Potentials of mean force for protein structure prediction vindicated, formalized and generalized

Hamelryck, Thomas; Borg, Mikael; Paluszewski, Martin; Paulsen, Jonas; Frellsen, Jes; Andreetta, Christian; Boomsma, Wouter Krogh; Bottaro, Sandro; Ferkinghoff-Borg, Jesper

Published in:
PLoS ONE

DOI:
10.1371/journal.pone.0013714

Publication date:
2010

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Abstract
Understanding protein structure is of crucial importance in science, medicine and biotechnology. For about two decades, knowledge-based potentials based on pairwise distances—so-called “potentials of mean force” (PMFs)—have been center stage in the prediction and design of protein structure and the simulation of protein folding. However, the validity, scope and limitations of these potentials are still vigorously debated and disputed, and the optimal choice of the reference state—a necessary component of these potentials—is an unsolved problem. PMFs are loosely justified by analogy to the reversible work theorem in statistical physics, or by a statistical argument based on a likelihood function. Both justifications are insightful but leave many questions unanswered. Here, we show for the first time that PMFs can be seen as approximations to quantities that do have a rigorous probabilistic justification: they naturally arise when probability distributions over different features of proteins need to be combined. We call these quantities “reference ratio distributions” deriving from the application of the “reference ratio method.” This new view is not only of theoretical relevance but leads to many insights that are of direct practical use: the reference state is uniquely defined and does not require external physical insights; the approach can be generalized beyond pairwise distances to arbitrary features of protein structure; and it becomes clear for which purposes the use of these quantities is justified. We illustrate these insights with two applications, involving the radius of gyration and hydrogen bonding. In the latter case, we also show how the reference ratio method can be iteratively applied to sculpt an energy funnel. Our results considerably increase the understanding and scope of energy functions derived from known biomolecular structures.
Despite the progress in methodology and theory, and the dramatic increase in the number of experimentally determined protein structures, the accuracy of the energy functions still remains the main obstacle to accurate protein structure prediction [22,30,31]. Recently, several groups demonstrated that it is the quality of the coarse grained energy functions [10], rather than inadequate sampling, that impairs the successful prediction of the native state [30,31]. The insights presented in this article point towards a new, theoretically well-founded way to construct and refine energy functions, and thus address a timely problem.

We start with an informal outline of the general ideas presented in this article, and then analyze two notable attempts in the literature to justify PMFs. We point out their shortcomings, and subsequently present a rigorous probabilistic explanation of the strengths and shortcomings of traditional pairwise distance PMFs. This explanation sheds a surprising new light on the nature of the reference state, and allows the generalization of PMFs beyond pairwise distances in a statistically valid way. Finally, we demonstrate our method in two applications involving protein compactness and hydrogen bonding. In the latter case, we also show that PMFs can be iteratively optimized, thereby effectively sculpting an energy funnel [24,32–36].

Results and Discussion

Overview

In order to emphasize the practical implications of the theoretical insights that we present here, we start with a very concrete example that illustrates the essential concepts (see Fig. 1). Currently, protein structure prediction methods often make use of fragment libraries: collections of short fragments derived from known protein structures in the Protein Data Bank (PDB). By assembling a suitable set of fragments, one obtains conformations that are protein-like on a local length scale. That is, these conformations typically lack non-local features that characterize real proteins, such as a well-packed hydrophobic core or an extensive hydrogen bond network. Such aspects of protein structure are not, or only partly, captured by fragment libraries.

Formally, a fragment library specifies a probability distribution \( Q(X) \), where \( X \) is for example a vector of dihedral angles. In order to obtain conformations that also possess the desired non-local features, \( Q(X) \) needs to be complemented with another probability distribution \( P(Y) \), with \( Y \) being for example a vector of pairwise distances, the radius of gyration, the hydrogen bonding network, or any combination of non-local features. Typically, \( Y \) is a deterministic function of \( X \); we use the notation \( Y(X) \) when necessary.

For the sake of argument, we will focus on the radius of gyration \( r_g \) at this point; in this case \( Y(X) \) becomes \( r_g(X) \). We assume that a suitable \( P(r_g) \) was derived from the set of known protein structures; without loss of generality, we leave out the dependency on the amino acid sequence for simplicity. The problem that we address in this article can be illustrated with the following question: how can we combine \( P(r_g) \) and \( Q(X) \) in a rigorous, meaningful way? In other words, we want to use the fragment library to sample conformations whose radii of gyration \( r_g \) are distributed according to \( P(r_g) \). These conformations should display a realistic local structure as well, reflecting the use of the fragment library. Simply multiplying \( P(r_g(X)) \) and \( Q(X) \) does not lead to the desired result, as \( X \) and \( r_g \) are not independent; the resulting conformations will not be distributed according to \( P(r_g) \).

