure alignments for the different couples of proteins. We have plotted the energy profiles shown in Figure 7 but considering these structure-structure alignments (Figure 9). In general, there is a very good correlation between the alignments and the energy profiles obtained using the ATL-MFP. In most cases the gaps of the alignment produce the superimposition of the energy profiles for both proteins.

Discussion

In this paper we described a new approach at the atomic level for the derivation of knowledge-based MFPs. The goal of this approach is the improvement of two aspects currently present in the MFPs at the amino acid level: the first one is the low observed frequency for many amino acid pairs and it is improved using common type definitions; the second one is the accuracy of the interaction being described and it is improved working at the atomic
level. The improvement of these aspects allows us to obtain much better defined energy functions.

We have shown that the frequency of observations using this approach is always between 10 and 15-fold higher, for all the sequence separations considered. Here arises an important point. Sippl has suggested that segments higher than ten residues should not be useful, because the current pool of segments cannot be a valid representation of the conformational space (Sippl, 1990). We also point out that in our definition at the amino acid level, using a sequence separation higher than 7, leads to a very low average number of observations. Of course this will depend on the maximum distance

Figure 4. Plots of the energy functions $\Delta E^i_j(l)$ for the atomic pair 6-6 (A, C, E and G) and the $\alpha$-carbons of the amino acid pair Ala-Ala (B, D, F and H) at the sequence separations of $k = 6$ (A and B), $k = 10$ (C and D), $k = 20$ (E and F) and $k = 30$ (G and H).
used in the MFP definition, but it is an aspect that has to be carefully considered in the MFP implementations.

The shape of the energy functions of the ATL-MFP is normally very well defined in the whole spectrum of sequence separations considered in this work. The energy curves exhibit repulsion at short distances, a deep minimum at a certain distance that depends on the atomic pair and zero energy for long distances. In this approach at the atomic level, we normally found that the shape of the energy functions does not change with the sequence separation, as is the case in the approach at the amino acid level. Possibly, this is due to the fact that interactions described at the atomic level are much more precise than those when considering an α-carbon or a virtual β-carbon for an amino acid side-chain, and the same precision will be present in the reference system. This is consistent with the atomic energy functions recently published by Sippl’s group (Sippl, 1996; Sippl et al., 1996).

The approach of using atom types sometimes fails in raising the frequency because of the existence of some atom types that represent unique atoms. Despite this, as we have shown in the case of a low number of observations, the accuracy of the energy functions is improved when compared to its amino acid equivalent, because the reference system is also more precise when working at the atomic level than at the amino acid level. The curves obtained in these cases contain a deep minimum but not very high repulsion terms at short distances, because the weighting system used in this work is based on the number of observations. The shape of the atomic energy curves reflects improvement of the accuracy. The existence of only one very well

Figure 5. Plots of the energy functions $\Delta E_{ij}(l)$ for the atomic pair 19-19 (A, C and E) and the α-carbons of the amino acid pair Cys-Cys (B, D and F) at the sequence separations of $k = 10$ (A and B), $k = 20$ (C and D) and $k = 30$ (E and F).
defined minimum implies a very precise description of each pair interaction optimum, which in turn implies an improvement in the discriminant power of the energy function.

The current MFPs used in fold recognition at the amino acid level are much more accurate than the simple definition used in this paper for the comparison with the ATL-MFP. They include all the backbone atoms, a $\beta$-carbon or a virtual $\beta$-carbon, a solvent exposure term, etc. These MFPs have been shown to be able to identify misfolded protein structures (Sippl, 1993b). Our AAL-MFP completely fails in the accomplishment of this task, mainly due to the low accuracy of their energy curves describing only the interaction between $\alpha$-carbons and, perhaps, due to the lack of solvent exposure terms. In spite of the existence of more accurate AAL-MFPs, the problem is the same in terms of frequency, because these potentials always label the atoms under an amino acid name. Thus, the argument of low observed frequency is the same for all the MFPs considering distance dependence between amino acids. On the other hand, our MFP can be also improved by including solvation terms and other parameters. The idea was to compare two potentials in terms of frequency and specificity of the pairwise interactions, using the most simple definitions. Also, the MFPs at the amino acid level do not consider all the side-chain atoms of the different amino acids and, just considering the whole amino acid as a central body of one interaction is perhaps a rough simplification of

**Figure 6.** Energy profiles of some proteins in their native (continuous line) and misfolded (broken line) conformation. The profiles were calculated using the MFP at the atomic level (A, C and E) and the MFP at the amino acid level (B, D and F). A and B, 1PPT; C and D, 1LH1; E and F, 5PAD. A window average of 11 residues was used to calculate the energy profiles.
the energetics of the real system. All the atoms of an amino acid are being considered as one unique set under the same name, even if these atoms belong to different chemical functional groups. The genetic evolution process changes a complete amino acid at once and does not change just some side-chain atoms of one residue. This does not mean that the natural selection on the stability of a protein will also work at the amino acid level.

