Statistical Methods in Medical Research 2009; **18**: 505–526

Probabilistic models and machine learning in structural bioinformatics

Thomas Hamelryck Bioinformatics Center, Department of Biology, University of Copenhagen, Copenhagen N, Denmark

Structural bioinformatics is concerned with the molecular structure of biomacromolecules on a genomic scale, using computational methods. Classic problems in structural bioinformatics include the prediction of protein and RNA structure from sequence, the design of artificial proteins or enzymes, and the automated analysis and comparison of biomacromolecules in atomic detail. The determination of macromolecular structure from experimental data (for example coming from nuclear magnetic resonance, X-ray crystallography or small angle X-ray scattering) has close ties with the field of structural bioinformatics. Recently, probabilistic models and machine learning methods based on Bayesian principles are providing efficient and rigorous solutions to challenging problems that were long regarded as intractable. In this review, I will highlight some important recent developments in the prediction, analysis and experimental determination of macromolecular structure that are based on such methods. These developments include generative models of protein structure, the estimation of the parameters of energy functions that are used in structure prediction, the superposition of macromolecules and structure determination methods that are based on inference. Although this review is not exhaustive, I believe the selected topics give a good impression of the exciting new, probabilistic road the field of structural bioinformatics is taking.

1 Introduction

The prediction the three-dimensional (3D) structure of RNA and protein from sequence is one of the main open problems in science today.[1] A routine solution of this problem would be of tremendous importance in science, medicine and biotechnology, due to its relevance for crucial applications such as understanding our genomes in molecular detail or the design of novel drugs and man-made enzymes. Despite some important recent breakthroughs,[2–5] prediction and design of macromolecular structure remains not routinely possible, and in the case of protein structure prediction, progress even seems to be stagnating.[6]

The problem of 3D structure prediction can be viewed as a problem in physics, where the solution is expected to come from a description of the water–protein or water–RNA system at the level of quantum mechanics. At present, such a description is essentially out of reach due to excessive computational demands, although progress is being made by using various approximations, such as molecular dynamics.[7] At the other end of the spectrum, one finds knowledge-based methods that try to translate the existing

Address for correspondence: Thomas Hamelryck, Bioinformatics Center, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark. E-mail: thamelry@binf.ku.dk

10.1177/0962280208099492

information on macromolecular structure into methods to predict structure. Most of the so-called *de novo* methods that predict the structure of proteins and RNA from sequence information alone are at present knowledge based. These methods essentially treat the structure prediction problem as a problem of statistical inference, even if this is not always clearly spelled out.

Here, I will give an overview of some of the special statistical challenges that are associated with the prediction and analysis of macromolecular structure, with an emphasis on proteins. I will focus in particular on methods that view the prediction, analysis and experimental determination of macromolecular structure as a problem in Bayesian inference, and point out some recent success stories. A second goal of this review is to point out that these probabilistic methods can be interpreted in the framework of statistical physics, and vice versa, that many physics based methods can be seen from a probabilistic viewpoint. After all, as ET. Jaynes pointed out many years ago, statistical mechanics is statistical inference applied to physical-problems, [8–10] and therefore it should not come as a surprise that methods developed from a purely probabilistic point of view can have a clear physical interpretation.

This review starts with a gentle introduction to macromolecular structure. Next, in Section 3, I introduce two probabilistic methodologies that have been found recently to be particularly useful in structural bioinformatics: Bayesian networks and directional statistics. In Section 4, several recently developed probabilistic solutions to problems in structural bioinformatics are presented, and I also give some alternative views on established methods. In the final section, I discuss probabilistic methods to infer macromolecular structure from experimental data.

2 Macromolecular structure

Proteins are the workhorses of the living cell. They are, for example, responsible for digesting food, transferring signals from the environment and protecting organisms against harmful infections. In medicine, proteins are extremely important since they form the target of virtually all drugs. In biotechnology, proteins are used to catalyse reactions that would be extremely difficult to perform efficiently by non-biological means.

Proteins are simple linear polymers of amino acids. One can think of them as a set of beads-on-a-string, in which the beads are individual amino acids. Brought in a watery environment, a typical, well-behaving protein will fold into a specific, compact 3D conformation or fold (Figure 1). Broadly speaking, which shape the protein will adopt depends entirely on its sequence of amino acids.[11] The driving force behind the formation of the compact fold is the hydrophobic effect, or the shielding of the hydrophobic amino acids from the water surrounding the protein. The same effect is also responsible for the fact that oil and water do not mix, for example. Although the hydrophobic effect is now quite well understood,[12] it is still not possible to predict protein structure from sequence. One of the reasons is that many other subtle interactions play a role as well, such as hydrogen bonding, electrostatic effects and van der Waals forces.

Until recently, RNA was seen as a mere information carrier between the genomic information encoded in DNA, and the chemical and structural functions performed

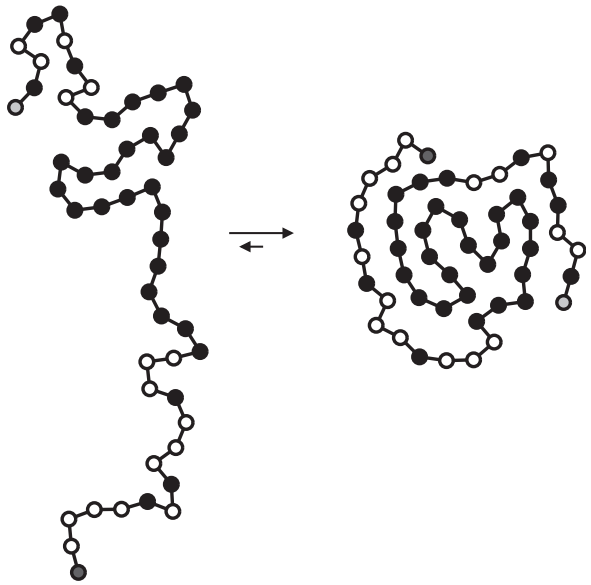

Figure 1 Schematic illustration of protein folding that is driven by the hydrophobic effect. In a watery environment, a protein spontaneously goes from an unfolded, extended state (on the left) to the compact, folded state (on the right). In the compact fold, the hydrophobic amino acids (shown as black circles) are in general shielded from the solvent, while the hydrophilic amino acids (white circles) are in general exposed.

by proteins. That view has changed radically,[13] with the emergence of RNA with enzymatic activities and a wide variety of functions performed by so-called non-coding RNAs. As for proteins, the 3D structure of an RNA molecule is of crucial importance for understanding its function, and as a result, the problem of the prediction of RNA structure in atomic detail has become of acute importance. In the case of RNA, the main driving force behind the formation of the folded conformation is base pairing, in which nucleotides pair through the formation of hydrogen bonds, and base stacking, which involves the aromatic rings of the bases.

The detailed 3D structure of RNA and protein can be determined by biophysical methods, such as nuclear magnetic resonance (NMR) and X-ray crystallography. However, while determining the sequence of RNA or protein is (relatively) easy, experimentally determining the 3D structure is typically expensive, time craving and difficult, and sometimes even impossible. Therefore, there is great interest in predicting protein structure from sequence. Despite some recent advancements, this is still not routinely possible at present.

2.1 Notation

Below, I will refer to a 3D structure as $\mathbf{x}$ and an amino acid sequence as $\mathbf{a} = \{a_1, a_2, \ldots, a_N\}$, where a_i refers to amino acid i in the sequence. A structure $\mathbf{x}$ is parameterised as a sequence of atom positions (3D vectors) or various angles, depending on the context.