The solution is given in Fig. 1; it involves the probability distribution \( Q_R(r_g) \), the probability distribution over the radius of gyration for conformations sampled solely from the fragment library. The subscript \( R \) stands for reference state as will be explained below. The solution generates conformations whose radii of gyration are distributed according to \( P(r_g) \). The influence of \( Q(X) \) is apparent in the fact that for conformations with a given \( r_g \), their local structure \( X \) will be distributed according to \( Q(X|r_g) \). The

![Figure 1. Illustration of the central idea presented in this article.](image)

In this example, the goal is to sample conformations with a given distribution \( P(r_g) \) for the radius of gyration \( r_g \), and a plausible local structure. \( P(r_g) \) could, for example, be derived from known structures in the Protein Data Bank (PDB, left box). \( Q(X) \) is a probability distribution over local structure \( X \), typically embodied in fragment library (right box). In order to combine \( Q(X) \) and \( P(r_g) \) in a meaningful way (see text), the two distributions are multiplied and divided by \( Q_R(r_g) \) (formula at the bottom); \( Q_R(r_g) \) is the probability distribution over the radius of gyration for conformations sampled solely from the fragment library (that is, \( Q(X) \)). The probability distribution \( P(X) \) will generate conformations with plausible local structures (due to \( Q(X) \)), while their radii of gyration will be distributed according to \( P(r_g) \), as desired. This simple idea lies at the theoretical heart of the PMF expressions used in protein structure prediction.

doi:10.1371/journal.pone.0013714.g001

Reference Ratio Distributions
latter distribution has a clear interpretation: it corresponds to sampling an infinite amount of conformations from a fragment library, and retaining only those with the desired \( r_g \). Note that even if we chose the uniform distribution for \( Q(X) \), the resulting \( Q_R(r_g) \) will not (necessarily) be uniform.

Intuitively, \( P(r_g) \) provides correct information about the radius of gyration, but no information about local structure; \( Q(X) \) provides approximately correct information about the structure of proteins on a local length scale, but is incorrect on a global scale (leading to an incorrect probability distribution for the radius of gyration); finally, the formula shown in Fig. 1 merges these two complementary sources of information together. Another viewpoint is that \( P(r_g) \) and \( Q(r_g) \) are used to correct the shortcomings of \( Q(X) \). This construction is statistically rigorous, provided that \( P(r_g) \) and \( Q(X) \) are proper probability distributions.

After this illustrative example, we now review the use of PMFs in protein structure prediction, and discuss how PMFs can be understood and generalized in the theoretical framework that we briefly outlined here.

Pairwise PMFs for protein structure prediction

Many textbooks present PMFs as a simple consequence of the Boltzmann distribution, as applied to pairwise distances between amino acids. This distribution, applied to a specific pair of amino acids, is given by:

\[
P(r) = \frac{1}{Z} e^{-\frac{F(r)}{kT}}
\]

where \( r \) is the distance, \( k \) is Boltzmann’s constant, \( T \) is the temperature and \( Z \) is the partition function, with \( Z = \int e^{-\frac{F(r)}{kT}} dr \).

The quantity \( F(r) \) is the free energy assigned to the pairwise system. Simple rearrangement results in the inverse Boltzmann formula, which expresses the free energy \( F(r) \) as a function of \( P(r) \):

\[
F(r) = -kT \ln P(r) - kT \ln Z
\]

To construct a PMF, one then introduces a so-called reference state with a corresponding distribution \( Q_R \) and partition function \( Z_R \), and calculates the following free energy difference:

\[
\Delta F(r) = -kT \ln \frac{P(r)}{Q_R(r)} - kT \ln \frac{Z}{Z_R}
\]

The reference state typically results from a hypothetical system in which the specific interactions between the amino acids are absent [7]. The second term involving \( Z \) and \( Z_R \) can be ignored, as it is a constant.