When one mutation is introduced in a gene, the change in the stability of the mutant protein involves all the atom interactions between the new residue atoms and the atoms of the amino acids that potentially can interact with them.

The current MFPs used in fold recognition have shown to be successful in the identification of a compatible fold for a specific sequence, when the fold is present in the database of known protein

**Figure 7.** Energy profiles of some couples of structurally related proteins using the MFP at atomic level. A, 1mol-A; B, 1stf-I; C, 1ash; D, 1hp; E, 3ink-M; F, 1rcb; G, 2btf-A and H, 1hpm. A window average of 11 residues was used to calculate the energy profiles.
structures (Lemer et al., 1995). However, they normally fail in producing the correct alignment between the sequence and the structure, when compared with the structure-structure alignment. Thus, most of the time, these MFPs are not successful in producing a refined structural model of the protein, but they rather suggest a list of possible folds for the query sequence. We propose that using this new MFP over some already available candidates, previously obtained with current threading methods, could improve the alignment between sequence and structure and contribute in obtaining a tentative model of the protein. In that way, both approaches could be used complementary in the different stages of the fold recognition procedure. When an amino acid dependence is not

![Image](image-url)

**Figure 8.** Energy profiles of some couples of structurally related proteins using the MFP at the amino acid level. A, 1mol-A; B, 1stf-I; C, 1ash; D, 1lp-f; E, 3ink-C; F, 1rcb; G, 2btf-A and H, 1hpm. A window average of 11 residues was used to calculate the energy profiles.
discriminant, a more accurate atomic dependence could be used.

We have shown that the ATL-MFP is able to produce very similar energy profiles for couples of proteins very close at the structural level, but sharing a very low sequence identity. Also, the energy profiles correlate very well with the structure-structure alignments of these proteins. This finding suggests that this MFP should be very accurate in the identification of correct sequence-structure alignments.

However, there is still one problem that could arise when working at the atomic level in fold recognition. We must place the side-chain atoms of the query sequence on the target backbone. There are many different rotamer databases available and

Figure 9. Energy profiles of Figure 7 superimposed using the structure-structure alignments from the FSSP database, for the different couples of structurally related proteins. A, 1mol-A (heavy line) and 1stf-I; B, 1ash (heavy line) and 1lp; C, 3ink-C (heavy line) and 1rcb; and D, 2btf-A (heavy line) and 1hpm. The proteins plotted in the heavy line conserve the real position of their amino acids in the graph.
### Methods

**Atom type definition**

The first step in our approach was to define the different atom types for all the heavy atoms of the 20 amino acids. In a strict physico-chemical point of view, all the atoms with different environments, connectivity and chemical nature, would be different. In the 20 amino acids the total number of heavy atoms is 167 and the number of non-equivalent heavy atoms is 98. In order to reduce this number and raise the observed frequencies, we decided to do some approximations and we have finally defined a total number of 40 different atom types for all the heavy atoms of the 20 amino acids (Figure 1). The atom type definition is based on its connectivity, chemical nature and location level (side-chain or backbone). For example, one type of heavy atom is the carbon of a methyl group bonded to a carbon with $sp^3$ hybridization. This type represents the $\beta$-carbon of alanine side-chain, the $\gamma_1$ and $\gamma_2$-carbons of valine, the $\delta_1$ and $\delta_2$-carbons of leucine, etc. All the atoms belonging to each atom type definition are shown in Table 1.

