Chapter XX

For submission to "Protein Structure Prediction: a Bioinformatic Approach", ed. Igor Tsigelny, International University Line, La Jolla USA, 2001.

Protein structure prediction by threading: force field philosophy, approaches to alignment

Manuscript date: 31 July 2001

Thomas Huber and Andrew E. Torda*

Department of Mathematics
The University of Queensland
Brisbane Qld 4072
Australia
phone: +61-7-3365 7060
facsimile: +61-7-3365 6136
huber@maths.uq.edu.au

and
*Research School of Chemistry
Australian National University
Canberra ACT 0200
Andrew.Torda@anu.edu.au

Introduction

If you are given a protein's sequence, you might have all the information you need to predict its structure. You have the composition and (bond) topology of the system, so you only have to rearrange its atoms so they are somewhere in the major free energy basin and the problem is solved. There might be some problems with this approach. The search space grows exponentially with the number of particles. If you are able to search the conformational space for a five residue peptide this year, it might be another year or two until you can tackle six residues when your computer is several fold faster. Then, you have to have an energy or score function which really can discriminate between correct and incorrect conformations. Any score function which is fast enough to apply to more than a few hundred atoms will be full of approximations and no longer close to the best level of theory. It is also worth remembering that we are only assuming that the native protein conformation really is a free energy minimum and that some native conformations may only be of low energy because of prosthetic groups or unusual interactions (with ligands, ions or other proteins). Finally, if we believe that proteins find a free energy minimum, then we should remember that while potential energy is a property of a conformation, free energy is a function of many conformations. Considering all these problems, it would be fair to say that the history of protein structure prediction is one of approximation, optimism and cunning heuristics.

Our particular interest is in the set of approximations and heuristics that underpin the methodology or set of methods known as protein threading. Specifically, we are most interested in a class of score functions built for this application, how they may be constructed to ease the conformational search problem and any other devices which we can employ within a threading framework.

Protein threading grew out of the observation that often when a protein structure is solved, it is remarkably similar to one already in the protein data bank - even when it would not be expected from sequence similarity. Perhaps, it was reasoned, it would be a major advance if you could take a sequence of interest and just find the most compatible structure from the protein data bank. Looking back, it seems that the threading is the child of two camps:

1. The biologist's approach to structure prediction:

   by comparison and induction - if sequence1 is similar to sequence2 then structure1 is similar to structure2 and there is probably an evolutionary explanation

2. The physicist's approach to structure prediction

   proteins form structures according to fundamental rules which we call energies or free energies

It is clear why the schools merged. Biologists learnt that protein structures are much more conserved than protein sequences, so comparing sequences is not sufficient. Physicist's, on the other hand, saw that the set of known conformations is much smaller than the set of all possible conformations. Threading as we know it now contains elements from both camps. From the physicists side come elements such as low-resolution or coarse-grained force fields. Classical biology / bioinformatics has provided techniques such as sequence to structure alignment methods based on dynamic programming.

Common methodology

For the rest of this chapter, we repeatedly refer to threading and its properties. We assume several elements:

- One has a sequence whose structure is to be predicted and this sequence does not have significant homology to anything of known structure, otherwise the predictor would probably be better off with some other method.

- There is a library of known structures / templates. The sequence will be aligned to and tested against each in turn. The original sequence of the template is known, but may not be used.

- There is some kind of score function or force field which is capable of returning the happiness of a sequence residue at any position on a template

- One has a method for calculating sequence to structure alignments with gaps. This must be able to handle gaps and insertions. The most common methods are dynamic programming and Monte Carlo.

Now, some points need some expanding. There are several ways to align protein sequences to structures (templates). We describe one diagrammatically. Figure 1 shows a sequence whose structure we do not know at the start (left of diagram) and some template structure from the library (right of diagram). This template is not exactly the same shape or size as the unknown answer, but it is the best library member. The library has hundreds or thousands such templates, most of which are completely wrong for the sequence. The computational problem is summarised in Figure 2. On the one side we have the sequence (unknown structure) and on the right, the template. We then construct a matrix which gives the score that every sequence residue would have if it were placed at any numbered position from the template (Figure 3). A dynamic programming method is used to find the path through the matrix which preserves sequence ordering, while allowing gaps and insertions. In this example, there is a gap, omitting residues 7, 8 and 9 from the template. Having calculated the

sequence to structure alignment, we now have a prediction for the sequence shown in Figure 3. It is not a perfect guess (left hand side), but it is as good as one can do using the available template.

With this background, we can be more specific about the contents of this chapter. We are interested in methods which use score functions which look (mathematically) like force fields, although they may not model real physics very closely. We are also interested in the approximations and methods one uses to find a sequence to structure alignment. Lastly, we discuss some of the methods which could be called elegant heuristics or computational tricks, but seem particularly tied to protein threading.

Force field based scoring

As described above, one is going to need a function which can score a sequence residue at a position on a template. For the moment, we concentrate on what could be called pair-wise, through space interaction functions. By this, we mean the situation shown in Figure 4. We want the score associated with a residue "A" at the position on the template and it will be calculated by considering interactions with its neighbours through space as shown by the arrows. This kind of problem is not unique to threading, but has been at the heart of every method to model or simulate molecules. Before considering threading specific score functions, they can be placed in context by considering the history of through-space force fields in general.

The difference between methods lies in the choice of elementary unit (i.e. particle with no explicit internal degrees of freedom), how an environment is incorporated, what empirical functional forms are used and how they are parameterised. In purely ab initio methods, for example, elementary units are wave functions of electrons from the isolated system in vacuo, and the only "empirical parameters" are fundamental constants, such as Plancks constant. Semi-empirical ab inito methods still work at an electronic resolution, but use parameterised integrals to reduce the computations. A further step of approximations in the representation is made in molecular mechanics force fields. The smallest entity is an atom and electronic degrees of freedom are (in a time-average sense) implicitly considered in empirically parameterised atomic interaction functions. After omitting electronic detail, there is a significant gain in computational simplicity and it often becomes feasible to include the local environment explicitly in calculations.

An atomic representation is intuitively appealing for modelling molecular properties, but a much lower resolution description is required to simulate macroscopic behaviour of complex systems.[1] This leads to the concept of smooth particle simulation, in which the properties of a mesoscopic volume element of a system are captured by a single particle-like object. Just like in a molecular mechanics

force field, inter-particle interactions are then based on physical concepts and their parameters are fine tuned empirically.

Regardless of these details, whether a method uses a sub-atomistic or a mesoscopic description, it will contain approximations of some kind. The best method for a given application domain is the result of a trade-off between computational expense and performance.

In the particular application domain of modelling the overall structure of proteins, the representation of an amino acid residue by a single (or few) interaction sites has become popular for several reasons. Firstly, there is a rationalisation based on beliefs of protein folding. It is often believed that a protein chain collapses to its correct fold before an annealing of side chain conformations forms the exact native structure. If this is true, then it may not be necessary to model the details of side chain interactions and it could be adequate to treat them in a mean field manner. Secondly, there is a purely pragmatic justification. Omitting side chain degrees of freedom greatly simplifies calculations. With this simplification, however, one is faced with the problem of finding a score or potential energy functions which summarises mean-field, many-body interactions from amino acids into a few pair-wise interactions.

Given an idea of what level of force field resolution one wants, what methods are available to build and parameterise force fields ?

Parameterising force fields

It is useful to consider three kinds of force field:

1. Physically-based potential energies
2. Potentials of mean force
3. Optimised force fields

These are discussed in turn.

Physically-based potential energies

By physically-based, we mean force fields which try to mimic the true physics of a system. If we follow physics, the answers will be correct and the methods will be transferable from system to system. At the risk of being too simple, consider a score function with atom-atom bonds. We can find bond lengths from X-ray crystallography, model the bond as a harmonic spring and say the energy for a single bond between particles at positions $\vec{r}_i$ and $\vec{r}_j$ is given by

$$V^{bond}\left(\vec{r}_i,\vec{r}_j\right)=k_{ij}^{bond}\left(\left|\vec{r}_i-\vec{r}_j\right|-r_0\right)^2 \quad (1)$$

where k_{ij}^{bond} is the spring constant appropriately chosen for the bond between particles *i* and *j*, and r_0 is the ideal bond length. To find the energy of the system as a whole, we sum over all V^{bond} terms for each bonded *ij* pair. Similarly, if we know the partial charges on atoms, we could just use Coulombs law to calculate the electrostatic energy. Continuing in this vein, one could add in Lennard-Jones terms, bond angles, maybe dihedral angles and come up with a full force field, suitable for an application such as molecular dynamics calculations or energy minimisation.[2-4]

Because we believe in physics, one must ask why this kind of force field is not more often used in protein threading. Firstly, there has been a widespread belief that these empirical, atomistic force fields are not very good at discriminating between correct and incorrect models for a protein whenever the incorrect model is basically well folded and reasonably packed.[5] This has lead some groups to incorporate extra pair-wise terms to at least account for solvation.[6-9] Next, there is an issue of computational expense. Calculating the energy of a protein conformation is fast, but sequence to structure alignment calculations can involve huge numbers of these calculations. There is also the issue of potential versus free energy. Classical force fields are designed to yield potential energies, but in the introduction, we stated that proteins are more interested in free energy minima. One could search for approximations to entropic effects, but there will be fundamental limitations as to how well a single conformation can approximate an ensemble property such as free energy. [10] Lastly, there is a very good reason these force fields will always be problematic. To work at atomic resolution, one must know where all the atoms are. In protein sequence to structure threading and alignment, one usually does not know the location of the side-chain of any residue, let alone the neighbours it interacts with. This is discussed at length under sequence to structure alignments.