In practice, \( P(r) \) is estimated from the database of known protein structures, while \( Q_R(r) \) typically results from calculations or simulations. For example, \( P(r) \) could be the conditional probability of finding the \( C\beta \) atoms of a valine and a serine at a given distance \( r \) from each other, giving rise to the free energy difference \( \Delta F \). The total free energy difference of a protein, \( \Delta F_{TOT} \), is then claimed to be the sum of all the pairwise free energies:

\[
\Delta F_{TOT} = \sum_{i<j} \Delta F(r_{ij}|a_i,a_j)
\]

where the sum runs over all amino acid pairs \( a_i,a_j \) (with \( i<j \)) and \( r_{ij} \) is their corresponding distance. It should be noted that in many studies \( Q_R \) does not depend on the amino acid sequence [11].

Intuitively, it is clear that a low free energy difference indicates that the set of distances in a structure is more likely in proteins than in the reference state. However, the physical meaning of these PMFs have been widely disputed since their introduction [2,12–15]. Indeed, why is it at all necessary to subtract a reference state energy? What is the optimal reference state? Can PMFs be generalized and justified beyond pairwise distances, and if so, how? Before we discuss and clarify these issues, we discuss two qualitative justifications that were previously reported in the literature: the first based on a physical analogy, and the second using a statistical argument.

PMFs from the reversible work theorem

The first, qualitative justification of PMFs is due to Sippl, and based on an analogy with the statistical physics of liquids [37]. For liquids [8,9,13,14,37], the potential of mean force is related to the pair correlation function \( g(r) \), which is given by:

\[
g(r) = \frac{P(r)}{Q_R(r)}
\]

where \( P(r) \) and \( Q_R(r) \) are the respective probabilities of finding two particles at a distance \( r \) from each other in the liquid and in the reference state. For liquids, the reference state is clearly defined; it corresponds to the ideal gas, consisting of non-interacting particles. The two-particle potential of mean force \( W(r) \) is related to \( g(r) \) by:

\[
W(r) = -kT \log g(r) = -kT \log \frac{P(r)}{Q_R(r)}
\]

According to the reversible work theorem, the two-particle potential of mean force \( W(r) \) is the reversible work required to bring two particles in the liquid from infinite separation to a distance \( r \) from each other [8,9].

Sippl justified the use of PMFs – a few years after he introduced them for use in protein structure prediction [7] – by appealing to the analogy with the reversible work theorem for liquids [37]. For liquids, \( g(r) \) can be experimentally measured using small angle X-ray scattering; for proteins, \( P(r) \) is obtained from the set of known protein structures, as explained in the previous section. The analogy described above might provide some physical insight, but, as Ben-Naim writes in a seminal publication [13]: “the quantities, referred to as ‘statistical potentials,’ ‘structure based potentials,’ or ‘pair potentials of mean force,’ as derived from the protein database, are neither ‘potentials’ nor ‘potentials of mean force,’ in the ordinary sense as used in the literature on liquids and solutions.”

Another issue is that the analogy does not specify a suitable reference state for proteins. This is also reflected in the literature on statistical potentials; the construction of a suitable reference state continues to be an active research topic [3,22,38–41]. In the next section, we discuss a second, more recent justification that is based on probabilistic reasoning.

PMFs from likelihoods

Baker and co-workers [18] justified PMFs from a Bayesian point of view and used these insights in the construction of the coarse
grained ROSETTA energy function; Samudrala and Moult used similar reasoning for the RAPDF potential [42]. According to Bayesian probability calculus, the conditional probability \( P(X|A) \) of a structure \( X \), given the amino acid sequence \( A \), can be written as:

\[
P(X|A) = \frac{P(A|X)P(X)}{P(A)} \propto P(A|X)P(X)
\]

\( P(X|A) \) is proportional to the product of the likelihood \( P(A|X) \) times the prior \( P(X) \). By assuming that the likelihood can be approximated as a product of pairwise probabilities, and applying Bayes' theorem, the likelihood can be written as:

\[
P(A|X) \approx \prod_{i<j} P(a_i,a_j|r_{ij}) \propto \prod_{i<j} \frac{P(r_{ij}|a_i,a_j)}{P(r_{ij})}
\]  

(5)

where the product runs over all amino acid pairs \( a_i,a_j \) (with \( i<j \)), and \( r_{ij} \) is the distance between amino acids \( i \) and \( j \). Obviously, the negative of the logarithm of expression (5) has the same functional form as the classic pairwise distance PMFs, with the denominator playing the role of the reference state in Eq. 1. The merit of this explanation is the qualitative demonstration that the functional form of a PMF can be obtained from probabilistic reasoning. Although this view is insightful – it rightfully drew the attention to the application of Bayesian methods to protein structure prediction – there is a more quantitative explanation, which does not rely on the incorrect assumption of pairwise decomposability [12–14,43], and leads to a different, quantitative conclusion regarding the nature of the reference state. This explanation is given in the next section.