3 Probabilistic models for molecular structure

3.1 Directional statistics

In order to develop probabilistic models of molecular structure, one needs to have a proper mathematical language available. The 3D shape of an RNA or protein molecule can be conveniently described for many purposes in terms of angles (in $[0, \pi]$) and dihedral angles (in $[-\pi, \pi[$) (Figure 2). Naively applying the usual statistical machinery to such data often leads to problems. For example, using a Gaussian distribution for dihedral angle data can run into problems due to the wrap-around property at $-\pi/\pi$. In other words, it is often necessary to work on the right manifold, which in the latter case is the circle, and not the real line. In this case, the von Mises distribution on the circle is a suitable choice.

The statistics of data that inhabit manifolds such as spheres, tori or more exotic manifolds such as the real projective plane is the realm of directional statistics.[14] Examples of such data are angles, unit vectors, rotations and axes (lines through the origin in $\mathbb{R}^n$). As these types of data are ubiquitous in structural bioinformatics, the potential of directional statistics in this field is enormous. Surprisingly, applications of directional statistics in structural bioinformatics are few and far between, and certainly not part of the main stream. However, I expect that this situation will change drastically in the near future. In Section 4.1, I will discuss some of our recent work in this area.

I will end this section with giving some examples of probability distributions that are particularly useful for describing molecular structure. The Kent distribution, proposed by John T. Kent in 1982,[15] is a distribution on the 2D sphere (Figure 3). In order to avoid some common confusion, let me point out that the 2D sphere is the surface of the 3D ball. A point on the sphere can be specified by a 3D unit vector, or equivalently, a set of polar coordinates (one angle, and one dihedral angle).

The Kent distribution is the equivalent the the general bivariate Gaussian distribution, and has the following probability density function:

$$f(\mathbf{v}) = \frac{1}{C(\kappa, \beta)} \exp\left(\kappa \gamma_1 \cdot \mathbf{v} + \beta \left[(\gamma_2 \cdot \mathbf{v})^2 - (\gamma_3 \cdot \mathbf{v})^2\right]\right) \tag{1}$$

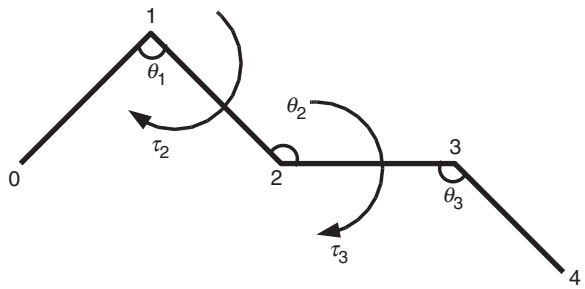

Figure 2 A set of four connected lines illustrates the concept of a dihedral angle, and the geometry of the Cα trace representation. The individual points 0, 1, 2, 3 and 4 correspond to the Cα atoms of five consecutive amino acids. The θ angles are 'ordinary' angles (in $[0, \pi]$) formed by two consecutive line segments. The dihedral angles τ (in $[-\pi, \pi[$) are defined by three consecutive segments. For example, τ_2 is the angle formed by segments 01 and 23 when looking along segment 12.

where $\cdot$ is the dot product of two vectors, $\mathbf{v}$ is a unit vector, $C(\kappa, \beta)$ is a normalising constant, $\kappa \geq 0$ is a concentration parameter, β (with $0 \leq 2\beta < \kappa$) determines the ellipticity of the contours of equal probability and $(\gamma_1, \gamma_2, \gamma_3)$ are three unit vectors that determine the position and orientation of the equiprobability contours on the sphere. In section 4.1, I will describe some important application of this distribution. The distribution is also known as the five-parameter Fisher–Bingham distribution, as it has five independent parameters (κ, β, and a 3×3 orthogonal matrix containing the γ vectors) and belongs to a wider family of distributions. The use of this distribution in modelling the local structure of proteins is discussed in section 4.1. A distribution from the same family, the von Mises–Fisher distribution,[14,16] generalises the spherical (isotropic) Gaussian distribution on the N-dimensional sphere.

Other distributions include the recently proposed bivariate von Mises distribution on the torus[17] (discussed in section 4.1), the Bingham distribution on the sphere with the antipodes identified,[18] and the Matrix–Fisher distribution on Stiefel manifolds.[19] The latter distribution can be used as a probability distribution over the space of rotations SO(3), which has found applications in crystallography. [20] For a thorough discussion of directional statistics, I refer to the book of Mardia and Jupp.[14]

3.2 Bayesian networks

Probabilistic models such as mixture models, hidden Markov models, Bayesian networks and Markov random fields can all be seen as examples of graphical models.

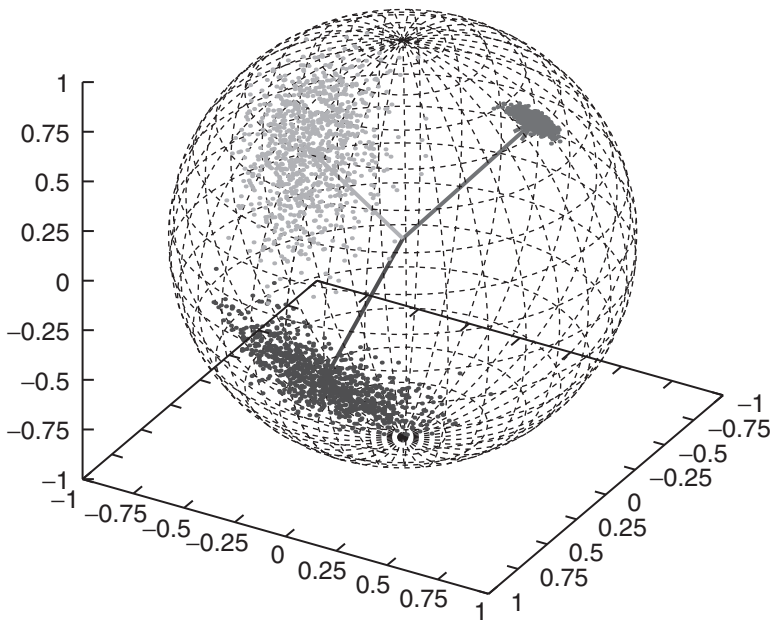

Figure 3 Three points sets sampled from the Kent distribution on the sphere. The mean directions are shown with arrows. The κ parameter is highest for the rightmost, resulting in a concentrated distribution. The β parameter is highest for the point set at the bottom, resulting in a distribution with high ellipticity. The position on the sphere, and the orientation of the equiprobability contours, are determined by the $\gamma_1, \gamma_2, \gamma_3$ vectors.

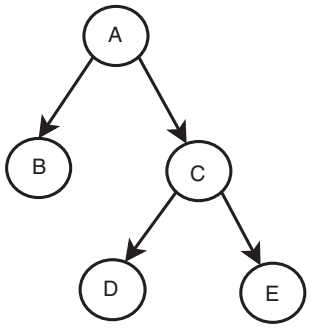

Figure 4 A simple Bayesian network. The circles are variables, and the edges of the graph encode the conditional independencies between the variables.

Essentially, graphical models are graphs that represent a joint probability distribution, in which the nodes are variables (which can be parameters, random variables or hypotheses) and the graph structure determines the possible factorisations of the joint distribution. Here, I will focus on graphical models specified by directed, acyclic graphs, which are called Bayesian networks.[21] Interestingly, factor graphs have recently provided an elegant, unified approach to both directed (Bayesian networks) and undirected (Markov random fields) graphical models.[22]

Let us consider Figure 4 as a representative example of a Bayesian network.

As mentioned before, the network's graph structure determines the possible factorizations of the joint probability distribution of the variables in the graph. The joint probability distribution represented by the graph in Figure 4 can be factorised as:

$$P(A, B, C, D) = P(A)P(B \mid A)P(C \mid A)P(D \mid C)P(E \mid C)$$

A Bayesian network is a carrier of conditional independence relationships. The absence of an edge between two nodes guarantees that there is a set of nodes that renders them conditionally independent. In the example network, this is for example the case for nodes B and E:

$$P(B, E \mid A, C) = P(B \mid A)P(E \mid C)$$

Sequential data can be easily handled by so-called *dynamic* Bayesian networks,[a] in which the networks structure is 'unrolled' along the length of the sequence.[23] Bayesian networks are particularly attractive because they are generative: one can generate samples from them, which is often of crucial importance. In Section 4.1, I will discuss how a combination of directional statistics and dynamic Bayesian networks leads to an elegant, generative model of the local structure of proteins.

