THE PROTEIN FOLDING PROBLEM: FINDING A FEW MINIMUMS
IN A NEAR INFINITE SPACE

Patrick Argos and Ruben Abagyan

European Molecular Biology Laboratory
Postfach 10 22 09
Meyerhofstr. 1
69012 Heidelberg
Germany

Tel:    +49-6221-387275 (office, direct)
       +49-6221-387361 (secretary)
       +49-6221-387517 (fax)
ABSTRACT
Folding a protein from only a knowledge of its amino acid sequence is a formidable many-body problem. Since it is computationally impossible to test all possible atomic conformations to determine the global minimum representing the compact state, methods need to be developed to sample only a small part of the configurational space and yet delineate the free energy optimum (or nearly so). This article reviews such techniques.

The Problem
Proteins are known to fold from an extended-like structure to a compact entity where atoms from the mainchain backbone (a repeating N-C\(_\alpha\)-C) and the sidechains of the particular amino acid types (C\(_\alpha\)-R), constituting the sequence of the biomolecule, are in each other’s proximity, almost to the extent of hexagonally close packed (for basics see Argos (1992)). The folded protein thus resembles a cluster of electron-filled spheres with some overlap due to covalent bonding. The overall shape is roughly that of an ellipsoid or, at times, a near sphere. Some proteins have been known to accomplish this folding feat virtually on their own. All that is required is (a) the presence of a polypeptide with a specific primary sequence taken from the 20 amino acids (eg, phenylalanine, arginine, histidine etc.); (b) a test tube containing a solution that reasonably mimics the characteristics of the cellular environment; and (c) a little bit of time (Anfinsen et al., 1961). In vivo, it’s sometimes more complex, where the folding is often accomplished at the ribosome as the protein is being chemically assembled with the use of the DNA template or in an intricately constituted region of the cell or with the aid of some already folded helper molecules, the so-called chaperonins (Lorimer, 1992). The crucial information dictating the three-dimensional structure is in the protein sequence where hydrophobic amino acids cluster to form a central core and hydrophilic residues meet the aqueous solvent. Certain residue types
prefer to be in amphipathic $\alpha$-helical substructures where a consecutive set of backbone $C_\alpha$ atom (to which the sidegroups are attached) follow a spiral. Other residues like to be in $\beta$-strands which are extended; yet others are in reverse turns, usually composed of 4 consecutive amino acids in the sequence and folding along a U. The remaining mainchain path is left to the category coil with no especial pattern. These so-called secondary structural types are driven by the formation of successive hydrogen bonds where, in helices, the carbonyl oxygen of the (i+4)th residue joins the peptide nitrogen of the ith amino acid. In strands, the H-bonds occur across extended segments and, in U-turns, the mainchains of residues i and i+3 join up. For a primer in these protein structural matters, see Argos (1992) and references therein. Figures 1 and 3(b) exemplify molecular conformations in stereo stick diagrams.

The molecular biological feat of this century (and there isn't much left of it) will be the theoretical ability to predict the tertiary fold (position of every atom) of the protein from the sequence alone; after all, it's just a matter of physical and chemical forces. This would not be a problem for the theoreticians if the experimentalists could determine all the protein structures in rapid fire; but this is not the case: there are over 50,000 sequences presently known (Bairoch & Boeckmann, 1992; Barker et al., 1993) and these are complimented by only 200 or so determined structures with unique topologies (Bernstein et al., 1977). The shame of it all is that biological engineering, drug design (eg, Betz et al., 1993), and the like cannot be well understood without knowledge of the tertiary structure. So, it is a "real must" to solve the problem theoretically.

A typical protein has its share of degrees of freedom. Just consider 400 residues, each with about ten atoms (carbon, nitrogen, sulphur, oxygen) and each atom with three ($x$, $y$, $z$) co-ordinates: the total, 12,000. Even allowing for the covalent bonding
constraint, the possibilities in potential conformation seem limitless (van Gunsteren et al. (1990); Desmet et al., 1992). Just imagine the number of combinations of positions (say, each separated by a few tenths of an angstrom (10^{-8} \text{ cm})) each atom could take in making a potential fold. But there are almost invariably only a few right answers as known from the experiments: one sequence and basically one fold. To calculate all the free energies for all the possible folds is computationally out-of-the-question; even the energy terms are not always well formulated. The key (at least, the presently explored track) is to hop around this near infinite conformational space to land happily and quickly on the spot (or in its vicinity) with minimum global energy (sometimes, several minimums are possible). The topic of this little "treatise" will be to discuss ways that researchers have tried to do the "hopping". We will concentrate on the work done with colleagues in Argos' research group at EMBL but definitely not to the exclusion of the work of others.