Potentials of mean force

Potential of mean force are certainly the most popular kind of interaction function for protein threading and come from, literally, textbook statistical mechanics.[11] Basically, the philosophy is that we can look at some property of a system, like the typical distance between two particles. If the particles are always close to each other, we might think there is some attractive force between them. If some other pair of particles are never near each other, we might think that they repel each other. This can be formalised by considering the radial distribution function, *g(r),* which really tells you the ratio of the probability seeing two particles at distance *r* and what one would expect from random placement. Then, one simply follows the Boltzmann relation to say

$$A(r) = -k_B T \ln(g(r)) \quad (2)$$

where *A(r)* is a potential of mean force, k_b is Boltzmanns constant and *T* is the temperature. The result of this calculation is usually a tabulated interaction function. Most implementations collect observations in distance (*r*) bins and have corresponding values for *A(r)* at each discrete distance. Philosophically, the results of the method are more interesting. Obviously, there are entropic contributions, but as the name implies, there are mean contributions from every possible source. For example, the interaction between two particles may be physically very influenced by solvent. In this formulation, the solvent (along with other contributions) is present in an average sense.

Potentials of mean force do not rely on any particular representation of the system. In the statistical mechanical literature, they are usually atomistic, but there is nothing to stop someone collecting data at the level of whole amino acids. This is exactly what Tanaka and Scheraga did more than 25 years ago.[12] They treated a set of protein structures as if it were a statistical mechanical ensemble and extracted interaction functions between whole residues in proteins. The philosophy was applied to a larger set of proteins nearly a decade later[13] and by the 1990's there were many more implementations of the principle.[14-16]

There are many variations on the formulations for the collection of potentials of mean force and they are covered in this volume by authors that use them. While they are the most popular, they are not totally without critics.[17,18] Fundamentally, protein structures from the protein data bank are not strictly a real statistical mechanical ensemble, although many would argue that they are a good enough approximation. Furthermore, free energies are definitely not additive quantities.[19] Perhaps, one can avoid this debate by considering the philosophy of the next section.

Optimised force fields

Force fields based on physics have useful properties. Because they are in real units, they can be compared against properties that depend on time or temperature. Sometimes, however, one does not care about kinetic or energetic effects. For protein structure prediction, you only need a function which takes a sequence and structure and returns the best score when the coordinates are correct (native). There is absolutely no need for this function to work in conventional energy units or for its values to scale to energy in any way. This has led many groups to pursue a direction which is not bound by conventional physics. Can one build a score function which is simply able to tell good from bad structures or reliably pick the best candidate from a set of trial configurations ? You could call this a fold recognition, sequence-structure compatibility or discrimination function. Under duress, it could even be called a quasi-energy function or force field. To build this kind of score function, you

do need (a) a set of native proteins (sequences with their correct structures) and
(b) for each native protein structure, some set of non-native, decoy conformations
(c) some set of energy / score functions with some adjustable parameters.

Now, we consider that class of score functions which are built by optimising their score parameters, rather than by following any rules from physics or statistical mechanics. As an example, consider the approach suggested more than a decade ago in various forms by Crippen and co-workers.[20-22] First, they picked some general form for interactions. This may have been of an almost Lennard-Jones form like

$$V(r_{ij}, a_{ij}, b_{ij}) = a_{ij} r_{ij}^{-12} + b_{ij} r_{ij}^{-10} \tag{3}$$

where r_{ij} refers to the distance between some point on each residue, i and j and a_{ij} and b_{ij} are some adjustable parameters depending on the types of residues i and j. The force field may typically contain tens or hundreds of these a and b parameters. Assuming additive interactions, one can make the obvious summation over all pairs ij to get a total energy

$$V^{tot}\left(\vec{r}, \vec{a}, \vec{b}\right) = \sum_{j>i} V\left(r_{ij}, a_{ij}, b_{ij}\right) \tag{4}$$

where we note that the final energy is a function of all parameter values a and b as well as coordinates. Crippen and co-workers then asked whether they could adjust the various a and b values of eq. 3 so that the energy of a native protein is always lower than that of a decoy conformation. If we think of a and b as variables instead of parameters, we could say that we want to solve an inequality

$$V_{nat}^{tot}(\vec{r}, \vec{a}, \vec{b}) < V_{dec}^{tot}(\vec{r}, \vec{a}, \vec{b}) \tag{5}$$

with respect to a and b. Of course, there would be a very large set of inequalities to be solved, considering one protein in the parameterisation set and its decoy structures, then every protein in the parameterisation set and its corresponding decoys.

In this description, there has been no discussion of what the exact form of eq. (3) should be, nor details such as what atoms (interaction centres) are used for calculating the r_{ij} values. These are just examples to make the point that this kind of score function can be a very artificial creation. It may be possible to explain the final a and b values in eq. 3 in physical terms, but this would be a rationalisation after the fact. It is not implicit in the procedure.

Continuing in this vein, there is now a volume of work along these lines which we can try to view in some unified way. Firstly, assume there is a set of native structures, each with decoys. These will not be changed, so we do not explicitly write the vectors *r*, of coordinates. Similarly, we will not bother to label total energies *tot* since energies are assumed to be totals over a protein. Instead, we will consider energies as functions of parameters and use terms like $V_{nat}(\vec{p})$ and $V_{dec}(\vec{p})$ for native and decoy energies respectively. If a protein has N_{dec} decoy structures, then perhaps

$$c(\vec{p}) = \sum_{i=1}^{N_{dec}} V_{nat}(\vec{p}) - V_{dec}^{i}(\vec{p}) \tag{6}$$

would be a useful cost function to minimise. To work on many proteins, we should, of course, have

$$C(\vec{p}) = \sum_{j=1}^{N_{nat}} c_j(\vec{p}) \tag{7}$$

where we sum the cost function over N_{nat} native proteins. Now, it appears that one has the ingredients for building a force field. For some set of interaction functions, one wants to minimise eq. 7 by adjusting the parameter vector, $\vec{p}$. One way to approach this is to see that for many types of interaction function, including those of eq. 3, one can take the partial derivatives with respect to parameters. This could lead to using a minimiser such as steepest descents or conjugate gradients, but perhaps there are multiple minima on the cost function surface. In that case, it would be better to use an even more powerful minimiser such as simulated annealing.[23] This approach was taken to extremes in one piece of work which borrowed a method from density functional theory.[24] The parameters were given fictitious masses and velocities in parameter space. Then, by analogy with molecular dynamics, a function like eq. 7 was treated as if it were a potential energy and parameter dynamics was used in parameter space.[25]

This approach is entertaining, but perhaps somewhat intractable. The quasi-forces experienced by a parameter depend on a potentially huge number of native protein conformations, decoy conformations and all the interactions within each structure. A better method would be to try to gather the relevant properties of native and decoy structures into a precalculated set of properties. Consider the energy of misfolded decoy structures for some sequence. Under certain conditions, their energy will follow a normal distribution as shown in Figure 5. This is characterised by the mean, $\langle E_{dec} \rangle$ and standard deviation, σ. We want a score function which gives an energy for the native structure that is much lower than the mean energy of the decoy structures. One could maximise the difference by simply scaling the energies, but this would not be very helpful. Instead, one wants to

adjust parameters so as to move the energy of the native structure to the left of the picture, while simultaneously minimising the standard deviation. This can be seen as optimising a general statistical property, the z-score which just measures how many standard deviations an observation is from the mean. In the case of building score functions, we are not interested in optimising the z-score for one protein, but in optimising the z-score for many proteins simultaneously. Typical numbers would be 300-400 native structures and 10^7 total misfolded decoys.[26]

Z-score optimised force fields appeared from several groups, almost simultaneously, with differences in functional forms and implementation.[26-29] Compared to other methods for optimising force fields, the methodology has several attractive features. Firstly, it can be quite fast. Many properties depending on the coordinates can be calculated before any minimisation is applied. Next, the approach is, in principle, capable of finding something which could be called the best possible score function. The parameters should be sent to the best possible values without the restriction of following some assumed distribution.