A general statistical justification for PMFs

Expressions that resemble PMFs naturally result from the application of probability theory to solve a fundamental problem that arises in protein structure prediction: how to improve an imperfect probability distribution \( Q(X) \) over a first variable \( X \) using a probability distribution \( P(Y) \) over a second variable \( Y \) (see Fig. 2, Fig. 1 and Materials and Methods). We assume that \( Y \) is a deterministic function of \( X \); we write \( Y(X) \) when necessary. In that case, \( X \) and \( Y \) are called fine and coarse grained variables, respectively. When \( Y \) is a function of \( X \), the probability distribution \( Q(Y) \) automatically implies a probability distribution \( Q(X,Y(X)) \). This distribution has some unusual properties: \( Q(X,Y(X)) = Q(X) \); and if \( Y' \neq Y(X) \), it follows that \( Q(X,Y') = 0 \).

Typically, \( X \) represents local features of protein structure (such as backbone dihedral angles), while \( Y \) represents nonlocal features (such as hydrogen bonding, compactness or pairwise distances). However, the same reasoning also applies to other cases; for example, \( P(Y) \) could represent information coming from experimental data, and \( Q(X) \) could be embodied in an empirical force field as used in molecular mechanics [2,44] (see Fig. 2).

Typically, the distribution \( Q(X) \) in itself is not sufficient for protein structure prediction: it does not consider important nonlocal features such as hydrogen bonding, compactness or favorable amino acid interactions. As a result, \( Q(X) \) is incorrect with respect to \( Y \), and needs to be supplemented with a probability distribution \( P(Y) \) that provides additional information. By construction, \( P(Y) \) is assumed to be correct (at least useful).

The above situation arises naturally in protein structure prediction. For example, \( P(Y) \) could be a probability distribution over the radius of gyration, hydrogen bond geometry or the set of pairwise distances, and \( Q(X) \) could be a fragment library [18] or a

---

**Figure 2. General statistical justification of PMFs.** The goal is to combine a distribution \( Q(X) \) over a fine grained variable \( X \) (top right), with a probability distribution \( P(Y) \) over a coarse grained variable \( Y(X) \) (top left). \( Q(X) \) could be, for example, embodied in a fragment library \( \mathcal{F} \), a probabilistic model of local structure \( \mathcal{T} \) or an energy function \( \mathcal{E} \); \( Y \) could be, for example, the radius of gyration, the hydrogen bond network, or the set of pairwise distances. \( P(Y) \) usually reflects the distribution of \( Y \) in known protein structures (PDB), but could also stem from experimental data \( D \). Sampling from \( Q(X) \) results in a distribution \( Q_{R}(Y) \) that differs from \( P(Y) \). Multiplying \( P(Y) \) and \( Q(Y) \) does not result in the desired distribution for \( Y \) either (red box); the correct result requires dividing out the signal with respect to \( Y \) due to \( Q(X) \) (green box). The reference distribution \( Q_{R}(Y) \) in the denominator corresponds to the contribution of the reference state in a PMF. If \( Q_{R}(Y) \) is only approximately known, the method can be applied iteratively (dashed arrow). In that case, one attempts to iteratively sculpt an energy funnel. The procedure is statistically rigorous provided \( Q(X) \) and \( P(Y) \) are proper probability distributions; this is usually not the case for conventional pairwise distance PMFs.

doi:10.1371/journal.pone.0013714.g002
A generalized PMF: radius of gyration

As a first application of the reference ratio method, we consider the task of sampling protein conformations with a given probability distribution \( P(\mathbf{r}_g) \) for the radius of gyration \( \mathbf{r}_g \). For \( P(\mathbf{r}_g) \), we chose a Gaussian distribution with mean \( \mu = 22 \text{ Å} \) and standard deviation \( \sigma = 2 \text{ Å} \). This choice is completely arbitrary; it simply serves to illustrate that the reference ratio method allows imposing an exact probability distribution over a certain feature of interest. Applying Eq. 6 results in:

\[
P(X|A) = \frac{P(\mathbf{r}_g(X))}{Q(\mathbf{r}_g(X)|A)} Q(X|A)
\]

For \( Q(X|A) \), we used TorusDBN – a graphical model that allows sampling of plausible backbone angles [45] – and sampled conditional on the amino acid sequence \( A \) of ubiquitin (see Materials and Methods). \( Q(\mathbf{r}_g|A) \) is the probability distribution of the radius of gyration for structures sampled solely from TorusDBN, which was determined using generalized multihistogram MCMC sampling (see Materials and Methods).