Hopping strategies need not be tested on large polypeptides; finding good approaches is too risky to make all these calculations for many atoms. Anyway, it is almost impossible for large proteins. So we (Abagyan & Argos, 1992) chose a pentapeptide (met-enkephalin) to test conformational space sampling for \textit{ab initio} folding; i.e., from a knowledge of the sequence alone and with a computer. However, nature is sometimes kind and there are ways to tackle (at least try to tackle) bigger molecules. Sequences with 50\% or more identical residues and similar biological functions have a reasonably similar folding topology in the mainchain (Chothia & Lesk, 1986) but with the sidechains largely in different positions, especially the ones that are substituted but nonetheless equally well packed in the protein core. Since some folds are known empirically, then the conformational task is to place the sidegroups correctly in the homologous molecule using the backbone fold of the experimentally examined one. This reduces the conformational space and makes
computation feasible. We will also discuss in this treatise this so-called homology modelling approach, "homology" in the primary sequences and "modelling" for the amino acid sidechains.

**Energy Calculations**

When calculating the free energy of a protein molecule, there are many terms to consider: van der Waals, hydrogen bonding, torsion, entropic, solvation, electrostatic, etc (see (Creighton, 1993) for an introduction). The well formulated components and those that give a basic and minimum representation (Nemethy et al., 19B) include (i) van der Waals which is the sum of attraction or repulsion energy for all atom pairs, each with given spacing \( d \); (ii) hydrogen bonding which also depends on \( d \) but with different power laws; and (iii) torsional which depends on the state of the mainchain dihedral angles. It helps greatly to formulate the molecular geometry in terms of main and sidechain dihedral angles which reduces substantially the number of variables (Momany et al., 1985; Abagyan & Mazur, 1989; Mazur & Abagyan, 1989). These angles are simply the rotational status of bonded vectors between two atoms relative to atom groups on either side. For instance, the sidechain of a phenylalanine consists of a \( \beta \) atom which is bonded to the backbone \( \alpha \) on one side and to \( \gamma \) carbon of a six-membered benzene ring on the other side. Since the ring is planar and the atoms within it can be constrained to given bond distances and directions, the sidegroup can be spatially characterized by the benzene's rotational state about the \( \alpha-\beta \) and \( \beta-\gamma \) bonds. The dihedral angles of all the amino acid types have been characterised by accepted conventions (Clauwaert & Xia, 1993).

It is also useful to keep track of the lowest energies visited in any conformational search. These can be kept in "stacks" which represent a memory of minimum energies visited, then respective conditions, and within given ranges (Abagyan & Argos, 1992).
Visualisation of the visitation distribution in the multidimensional space is also important, especially when the conformational minima are known from experiment and several search protocols are under assessment. A principal component analysis can be performed over the known minimums for a particular polypeptide and then the two axes of a spatial projection containing the most information (i.e., greatest separation of the minimum points) can be determined for a handy visualisation (Abagyan & Argos, 1992). Projections can be found which contain nearly 75% of the distinguishability amongst minimums in the massively dimensional conformational space.

**Ab Initio Folding Protocols**

There are many approaches to conformational space hopping in hopes of landing upon a global minimum(s) without testing all configurational possibilities. The simplest is to select randomly the positions of all the biomolecular atoms (of course with expected stereochemical constraints such as bond distance), minimize the energy $E$ of the starting configuration, and then change randomly the variables and once again minimize, and repeat. The alteration step can be modified to improve the probability to find a minimum by (i) only accepting a new trial conformation if the preminimized energy is less than the previously determined minimum; (ii) allowing only modifications in the degrees of freedom within certain ranges relative to the last tested conformation; (iii) controlling the extent of randomization according to a temperature ($T$) factor (i.e., accept a new trial conformation if a randomly selected number between zero and one is less than $\exp\left(-\frac{\Delta E}{RT}\right)$ where $\Delta E = E_{\text{new}} - E_{\text{previous}}$); (iv) reducing the temperature according to a protocol, often referred to as simulated annealing; (v) selecting only sidechain configurations that are known to occur frequently in proteins with known structure (rotameric states); and (vi) only partially minimizing the trial structure (fewer function calls $M_{\text{calls}}$ in the iterative procedure)
before acceptance or rejection to save on computational time and allow more molecular conformations to be searched. Abagyan & Argos (1992) review the methods and provide references.