There are now many score functions in some way for protein fold recognition although they differ widely in what the type of interaction function they use and the quantity they optimise[30-35]. Some are based on optimising a penalty function and some on constraint satisfaction. Interaction functions include Lennard-Jones like interactions,[20,22,25] tabulated distance-dependent terms[32,36], sigmoidal contact functions[21,26] and combinations of very general, almost polynomial basis functions.[37] The score functions also differ in their level of detail (one or more than one site per residue), where they place the interactions (backbone C^{α}, sidechain centre of mass, C^{β}) and how they treat distance along the protein backbone. For example, should one parameterise interactions between residues very near in the sequence separately from those with many intervening residues ?

It is certain that not all of these score functions are equally good, but it is not possible to say which are best. There is no agreed measure for testing, nor consensus as to what they should be able to do. Most workers would like to be able to recognise a structural homologue for a protein given only its sequence, but there is no accepted definition of structural homologue. It could be defined in terms of structural difference and the amount of structural overlap. Furthermore, score functions are rarely compared in isolation. The results that a group observes will depend on the alignment or other testing method they use.

Rather than say what the best score function is, one might evade the question by saying there is probably no ideal force field. Different formulations will probably perform best on different problem domains.[37] For example, a force field may be trained on a set of native structures and decoys where the decoys can be quite close to the native structure. This could be too difficult, so one may ignore

decoys if they are similar to the native[25] or be more sophisticated and ask that the energy function be sensitive to just how similar a decoy is to the native.[21] One could simply ignore the issue and hope that if a few decoys are similar to native structures, then they disappear in the statistical noise.[26]

Ultimately, nobody knows the limits of this kind of approach to parameterising force fields. If people cannot reliably predict protein structure, it could be because they have used poor interaction functions, they have optimised the wrong penalty function or satisfied the wrong constraints. Most ominously, it has even been shown that some forms of interaction function will never be able to distinguish native structures from certain decoys ![38]

Alignment philosophy

Common alignment and score methods

In the introduction, we stated that one needs a means to find sequence to structure alignments. In its most general form, this problem is surprisingly difficult and is actually NP-complete.[39] This means one can probably say that there is no deterministic method, guaranteed to find the best alignment is reasonable time. Instead, there is a selection of approximations in the literature.

Although we described the problem in terms of a score matrix in Figure 2, there are other heuristics one could try. One could place sequence residues on the template, take random steps and apply a conventional Monte Carlo / simulated annealing method to find a good alignment.[40] One could just as well use a genetic algorithm.[41,42] There are two reasons we do not use these methods. First, in such a complex search space, one cannot guarantee finding the optimal solution. Second, making Monte Carlo tractable requires putting restrictions on the placement of gaps and insertions. These methods may work well enough in practice, but in the pursuit of elegance we would prefer a method which can put a gap or insertion of any length at any position.

The second major class of alignment methods relies on dynamic programming. It is deterministic and guaranteed to find the best possible alignment for the problem as posed. The price however, is that restrictions are placed on the scoring. The most popular methods have their roots in sequence comparison and can only use what one would call a single body term. Through space scoring functions require a scheme as in Figure 4, but this cannot be used. In the figure, we know where the residue of type "A" is, but not its neighbours. They have not yet been aligned. One approximation is a two-level dynamic programming method, but it is computationally very expensive.[16] Alternatively, one could remember that we do know the residues which were present on the original template structure. In the diagram, we could label the neighbours of "A" with their residue types from the

original template and then calculate a score directly.[15,43,44] The quality of this so-called frozen approximation is obviously good in the case of homologous proteins with high sequence similarity. How good this approximation is for proteins with very distant homology or for orthologous proteins is, however, debatable. In the next section, we describe a different approximation which avoids the problem of sequence memory of the template.

A new class of methods has been proposed which are closer in spirit to branch and bound algorithms.[45-47] Despite their elegance and potential power, they have not become popular, probably due to difficulty in implementation.

Sausage alignments

Our code, travelling under the name of sausage, attempts to operate at the best possible trade-off between scoring and searching methods, while simultaneously remaining as simple as possible.[48,49] The approach is to split the prediction into two very separate steps of sequence to structure alignments and ranking the models that are produced. These are not just conceptually separate, but usually done with different force fields and different gap penalties. There are distinct advantages to the approach. The tasks of aligning sequences and ranking models are fundamentally different and the best score function for one may not be the best for the other. Significantly, it is easy to fine tune the details for each step and each task can be optimised independent of the other.

The most unusual feature of sausage is a force field approximation in the first step which allows an optimal alignment to be calculated in polynomial time. The approximation is only used for alignments and final ranking scores are calculated with a force field with full pair-wise interactions. The aim is to find a function which allows scoring a single residue at any position on a template as shown in Figure 4. Looking at the picture, we know the location of the residue's neighbours, but not their identity. This leads naturally to building a special score function which uses the identity and coordinates of one particle ("A" in the diagram), but only the coordinates of its neighbours. In other words, the particle interacts with neighbours with some kind of average particle type.

For example, a conventional neighbour specific score function for three amino acid types would have parameters for pairs AA, AB, AC, BB, BC, CC. In a neighbour non-specific (alignment) score function there are only 3 parameters for pairs AX, BX, CX, where X is a generic amino acid type. The X residue is conceptually an average amino acid, but its parameters result from numerical optimisation and not averaging over an existing score function. Despite the rather drastic approximation of treating all neighbours uniformly, the method works remarkably well. It is

particularly attractive in the case when one thinks of proteins which adopt similar structures due to convergent evolution and may have no similarity at the amino acid level.

Beyond pair-wise terms

Most threading force fields are dominated by pair-wise terms and approximations to solvation terms which one could call single-body terms. Maybe these score functions get incrementally better every year, but maybe they will never be sufficient to recognise protein folds.[38] Given the startling array of functional forms listed above, there is no reason to think that any group has found the correct way to encode certain kinds of information. For example, we know that there are statistical propensities for small stretches of amino acids to prefer certain secondary structures. Looking at typical interaction functions formulated in terms of distances, is there any reason to believe that they will (statistically) drive backbone angles to the correct conformations ? Perhaps it would be better to find some other way of encoding secondary structure preferences into a protein prediction scheme. Similarly, some of the most adept fold recognition methods do not use any pair-wise, distance dependent interaction terms at all. Hidden Markov[50-58] models are well described in this volume and they, along with psi-blast[59] work entirely on proteins which have been reduced to one-dimensional strings. Perhaps a computational chef will find a delicate blend of terms from different fields which most economically captures available information. For the moment, we consider a bucket-chemist's approach of throwing terms together and stirring with optimism. Specifically, what issues arise when one tries to add different kinds of information together ?

First, consider the case of sequence similarity and the toy example of Figure 6. We know our sequence of interest, but we also know the original sequence from the template. Looking at the example, it is easy to calculate a sequence similarity score by looking up the AH element in an amino acid substitution matrix, then the DK element and so on. We now suggest that this can be added into an existing, pair-wise, through-space score function. This is never going to be an elegant process. If we use the language of force fields, then our score function is returning some kind of energy and any term we add is also an energy. To do this, is as outrageous as saying you can measure sequence similarity in the same units as steric overlap.

Aside from this offence against scientific aesthetics, there are technical reasons to be careful. When adding terms together, we would like them to have the property we might call orthogonality. That is, a through-space score function term should provide different and independent information to something like a sequence comparison term. We can say, in advance, that this is certainly not the case and make the point with a simple example. Most through-space, pair-wise score functions encode something like surface exposure / burial preferences. Amino acid substitution matrices encode the

hydrophobicity by saying that similar residues can be swapped for each other. Clearly, both the bare score function and a substitution matrix are going to contain some similar information and the point could have been made with other properties such as size, aromaticity, or polarity. Now, we do not know in advance the extent of the overlap or independence, so there is little guide as to how the different components should be combined. One idea is simple empiricism or trial and error.[60,61]

An example of this trial and error was proposed by Panchenko et al who added a sequence profile-based similarity term to a Boltzmann-based, pair-wise set of interaction functions.[61] In order to scale the relative contributions, they simply titrated the sequence term against the rest of the score function, measuring success across a set of test proteins. When mixing terms, one can do better than this. If one has a rapid way of scoring force field quality, we can use straightforward numerical optimisation to find the weight given to the sequence similarity term. This can be shown by example.[62] We can define a sequence similarity score for a site simply by the score from a BLOSUM62 amino acid substitution matrix.[63] Next, we construct a measure of alignment quality based on 572 pairs of proteins with structural similarity, but no significant sequence identity.[64] Basically, we align the sequence of one member of each pair to the structure of the other and use a measure of structure quality to judge the alignment. This many alignments can be calculated quickly and a simplex optimisation method can be used to adjust the weight on the sequence similarity term. Figure 7 shows the result of this kind of calculation. One should define the parameterisation set and quality function formally, but the figure is sufficient to make several points.