In Fig. 3, we contrast sampling from Eq. 8 with sampling from \( P(\mathbf{r}_g(X)) Q(X|A) \). In the latter case, the reference state is not properly taken into account, which results in a significant shift towards higher radii of gyration. In contrast, the distribution of \( \mathbf{r}_g \) for the correct distribution \( P(X) \), given by Eq. 8, is indistinguishable.
able from the target distribution. This qualitative result is confirmed by the Kullback-Leibler divergence [50] – a natural distance measure for probability distributions expressed in bits – between the target distribution and the resulting marginal distributions of \( r_g \). Adding \( Q_{\theta}(r_g|X)/A \) to the denominator diminishes the distance from 0.08 to 0.001 bits. For this particular PMF, the effect of using the correct reference state is significant, but relatively modest; in the next section, we discuss an application where its effect is much more pronounced.

Iterative optimization of PMFs: hydrogen bonding

Here, we demonstrate that PMFs can be optimized iteratively, which is particularly useful if the reference probability distribution \( Q_g(H|A) \) is difficult to estimate. We illustrate the method with a target distribution that models the hydrogen bonding network using a multinomial distribution.

We describe the hydrogen bonding network \( H \) with eight integers \((n_z,n_b,n_d)\) representing the number of residues that do not participate in hydrogen bonds in \( z \)-helices, \( \beta \)-sheets and coils, respectively. The five remaining integers \((n_z,n_b,n_d,n_c,n_r)\) represent the number of hydrogen bonds within \( z \)-helices, within \( \beta \)-strands, within coils, between \( z \)-helices and coils, and between \( \beta \)-strands and coils, respectively.

As target distribution \( P(H) \) over these eight integers, we chose a multinomial distribution whose parameters were derived from the native structure of protein G (see Materials and Methods). \( P(H) \) provides information, regarding protein G, on the number of hydrogen bonds and the secondary structure elements involved, but does not specify where the hydrogen bonds or secondary elements occur. As in the previous section, we use TorusDBN as the sampling distribution \( Q(X|A) \); we sample backbone angles conditional on the amino acid sequence \( A \) of protein G. Native secondary structure information was not used in sampling from TorusDBN.

The reference distribution \( Q_{\theta}(H|A) \), due to TorusDBN, is very difficult to estimate correctly for several reasons: its shape is unknown and presumably complex; its dimensionality is high; and the data is very sparse with respect to \( \beta \)-sheet content. Therefore, \( Q_{\theta}(H|A) \) can only be approximated, which results in a suboptimal PMF. A key insight is that one can apply the method iteratively until a satisfactory PMF is obtained (see Fig. 2, dashed line). In each iteration, the (complex) reference distribution is approximated using a simple probability distribution; we illustrate the method by using a multinomial distribution, whose parameters are estimated by maximum likelihood estimation in each iteration, using the conformations generated in the previous iteration. In the first iteration, we simply set the reference distribution equal to the uniform distribution.

Formally, the procedure works as follows. In iteration \( i+1 \), the distribution \( P(H|A) \) is improved using the samples generated in iteration \( i \):

\[
P_{i+1}(X|A) = \frac{P(H(X))}{P_R(H(X)|A)} P_A(X|A)
\]

where \( P_R(H|A) \) is the reference distribution estimated from the samples generated in the \( i \)-th iteration, \( P_0(X) = Q(X|A) \) stems from TorusDBN, and \( P_{R,0}(H|A) \) is the uniform distribution. After each iteration, the set of samples is enriched in hydrogen bonds, and the reference distribution \( P_{R,i}(H|A) \) can be progressively estimated more precisely. Note that in the first iteration, we simply use the product of the target and the sampling distribution; no reference state is involved.