A small pentapeptide (Met-enkephalin, with sequence Tyrosin(T)-Glycine(G)-Glycine(G)-Phenylalanine(F)-Methionine(M), and referred to as Met-E) was selected for various search runs based on the previously listed protocols (Abagyan & Argos, 1992). Over the years, Met-E has been shown or suggested to occupy about 8 different basic conformations with associated energy minimums, all within a few kcal/mol of each other with the global minimum at -11.84 kcal/mol (conformation illustrated in Figure 1). The two principal axes for the projection based on these minima (and 15 more in their vicinities) was determined and found to represent 75% of the possible separation information in the larger configurational space. Various of the protocols and associated parametric ranges were tested and assessed according to the mean number of times one of the 8 minimum were found from various runs and the frequency with which the global minimum was reached. A run is defined by a starting conformation followed by various trail configurations for a particular protocol and for a set number of computer function calls; i.e., a set execution time. Different constant temperatures (1000-2530K) were tested for controlling the sampling distribution. Also attempted were various step sizes in selecting randomly the dihedral angles ($\Delta \theta = 45\text{-}180^\circ$), levels of energy minimization through iteration for a given step measured in function calls (1\text{-}150), simulated annealing cooling schedules, and rotameric states. Figure 2 shows the projected visualization of space sampling relative to the known Met-E minimums for two different temperatures; it is clear that $T=600^\circ\text{K}$ is optimal. Table I exemplifies the results for various temperatures, $\Delta \theta$ steps and function calls allowed to achieve a minimum. In each trial run from a starting configuration, the same total computing time was used. The
conclusions reached are that T=600°K is optimal with full energy minimization (150 calls) for each trial conformation and random step sizes set at the maximum (180°). Use of rotameric states and simulated annealing cooling worsened or at least did not improve the ability to discover the minima. So, with the use of a typical scientific computer workstation, it is possible to explore \textit{ab initio} the minimum energy conformation of an oligopeptide, not by sampling over the entire configurational space but by visiting only certain regions and minimizing fully a given conformation in the locale and then moving on to another region. It is not effective to sample many configurations within a locale or without near completion of the energy minimizing iterations. Since repeat visitation should not be pursued, a log of energy stacks (i.e., best minimum energy values within certain ranges and their associated conformational states) must be kept.