Most clearly, the performance of the bare force field (left hand side of plot) is improved by adding a sequence similarity term. This suggests that we are adding information which was not encoded in the original score function. At the same time, the term cannot be too high, otherwise performance decreases again. We deliberately do not give any more details on this calculation, because the exact result will be very force field specific. From the point of view of force field construction, there are other lessons to be learnt. If we had a proper, self-consistent approach, the force field would have been optimised with the sequence similarity term built in and not added as an accessory at the end. At the same time, it could be that there is no single ideal weight for the extra term. If a sequence has significant sequence similarity to a template, then this term should be weighted very high (this is the domain where sequence alignments are reliable). As similarity between sequence and template decreases, the term will add noise and its weight should be gradually decreased. To this end, we note that one popular threading code already incorporates a sequence term which is switched on or off, depending on the degree of sequence/template similarity.[65]

Another source of extra information is the secondary structure known or predicted for a sequence. This should not be necessary since a perfect force field would naturally prefer the correct secondary

structure for a residue. In practice, it appears that no such perfect score function exists so it is common to bias a score function with the output from some secondary structure prediction method.[66] The reason is probably a matter of how the encoding is performed. Popular and successful methods for predicting secondary structure are usually based on neural networks which consider a window of residues centred on the site to be predicted.[67-72] Through-space score functions have not, generally, been parameterised in these terms.

If one is going to include secondary structure information, there are several forms one might try. We think of secondary structure in terms of backbone φ, and ϕ angles, but these could be used to put restraints on some interatomic distances. One could try simply matching the type (α-helix, β-sheet) predicted for the sequence to the type observed in the structure and constructing a score term which rewards a correct match. In our experience, neither of these approaches work well with real structures.[62] α-helices are easy to encode in terms of specific inter-residue distances, but β-sheets are not (the neighbour is not specified). Simple switching functions have been similarly unsuccessful since they require some threshold for recognising a type of secondary structure, but backbone angles from real structures too often sit near the borders of classic secondary structure ranges. A far better approach is to use a continuous function which gently rewards a residue for agreeing with a prediction. In our code, we have used[73]

$$V^{\text{sec}}(\psi) = -\cos(\psi_0 - \psi) - 1 \qquad (8)$$

where ψ is the backbone angle at the site on the template and ψ_0 is the literature value for the predicted secondary structure for the residue (ψ_0=-47° for α-helix and ψ_0=124° for β-sheet). One probably should use an extra weighting depending on the confidence in the secondary structure assignment, but in our implementation, we have used only high confidence predictions and added a single weighting coefficient to balance this term against the rest of the scoring function. As with the sequence similarity term described above, we have used numerical optimisation to balance this term against the rest of the scoring function and find a similar pattern. First, we do find a significant improvement in alignment quality after incorporating predictions from a popular server.[74] Second, this term should not be weighted too high, otherwise performance drops. While predictions are sometimes wrong, they do provide information which was not present in the original score function.

Given the utility of other forms of information, it would seem that protein fold recognition tools should be able to incorporate more terms from predictions or experiment. Correlated mutations may provide information about residues near in space.[75,76] Perhaps predictions of solvent accessibility are better than the hydrophobic preferences in current score functions.[77] Probably it is more exciting to

consider what can be done to take advantage of experimental data. If an experiment is going to produce a reliable structure, it might be best to avoid theoretical speculation. If, however, an experiment provides sparse and perhaps even noisy information, then it may be very valuable to combine it with protein fold recognition / threading. Nuclear magnetic resonance (NMR) spectroscopy is a prime example. Although NMR is known for producing final structures, there is a wealth of data which is never used to produce a final structure. Even if a protein is too large for a structure determination, the early resonance assignments may be enough to predict secondary structure at more than 90 % reliability, far in excess of the best prediction methods.[78-81] It was shown some years ago with early fold recognition codes that even a relatively small amount of reliable secondary structure information from NMR can make an enormous difference to the reliability of fold recognition.[73]

Certain kinds of experiments can also yield sparse distance information. Conventional protein chemistry may locate disulfide bonds and cross-linking combined with advances in protein mass-spectrometry will put limits on the distance between sites in proteins.[82] Currently, the challenge is to computationally take advantage of this data. It is easy to filter proteins as to whether they agree with distances, but it may be harder to use the information directly. Young et al checked threading predictions for agreement with mass-spectrometry data, but there was no attempt to incorporate the information into sequence to structure alignments.[83]

Template libraries

Throughout this chapter we have spoken of aligning a sequence to a template structure from a library. The library itself will have a distinct impact on the success of a method. It must include all possible candidate structures and maybe it should avoid duplication of very similar proteins. This requires some sampling or clustering method which can be based on either sequences or structures. This is not a simple process since there is tremendous redundancy (there are now more than 400 variations on T4 lysozyme) and the selection should be biased towards high quality structures. Furthermore, a particular threading implementation may work best if the library is biased to prefer larger or smaller structures. At the more detailed level, there are possible improvements to be made to the actual coordinates or representations used in the library. Here we consider two possibilities. First, can one average over structures and second, can one use numerical optimisation of the structures so as to make them more suitable for recognition by a sequence ?

Despite the use of coarse-grained models and simple interaction functions, a major problem with threading scores is their sensitivity to small changes in structure and/or sequence. It is very easy to

recognise the correct structure for a sequence. It is usually difficult to recognise a structure which looks similar, by eye, to the correct one, but which differs in many details. In the case of sequence changes, the problem has been well studied. Basically, one should perform calculations on sets of aligned sequences whenever possible. Taking advantage of structural similarities is much more challenging and has proved less popular.

There are good technical reasons for this. First, it is not clear what constitutes a similar structure. Should one require that two proteins are superimposable to within a coordinate RMSD of 3 Å for 90 % of their residues or should one say that a helix-strand-helix motif provides a similar environment to the residues within ? Next, there is the problem of using aligned three-dimensional structures. Averaging Cartesian coordinates results in structures with disturbed bonds, while averaging in internal angles or even coordinates in transformed space (such as Fourier space routinely used in X-ray crystallography) results in similar problems. One way around the problem is to largely ignore gapped alignments. Finkelstein et al have shown several times that one can average scores from different structures, but have not managed to put this in the context of sequence to structure alignments using dynamic programming.[84-86] In order to tackle this problem, we have taken a somewhat indirect approach and asked if we can average environments within a protein. First, we have taken protein structural alignments from the literature[87-90] and from each set of aligned structures, declared one to be the parent or representative protein. For each site on this protein, one can calculate the score that a residue of each of the 20 types would experience. One can then look at corresponding sites on other proteins from the structural alignments and calculate the analogous scores. These can now be directly averaged for each type of amino acid. The approach requires a score function which can be calculated for a sequence residue at an arbitrary position on a template without knowing the final alignment, but this is straightforward. One can either use a neighbour non-specific score function[91] described above, or even the frozen approximation favoured by other groups.[15,43,44]

As an informal example, we can show some results for alignment quality, using the same quality function and test set as in Figure 7. Figure 8 shows the results using the bare, z-score optimised force field, the same force field with structure averaging and the same force field, but using a simple averaging over sequences aligned by blast[92] searches. Structure averaging provides a clear improvement in the quality of models produced (published elsewhere). In this specific implementation, the benefit is not as much as with multiple sequences. This might be a general finding or it may simply reflect more experience setting thresholds for multiple sequence alignments.

The second idea for manipulating templates in a library is to change the protein structures so as to support the scoring function in discriminating a family of native sequences from these same sequences aligned on any other structures. Conceptually, the idea is very simple. One could, for example, have a library of two different proteins with (families of) sequences A1, A2 and their corresponding structures S1, S2. What one would like to do is to change the coordinates of the first structure S1 in order to maximise the score of sequence(s) A1 on structure S1 and minimise the score of sequence(s) A1 aligned on the other structure S2, while simultaneously performing a similar optimisation for the second sequence(s) A2. One then could think of optimising the two structures independent from each other, trying to increase the gap between sequences A2 and A1 on structure S1, and the gap between A1 and A2 on structure S2.