Fig. 4 shows the evolution of the fractions versus the iteration number for the eight hydrogen bond categories; the structures with minimum energy for all six iterations are shown in Fig. 5. In the

![Figure 3. A PMF based on the radius of gyration.](image-url)

The goal is to adapt a distribution \( Q(X|A) \) – which allows sampling of local structures – such that a given target distribution \( P(r_g) \) is obtained. For \( A \), we used the amino acid sequence of ubiquitin. Sampling from \( Q(X|A) \) alone results in a distribution with an average \( r_g \) of about 27 Å (triangles). Sampling using the correct expression (open circles), given by Eq. 8, results in a distribution that coincides with the target distribution (solid line). Not taking the reference state into account results in a significant shift towards higher \( r_g \) (black circles).

DOI:10.1371/journal.pone.0013714.g003
first iteration, the structure with minimum energy (highest probability) consists of a single α-helix; β-sheets are entirely absent (see Fig. 5, structure 1). Already in the second iteration, β-strands start to pair, and in the third and higher iterations complete sheets are readily formed. The iterative optimization of the PMF quickly leads to a dramatic enrichment in β-sheet structures, as desired, and the fractions of the eight categories become very close to the native values (Fig. 4).

Conclusions

The strengths and weaknesses of PMFs can be rigorously explained based on simple probabilistic considerations, which leads to some surprising new insights of direct practical relevance. First, we have made clear that PMFs naturally arise when two probability distributions need to be combined in a meaningful way. One of these distributions typically addresses local structure, and its contribution often arises from conformational sampling. Each conformational sampling method thus requires its own reference state and corresponding reference distribution; this is likely the main reason behind the large number of different reference states reported in the literature [3,22,38–41]. If the sampling method is conditional upon the amino acid sequence, the reference state necessarily also depends on the amino acid sequence.

Second, conventional applications of pairwise distance PMFs usually lack two necessary features to make them fully rigorous: the use of a proper probability distribution over pairwise distances in proteins for $P(Y|A)$, and the recognition that the reference state is rigorously defined by the conformational sampling scheme used, that is, $Q(X|A)$. Usually, the reference state is derived from external physical considerations [11,51].

Third, PMFs are not tied to pairwise distances, but generalize to any coarse grained variable. Attempts to develop similar quantities that, for example, consider solvent exposure [52,53], relative side chain orientations [54], backbone dihedral angles [55,56] or hydrogen bonds [37] are thus, in principle, entirely justified. Hence, our probabilistic interpretation opens up a wide range of possibilities for advanced, well-justified energy functions based on sound probabilistic reasoning; the main challenge is to develop proper probabilistic models of the features of interest and the estimation of their parameters [49,57]. Strikingly, the example applications involving radius of gyration and hydrogen bonding that we presented in this article are statistically valid and rigorous, in contrast to the traditional pairwise distance PMFs.

Finally, our results reveal a straightforward way to optimize PMFs. Often, it is difficult to estimate the probability distribution that describes the reference state. In that case, one can start with an approximate PMF, and apply the method iteratively. In each iteration, a new reference state is estimated, with a matching probability distribution. In that way, one iteratively attempts to sculpt an energy funnel [24,32–36]. We illustrated this approach with a probabilistic model of the hydrogen bond network. Although iterative application of the inverse Boltzmann formula has been described before [24,35,58,59], its theoretical justification, optimal definition of the reference state and scope remained unclear.

As the traditional pairwise distance PMFs used in protein structure prediction arise from the imperfect application of a statistically valid and rigorous procedure with a much wider scope, we consider it highly desirable that the name “potential of mean force” should be reserved for true, physically valid quantities [13]. Because the statistical quantities we discussed invariably rely on the use of a ratio of two probability distributions, one concerning protein structure and the other concerning the (now well defined) reference state, we suggest the name “reference ratio distribution” deriving from the application of the “reference ratio method”.

Pairwise distance PMFs, as used in protein structure prediction, are not physically justified potentials of mean force or free energies.
[2,13] and the reference state does not depend on external physical considerations; the same is of course true for our generalization. However, these PMFs are approximations of statistically valid and rigorous quantities, and these quantities can be generalized beyond pairwise distances to other aspects of protein structure. The fact that these quantities are not potentials of mean force or free energies is of no consequence for their statistical rigor or practical importance – both of which are considerable. Our results thus vindicate, formalize and generalize Sippl’s original and seminal idea [7]. After about twenty years of controversy, PMFs – or rather the statistical quantities that we have introduced in this article – are ready for new challenges.