I end with a note on the joint probability distribution of Bayesian networks that are trees or polytrees [b] (see Pearl,[21] Section 2.3.4), as this will be useful later to understand

---

[a] *The word dynamic is a bit confusing here: it originates from the fact that sequences modelled by these networks are often temporal in character, such as, for example, speech signals. In the case of protein or RNA sequences, this is of course not the case.*

[b] *A polytree is a graph with at most one undirected path between any two nodes.*

the construction of a likelihood function used in protein structure prediction. First, note that any joint probability distribution can be written in the following form:

$$P(A, B, C, D, E) = P(A)P(B \mid A)P(C \mid A, B)P(D \mid A, B, C)P(E \mid A, B, C, D)$$

When the conditional independencies encoded in our example network are taken into account, this can be written this as:

$$P(A, B, C, D, E) = \frac{P(A, B)P(A, C)P(C, D)P(C, E)}{P(A)P(C)^2} \tag{2}$$

Thus, for a directed tree, the joint probability distribution can be written as a product of the joint probability of the edges, divided by the probabilities of the individual nodes, raised to the power $q - 1$, where q is the number of edges attached to the node. We will return to this result in Section 4.2.

4 Structure prediction and analysis

4.1 Models of local structure for conformational sampling

In this section, I will give an overview of some of the joint work done at the Bioinformatics center, University of Copenhagen and the Department of Statistics, University of Leeds. Much of the work has been focussed so far on the first necessary ingredient of any structure prediction method: the exploration of the conformational space.

The current state-of-the-art method to generate protein-like conformations is the so-called fragment assembly method. Conceptually, the method is very simple. One simply constructs a library of short, linear fragments from existing protein structures, and constructs novel conformations by tying a set of fragments together. Which fragments are chosen depends on the similarity, on a local scale, between the sequence that is to be modelled and the sequence of the fragment. This approach was first proposed by Jones and Thirup in 1986[24] for the construction of protein structures from X-ray crystallography data. The fragment assembly method was subsequently adopted for *de novo* structure prediction by David Baker and his co-workers in 1997,[25] and it is currently still the method of choice for most successful *de novo* protein structure prediction methods.

The reason for its success lies in the fact that there is a strong influence of local sequence on local structure.[26] There are strong indications that the native structure corresponds to the compact structure that is compatible with the structural preferences encoded on a local scale.[27,28]

Despite the success of these fragment assembly methods, they have many disadvantages.[29] Most importantly, the fragment approach is non-probabilistic, which makes it difficult to use the method rigorously in a Markov chain Monte Carlo (MCMC) context. In addition, combining fragments together causes edge-effects where the fragments are combined, which leads to the proposal of conformations that are not found in native proteins. The use of fragments also discretizes conformational space, which in reality of course is of a continuous nature. Much of our work done in the last 4 years

has been focussed on the development of fully probabilistic methods to capture local sequence/local structure signals in continuous space.

In order to do this, one first needs a good parameterisation for the local structure of proteins. Two such parameterisations are well established (Figure 5). The first is called the Cα-trace representation of a protein, in which each amino acid is represented as a single point (its Cα atom), and the overall shape of the protein is outlined by a series of connected segments[30,31] (Figure 2). To a good approximation, the length of these segments can considered to be fixed (about 3.8 Å). As a result, the geometry of the Cα trace of n amino acids can be expressed as a series of $n-3$ dihedral angles (called τ, with $\tau \in [-\pi, \pi[$) and $n-2$ angles (called θ, with $\theta \in [0, \pi]$). An (θ, τ) angle pair can be interpreted as a set of polar coordinates, which leads to the insight that the Cα trace of a protein can be parameterised as a sequence of unit vectors, or, equivalently, a sequence of points on the unit sphere S^2 in $\mathbb{R}^3$.

A considerably more detailed representation is the so-called backbone representation of a protein, in which each amino acid is represented by three of its atoms (C, N and Cα). This representation captures the geometry of the protein backbone in atomic detail. Again, this representation can be parameterised as a sequence of angle pairs, in this case $n-1$ dihedral angle pairs. These angles are called ϕ and ψ (both $\in [-\pi, \pi[$), and are the angles plotted in the celebrated Ramachandran plot.[32] This plot is widely used by structural biologists to judge the quality of an experimental protein structure, as these angles are expected to avoid certain values in high-quality structures. A dihedral angle pair defines a point on the torus T^2 (the surface of a doughnut or a tyre), and hence the backbone representation can be parameterised as a series of points on the torus.

Armed with these insights, it now becomes clear that the statistical challenge is to develop a probabilistic model that represents a string of symbols (the amino acid sequence) and a sequence of points on two different manifolds (the sphere or the torus). The first aspect of the problem, dealing with sequential observations, can be easily solved

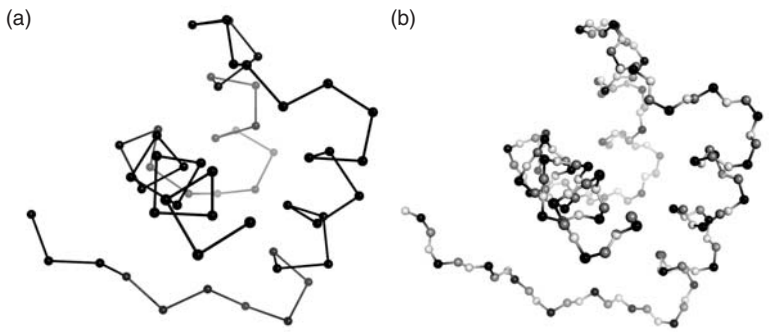

Figure 5 Two simplified representations of the same protein (the engrailed homeodomain, with Protein Data Bank code 1ENH). **(a)** The Cα trace representation, in which uses a single point (at the Cα atom) for each amino acid. The edges between the consecutive points are not real chemical bonds, but span several real chemical bonds. **(b)** The full backbone representation, which uses three points (Cα, C, N) for each amino acid. In this case, the edges do correspond to real chemical bonds between atoms. The side chains, which are not shown for clarity, are attached to the Cα atoms and determine the amino acid type.

by using a graphical model such as a hidden Markov model or a dynamic Bayesian network (Section 3.2). For the second aspect of the problem, the field of directional statistics comes into play (Section 3.1). Observations on the sphere can be modelled by the Kent distribution,[15] while observations on the torus can be handled by the bivariate von Mises distribution.[17]

The final model is shown in Figure 6. It consists of a single first-order Markov chain of hidden nodes, each with three observed nodes attached. One amino acid is represented by one hidden node, where the attached observed nodes represent the amino acid symbol, secondary structure (α-helix, β-sheet or coil) and angle pair. Two models were developed, respectively dealing with Cα-trace and full backbone geometry.[28,33–35] The first model uses the Kent distribution to represent the angle pairs as points on the sphere, while the second model uses the bivariate von Mises distribution to represent the angle pairs as points on the torus. The parameters of both models were optimised with stochastic expectation maximisation,[36] using a large training set of experimental protein structures. The optimal number of hidden node states was determined using the Bayesian information criterion. The joint probability distribution of amino acid sequence **a**, secondary structure sequence **s** and angle pair sequence $\mathbf{x}_L$ (where the subscript L stands for 'local') is obtained by marginalising over all possible hidden node sequences **h**:

$$P(\mathbf{a}, \mathbf{s}, \mathbf{x}_L) = \sum_{\mathbf{h}} P(\mathbf{a} \mid \mathbf{h}) P(\mathbf{s} \mid \mathbf{h}) P(\mathbf{x}_L \mid \mathbf{h}) P(\mathbf{h}) \tag{3}$$