### Non-redundant fold data set

A set of 180 protein chains with complete atomic co-ordinates was used to perform the calculations. This set was derived from the October 1995 release of PDB representative list (Hobohm et al., 1992; Hobohm & Sander, 1994), excluding all the proteins with duplicated or missing atoms, structural gaps, and with a number of residues lower than 300. All these proteins share a sequence identity lower than 25% and have a resolution below 3.0 Å. The list of this data set, using the PDB chain identifiers, is as follows: 1AJJ, 1ADD, 1AEF, 1AGX, 1ALKA, 1AGM, 1AMP, 1AORA, 1AOZA, 1ARB, 1ARS, 1ASZA, 1ATNA, 1AIZCA, 1ABB, 1BBP, 1BBT2, 1BBT5, 1BMTA, 1BVP1, 1CAUA, 1CAUB, 1CCR, 1CELA, 1CEWI, 1CFB, 1CHMA, 1CID, 1CMCA, 1COLA, 1CPCA, 1PCPB, 1PRL, 1CSN, 1CTM, 1CTN, 1CTT, 1DAA3, 1DAA4, 1DBA, 1DBA6, 1DYNA, 1DEE, 1DEE1, 1DIM1, 1ERIA, 1FBAA, 1FCD2, 1FCFDA, 1FCDC, 1FKF1, 1FNC, 1FRUB, 1GBS, 1GSHA, 1G42, 1GMFA, 1GOF, 1GPH1, 1GPR, 1GQ2, 1HDCA, 1HEX, 1HFI, 1HJR, 1HMY, 1HNF, 1HSLA, 1HTMD, 1HUCB, 1HVD, 1IAE, 1IAG, 1IGP, 1IKB, 1IB3, 1ILB, 1ILK, 1ILD, 1IDLA, 1ILFE, 1ILTS, 1ILSD, 1MSAA, 1MSC, 1N1L, 1NAR, 1NBAA, 1NDH, 1NNT, 1NPK, 1PBE, 1PPB, 1PCRH, 1PHP, 1PHR, 1PLQ, 1PMY, 1PNP, 1POA, 1POE, 1POX, 1PPI, 1PRCC, 1PRCL, 1PRCT, 1PRTD, 1SPSA, 1SRML, 1SRCB, 1SRIBA, 1RTP1, 1RVAA, 1SACA, 1SCUA, 1SCUB, 1SRYA, 1STD, 1TADC, 1TCA, 1THV, 1TIB, 1TLC, 1TLP, 1TNA, 1TPHI, 1TRKA, 1TTSA, 1TLA, 1ULAA, 1UNX, 1YPTB, 1YRTIA, 2ACG, 2BBKH, 2BAPI, 2BPA2, 2CAS, 2CHR, 2CPL, 2CYT, 2DKB, 2END, 2EFR, 2GSTA, 2HNO, 2HPDA, 2KAUB, 2KAUC, 2LGA, 2LIV, 2MAAD, 2MEVI, 2MEG, 2MN, 2MTAC, 2MTA, 2OHFXA, 2PGD, 2PIA, 2RN2, 2SAS, 2SCPA, 2SIL, 2SNV, 2STV, 2TGI, 2TMVP, 3AAHA, 3MDDA, 3SC1, 4BLMA, 4ENL, 4FENX, 4RHV1, 4RHV3, 5P21, 6FAB, 6TAA, 8ACN, 8ATCA, 8CAT.

### Mean force potentials

The potentials have been calculated using 20 different classes of distance ($l$), ranging from 0 to 10 Å at 0.5 Å intervals. All pairwise occurrences out of this range were excluded. The sequence separation or topological factor ($k$) was always considered at the amino acid level for some very fast methods for the placement of the side-chain atoms on a given backbone, but a perfect method is not available (Holm & Sander, 1991; Lee & Subbiah, 1991; Bassolino-Klimas & Bruccoleri, 1992; Desmet et al., 1992; Dunbrack & Karplus, 1993; Koehl & Delarue, 1994; Hwang & Liao, 1995). We do not know how sensitive the ATL-MFP is to the wrong placement of the side-chain atoms and because of that we do not know if this MFP will work as well on fold recognition as it does performing energy profiles on native proteins.

The definition of atom types could also be considered for the derivation of MFPs to be used in comparative modelling, ab initio structure prediction and molecular recognition techniques. In these procedures an atomic level approach with accurate energy functions would be very useful.

