Journal of Structural Biology 184 (2013) 75–82

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

# Probabilistic determination of probe locations from distance data

Xiao-Ping Xu [a], Brian D. Slaughter [b], Niels Volkmann [a,*]

[a] Sanford-Burnham Medical Research Institute, 10901 N. Torrey Pines Rd., La Jolla, CA 92037, United States
[b] The Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, United States

ARTICLE INFO

*Article history:*
Available online 13 June 2013

*Keywords:*
FRET
Electron microscopy
Probability function
Distance data

ABSTRACT

Distance constraints, in principle, can be employed to determine information about the location of probes within a three-dimensional volume. Traditional methods for locating probes from distance constraints involve optimization of scoring functions that measure how well the probe location fits the distance data, exploring only a small subset of the scoring function landscape in the process. These methods are not guaranteed to find the global optimum and provide no means to relate the identified optimum to all other optima in scoring space. Here, we introduce a method for the location of probes from distance information that is based on probability calculus. This method allows exploration of the entire scoring space by directly combining probability functions representing the distance data and information about attachment sites. The approach is guaranteed to identify the global optimum and enables the derivation of confidence intervals for the probe location as well as statistical quantification of ambiguities. We apply the method to determine the location of a fluorescence probe using distances derived by FRET and show that the resulting location matches that independently derived by electron microscopy.



## 1. Introduction

The multi-dimensional scoring-function landscapes encountered in structural biology tend to be rugged and complex (Frauenfelder et al., 1991; Frauenfelder and Leeson, 1998). As a consequence, simple optimization algorithms tend to get trapped in local minima, unable to locate the globally optimal solution. A multitude of search strategies such as, for example, simulated annealing (Kirkpatrick et al., 1983) or genetic algorithms (Holland, 1975) have been developed over the years in order to overcome this problem and these types of methods have been adapted to a large variety of optimization problems in structural biology (see for example Brünger et al., 1998, 2011; Webb et al., 2011; Das and Baker, 2008; Chacón et al., 1998). Though superior to simple direct optimization in that they can escape local minima, these methods are neither guaranteed to identify the global optimum nor do they provide simple ways to put the found optimum into the context to other existing optima because only a small sub-set of the global scoring landscape is sampled. In contrast, if the scoring landscape can be explored in its entirety, not only the global optimum can be identified without ambiguity, it can also be related to the rest of the scoring landscape so that confidence intervals and significant levels can be obtained. It is then possible, for example, to determine whether the global optimum is actually significantly better at some desired confidence level than the second-best local optimum or whether the second-best optimum needs to be considered a viable alternative solution at that confidence level.

Recently, methods for a more complete evaluation of scoring function space based on probability theory and Bayesian inference have been introduced in the context of structure determination by NMR (Rieping et al., 2005) or with sparse data (Habeck, 2011). Bayesian inference incorporates prior information and accounts for ambiguities through probability distributions and is especially useful when the observable to parameter ratio is low and the data is insufficient to determine the parameters unambiguously. Here, we consider the location determination of a probe from distance constraints and known attachment sites. We propose a method based on probability calculus that allows exploration of the entire scoring function space by directly combining probability functions representing the data without the construction of a likelihood function as is required for Bayesian inference.

## 2. Methodology

In order to use probability calculus, we first need to convert the experimental information into adequate probability functions $p(x)$ with

$$p(x) \geqslant 0 \text{ for all } x \tag{1}$$

and

$$\int p(x)dx = 1 \tag{2}$$

* Corresponding author. Fax: +1 858 646 3195.
*E-mail address:* niels@burnham.org (N. Volkmann).

1047-8477/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2013.05.020

where the integral extends over the entire range of $x$, usually taken as $-\infty$ to $\infty$. In the current context, $x$ is a three-dimensional coordinate $[x_1, x_2, x_3]$.

In the context of location determination from distance constraints we use two types of probability functions: (i) functions that describe the location of a known entity, for example the attachment site of a GFP label on a macromolecule or the location of the label itself and (ii) functions that describe the distance between two entities such as distances derived from fluorescence resonance energy transfer (FRET) between two fluorophores. Both these functions are continuous in nature, justifying the use of probability functions for continuous variables.