Although this expressions looks daunting due to the enormous number of possible hidden node sequences, it is in fact quite easy to evaluate this probability using the forward algorithm.[37]

How are these models used in practice? The aim of the models is not prediction, but sampling (Figure 7). They are used to generate plausible protein structures that are subsequently accepted or rejected in an MCMC procedure, using some kind of energy function $E_\theta(\mathbf{x}, \mathbf{a})$, with parameter vector θ, that takes into account local and nonlocal interactions, and possibly also experimental data. The sampling is done in

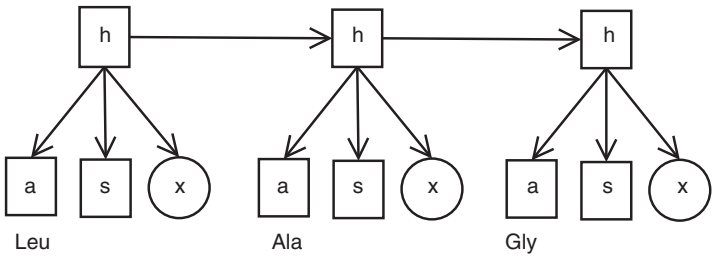

Figure 6 A probabilistic model of local protein structure. The model is a dynamic Bayesian network, that can be seen as an hidden Markov model with multiple outputs. The nodes of the graph represent variables, and the edges encode the conditional independencies between the variables. The boxes represent discrete nodes, and the circles represent continuous nodes. *h*: hidden node; *s*: secondary structure (helix, strand or coil); *a*: amino acid symbol (which can adopt 20 possible values); *x*: angle pair (Kent or bivariate von Mises distribution). Example sequence input is indicated (leucine, alanine, glycine) – the values of the other nodes can be sampled conditional upon the input values in an efficient way.

the following way: first, the observed nodes that represent the amino acid symbols are initialised with the values specified by the protein sequence of interest. Then, a sequence of hidden node values is sampled from $P(\mathbf{h} \mid \mathbf{a})$. This can be done very efficiently using the forward-backtrack algorithm.[38] Once the hidden node sequence is obtained, sampling a sequence of angles from $P(\mathbf{x}_L \mid \mathbf{h})$ is trivial. If information on secondary structure content is also available, sampling of the hidden node sequence can be done using $P(\mathbf{h} \mid \mathbf{a}, \mathbf{s})$.

There are many advantages attached to the use of these models. Unlike with fragment assembly, each sampled conformation comes with an attached probability, which makes rigorous Metropolis-Hastings sampling (respecting detailed balance) possible. It is also possible to resample a part of a conformation seamlessly, while talking the rest of the structure into account, which is impossible using fragment libraries. The models are flexible, and can also be used in structure design by sampling possible amino acid sequences that are compatible with a given local protein structure. In addition, sampling is fast: about 10 times faster than fragment assembly. In conclusion, given the advantages of this approach to conformational sampling, it can be expected that probabilistic methods will become predominant for this purpose in the near future.

4.2 Learning the parameters of energy functions

As mentioned in the introduction, the parameters of knowledge-based energy functions are optimised using the structural information present in the database of known structures. This is an important problem, which is currently only partly solved. Of the many approaches to solve it, two are by far the most popular.

First, one can optimise the energy difference between the native structure and a set of alternative conformations (typically called 'decoys'). This can be done using methods such as funnel sculpting,[39] Z-score optimisation,[40] linear programming [41] or even a brute force trial-and-error approach. The weak points of these methods are that the

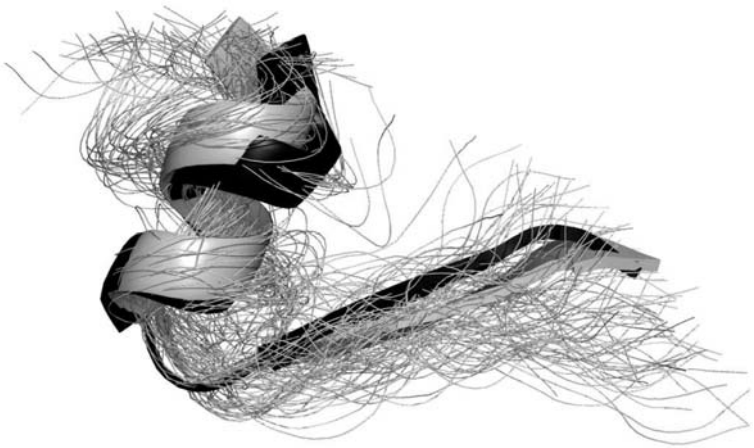

Figure 7 A set of samples generated using the full backbone model. The true structure is shown in black. The samples are shown as curves, and the consensus structure (the centroid of the set of samples) is shown in grey. The samples were generated using amino acid sequence and native secondary structure as input.

parameters might depend on the decoy set that is used, that the resulting energy function might have the shape of a golf course instead of the desired funnel shape,[42] or that the optimisation criteria are to a certain extent *ad hoc*.

One promising way to avoid some of these problems is to make use of Boltzmann learning, as recently proposed by Winther and Krogh.[43] Suppose the energy of a structure $\mathbf{x}$ for a certain sequence $\mathbf{a}$ is given by a parameterised energy function $E_\theta(\mathbf{x}, \mathbf{a})$, with θ being a vector of parameters. At thermal equilibrium, the probability for a protein of being in the native state $\mathbf{x}_N$ is given by:

$$\mathrm{P}(\mathbf{x}_N \mid \mathbf{a}, \theta) = \frac{\int_{\mathbf{x}_N} \exp\left(-\beta E_\theta(\mathbf{x}, \mathbf{a})\right) \mathrm{d}\mathbf{x}}{\int \exp\left(-\beta E_\theta(\mathbf{x}, \mathbf{a})\right) \mathrm{d}\mathbf{x}}$$

where $\beta = 1/kT$, with k being Boltzmann's constant and T the absolute temperature. The integral in the numerator runs over all conformations that make up the native state $\mathbf{x}_N$, while the integral in the denominator runs over the whole conformational space. Winther and Krogh[43] propose to optimise the parameters of the energy function by maximum likelihood:

$$\theta_{ML} = \operatorname{argmax} \sum_i \log P(\mathbf{x}_{N,i} \mid \mathbf{a}_i, \theta)$$

where the sum runs over all proteins in a training set. If the energy function is differentiable, one can optimise θ by simple gradient ascent:

$$\theta' = \theta + \eta \bigtriangledown_\theta \sum_i \log P(\mathbf{x}_{N,i} \mid \mathbf{a}_i, \theta) \tag{4}$$

where $\bigtriangledown$ is the gradient operator. The second term can be written as:

$$\eta\beta \sum_i \left\langle \bigtriangledown_\theta E_\theta(\mathbf{x}_i, \mathbf{a}_i) \right\rangle - \left\langle \bigtriangledown_\theta E_\theta(\mathbf{x}_i, \mathbf{a}_i) \right\rangle_{\mathbf{x}_{N,i}} \tag{5}$$

where $\langle\cdot\rangle$ is the expectation is over the whole conformational space, and $\langle\cdot\rangle_{\mathbf{x}_{N,i}}$ is the expectation over the native state of protein i. Although conceptually elegant, the method has some inherent problems. First, one needs to define the native conformational space, and that requires some *ad hoc* decisions. In the article where the method was proposed,[43] all structure within 1 Å root mean square deviation of the experimental structure $\mathbf{x}_E$ were considered to belong to the native state. Second, the first expectation requires sampling the whole conformational space, and that is very computationally expensive. Nonetheless, it was shown that good results can indeed be obtained.[43]

Podtelezhnikov and co-workers[44] propose a computationally efficient, approximative variant of the Krogh and Winter's approach. They approximate the gradient by starting from the experimental state $\mathbf{x}_E$, and letting the system evolve to new conformation $\mathbf{x}_K$

by a Monte Carlo method, using the current setting of the parameters of the energy function. The gradient can then be approximated by:

$$\nabla_\theta \log P(\mathbf{x} \mid \mathbf{a}, \theta) \simeq \nabla_\theta E_\theta(\mathbf{x}_K, \mathbf{a}, \theta) - \nabla_\theta E_\theta(\mathbf{x}_E, \mathbf{a}, \theta) \tag{6}$$

This approximation to full blown Boltzmann learning is due to Hinton,[45] and is called contrastive divergence. Contrastive divergence abolishes both the need for extensive sampling of the whole conformational space, which is computationally expensive, and avoids an *ad hoc* definition of the native state. Podtelezhnikov and co-workers[44] applied this technique to the optimisation of the parameters of a hydrogen bond energy, with convincing results.