**Materials and Methods**

**Outline of the problem**

We consider a joint probability distribution \( Q(X,Y) \) and a probability distribution \( P(Y) \) over two variables of interest, \( X \) and \( Y \), where \( Y \) is a deterministic function of \( X \); we write \( Y(X) \) when relevant. Note that because \( Y \) is a function of \( X \), it follows that \( Q(X) = Q(X,Y(X)) \); and if \( Y \neq Y(X) \), then \( Q(X,Y) = 0 \).

We assume that \( P(Y) \) is a meaningful and informative distribution for \( Y \). Next, we note that \( Q(X,Y) \) implies a matching marginal probability distribution \( Q_R(Y) \) (where the subscript \( R \) refers to the fact that \( Q_R(Y) \) corresponds to the reference state, as we will show below):

\[
Q_R(Y) = \int Q(X,Y) dX
\]

We consider the case where \( Q_R(Y) \) differs substantially from \( P(Y) \); hence, \( Q_R(Y) \) can be considered as incorrect. On the other hand, we also assume that the conditional distribution \( Q(X|Y) \) is indeed meaningful and informative (see next section). This distribution is given by:

\[
Q(X|Y) = \begin{cases} 0 & \text{if } Y \neq Y(X) \\ \frac{Q(X)}{\int Q(X')\delta(Y(X')-Y)dX'} & \text{if } Y = Y(X) \\ \end{cases}
\]

where \( \delta(\cdot) \) is the delta function. The question is now how to combine the two distributions \( P(Y) \) and \( Q(X) \) – each of which provide useful information on \( X \) and \( Y \) – in a meaningful way. Before we provide the solution, we illustrate how this problem naturally arises in protein structure prediction.

**Application to protein structure**

In protein structure prediction, \( Q(X,Y) \) is often embodied in a fragment library; in that case, \( X \) is a set of atomic coordinates obtained from assembling a set of polypeptide fragments. Of course, \( Q(X,Y) \) could also arise from a probabilistic model, a pool of known protein structures, or any other conformational sampling method. The variable \( Y \) could, for example, be the radius of gyration, the hydrogen bond network or the set of pairwise distances. If \( Y \) is a deterministic function of \( X \), the two variables are called coarse grained and fine grained variables, respectively. For example, sampling a set of dihedral angles for the protein backbone uniquely defines the hydrogen bond geometry between any of the backbone atoms.

Above, we assumed that \( Q(X|Y) \) is a meaningful distribution. This is often a reasonable assumption; fragment libraries, for example, originate from real protein structures, and conditioning on protein-like compactness or hydrogen bonding will thus result in a meaningful distribution. Of course, sampling solely from \( Q(X,Y) \) is not an efficient strategy to obtain hydrogen bonded or compact conformations, as they will be exceedingly rare. We now provide the solution of the problem outlined in the previous section, and discuss its relevance to the construction of PMFs.

---

Figure 5. Highest probability structures for each iteration. The structures with highest probability out of 50,000 samples for all six iterations (indicated by a number) are shown as cartoon representations. The N-terminus is shown in blue. The figure was made using PyMOL [64].

doi:10.1371/journal.pone.0013714.g005
Solution for a proper joint distribution

A first step on the way to the solution is to note that the product rule of probability theory allows us to write:

\[ P(X, Y) = P(Y)P(X|Y) \]

As only \( P(Y) \) is given, we need to make a reasonable choice for \( P(X|Y) \). We assume, as discussed before, that \( Q(X|Y) \) is a meaningful choice, which leads to:

\[ P(X, Y) = P(Y)Q(X|Y) \]

In the next step, we apply the product formula of probability theory to the second factor \( Q(X|Y) \), and obtain:

\[ P(X, Y) = P(Y) \frac{Q(X,Y)}{Q_R(Y)} \]  (11)

The distribution \( P(X, Y) \) has the correct marginal distribution \( P(Y) \).

In the next two sections, we discuss how this straightforward result can be used to great advantage for understanding and generalizing PMFs. First, we show that the joint distribution specified by Eq. 11 can be reduced to a surprisingly simple functional form. Second, we discuss how this result can be used in MCMC sampling. In both cases, expressions that correspond to a PMF arise naturally.

