Looking for central tendencies in the conformational freedom of proteins using NMR measurements

Fabrizio Clarelli¹,² and Luca Sgheri¹

1 Istituto per le Applicazioni del Calcolo (CNR), sede di Firenze, Via Madonna del Piano, 10, 50019 Sesto Fiorentino (FI), Italy
2 Department of Pharmacy, UiT, The Arctic University of Norway, 9037 Tromsø, Norway

E-mail: l.sgheri@iac.cnr.it

Received 5 April 2016, revised 25 November 2016
Accepted for publication 20 December 2016
Published 16 January 2017

Abstract
We study the conformational freedom of a protein made by two rigid domains connected by a flexible linker. The conformational freedom is represented as an unknown probability distribution on the space of allowed states. A new algorithm for the calculation of the maximum allowable probability is proposed, which can be extended to any type of measurements. In this paper we use pseudo contact shifts and residual dipolar coupling. We reconstruct a single central tendency in the distribution and discuss in depth the results.

Keywords: protein folding, paramagnetic NMR, underdetermined problems

Supplementary material for this article is available online
(Some figures may appear in colour only in the online journal)

1. Introduction

Flexibility is a key point in the functioning of macromolecules such as proteins [12, 14]. One of the few techniques which permit one to extract information about the conformational freedom of proteins in physiological condition is NMR spectroscopy. In the last decade a vast literature has flourished on this topic, we refer the reader to the two recent review papers [27, 29] for a general discussion on the available techniques.

Since the temporal scale of the fluctuations in the fold of a protein is several orders smaller than the time needed to take NMR measurements [24], information about the flexibility can only be recovered as a probability function on the set of allowed states. This set can be parametrized using for instance Cartesian coordinates of atoms involving measurements, dihedral
angles of the backbone or Euler transformations determining the position of rigid protein domains.

The recovery of this probability distribution is an under-determined ill-posed problem independently of the chosen parameters. The number of constraints is in fact far too small to determine a unique solution, except in trivial cases. Without further assumptions, any set of values of the parameters which is compatible with the measurements may be seen as a solution, and in principle there is no way of telling if a solution is better than another.

The lack of uniqueness, combined with the heterogeneity of the NMR measurement scenarios, led to a plethora of different techniques with different acronyms all trying to determine the ‘best’ solution. The two cited review papers try to classify (each from a different point of view) these approaches.

From the mathematical point of view it may be observed that there are two limit cases for the solutions.

- A first approach is to find the solution which minimizes the additional hypotheses on the data, thus using either the maximum entropy principle (MEP) [15], or the Kullback–Leibler divergence [17], which is the relative entropy. The MEP solution maximizes the uncertainty on the data, so each feature shown by the MEP solution is relevant. On the other hand the MEP solution is in general a continuous probability distribution, so that a large number of states is normally needed in order to approximate it [11]. The large number of variables involved raises not easy computational issues.

- A second approach is to find if the measurements carry some preference for certain states. The maximal allowable probability (MAP) is the largest weight of a given state in a probability distribution which is a solution. As a function of the state, the MAP is not a solution but a sharp bound from above. The zones with a large MAP are the positions favoured by the data. On the other hand it is not possible to establish to what extent the true distribution shows these asymmetries. We can only say that the largest asymmetries should be in favour of the zones indicated by the MAP technique.

Both approaches are equally able to recover the unknown probability distribution in the extremal cases. When there is very little conformational freedom the physical situation can be thought as a series of oscillations around a central state. In this case the MEP solution tends to a Dirac function of the central state, while the MAP estimate tends to be 1 for the central state and 0 for the other states. On the other hand when there is a very large conformational freedom the MEP solution tends to the uniform distribution in the space metric and the spread of the MAP estimate is minimal.

In the in-between cases the two approaches diverge. The MEP solution is obtained as the solution with the minimal spread in the probability density. On the other hand the MAP estimate for each state is obtained via solutions with the largest possible spread between the probability of the estimated state and the probabilities of the states needed to complete the solution. Since the problem is underdetermined, both approaches are consistent with the data. They focus on different aspects of the problem and they are in some sense complementary. For a deeper discussion of this topic, we refer the reader to [27].

In this paper we show a development of the MAP approach which permits the combination of different NMR constraints. The MAP approach has been inspired by [6], and the rigorous mathematical definition of the MAP has been given in [13], though different names have been subsequently used for this same bound of the probability. A geometrical algorithm has been developed in [20] to calculate the MAP estimate when only residual dipolar coupling (RDC) [33] are considered, using the linearity properties of these measurements. Residual dipolar coupling and pseudo contact shift (PCS) [18], which are frequently obtained together, have
been used to analyse the conformational freedom of calmodulin via the MAP estimate in [9],
using a complicated and time-consuming numerical procedure. The main difficulty is that
PCS (as is the case for most sets of data) do not possess the linearity properties of RDC, which
permits working on averaged tensors.