Without any doubt, the most popular way to construct an energy function is to view the database of known structures as a Boltzmann distribution of interacting amino acids, and use this assumption to construct an energy potential.[46–48] The Boltzmann distribution applies to a system of particles, at a certain equilibrium temperature, which can adopt a number of different states, each with a certain energy. Statistical mechanics tells us that in this case, the chance of the occurrence of a state r with energy $E(r)$ is:

$$P(r) = \frac{1}{Z} \exp\left(\frac{-E(r)}{kT}\right) \tag{7}$$

and Z is a normalisation constant (the partition function) given by:

$$Z = \sum_r \exp\left(\frac{-E(r)}{kT}\right) \tag{8}$$

where the sum runs over all possible states. In simple words, for such a system, Boltzmann's equation provides us with a relation between the energy and the probability of a given state. Now, if one assumes that the pairwise interactions seen in the database of known structures for two given amino acid types follow a Boltzmann distribution, the following relation applies:

$$E(r; t_1, t_2) = -kT \log P(r \mid t_1, t_2) - kT \log Z \tag{9}$$

where $P(r \mid t_1, t_2)$ is the probability of observing a certain state r for an amino acid pair consisting of two fixed types t_1, t_2 (for example, $t_1 =$ alanine and $t_2 =$ glycine) in proteins. Then, one subtracts an 'average energy' E^* (with corresponding partition function Z^*) for all amino acid types,

$$E^*(r) = -kT \log P(r) - kT \log Z^* \tag{10}$$

which finally leads to the following expression for the energy difference:

$$\begin{aligned} \Delta E(r; t_1, t_2) &= E - E^* \\ &= -kT \log \frac{P(r \mid t_1, t_2)}{P(r)} - kT \log \frac{Z}{Z^*} \end{aligned} \tag{11}$$

The second term is a constant, and hence one ends up with a convenient way to estimate energy differences from observed amino acid pair geometries in known protein structures. The reasons for the subtraction of the reference energy are rather obscure from a physical point of view, and related to the estimation of a 'potential of mean force' from a radial distribution function,[49] but there is also a probabilistic explanation of this procedure (see below). Finally, the total energy of a protein conformation is then simply the sum of the pairwise energies:

$$\begin{aligned} \Delta E(\mathbf{x}, \mathbf{a}) &= \sum_{i<j} \Delta E(r_{ij}; t_1 = a_i, t_2 = a_j) \\ &= -kT \sum_{i<j} \log \frac{P(r_{ij} \mid t_1 = a_i, t_2 = a_j)}{P(r_{ij})} \end{aligned} \tag{12}$$

where r_{ij} somehow describes the state of the pair of amino acids a_i, a_j in a structure. The sum runs over all amino acid pairs. For the state r_{ij}, one typically chooses pairwise distances between amino acids at positions i and j. Typically, these distances are binned to discretise the problem, that is, to end up with a small, finite number of states. More elaborate descriptors can be used as well, for example including relative orientation.[50] The efficiency of the method greatly depends on how $P(r_{ij})$ is defined, which depends highly on a so-called *reference state* (see for example Liu *et al*).[51]

The weak point of the method is that the pairwise amino acid contacts in the PDB are viewed as some kind of Boltzmann distribution, which is clearly unjustified. Amino acid pairs in protein structures are not snapshots of a single system in thermal equilibrium. Discussions of the theoretical and practical problems associated with potentials of mean force are manifold.[49,52–54] Nonetheless, these potentials can be quite successful, although their success critically depends on the choice of the reference state, which is far from straightforward.

As shown by Simons *et al.*,[25] expressions that are very similar to the potentials of mean force described above can be obtained by formulating the problem using Bayes' theorem:

$$P(\mathbf{x} \mid \mathbf{a}) \propto P(\mathbf{a} \mid \mathbf{x}) P(\mathbf{x})$$

The prior, $P(\mathbf{x})$, can take into account features such as radius of gyration, packing of secondary structure elements and the presence of sterical clashes.[55] For the likelihood, an unfounded but convenient assumption of amino acid pair independence leads to the expression:

$$\begin{aligned} P(\mathbf{a} \mid \mathbf{x}) &= \prod_{i<j} P(a_i, a_j \mid r_{ij}) \\ &\propto \prod_{i<j} \frac{P(r_{ij} \mid a_i, a_j)}{P(r_{ij})} \end{aligned} \tag{13}$$

where r_{ij} is the distance between amino acids at positions i and j. This expression is equivalent to the expressions obtained by the potential of mean force method, and explains the use of the subtraction of the average energy which introduces the factor $P(r_{ij})^{-1}$.

For the Rosetta protein structure prediction method,[25] Baker and co-workers propose the following expression for the likelihood $P(\mathbf{a} \mid \mathbf{x})$:

$$P(\mathbf{a} \mid \mathbf{x}) \approx \prod_i P(a_i \mid e_i) \prod_{i<j} \frac{P(a_i, a_j \mid e_i, e_j, r_{ij})}{P(a_i \mid e_i, e_j, r_{ij}) P(a_j \mid e_i, e_j, r_{ij})} \tag{14}$$

where e_i is the solvent exposure of amino acid i. Although it is not mentioned in the article (where it is called an 'expansion'), this expression for the likelihood can be somewhat justified by viewing the graph of pairwise interactions in a protein as a polytree. As pointed out in section 3.2 (Equation (2)), in the case of a polytree, the likelihood is the product of the marginal pairwise probabilities (the edges of the tree, in this case amino acid pairs), divided by the product of the marginal individual probabilities (the nodes of the tree, in this case amino acids). This is essentially the form of the expression used by Rosetta[25] for the calculation of the conditional likelihood, as the first product and a subset of the factors in the denominator will almost cancel. An expression with a similar shape is obtained for pairwise Markov random fields, when using the Bethe free energy approximation (which becomes exact in this case).[56]

I will finish this section with an outlook on the future. A protein is essentially a graph that represents interacting amino acids, and many problems arising in the context of structure prediction can be seen as inference problems on graphs.[57–59] This brings the problem into the realm of graphical models such as Bayesian networks and Markov random fields. Recently, great progress has been made in understanding these models in a general framework,[22] and efficient, theoretically justified methods such as the Bethe and Kikuchi free energy approximations are now available for inference problems that were long regarded as intractable.[56,60,61] For example, the classic problem of minimum energy side chain placement on a given backbone structure can be conveniently reformulated as a problem of inference in a Markov random field, leading to improved solutions.[57,58] In general, methods based on generalised belief propagation can be used to calculate free energy estimates,[59] which makes them an attractive replacement of the currently used, poorly justified knowledge-based potentials. It will be exciting to see in the future how graph-based inference methods will be extended to the problem of *de novo* structure prediction.