We have performed such structure optimisations for a complete fold library of 893 non-redundant structures. For each protein, a psi-blast search was performed to find a family of up to 20 homologous sequences with less than 95% sequence identity to each other. To speed up otherwise intractable calculations, not all 893 aligned sequences, but only the 20 best scoring alignments from structurally unrelated proteins were included as the unwanted negative cases in the calculation. During the optimisation it was necessary to include other additional scoring terms. The sausage score function is based on a simple smoothed contact term and was parameterised for fold recognition only. This means that it lacks conventional force field terms which prevent residues overlapping or adopting conformations not found in proteins. To repair this deficiency, a molecular mechanics (GROMOS[2]) force field term for a generic poly-Ala chain was used to keep structures protein-like. Furthermore, a harmonic restraint was used to keep proteins within 1 Å of their native coordinates. These terms were brought together into an energy-like score which was then minimised with respect to structure coordinates using 100 steps of quasi-Newton minimisation followed by 500 steps of momentum biased (molecular dynamics like) optimisation. During this procedure, the gap between homologous sequences and alternative sequences generally widens, as one would have hoped from the design of the calculation.

The effect of the optimisation is demonstrated in Figure 9 for the example of lyzozyme (pdb code 153l). The left side of the figure shows the energy spectrum of sequences placed on the native X-ray structure before the optimisation. The green bars correspond to 20 homologous sequences, whereas the red bars indicate the 20 best scoring alignments of structurally unrelated proteins. The right side of Figure 9 shows the recalculated energy spectrum of the same sequences on the optimised structure. Clearly, our aim was achieved and now homologous sequences are well separated from alternative sequences. In our experience, fold recognition with the optimised library generally performs better than using a library of X-ray structures. This could be due to regularising structures from the protein

databank, since they do contain errors,[93] and the molecular mechanics term will help remove these. Much more significantly, the optimisation intensifies features particular to each fold, at least in terms of this score function, while simultaneously deprecating common features. Furthermore, the coordinates can even adjust to compensate for weaknesses in the score function such as cut-off effects. In this scheme there is a mutual dependence of the scoring function (which is derived from structures) and the optimised library (optimised with respect to some quality of the scoring function). It would therefore be desirable to bring both parts into harmony by iterating the procedure to self-consistency. Until now we have limited the optimisation to only one cycle due to the high computational costs of force field generation and library optimisation. In future we will build a new fold library and force field which will be iteratively refined to self-consistency.

So far, this description has overlooked a flaw in the structure optimisation. There is no term which enforces absolute values for structures. Imagine that, during optimisation, the gap between homologous and alternate scores increases as it should. Within the framework, there is nothing to stop the absolute set of scores on a structure shifting together. This could mean that some structure ends up as very preferred by its native sequence, but also preferred by every other sequence. So far, this phenomenon has never been seen and we mainly see changes in the energy/score gaps between correct / incorrect sequences. Ultimately, it should be possible to use a scheme where structures are optimised simultaneously and connected to each other. This will be even more computationally expensive, but will guarantee that, for all homologous sequences, the native structure is higher in score than the same sequence aligned to any other structure.

Further outlook and speculation

The most remarkable feature of protein threading is its popularity and immaturity. By this, we mean the number of publications, often proposing new methods. This is quite different to other fields. Consider protein sequence analysis where there is a number of accepted methods for fast or accurate alignment and the statistics are relatively well understood. Score matrices do not change much from year to year. Consider molecular dynamics (MD) simulations. They are longer every year and there are regular proposals for improvements to terms such as electrostatics or treatment of solvent. Nobody, however, expects that MD simulators will abandon their classic integrators or basic pair-wise force fields. In threading, there is no such consensus. There are alignments calculated by dynamic programming and some by Monte Carlo. There are force fields based on Boltzmann statistics and others on numerical optimisation. Despite public comparisons of results, it is very difficult to say what the best approach is. Rarely, does one group's sequence to structure alignment

method get used with another group's score function and a different group's method for assessing reliability.

Can one at least guess where the biggest weaknesses in current methods are ? It is not clear if the best combination of multiple sequences, structures and score functions has been found. Most of the work in this direction has come from groups with through-space score functions adding terms for sequence similarity. Given the success of sequence analysis, without any use of coordinates, it is quite possible that they will improve further as they make better use of through space relations.

Boltzmann-based or pure physically-based force fields are changing relatively slowly, but score functions which come from optimisation or constraint satisfaction still appear in different forms. One obvious change would be a more holistic approach where terms such as secondary structure prediction and sequence similarity are cast as part of the original score function construction problem. Currently, these kinds of terms are treated as decoration on the main score function.

Finally, threading may not be the best approach at all. There may be a simple force field waiting to be built which will allow proteins to swiftly find their native states with a method such as dynamics simulation, genetic algorithm or other search method. Quite possibly, a fragment-based approach will end up as the most successful.[66]

More modestly, some improvement may come without much change to the score functions or search methods, but simply from better estimates of reliability. Sequence analysis has a good statistical basis and the statistics of structure prediction are the basis of another chapter in this volume.

The biggest threat to protein predictors comes from experimentalists. Advances in automation, robotics and chemistry mean that structures are solved at an ever increasing rate. The question is whether protein structure predictors will be able to make a useful contribution before they are outdated by the flood of experimental data.

References

1. Kremer K, Müller-Plathe F: **Multiscale problems in polymer science: Simulation approaches.** *MRS Bull* 2001, **26**:205-210

2. van Gunsteren WF, Billeter SR, Eising AA, Huenenberger PH, Krueger P, Mark A, Scott WRP, Tironi IG: **Biomolecular Simulation: The GROMOS96 manual and user guide**. Zurich and Groningen:vdf Hochschulverlag AG an der ETH Zurich and BIOMOS b.v., 1996.

3. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M: **All-atom empirical potential for molecular modeling and dynamics studies of proteins.** *J Phys Chem B* 1998, **102**:3586-3616

4. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA: **A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.** *J Am Chem Soc* 1995, **117**:5179-5197

5. Novotny J, Bruccoleri R, Karplus M: **An analysis of incorrectly folded protein models - implications for structure predictions.** *J Mol Biol* 1984, **177**:787-818

6. Lazaridis T, Karplus M: **Discrimination of the native from misfolded protein models with an energy function including implicit solvation.** *J Mol Biol* 1999, **288**:477-487

7. Janardhan A, Vajda S: **Selecting near-native conformations in homology modeling: the role of molecular mechanics and solvation terms.** *Protein Sci* 1998, **7**:1772-1780

8. Wang YH, Zhang H, Scott RA: **A New Computational Model for Protein-Folding Based on Atomic Solvation.** *Protein Sci* 1995, **4**:1402-1411

9. Wang YH, Zhang H, Li W, Scott RA: **Discriminating Compact Nonnative Structures from the Native Structure of Globular-Proteins.** *Proc Natl Acad Sci USA* 1995, **92**:709-713

10. Reith D, Huber T, Müller-Plathe F, Torda AE: **Free energy approximations in simple lattice proteins.** *J Chem Phys* 2001, **114**:4998-5005

11. Chandler D: **Introduction to modern statistical mechanics**. New York:Oxford University Press, 1987.

12. Tanaka S, Scheraga HA: **Statistical mechanical treatment of protein conformation. 1. Conformational properties of amino-acids in proteins.** *Macromolecules* 1976, **9**:142-159

13. Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules* 1985, **18**:534-552

14. Sippl MJ: **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* 1990, **213**:859-883

15. Sippl MJ: **Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.** *J Comput Aided Mol Des* 1993, **7**:473-501

16. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition.** *Nature* 1992, **358**:86-89

17. Ben-Naim A: **Statistical potentials extracted from protein structures: are these meaningful potentials ?** *J Chem Phys* 1997, **107**:3698-3706

18. Thomas PD, Dill K: **Statistical potentials extracted from protein structures: how accurate are they ?** *J Mol Biol* 1996, **257**:457-469

19. Dill KA: **Additivity principles in biochemistry.** *J Biol Chem* 1997, **272**:701-704

20. Crippen GM, Snow ME: **A 1.8 Angstrom resolution potential function for protein folding.** *Biopolymers* 1990, **29**:1479-1489

21. Maiorov VN, Crippen GM: **Contact Potential That Recognizes the Correct Folding of Globular-Proteins.** *J Mol Biol* 1992, **227**:876-888

22. Seetharamulu P, Crippen GM: **A Potential Function for Protein Folding.** *J Math Chem* 1991, **6**:91-110

23. Kirkpatrick S, Gelatt Jr. CD, Vecchi MP: **Optimization by simulated annealing.** *Science* 1983, **220**:671-680

24. Car R, Parrinello M: **Unified approach for molecular dynamics and density-functional theory.** *Phys Rev Lett* 1985, **55**:2471-2474