\textbf{Homology Modelling Approaches}

Homology modelling reduces the conformational space to a search for the optimal dihedral angles of only the sidechains. It is known that two proteins with aligned sequences that reach the 50\% or greater residue identity level have similar mainchain topology in almost all regions (Chothia \textit{et al.}, 1986) where there are no insertions or deletions, usually confined to exposed loops connecting secondary structural elements as \(\alpha\)-helices or \(\beta\)-strands. If the three-dimensional structure has been experimentally determined for one protein, then its backbone fold can be applied to the homologous protein for which no tertiary structure is known and the sidechains, substituted and otherwise, must be positioned. There are numerous methods that attempt this task; and, since structures for homologous proteins are known (Bernstein \textit{et al.}, 1977), standards-of-truth exist to check for their effectiveness. Once again, as with the \textit{ab initio} folding techniques, the goal is to reduce the search space.
Lee & Subbiah (1991) used a discrete Monte Carlo algorithm with a specific simulated annealing technique to move randomly all angles in 10° steps for sidegroups; the mean low energy conformers are selected for each substituted residue in the presence of the others. Atomic overlap, resulting from placing the low energy sidechain conformers onto the given backbone at the end of the process, can be eradicated by a minimization procedure. Tuffery et al. (1991) and Holm & Sander (1991) use only preferred sidegroup dihedral states delineated from a database of known tertiary architectures (Ponder & Richards, 1987; Schrauber et al., 1993) (generally about 5 per residue type) and calculate energies for various tested conformations through Monte Carlo and simulated annealing procedures. Computational effort can be saved by precalculating all possible rotameric pair interactions (Holm & Sander, 1991) and by reducing the energy terms used to minimize (eg, only van der Waals). Tuffery et al. (1991) chose the rotameric states with the help of the genetic algorithm. Dunbrack & Karplus (1993) go further in rotamer resolution by selecting preferred sidechain states that correspond to particular mainchain dihedral conformations (eg, α-helical). Levitt (1992) takes all possible fragments with sequence length about 4 from the target structure (most of whose Cα atoms are positionally assumed from the homologous structure) and matches them to those segments with best energy compatibility (sequence and structure) from the set of known tertiary protein structures. The segments are successively added to the model from a random sequence start site and van der Waals energy minimized with each addition. The sidechain dihedrals for the ten best overall models selected from various and many start sites are averaged and followed by a final minimization to avoid atomic clashes. Eisenmenger et al. (1993) use only the mainchain atoms from the homologous protein and calculate, using standard atom-atom potentials and a continuous search procedure, the optimal state over various dihedral angle settings for each sidegroup of the target structure. After each residue is singly tested in the presence of only the backbone atoms (including Cβ
atoms), then optimal states are combined and a final full minimization performed over all sidegroups taken simultaneously. It is noteworthy that 88% of sidegroup atom contacts are with mainchain atoms over several proteins studied. Table II shows how the sidegroup predictions, using only the backbone atoms (GAP model) or all atoms (ALL model, excluding only those atoms of the sidegroup to be predicted for a given tertiary structure) are nearly as accurate for the two models. Figure 3 illustrates the predicted tertiary structure of poplar-leaf plastocyanin, utilizing the backbone fold of green algae plastocyanin. The sequences of the two homologous proteins are identical over 60% of their amino acids (Figure 3). The typical assessment of a technique's accuracy in sidechain conformational prediction involved the use of the backbone fold, taken directly from a particular protein tertiary structure, to predict the known sidegroup configurations of the same protein (eg., Eisenmenger et al., 1993). The root-mean-square distance (rmsd) in Å between the observed and modelled sidechain atoms can be calculated from an overall spatial superposition of the chemically and structurally equivalent groups. Typically the rmsd averaged over several different protein structures is about 1.7 to 1.8 Å for nearly all the methods. Eisenmenger et al. (1993) found that the rmsd is 1.7 Å in the GAP model (relying only on the backbone atom information) while utilizing all other atoms except the sidegroup in question yielded a 1.5 Å rmsd, a relatively small improvement. Given that experimental error is probably much less than 0.5 Å in positioning most atoms, it would appear that the energy terms must be better formulated, further (preferably all) energy types must be incorporated, and/or the configurational space sampling technique must be improved. The homology problem grows worse in that the assumed fixed mainchains can shift by some fraction of an Å which grows larger as sequence similarity diverges in the two proteins considered. Obviously the substituted sidegroups can then adopt more states than are presently accounted for;
correspondingly, the combinations grow further by adding the backbone degrees of freedom. Future research is required.
ACKNOWLEDGEMENTS

The authors are very grateful to their scientific colleagues, Frank Eisenmenger and Frank Eisenhaber, who contributed essential ideas and work to the topics discussed here. Without them, this article would never have appeared.
REFERENCES


### TABLE I
Comparison of conformational search efficiency using different sampling parameterization

<table>
<thead>
<tr>
<th></th>
<th>Number of runs*</th>
<th>Performance Score†</th>
<th>Global Minimum‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>temperature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000°K</td>
<td>7</td>
<td>2.7</td>
<td>0</td>
</tr>
<tr>
<td>800°K</td>
<td>7</td>
<td>3.6</td>
<td>1</td>
</tr>
<tr>
<td>600°K</td>
<td>10</td>
<td>5.8</td>
<td>6</td>
</tr>
<tr>
<td>400°K</td>
<td>7</td>
<td>3.0</td>
<td>2</td>
</tr>
<tr>
<td>253°K</td>
<td>9</td>
<td>1.7</td>
<td>3</td>
</tr>
<tr>
<td>dihedral step size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>180°</td>
<td>10</td>
<td>5.8</td>
<td>6</td>
</tr>
<tr>
<td>126°</td>
<td>6</td>
<td>3.2</td>
<td>1</td>
</tr>
<tr>
<td>90°</td>
<td>6</td>
<td>2.5</td>
<td>1</td>
</tr>
<tr>
<td>45°</td>
<td>4</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>number of minimization calls and temperature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150, 600°K</td>
<td>10</td>
<td>5.8</td>
<td>6</td>
</tr>
<tr>
<td>25, 600°K</td>
<td>5</td>
<td>3.0</td>
<td>1</td>
</tr>
<tr>
<td>25, 1000°K</td>
<td>5</td>
<td>2.2</td>
<td>0</td>
</tr>
<tr>
<td>1, 5000°K</td>
<td>5</td>
<td>0.0</td>
<td>0</td>
</tr>
</tbody>
</table>