4.3 Structure comparison

There are two problems in structural bioinformatics that have attracted an extraordinary amount of attention, and one of these is the optimal superposition of protein structures (the other problem is the prediction of secondary structure). Often, one

wants to compare two or more structures that for example differ in the presence of a bound ligand, the conformation of a flexible loop or, in the case of structures of different proteins, in chemical structure. The problem is a well-known topic in computational geometry: finding the translation and rotation to put two point sets on top of each other. Traditionally, one uses the unweighted least-squares method for this,[62] which assumes that the points are uncorrelated and that they have equal variances (that is, that they are homoscedastic). In this case, the optimal translation is obtained by moving all point sets to their centers of mass, and obtaining the optimal rotation using singular value decomposition. However, the correlation and variance assumptions are unfounded, as in the case of proteins the atoms are connected to each other with chemical bonds, and some regions in the protein may be highly mobile as compared to others.

Based on previous developments in shape theory,[81] Theobald and Wuttke[63,64] propose a statistical model where each structure $\mathbf{x}_i$ is seen as originating from a mean form $\mathbf{m}$. Conceptually, each point in $\mathbf{x}_i$ is obtained by perturbing the corresponding point in $\mathbf{m}$ with a Gaussian random error. Leaving out some details that are less important in the context of macromolecular structure superposition, the statistical model has the following form:

$$\mathbf{x}_i = (\mathbf{m} + \mathbf{E}_i)\mathbf{R}_i' - \mathbf{1}_k\mathbf{T}_i \tag{15}$$

where $\mathbf{R}_i$ is a 3×3 rotation matrix, $\mathbf{T}_i$ is a 1×3 row vector that represents the translational offset, and $\mathbf{1}_k$ is a $k \times 1$ vector of ones, with k equal to the number of atoms to be superimposed. The matrix $\mathbf{E}_i$ is distributed according to a *matrix normal distribution*[65,66] with zero mean, $N_{k,3}(0, \Sigma, I)$. The matrix Σ is a $k \times k$ covariance matrix, which described the variances of, and among, the atoms. Estimation of Σ requires constraining its parameters, which can be done by assuming that its eigenvalues are distributed according to an inverse gamma distribution, with parameters α and γ. The two parameters are set to a point estimate determined from the data, and the procedure can thus be interpreted as a shrinkage estimator for Σ. The maximum likelihood estimates for $\mathbf{R}_i, \mathbf{T}_i, \mathbf{m}, \Sigma$ and the parameters of the inverse gamma distribution α, γ can be found using a numerical algorithm.[64] The method is not only of theoretical importance, but leads to improved performance in real-life superposition problems. An implementation of the algorithm, called Theseus, is available.[64]

The above method assumes that the point sets to be superimposed have a simple one-to-one correspondence between their points, that is known in advance. This is often not the case, as proteins typically contain insertions or deletions. Simple one-to-one correspondence is also lacking in the related problem of matching of active sites of proteins. These sites are localised collections of atoms that fulfill the protein's chemical function, and similarities among these sites is an important way to infer a possible function for unknown proteins. Here, it is also important to decide whether the similarity between two sites is statistically significant. Green and Mardia[67] propose a Bayesian hierarchical model, to deal with this case. The model can be seen as a Bayesian network (Figure 8). Here, the two point sets $\mathbf{x} = \{x_1, \ldots, x_n\}$

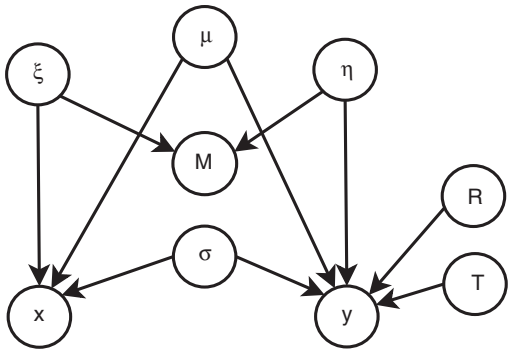

Figure 8 Green and Mardia's model[67] for active site superposition, represented as a Bayesian network. $\mathbf{x}$, $\mathbf{y}$: point sets to be matched; μ: the set of true locations from which $\mathbf{x}$ and $\mathbf{y}$ are derived; ξ, η: indexing arrays that match points in μ to points in $\mathbf{x}$, $\mathbf{y}$, respectively; $\mathbf{M}$: matching matrix, where $\mathbf{M}[j, k] = 1$ if x_j and y_k are derived from the same true location in μ, and zero otherwise; σ covariance of the Gaussian distribution that describes the noisy observation of $\mathbf{x}$ and $\mathbf{y}$ from μ; $\mathbf{R}$, $\mathbf{T}$: rotation and translation applied to all points in $\mathbf{y}$. Figure adapted from Green and Mardia.[67]

and $\mathbf{y} = \{y_1, \ldots, y_m\}$ (with $\mathbf{x}$ and $\mathbf{y}$ each representing a macromolecular structure, and $x_1, \ldots, x_n, y_1, \ldots, y_n$ being 3D vectors) are viewed as Gaussian perturbations of a set of true locations μ:

$$\begin{aligned} x_j &= N_3(\mu_{\xi[j]}, \sigma^2 I) \\ \mathbf{R}y_j + \mathbf{T} &= N_3(\mu_{\eta[j]}, \sigma^2 I) \end{aligned} \tag{16}$$

where N_3 is a 3D Gaussian distribution with a spherical covariance matrix, ξ and η are indexing arrays that match point in μ to points in $\mathbf{x}$ and $\mathbf{y}$, respectively. To the points in set $\mathbf{y}$, a rotation $\mathbf{R}$ and translation $\mathbf{T}$ is applied. The matching of the configurations is represented by the matching matrix $\mathbf{M}$, where $\mathbf{M}[j, k]$ is one if x_j and y_k are derived from the same point in μ, and zero otherwise.

The set of true locations μ is assumed to be generated by a *spatial Poisson process* with rate λ. In a spatial Poisson process, the number of points in a volume v follows a Poisson distribution with mean λv, and the point counts are independent for disjoint volumes.[68] The parameter λ can be interpreted as a point density.

Some of the true locations in μ give rise to points in $\mathbf{x}$ or $\mathbf{y}$, others to points in both $\mathbf{x}$ and $\mathbf{y}$, and some are not observed at all, with probabilities p_x, p_y, $1 - p_x - p_y - \rho p_x p_y$ and $\rho p_x p_y$, respectively. The parameter ρ controls the tendency for points to be matched: if close to zero, most true locations will match points in $\mathbf{x}$ or $\mathbf{y}$, or both. The joint model is:

$$P(\mathbf{M}, \mathbf{R}, \mathbf{T}, \sigma, \mathbf{x}, \mathbf{y}) \propto |\mathbf{R}|^n P(\mathbf{R}) P(\mathbf{T}) P(\sigma) \prod_{j,k:M[j,k]=1} \frac{\rho\phi\left(\frac{1}{\sigma\sqrt{2}}(x_j - \mathbf{R}y_k - \mathbf{T})\right)}{\lambda(\sigma\sqrt{2})^3} \tag{17}$$

where ϕ is the standard normal density in $\mathbb{R}^3$. Green and Mardia develop an MCMC approach to sample $\mathbf{M}$, $\mathbf{R}$, $\mathbf{T}$ and σ given point sets $\mathbf{x}$ and $\mathbf{y}$, for fixed ρ and λ. Some applications of this model, including partial labelling that causes some points to be matched more likely than others, are further explored in a series of articles.[67,69,70]

5 Inferential structure determination

High-resolution experimental structures of proteins and RNA are obtained from NMR and X-ray crystallography experiments. Typically, a structure is obtained from the data by minimisation of a so-called hybrid energy:

$$E(\mathbf{x}, \mathbf{a}) = E_\theta(\mathbf{x}, \mathbf{a}) + w_{\text{data}} E_{\text{data}}(\mathbf{x}, \mathbf{a}) \tag{18}$$

where the first energy term ensures that the model $\mathbf{x}$, $\mathbf{a}$ respects the chemical and physical constraints, while the second term measures the disagreement between the data and the model. The weight term w_{data} controls the relative contributions of the two energies, which needs to be determined by *ad hoc* methods or, more rigorously, by cross validation.[71]

Of course, macromolecular structure determination is a classic example of statistical inference, and should ideally be treated as such.[72] In this view, one needs a data likelihood $P(\mathbf{d} \mid \mathbf{x}, \xi, \mathbf{a})$ that quantifies the probability of observing the data $\mathbf{d}$ given the structural model $\mathbf{x}$, $\mathbf{a}$ and set of nuisance parameters ξ, a prior $P(\mathbf{x}, \mathbf{a})$ that brings in chemical, physical and any other information regarding the structure, and a prior over the nuisance parameters $P(\xi)$:

$$P(\mathbf{x}, \xi \mid \mathbf{d}, \mathbf{a}) \propto P(\mathbf{d} \mid \mathbf{x}, \xi, \mathbf{a}) P(\mathbf{x}, \mathbf{a}) P(\xi) \tag{19}$$

The nuisance parameters ξ typically deals with experimental errors and uncertainties, and can be integrated out, or included in the posterior.