25. Ulrich P, Scott W, van Gunsteren WF, Torda AE: **Protein Structure Prediction Force Fields - Parametrization With Quasi-Newtonian Dynamics.** *Proteins* 1997, **27**:367-384

26. Huber T, Torda AE: **Protein Fold Recognition Without Boltzmann Statistics or Explicit Physical Basis.** *Protein Sci* 1998, **7**:142-149

27. Hao MH, Scheraga HA: **How optimization of potential functions affects protein folding.** *Proc Natl Acad Sci USA* 1996, **93**:4984-4989

28. Koretke KK, Luthey-Schulten Z, Wolynes PG: **Self-consistently optimized statistical mechanical energy functions for sequence structure alignment.** *Protein Sci* 1996, **5**:1043-1059

29. Mirny LA, Shakhnovich EI: **How to derive a protein folding potential - a new approach to an old problem.** *J Mol Biol* 1996, **264**:1164-1179

30. Chiu TL, Goldstein RA: **Optimizing energy potentials for success in protein tertiary structure prediction.** *Fold Des* 1998, **3**:223-228

31. Xia Y, Levitt M: **Extracting knowledge-based energy functions from protein structures by error rate minimization: Comparison of methods using lattice model.** *J Chem Phys* 2000, **113**:9318-9330

32. Tobi D, Elber R: **Distance-dependent, pair potential for protein folding: results from linear optimization.** *Proteins* 2000, **41**:40-46

33. Tobi D, Shafran G, Linial N, Elber R: **On the design and analysis of protein folding potentials.** *Proteins* 2000, **40**:71-85

34. Micheletti C, Seno F, Banavar JR, Maritan A: **Learning effective amino acid interactions through iterative stochastic techniques.** *Proteins* 2001, **42**:422-431

35. Vendruscolo M, Najmanovich R, Domany E: **Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading?** *Proteins* 2000, **38**:134-148

36. Ayers DJ, Huber T, Torda AE: **Protein fold recognition score functions: unusual construction strategies.** *Proteins* 1999, **36**:454-461

37. Ohkubo YZ, Crippen GM: **Potential energy function for continuous state models of globular proteins.** *J Comput Biol* 2000, **7**:363-379

38. Vendruscolo M, Domany E: **Pairwise contact potentials are unsuitable for protein folding.** *J Chem Phys* 1998, **109**:11101-11108

39. Lathrop RH: **The protein threading problem with sequence amino acid interaction preferences is NP-complete.** *Protein Eng* 1994, **7**:1059-1068

40. Madej T, Gibrat JF, Bryant SH: **Threading a Database of Protein Cores.** *Proteins* 1995, **23**:356-369

41. Pedersen JT, Moult J: **Ab initio protein folding simulations with genetic algorithms: Simulations on the complete sequence of small proteins.** *Proteins* 1997, 179-184

42. Pedersen JT, Moult J: **Protein folding simulations with genetic algorithms and a detailed molecular description.** *J Mol Biol* 1997, **269**:240-259

43. Godzik A, Kolinski A, Skolnick J: **Topology fingerprint approach to the inverse protein folding problem.** *J Mol Biol* 1992, **227**:227-238

44. Wilmanns M, Eisenberg D: **Inverse protein folding by the residue pair preference profile method: estimating the correctness of alignments of structurally compatible sequences.** *Protein Eng* 1995, **8**:627-639

45. Lathrop RH, Smith TF: **Global optimum protein threading with gapped alignment and empirical pair score functions.** *J Mol Biol* 1996, **255**:641-665

46. Lathrop RH: **An anytime local-to-global optimization algorithm for protein threading in O(m(^{2})n(^{2})) space.** *J Comput Biol* 1999, **6**:405-418

47. Xu Y, Uberbacher EC: **A polynomial-time algorithm for a class of protein threading problems.** *Comput Appl Biosci* 1996, **12**:511-517

48. Huber T, Torda AE: **Sausage program.** 2001, http://www.rsc.anu.edu.au/~torda/sausage

49. Huber T, Russell AJ, Ayers D, Torda AE: **SAUSAGE: Protein threading with flexible force fields.** *Bioinformatics* 1999, **15**:1064-1065

50. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6**:361-365

51. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763

52. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: Extension and analysis of the basic method.** *Comput Appl Biosci* 1996, **12**:95-107

53. Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C: **Predicting protein structure using hidden Markov models.** *Proteins* 1997, 134-139

54. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies.** *Bioinformatics* 1998, **14**:846-856

55. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R: **Predicting protein structure using only sequence information.** *Proteins* 1999, 121-125

56. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284**:1201-1210

57. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D: **Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology.** *Comput Appl Biosci* 1996, **12**:327-345

58. Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains.** *Nucleic Acids Res* 1998, **26**:320-322

59. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402

60. Rost B, Schneider R, Sander C: **Protein fold recognition by prediction-based threading.** *J Mol Biol* 1997, **270**:471-480

61. Panchenko AR, Marchler-Bauer A, Bryant SH: **Combination of threading potentials and sequence profiles improves fold recognition.** *J Mol Biol* 2000, **296**:1319-1331

62. unpublished results

63. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89**:10915-10919

64. Torda AE: **List of proteins for alignment testing.** 2001, http://www.rsc.anu.edu.au/~torda/mult_strct/alignment_pairs.html

65. Jones DT: **GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences.** *J Mol Biol* 1999, **287**:797-815

66. Bonneau R, Baker D: **Ab initio protein structure prediction: progress and reports.** *Annu Rev Biophys Biomol Struct* 2001, **30**:173-189

67. Cuff JA, Barton GJ: **Evaluation and improvement of multiple sequence methods for protein secondary structure prediction.** *Proteins* 1999, **34**:508-519

68. Cuff JA, Barton GJ: **Application of multiple sequence alignment profiles to improve protein secondary structure prediction.** *Proteins* 2000, **40**:502-511

69. Chandonia JM, Karplus M: **New methods for accurate prediction of protein secondary structure.** *Proteins* 1999, **35**:293-306

70. Jones DT: **Protein secondary structure prediction based on position- specific scoring matrices.** *J Mol Biol* 1999, **292**:195-202

71. Rost B, Sander C: **Prediction of Protein Secondary Structure at Better Than 70- Percent Accuracy.** *J Mol Biol* 1993, **232**:584-599

72. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert GP, Lund O: **Prediction of protein secondary structure at 80% accuracy.** *Proteins* 2000, **41**:17-20

73. Ayers DJ, Gooley PR, Widmer-Cooper A, Torda AE: **Enhanced protein fold recognition using secondary structure information from NMR.** *Protein Sci* 1999, **8**:1127--1133

74. Rost B, Sander C, Schneider R: **PHD - an Automatic Mail Server For Protein Secondary Structure Prediction.** *Comput Appl Biosci* 1994, **10**:53-60

75. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins* 1994, **18**:309-317

76. Shindyalov IN, Kolchanov NA, Sander C: **Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?** *Protein Eng* 1994, **7**:349-358

77. Rost B, Sander C: **Conservation and prediction of solvent accessibility in protein families.** *Proteins* 1994, **20**:216-226

78. Wishart DS, Sykes BD: **The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data.** *J Biomol NMR* 1994, **4**:171-180

79. Wishart DS, Sykes BD: **Chemical shifts as a tool for structure determination.** *Method Enzymol* 1994, **239**:363-392

80. Wishart DS, Sykes BD, Richards FM: **The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy.** *Biochemistry* 1992, **31**:1647-1651

81. Wishart DS, Sykes BD, Richards FM: **Relationship between nuclear magnetic resonance chemical shift and protein secondary structure.** *J Mol Biol* 1991, **222**:311-333

82. Aebersold R, Goodlett DR: **Mass spectrometry in proteomics.** *Chem Rev* 2001, **101**:269-295

83. Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G: **High throughput protein fold identification by using experimental constraints derived from intramolecular cross- links and mass spectrometry.** *Proc Natl Acad Sci USA* 2000, **97**:5802-5806

84. Finkelstein AV: **3D protein folds: homologs against errors - a simple estimate based on the random energy model.** *Phys Rev Lett* 1998, **80**:4823-4825

85. Reva BA, Skolnick J, Finkelstein AV: **Averaging interaction energies over homologs improves protein fold recognition in gapless threading.** *Proteins* 1999, **35**:353-359

86. Dykunov D, Lobanov MY, Finkelstein AV: **Search for the most stable folds of protein chains: III. Improvement in fold recognition by averaging over homologous sequences and 3D structures.** *Proteins* 2000, **40**:494-501

87. Holm L, Sander C: **The FSSP database of structurally aligned protein fold families.** *Nucleic Acids Res* 1994, **22**:3600-3609

88. Holm L, Sander C: **The FSSP database: fold classification based on structure-structure alignment of proteins.** *Nucleic Acids Res* 1996, **24**:206-209.