### Table 1. Heavy atoms of the standard amino acids belonging to each atom type definition

<table>
<thead>
<tr>
<th>Atom type</th>
<th>Type definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C° (all amino acids, except Gly)</td>
</tr>
<tr>
<td>2</td>
<td>Gly-C°</td>
</tr>
<tr>
<td>3</td>
<td>N (all amino acids, except Pro)</td>
</tr>
<tr>
<td>4</td>
<td>C (all amino acids)</td>
</tr>
<tr>
<td>5</td>
<td>O (all amino acids)</td>
</tr>
<tr>
<td>6</td>
<td>Ala-C°, Ile-C°, Leu-C°, Thr-C°, Val-C°, Ile-C°, Leu-C°</td>
</tr>
<tr>
<td>8</td>
<td>Pro-N</td>
</tr>
<tr>
<td>9</td>
<td>Phe-C°, Trp-C°, Tyr-C°</td>
</tr>
<tr>
<td>11</td>
<td>Trp-C°</td>
</tr>
<tr>
<td>12</td>
<td>Trp-C°, Ser-C°</td>
</tr>
<tr>
<td>13</td>
<td>Ser-O°, Thr-O°</td>
</tr>
<tr>
<td>14</td>
<td>Thr-C°</td>
</tr>
<tr>
<td>15</td>
<td>Asn-N°, Gln-N°</td>
</tr>
<tr>
<td>16</td>
<td>Cys-S°</td>
</tr>
<tr>
<td>17</td>
<td>Lys-N°</td>
</tr>
<tr>
<td>18</td>
<td>Arg-C°</td>
</tr>
<tr>
<td>19</td>
<td>Arg-N°1, Arg-N°2</td>
</tr>
<tr>
<td>20</td>
<td>His-C°</td>
</tr>
<tr>
<td>21</td>
<td>His-C°, Trp-C°</td>
</tr>
<tr>
<td>22</td>
<td>His-N°</td>
</tr>
<tr>
<td>23</td>
<td>His-N°</td>
</tr>
<tr>
<td>24</td>
<td>Asp-C°, Glu-C°</td>
</tr>
<tr>
<td>25</td>
<td>Asp-O°1, Asp-O°2, Glu-O°1, Glu-O°2</td>
</tr>
<tr>
<td>26</td>
<td>Cys-C°, Met-C°</td>
</tr>
<tr>
<td>27</td>
<td>Met-C°</td>
</tr>
<tr>
<td>28</td>
<td>Tyr-C°</td>
</tr>
<tr>
<td>29</td>
<td>Pro-C°</td>
</tr>
<tr>
<td>30</td>
<td>Asn-C°, Gln-C°</td>
</tr>
<tr>
<td>31</td>
<td>Asn-O°1, Gln-O°1</td>
</tr>
<tr>
<td>32</td>
<td>Lys-C°</td>
</tr>
<tr>
<td>33</td>
<td>Arg-N°</td>
</tr>
<tr>
<td>34</td>
<td>Arg-C°</td>
</tr>
<tr>
<td>35</td>
<td>His-N°</td>
</tr>
<tr>
<td>36</td>
<td>Trp-N°</td>
</tr>
<tr>
<td>37</td>
<td>Tyr-O°</td>
</tr>
</tbody>
</table>

Some of the amino acids, such as Arg, Lys, and Cys, have several side-chain atoms on a given backbone, but a perfect method is not available. The potentials have been calculated using 20 different classes of distance ($l$), ranging from 0 to 10 Å at 0.5 Å intervals. All pairwise occurrences out of this range were excluded.
Table 2. Couples of structurally related proteins, sorted by RMSD, obtained from the FSSP database after applying some filters on it (see Methods)

<table>
<thead>
<tr>
<th>Proteins (PDB codes)</th>
<th>Sequence Lengths</th>
<th>RMSD (Å)</th>
<th>Z-score</th>
<th>Number of aligned residues</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1mol-A and 1stf-I</td>
<td>94 ; 98</td>
<td>1.9</td>
<td>10.7</td>
<td>85</td>
<td>11</td>
</tr>
<tr>
<td>1ash and 1hf</td>
<td>147 ; 142</td>
<td>2.3</td>
<td>15.1</td>
<td>132</td>
<td>14</td>
</tr>
<tr>
<td>3ink-C and 1rcb</td>
<td>121 ; 129</td>
<td>2.8</td>
<td>10.3</td>
<td>104</td>
<td>13</td>
</tr>
<tr>
<td>2btf-A and 1hpm</td>
<td>374 ; 378</td>
<td>2.9</td>
<td>25.0</td>
<td>289</td>
<td>14</td>
</tr>
</tbody>
</table>

Structurally related proteins

The selection of structurally related proteins sharing low sequence identity was performed applying some filters to the FSSP database (Holm et al., 1992; Holm & Sander, 1994b). We selected all the proteins sharing a sequence identity over equivalent positions lower than 15%, a protein length higher than 90 amino acids, a position RMSD of superimposed α-carbons lower than 3.0 Å, a Z-score higher than 10, a percentage of residues aligned over the sequence length higher than 75% and a similar length between the two related proteins. The most related couples of proteins obtained after applying these filters, using PDB identifiers, are shown in Table 2.

Energy profiles

The energy profiles were performed over the structures using the MFP and considering all the terms described on it. For each residue in the profile we used an average energy window of 11 residues, considering the five adjacent residues at the N-terminal and C-terminal extremes.

References


Edited by F. E. Cohen

(Received 10 October 1996; received in revised form 21 November 1996; accepted 13 December 1996)