* A run refers to a conformational search involving $7 \times 10^5$ total function calls (or their equivalent) and proceeding from an initial starting configuration randomly chosen.
† The average number of lowest energy conformations (8 in total for Met-E) found per simulation run.
‡ The total number of times in all the runs that the lowest energy conformation was reached.
TABLE II
Sidechain $X_1$ and $X_2$ angular deviations (in degrees) of low-energy conformers
from observed values according to residue type

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>$N$</th>
<th>GAP</th>
<th>ALL</th>
<th>GAP</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine</td>
<td>14</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>17</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Valine</td>
<td>22</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>25</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Threonine</td>
<td>26</td>
<td>13</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Leucine</td>
<td>29</td>
<td>10</td>
<td>5</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Serine</td>
<td>24</td>
<td>21</td>
<td>20</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Methionine</td>
<td>9</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Asparagine</td>
<td>18</td>
<td>8</td>
<td>9</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>Glutamine</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Lysine</td>
<td>24</td>
<td>15</td>
<td>12</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Aspartic</td>
<td>23</td>
<td>13</td>
<td>13</td>
<td>40</td>
<td>31</td>
</tr>
<tr>
<td>Glutamic</td>
<td>24</td>
<td>12</td>
<td>13</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Arginine</td>
<td>17</td>
<td>12</td>
<td>11</td>
<td>17</td>
<td>15</td>
</tr>
</tbody>
</table>

Mean

<table>
<thead>
<tr>
<th>Mean $\Delta X_1$</th>
<th>Mean $\Delta X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
</tr>
</tbody>
</table>

$N$ is the total number of sampled residues for each amino acid type as observed in six determined tertiary protein structures known at high resolution (Eisenmenger et al., 1993)

$\Delta X$ denotes the mean dihedral angle absolute value difference (in degrees) between observed and predicted values. Results are not given for all amino acid types due to the small sample size. The $X_1$ angle refers to rotation about the $C\alpha-C\beta$ band and $X_2$ about $C\beta-C\phi$ band where $C\alpha$ is a mainchain atom bonded to the sidechain $C\beta$ atom in turn bonded to the sidegroup $C\phi$. Values are not given for $X_3$ and $X_4$ angles (where possible for a given residue type) due to the small sample size.
FIGURE CAPTIONS

Figure 1. Stereo view of the lowest energy conformation of the Met-enkephalin pentapeptide which is a likely fold for receptor-membrane binding (Abagyan & Argos, 1992).

Figure 2. Energy profiles and conformational space search maps for two Met-enkephalin runs at a temperature T of 600°K (top) and 253°K (bottom). On the search map projections along the two principal axes of the search space, trajectory points are given over tested molecular configurations but are not connected for clarity. The positions of the eight principal low energy conformations are indicated by triangles (Δ) while the remainder of the 23 conformers from which they are selected (Abagyan & Argos, 1992) are indicated by crosses (+). Clearly, many more regions of the conformation space are visited at the higher temperature.

Figure 3. (A) Sequence alignment of the poplar leaf (1PCY) and green alga (7PCY) plastocyanins using procedures described by Eisenmenger et al. (1993). The single letter code is used to indicate amino acid types (Argos, 1992). An asterisk points to the identically conserved residues which constitute 60% of the aligned amino acids.

(B) A stereo illustration of the modelled (broken) and observed (continuous) main and sidechains for green alga plastocyanin (databank code 7PCY; Bernstein et al. (1977)) using the mainchain trace of poplar leaf plastocyanin (databank code 1PCY; Bernstein et al. (1977)). The drawing illustrates atoms with the stick representation where an atom is located at the intersection of two lines or at the tip of a line; the sticks connect bonded atoms. Sidechains 4, 18, 43, 59, 61, 85, 93 and 99 are not shown as they could not be experimentally determined.
FIGURE 3A

10  20  30  40  50
1PCY  IDVLLGADDGSALAFVPSEFSISPGEKIVFKNNAGFPHNIVFDEDSIPSG
7PCY  AAIVKLGDDGSLAFVPNNITVGAGKESIEFINNAGFPHNIVFDEDAVPAG
*  ***  **  *  **  **********  ***  *  **********  *  *
60  70  80  90  100
1PCY  VDASKISMSEEDLLNAKGETFEVALSNKGEYSFYCSPHQGAMGKVTIVN
7PCY  VDADAISA--EDYNNSKGQTVVRKTLTPGTVGYCQPHSGAMGKMTITVQ
***  **  ***  ***  *  *  **  ***  ***  ***  ***  **

FIGURE 3B

[Diagram of molecular structures]