89. Holm L, Sander C: **Dali/FSSP classification of three-dimensional protein folds.** *Nucleic Acids Res* 1997, **25**:231-234

90. Holm L, Sander C: **Touring protein fold space with dali/FSSP.** *Nucleic Acids Res* 1998, **26**:316-319

91. Huber T, Torda AE: **Protein sequence threading, the alignment problem and a two step strategy.** *J Comput Chem* 1999, **20**:1455-1467

92. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410

93. Hooft RWW, Vriend G, Sander C, Abola EE: **Errors in proteins structures.** *Nature* 1996, **381**:272

Figure Captions

Figure 1. Sequence of unknown structure and candidate library template

Figure 2. Aligning a sequence to a template by constructing a score matrix. Some example matrix elements are filled in and a path is marked for the best sequence to structure alignment.

Figure 3. Correct answer and predicted structure. The prediction corresponds to the template and alignment path from Figure 2.

Figure 4. Calculation of pair-wise, through-space scores

Figure 5. Distribution of energies of misfolded proteins.

Figure 6. Simple sequence alignment for scoring

Figure 7. Optimisation of weight for sequence similarity within overall force field

Figure 8. Effect of averaging over structures and sequences. *no average* refers to the bare score function, *struct average* to averaging over structures and *mult sequence* to averaging over sequences

Figure 9. Effect of structure optimisation on energy levels for lysozyme (153l)

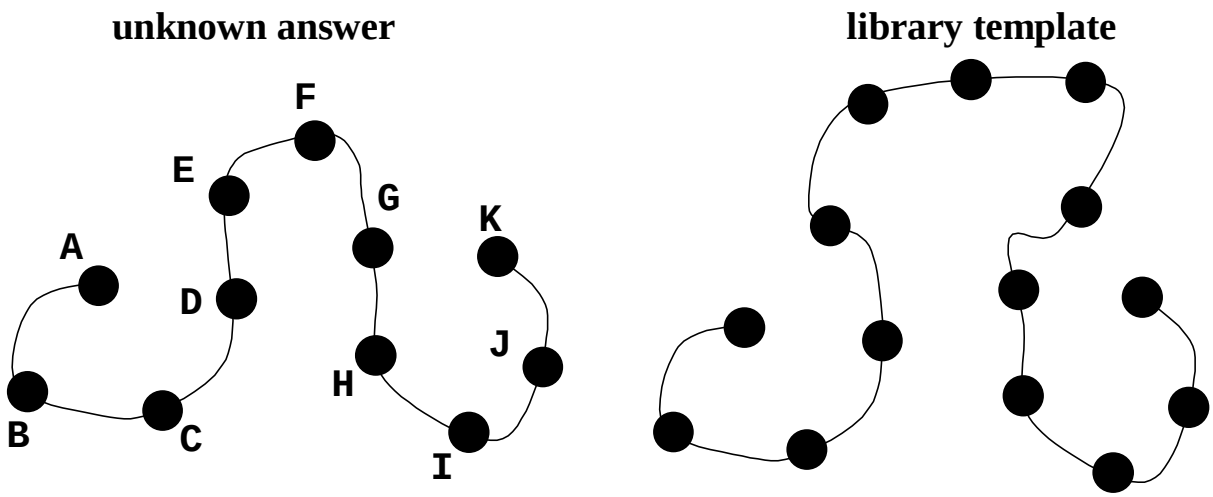


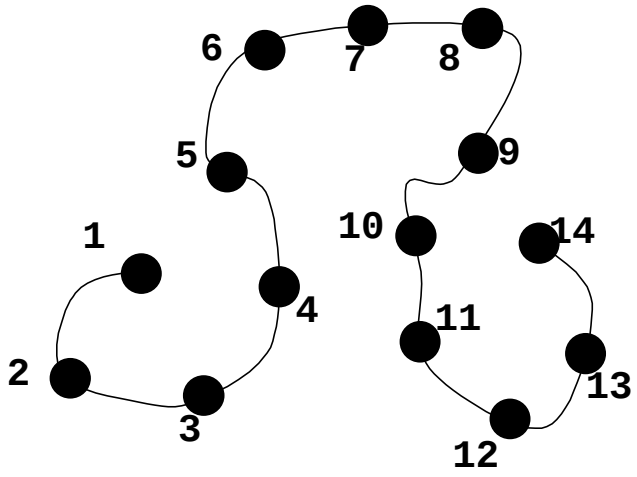
Figure 1

sequence

A B C D E F G H I J K

+

library template



score matrix

	A	B	C	D	E	F	G	H	I	J	K
1	23	8	22	7	…	…	…	…	…	…	…
2	7										
3	8										
4	9										
5	12										
6	7										
7	…										
8	…										
9	…										
10	…										
11	…										
12	…										
13	…										
14	…										

Figure 2

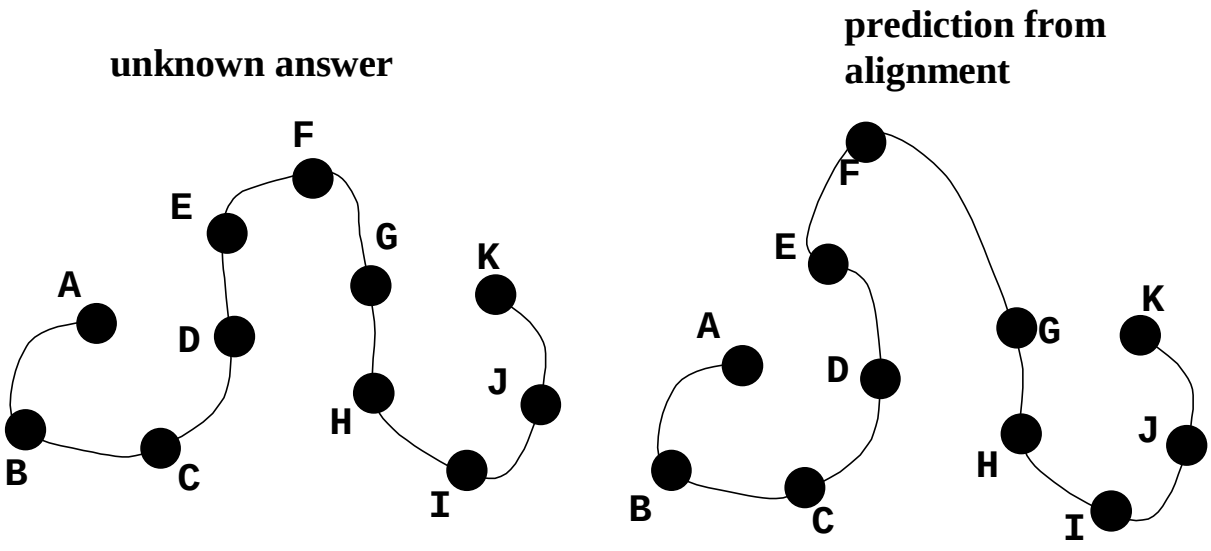

unknown answer
prediction from alignment
F
E
G
K
A
D
J
H
B
C
I
F
E
A
G
K
D
H
J
B
C
I

Figure 3

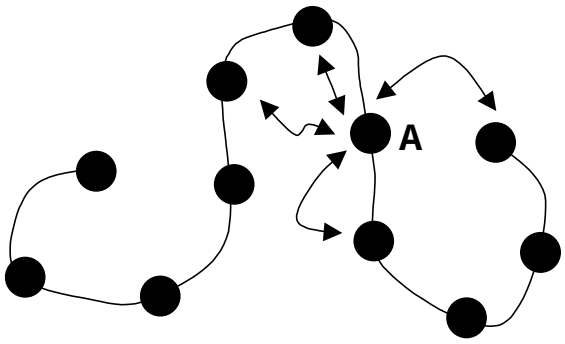


Figure 4

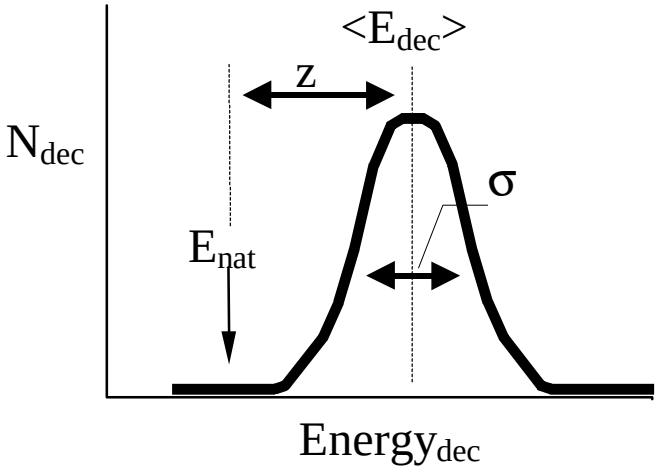


Figure 5. Distribution of decoy energies.