In NMR, this approach to structure determination has been especially fruitful.[72–74] In Rieping *et al.*,[72] a data likelihood is constructed by treating the data as independent measurements and model the deviations between observed and calculated distances using the log-normal distribution. In this case, the nuisance parameter ξ is the standard deviation σ of the log-normal distribution. For the prior on σ, Jeffreys' rule is applied, resulting in $P(\sigma) = \sigma^{-1}$. For the structure prior $P(\mathbf{x}, \mathbf{a})$, Boltzmann's equation is applied to a physical energy function E_θ:

$$P(\mathbf{x}, \mathbf{a}) \propto \exp\left(-\frac{1}{kT} E_\theta(\mathbf{x}, \mathbf{a})\right) \tag{20}$$

An ensemble of structures is obtained by generating samples from the posterior distribution using a Monte Carlo method. Using a Gibbs sampler scheme, samples of $\mathbf{x}$ and σ are drawn from the joint posterior in an iterative fashion. The method is implemented in the ISD program.[72–74]

In macromolecular X-ray crystallography, the situation is less favourable. Although there is in principle enough information available from physics, chemistry and diffraction data to solve the problem in a fully probabilistic framework as outlined above, such a method for X-ray crystallography has not yet been constructed. This is commonly attributed to the large number of correlations between the relevant variables.[75,76] Probabilistic methods are, however, widely applied to many subproblems in the structure determination process.[75,76]

Recently, small angle X-ray crystallography (SAXS) is becoming increasingly popular to obtain low resolution information on molecular structure.[77] In SAXS, the data is a diffraction pattern obtained from the macromolecule in solution. This type of data has a much lower information content than NMR or X-ray data, but is typically much easier to obtain. In simple words, the SAXS profile is the Fourier transform of the electron density distribution of the solvent/macromolecule system, and it contains information on the inter-atomic distances that are present in the macromolecule. The histogram of these distances can be obtained by an *indirect Fourier transform* (IFT) of the data, which requires the use of hyper parameters that ensure the smoothness of the histogram. Hansen proposed a Bayesian approach to the IFT calculation of this histogram, using a Bayesian treatment of the hyper parameters.[78,79] For protein structure determination from SAXS data, the non-probabilistic hybrid energy approach is typically used (see for example Petokhov *et al.*[80]). A direct Bayesian approach to structure determination from SAXS data, incorporating a prior on structure and a suitable likelihood in the spirit of the inferential structure determination method as outlined above, is as far as I know not developed yet.

6 Conclusions

From the developments highlighted above, it should be clear that probabilistic models and machine learning methods based on Bayesian principles are leading to substantial progress, and more success stories are to be expected. Let me end with highlighting a few of the many open challenges. A sound, computationally efficient probabilistic model of the nonlocal interaction in proteins and RNA is still lacking. The problem lies in capturing sufficient molecular detail, such as for example the interactions between multiple amino acids, in a tractable way. Such a model would be of tremendous potential for the prediction of macromolecular structure, and would probably also lead to additional insight in the protein folding process itself. In protein structure comparison, a probabilistic model that handles insertions and deletions in a rigorous and efficient way is high on the priority list. Finally, in inferential structure determination, methods are needed that go beyond the assumption of independence of the individual data observations (for the likelihood) or structural features (for the prior). The model of local structure discussed in 4.1, for example, would provide an excellent prior that obviously goes beyond the naive assumption of independence. Clearly, the field of structural bioinformatics provides a fertile hunting ground for the adventurous probabilist looking for challenging yet rewarding problems.

Acknowledgements

I acknowledge funding by the Danish *Program Commission on Nanoscience, Biotechnology and IT (NABIIT)* and the *Danish Research Council for Technology and Production Sciences (FTP)*. I thank my collaborators Wouter Boomsma, Jesper Ferkinghoff-Borg, John T. Kent, Anders Krogh and Kanti V. Mardia.

References

1 Maddox J. *What Remains to be Discovered*. London, UK, Macmillan; 1998.
2 Kuhlman B, Dantas G, Ireton G, Varani G, Stoddard B, Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science* 2003; **302**: 1364–8.
3 Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences USA* 2007; **104**: 14664–9.
4 Jiang L, Althoff E, Clemente F. *et al.* De novo computational design of retro-aldol enzymes. *Science* 2008; **319**: 1387–91.
5 Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 2008; **452**: 51–5.
6 Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. *Proteins* 2007; **69** Suppl 8: 57–67.
7 Karplus M, McCammon J. Molecular dynamics. *Nature Structural Biology* 2002; **9**: 646–52.
8 Jaynes E. Information theory and statistical mechanics. *Physical Review* 1957; **106**: 620–30.
9 Jaynes E. Information theory and statistical mechanics. II. *Physical Review* 1957; **108**: 171–90.
10 Jaynes E. *Probability Theory: The Logic of Science*. Cambridge, UK, Cambridge University Press; 2003.
11 Anfinsen C. Principles that govern the folding of protein chains. *Science* 1973; **181**: 223–30.
12 Chandler D. Interfaces and the driving force of hydrophobic assembly. *Nature* 2005; **437**: 640–47.
13 Eddy SR. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* 2001; **2**: 919–29.
14 Mardia K, Jupp P. *Directional Statistics*. New York, Wiley; 2000.
15 Kent J. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society B* 1982; **44**: 71–80.
16 Banerjee A, Dhillon I, Ghosh J, Sra S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research* 2006; **6**: 1345.
17 Mardia K, Taylor C, Subramaniam G. Bivariate von Mises densities for angular data with applications to protein bioinformatics. *Biometrics* 2007; **63**: 505–12.
18 Bingham C. An antipodally symmetric distribution on the sphere. *The Annals of Statistics* 1974; **2**: 1201–25.
19 Khatri C, Mardia K. The Von Mises-Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society B* 1977; **39**: 95–106.
20 Lassen N, Jensen D, Conradsen K. On the statistical analysis of orientation data. *Acta Crystallographica A* 1994; **108**: 741–48.
21 Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Fracisco, California, USA, Morgan Kaufmann; 1988.
22 Kschischang F, Frey B, Loeliger H. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* 2001; **47**: 498–519.
23 Ghahramani Z. Learning dynamic Bayesian networks. *Lecture Notes in Computer Science* 1997; **1387**: 168–97.
24 Jones T, Thirup S. Using known substructures in protein model building and crystallography. *European Molecular Biology Organization Journal* 1986; **5**: 819–22.
25 Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using

simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology* 1997; **268**: 209–25.
26 Bystroff C, Simons K, Han K, Baker D. Local sequence-structure correlations in proteins. *Current Opinion in Biotechnology* 1996; **7**: 417–21.
27 Chikenji G, Fujitsuka Y, Takada S. Shaping up the protein folding funnel by local interaction: Lesson from a structure prediction study. *Proceedings of the National Academy of Sciences USA* 2006; **103**: 3141–6.
28 Hamelryck T, Kent J, Krogh A. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology* 2006; **2**(9):e131.
29 Chikenji G, Fujitsuka Y, Takada S. A reversible fragment assembly method for de novo protein structure prediction. *Journal of Chemical Physics* 2003; **119**: 6895–903.
30 Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology* 1976; **104**: 59–107.
31 Oldfield T. A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. *Acta Crystallographica D* 2001; **57**: 82–94.
32 Ramachandran G, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* 1963; **7**: 95–9.
33 Kent J, Hamelryck T. Using the Fisher-Bingham distribution in stochastic models for protein structure. In Barber S, Baxter P, Mardia K, Walls R, eds. *Quantitative Biology, Shape Analysis, and Wavelets*, Vol. 24. Leeds, UK, Leeds University Press; 2005; L57–60.
34 Boomsma W, Kent J, Mardia K, Taylor C, Hamelryck T. Graphical models and directional statistics capture protein structure. In Barber S, Baxter P, Mardia K, Walls R, eds. *Interdisciplinary Statistics and Bioinformatics*, Vol. 25. Leeds, UK, Leeds University Press; 2006; 91–4.
35 Boomsma W, Mardia K, Taylor C, Ferkinghoff-Borg J, Krogh A, Hamelryck T. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences USA* 2008; **105**: 8932–7.
36 Diebolt J, Ip E. Stochastic EM: method and application. In Gilks WR, Richardson S, Spiegelhalter DJ, eds. *Markov Chain Monte Carlo in practice*, chap. 15. Chapman & Hall/CRC, 1996; 259–73.
37 Durbin R, Eddy SR, Krogh A, Mitchison G. *Biological sequence analysis*. Cambridge, UK, Cambridge University Press; 1998.
38 Cawley S, Pachter L. HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics* 2003; **19** Suppl 2: II36–II41.
39 Fain B, Levitt M. Funnel sculpting for in silico assembly of secondary structure elements of proteins. *Proceedings of the National Academy of Sciences USA* 2003; **100**: 10700–705.
40 Mirny L, Shakhnovich E. How to derive a protein folding potential? A new approach to an old problem. *Journal of Molecular Biology* 1996; **264**: 1164–79.
41 Loose C, Klepeis J, Floudas C. A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins* 2004; **54**: 303–14.
42 Bryngelson J, Wolynes P. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences USA* 1987; **84**: 7524–8.
43 Winther O, Krogh A. Teaching computers to fold proteins. *Physical Reviews E* 2004; **70**: 030903.
44 Podtelezhnikov A, Ghahramani Z, Wild D. Learning about protein hydrogen bonding by minimizing contrastive divergence. *Proteins* 2007; **66**: 588–99.
45 Hinton G. Training products of experts by minimizing contrastive divergence. *Neural Computation* 2002; **14**: 1771–800.
46 Tanaka S, Scheraga H. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 1976; **9**: 945–50.
47 Miyazawa S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 1985; **18**: 534–52.
48 Sippl M. Calculation of conformational ensembles from potentials of mean force. *Journal of Molecular Biology* 1990; **213**: 859–83.
49 Koppensteiner W, Sippl M. Knowledge-based potentials – back to the roots. *Biochemistry (Moscow)* 1998; **63**: 247–52.

50 Buchete N, Straub J, Thirumalai D. Development of novel statistical potentials for protein fold recognition. *Current Opinion in Structural Biology* 2004; **14**: 225–32.

51 Liu S, Zhang C, Zhou H, Zhou Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins* 2004; **56**: 93–101.

52 Thomas P, Dill K. Statistical potentials extracted from protein structures: how accurate are they? *Journal of Molecular Biology* 1996; **257**: 457–69.

53 Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. *Journal of Chemical Physics* 1998; **109**: 11101–108.

54 Moult J. Comparison of database potentials and molecular mechanics force fields. *Current Opinion in Structural Biology* 1997; **7**: 194–99.

55 Simons K, Ruczinski I, Kooperberg C, Fox B, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 1999; **34**: 82–95.

56 Yedidia J, Freeman W, Weiss Y. Bethe free energy, Kikuchi approximations and belief propagation algorithms. Tech. Rep., Mitsubishi Electric Research Laboratories Technical Report, TR-2001-16, 2001.

57 Yanover C, Weiss Y. Approximate inference and protein-folding. In Becker S, ST, Obermayer K, eds. *Advances in Neural Information Processing Systems 15*, Cambridge, MA, MIT Press; 2003; 1457–64.

58 Yanover C, Meltzer T, Weiss Y. Linear programming relaxations and belief propagation – an empirical study. *The Journal of Machine Learning Research* 2006; **7**: 1887–907.

59 Kamisetty H, Xing E, Langmead C. Free energy estimates of all-atom protein structures using generalized belief propagation. *Journal of Computational Biology* 2008; **15**: 755–66.

60 Yedidia J, Freeman W, Weiss Y. Generalized belief propagation. Tech. Rep., Mitsubishi Electric Research Laboratories Technical Report, TR-2000-26, 2000.

61 Yedidia J, Freeman W, Weiss Y. Understanding belief propagation and its generalizations. Tech. Rep., Mitsubishi Electric Research Laboratories Technical Report, TR-2001-22, 2002.

62 Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica A* 1976; **32**: 922–3.

63 Theobald D, Wuttke D. Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proceedings of the National Academy of Sciences USA* 2006; **103**: 18521–7.

64 Theobald D, Wuttke D. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* 2006; **22**: 2171–2.

65 Dutilleul P. The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* 1999; **64**: 105–23.

66 Dutilleul P, Pinel-alloul B. A doubly multivariate model for environmental data. *Environmetrics* 1996; **7**: 551–65.

67 Green P, Mardia K. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* 2006; **93**: 235–54.

68 Baddeley A. Spatial point processes and their applications. *Lecture Notes in Mathematics* 2007; **1892**: 1–75.

69 Mardia K, Green P, Nyirongo V, Gold N, Westhead D. Bayesian refinement of protein functional site matching. *BMC Bioinformatics* 2007; **8**: 257.

70 Davies J, Jackson R, Mardia K, Taylor C. The Poisson index: A new probabilistic model for protein-ligand binding site similarity. *Bioinformatics* 2007; **23**: 3001–08.

71 Brünger A. Free R value: A novel statistical quantity for assessing the accuracy of structures. *Nature* 1992; **355**: 472–5.

72 Rieping W, Habeck M, Nilges M. Inferential structure determination. *Science* 2005; **309**: 303–06.

73 Habeck M, Nilges M, Rieping W. Bayesian inference applied to macromolecular structure determination. *Physical Review E* 2005; **72**: 31912.

74 Habeck M, Rieping W, Nilges M. Weighting of experimental evidence in macromolecular structure determination. *Proceedings of the National Academy of Sciences USA* 2006; **103**: 1756–61.

75 McCoy A. New applications of maximum likelihood and Bayesian statistics in macromolecular crystallography. *Current Opinion in Structural Biology* 2002; **12**: 670–73.

76 McCoy A. Liking likelihood. *Acta Crystallographica D* 2004; **60**: 2169–83.
77 Svergun D, Koch M. Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics* 2003; **66**: 1735–82.
78 Hansen S. Bayesian estimation of hyperparameters for indirect Fourier transformation in small-angle scattering. *Journal of Applied Crystallography* 2000; **33**: 1415–21.
79 Vestergaard B, Hansen S. Application of Bayesian analysis to indirect Fourier transformation in small-angle scattering. *Journal of Applied Crystallography* 2006; **39**: 797–804.
80 Petoukhov M, Eady N, Brown K, Svergun D. Addition of missing loops and domains to protein models by X-ray solution scattering. *Biophysical Journal* 2002; **83**: 3113–25.
81 Dryden IL, Mardia K. *Statistical Shape Analysis*. New York, Wiley, 1998.