PMFs from combining distributions

Using the product rule of probability theory, Eq. 11 can be written as:

\[ P(X, Y) = P(Y) \frac{Q(Y|X)Q(X)}{Q_R(Y)} \]

Because the coarse grained variable \( Y \) is a deterministic function of the fine grained variable \( X \), \( Q(Y|X) \) is the delta function:

\[ P(X, Y) = P(Y) \frac{\delta(Y - Y(X))Q(X)}{Q_R(Y)} \]  (12)

Finally, we integrate out the, now redundant, coarse grained variable \( Y \) from the expression:

\[ P(X) = \int P(X, Y)dY \]

\[ = \int P(Y) \frac{\delta(Y - Y(X))Q(X)}{Q_R(Y)}dY \]

\[ = P(Y(X)) \frac{Q(X)}{Q_R(Y)} \]

and obtain our central result (Eq. 6). Sampling from \( P(X) \) will result in the desired marginal probability distribution \( P(Y) \). The influence of the fine grained distribution \( Q(X,Y) \) is apparent in the fact that \( P(X|Y) \) is equal to \( Q(X|Y) \). The ratio in this expression corresponds to the usual probabilistic formulation of a PMF; the distribution \( Q_R(Y) \) corresponds to the reference state. In the next section, we show that PMFs also naturally arise when \( P(Y) \) and \( Q(X,Y) \) are used together in Metropolis-Hastings sampling.

PMFs from Metropolis-Hastings sampling

Here, we show that Metropolis-Hastings sampling from the distribution specified by Eq. 11, using \( Q(X,Y) \) as a proposal distribution, naturally results in expressions that are equivalent to PMFs. The derivation is also valid if the proposal distribution depends on the previous state, provided \( Q(X,Y) \) satisfies the detailed balance condition.

According to the standard Metropolis-Hastings method [60], one can sample from a probability distribution \( \Pi(X,Y) \) by generating a Markov chain where each state \( X', Y' \) depends only on the previous state \( X, Y \). The new state \( X', Y' \) is generated using a proposal distribution \( \pi(X', Y'|X,Y) \), which includes \( \pi(X', Y'|X,Y) = \pi(X', Y') \) as a special case. According to the Metropolis-Hastings method, the proposal \( X', Y' \) is accepted with a probability \( \alpha \):

\[ \alpha(X', Y'|X,Y) = \min(1, p) \]

\[ p = \frac{\Pi(X', Y') \pi(X, Y|X', Y')}{\Pi(X, Y) \pi(X', Y'|X, Y)} \]  (13)

where \( Y, X \) is the starting state, and \( Y', X' \) is the next proposed state. We assume that the proposal distribution \( \pi(X', Y'|X,Y) \) satisfies the detailed balance condition:

\[ \pi(X', Y'|X,Y)\pi(X,Y) = \pi(X,Y|X', Y')\pi(X', Y') \]

As a result, we can always write Eq. 13 as:

\[ \frac{\Pi(X', Y') \pi(X,Y|X', Y')}{\Pi(X, Y) \pi(X', Y'|X, Y)} = \frac{\Pi(X', Y') \pi(X,Y|X', Y')}{\Pi(X', Y') \pi(X,Y|X', Y')} \]

The Metropolis-Hastings expression (Eq. 13), applied to the distribution specified by Eq. 11 and using \( Q(X', Y') \) or \( Q(X', Y'|X,Y) \) as the proposal distribution, results in:

\[ \frac{P(Y')Q_R(Y')}{P(Y)Q_R(Y)} \frac{Q(X,Y)}{Q(X',Y')} \]

which reduces to:

\[ \frac{P(Y')}{P(Y)} \frac{Q_R(Y')}{Q_R(Y)} \]  (14)

Hence, we see that the Metropolis-Hastings method requires the evaluation of ratios of the form \( P(Y)/Q_R(Y) \) when \( Q(X',Y') \) or \( Q(X',Y'|X,Y) \) is used as the proposal distribution; these ratios correspond to the usual probabilistic formulation of a PMF. Finally, when \( Y \) is a deterministic function of \( X \), the proposal distribution reduces to \( Q(Y') \) or \( Q(X'|X) \), and Eq. 14 becomes:

\[ \frac{P(Y(X'))}{P(Y(X))} \frac{Q_R(Y(X'))}{Q_R(Y(X))} \]

Application to radius of gyration and hydrogen bonding

Conformational sampling from a suitable \( Q(X|A) \) was done using TorusDBN [45] as implemented in Phaistos [61]; backbone angles \( (\phi, \psi \text{ and } \omega) \) were sampled conditional on the amino acid sequence. We used standard fixed bond lengths and bond angles in...
References