Once the location and distance information is converted into probability functions, probability calculus can be invoked to combine the information (Kolmogorov, 1950). For independent information defined in the same coordinate frame, probabilities need to be point-wise multiplied at every $x$:

$$p_{ij}(x) = p_i(x) \cdot p_j(x) \tag{3}$$

For independent information that has to be combined such as finding the probability of a distance sphere $p_j(y)$ at every point in space with probability $p_i(x)$, we are seeking the probability of the sum of two independent variables and a convolution operation needs to be performed:

$$p_{ij}(z) = (p_i * p_j)(z) = \int p_i(x) \cdot p_j(z-x)dx, \tag{4}$$

where $*$ denotes convolution. The convolution theorem states that a convolution between two functions is equivalent to the reverse Fourier transform of the product of their Fourier transforms:

$$(p_i * p_j) = F^{-1}(F[p_i] \cdot F[p_j]) \tag{5}$$

For example, in a typical application where FRET between several labels is used to determine the location of a probe, location information from two different label pairs would be defined in the same coordinate frame and the combined probability can be constructed by simply multiplying the probability functions describing the location information of each label pair voxel by voxel in real space. In case of distance information between two labels and information about the attachment site of one of the labels, the probability functions describing these two information sources need to be convoluted.

While distance and location parameters are continuous variables and thus continuous probability functions need to be used, it is advantageous for practical purpose to define those continuous functions on a discreet grid. In particular this strategy allows the use of sums instead of integrals so that, for example, Eq. (2) simplifies to

$$\sum p(x) = 1 \tag{6}$$

where $x$ is summed over all possible values. The choice of the grid size should be fine enough to account for the ruggedness of the scoring landscape so the global optimum can be determined accurately.

### 2.1. Practical recipe

The rules described above give rise to a simple and general practical recipe for obtaining probabilities of label locations in respect to macromolecules in the presence of distance information (see also Fig. 1):

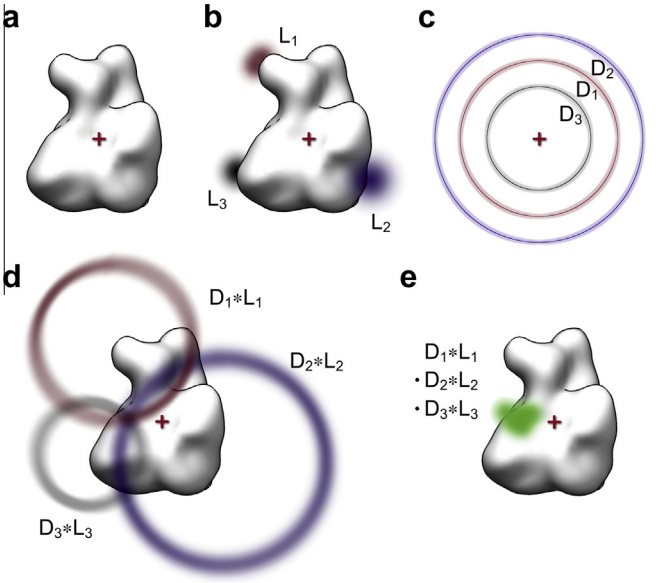


**Fig.1.** Schematic overview of methodology. (a) The common coordinate system is defined with the origin (red cross) set to the center of the macromolecular structure (white surface representation). (b) The probability functions for the locations of attachment points ($L_1$, $L_2$, $L_3$) are derived in the common coordinate system. (c) The experimental distance constraints are mapped onto spheres centered at the origin to derive the corresponding probability functions ($D_1$, $D_2$, $D_3$). (d) The distance and location probability functions are convoluted ($L_N * D_N$) to derive the probe location probability functions for each location/distance pair. (e) All location/distance pair probability functions are multiplied voxel-wise to derive the final probability function for the probe location (green).

1. Define the common coordinate system. This will usually be connected to structural information about a macromolecular assembly, for example the coordinate system where the density of an electron microscopy (EM) reconstruction is defined.
2. Define the bounding box. This step will provide a convenient way for normalizing all probabilities and should be large enough to contain all possible probability values appreciably larger than zero so that $\int p(x)\,dx = 1$ inside the bounding box and $\int p(x)\,dx = 1$ from $-\infty$ to $+\infty$ (Eq. (2)) are equivalent for practical purposes. It is convenient to define the origin of the coordinate system at the bounding box center.
3. Define grid size for calculations. This will define how the continuous probability function is discretized. The grid size should be small enough to capture the ruggedness of the probability landscapes faithfully. Normalization is then easily achieved for each probability function by enforcing $\Sigma p(x) = 1$ (Eq. (6)), summing over all $x$ in the bounding box.
4. Construct probability functions for known label locations in respect to coordinate system within bounding box. This should incorporate all anticipated uncertainties such as those coming from flexible linkers or from docking of atomic structures into lower resolution densities or envelopes. Also this step depends very heavily on the type of information available. It could be as simple as taking the coordinate of a known label location in a crystal structure, using its B-factor to get a Gaussian probability function. Some more involved cases will be presented in the application example.
5. Construct probability functions for distances between entities (i.e. labels) inside of bounding box, centered at origin. How this is exactly achieved will depend strongly on the method employed for determining the distances. For simple, relatively rigid cross-linkers, a Gaussian approximation would suffice, accounting for limited flexibility. A strategy for FRET distances is described in the application example.