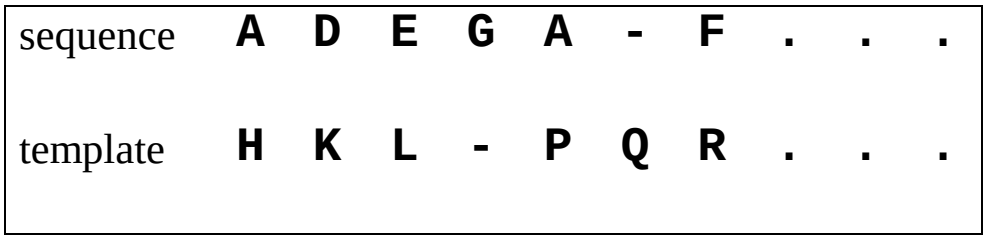


Figure 6

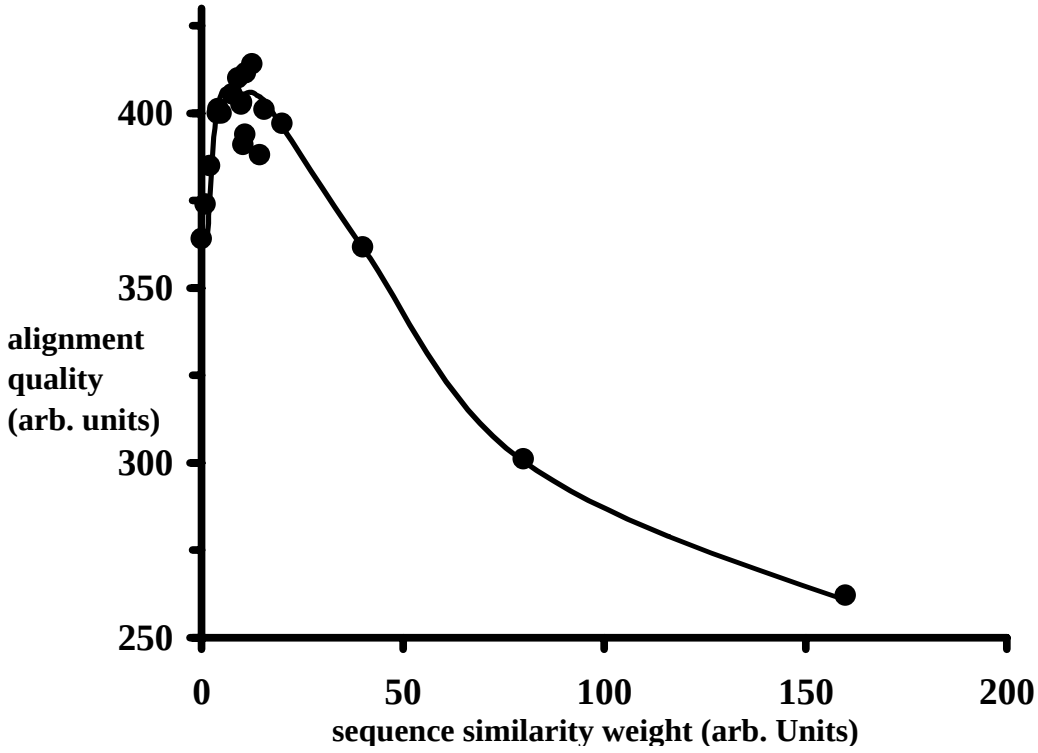

400
350
alignment
quality
(arb. units)
300
250
0
50
100
150
200
sequence similarity weight (arb. Units)

Figure 7

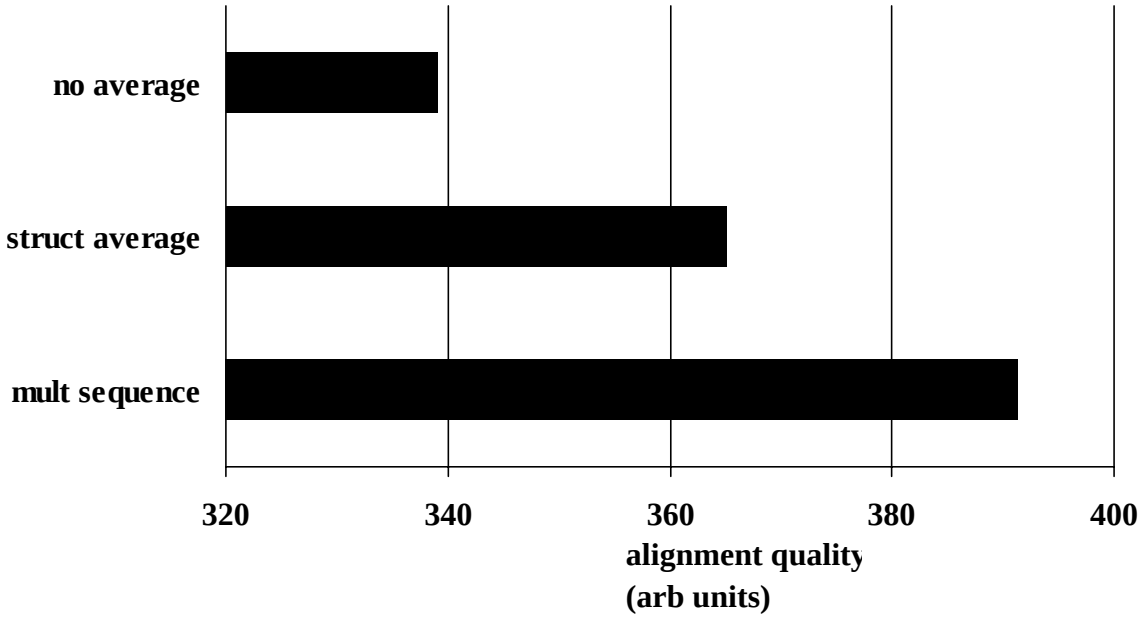

no average
struct average
mult sequence
320
340
360
380
400
alignment quality
(arb units)

Figure 8

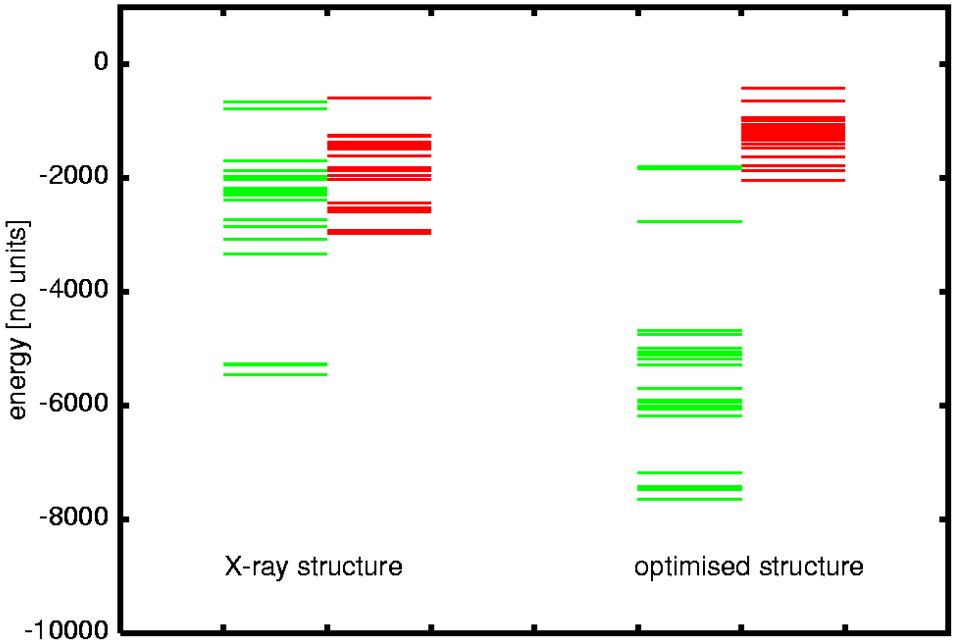

0
-2000
-4000
-6000
-8000
-10000
energy [no units]
X-ray structure
optimised structure

Figure 9