The maximum occurrence (MO) algorithm [8] uses a predetermined pool of conformers
to calculate the maximal probability. The choice of a finite number of conformers simplifies
the algorithm and reduces the time needed for the calculations. With this choice the posi-
tions which would cause physical violations of the atoms of the two domains may be directly
eliminated from the sample. The MO and MAP approaches are compared in figure 9 of [27],
bottom right panel. While the MAP value is always larger (hence better) than the MO value,
the geometrical algorithm for the determination of the MAP estimate was up to now only
applicable to the RDC measurements.

The SES (sparse ensemble selection) method has been developed in [5], and is focused on
recovering a small set of conformers with large probabilities. A recent paper compares the two
approaches and shows the information content of RDC and PCS [2].

A widely used method to determine the conformational freedom of protein domains is the
ensemble averaging (EA) [10], see for instance [19]. In the EA approach different conformers
are added to the solution until some criterion on the agreement of the measurements is met. In
[28] experimental RDC and PCS are used to determine an ensemble of conformers in the case
of calmodulin, the same molecule that we use as a test case in the paper.

In this paper we extend the geometrical algorithm of [20] to the case of PCS. Indeed, since
we drop the linearity requirement only fulfilled by paramagnetic RDC, the approach can be
extended to any set of measurements.

2. Theory

We use the calmodulin measurement scenario [9] as our test case. Calmodulin is a protein
made by two rigid domains (the N and C terminals) connected by a flexible linker, see figure 2
in section 4. A paramagnetic ion may be inserted in the binding sites of the N terminal, which
is also called the metal domain. We can then measure NMR data for atoms belonging to both
the N and C terminals.

The RDC measurements [33] are defined by

\[ \delta_{RDC,j} = \frac{c_{RDC}}{||P_{j}||^5} P_{j}' \chi P_{j} \]  \hspace{1cm} (1)

where \( P_{j} = P_{j_1} - P_{j_2} \) is the vector connecting selected pairs \( j_1 \) and \( j_2 \) of chemically linked
atoms and \( c_{RDC} \) is a constant. The paramagnetic tensor \( \chi \) is a symmetric and null-trace 3 \times 3
matrix, thus depending on 5 coefficients. Since the atoms are chemically bound, their distance
may be considered fixed, so that \( ||P_{j}|| \) is constant, and the only dependence is on the orientation
of the vector \( P_{j} \). If the atoms belong to the same domain as the metal, the tensor \( \chi \) and the
vector \( P_{j} \) belong to the same rigid structure. The RDC of the metal domain can be used to fit
the numerical values of the \( \chi \) tensor using (1).

The PCS measurements [18] are given by

\[ \delta_{PCS,j} = \frac{c_{PCS}}{||P_{j}||^3} P_{j}' \chi P_{j} \]  \hspace{1cm} (2)
a formula very similar to (1), with a different constant. In this case $P_j$ is however the vector connecting the metal and selected atoms $j$. When the atom belongs to the metal domain there is no difference between RDC and PCS. In fact the atom does not move with respect to the metal ion, so $P_j$ is fixed. Indeed RDC and PCS of atoms of the metal domain can be coupled to obtain a better fit for the paramagnetic tensor $\chi$ (and possibly for the location of the paramagnetic metal ion), see for instance [21]. From now on we suppose that this is the case, so that the paramagnetic tensor is known and only RDC and PCS for atoms belonging to the $C$ terminal are considered. Since the $C$ terminal moves with respect to the metal ion, the NMR measurements are averages of different states of the molecule, so that we may speak about mean PCS or RDC.

The RDC and PCS are in fact the average of the values obtained for different positions of the $C$ terminal (also called conformers). Each conformer is identified by an Euler transformation $E \equiv (R, t)$, where $R$ is a rotation and $t$ a translation. Note however that the $P_j$ of formulas (1) for RDC are difference of coordinates, so that the translations cancel, and we have

$$E(P_j) = R(P_j - t) \text{ for PCS}, \quad E(P_j) = RP_j \text{ for RDC}. \quad (3)$$

Note also that $\|E(P_j)\| = \|P_j\|$ does not depend on $(R, t)$ in the case of RDC. Because of the linker we can always suppose $t_{\text{min}} \leq \|t\| \leq t_{\text{max}}$, so that the space of allowed Euler transformations is compact.

Let $D$ be the space of probability distributions on this compact space. Each $d \in D$ is identified by the probability density $p(R, t) \geq 0$, such that $\int_{R, t} p(R, t) dR dt = 1$. Then

$$\delta_{\text{RDC}, j} = \frac{c_{\text{RDC}}}{\|P_j\|} \int_{R, t} p(R, t)(RP_j)'(RP_j) dR dt. \quad (4)$$

Since $p$ does not depend on $t$ for the RDC, we have $p = p(R)$ and $\int_R p(R) dR = 1$. Using the mean paramagnetic tensor

$$\bar{\chi} = \int_R p(R)R'\chi(R)dR. \quad (5)$$

Equation (4) becomes:

$$\delta_{\text{RDC}, j} = \frac{c_{\text{RDC}}}{\|P_j\|} P_j'\bar{\chi}P_j. \quad (6)$$

The same technique cannot be used for PCS because of the term $E(P_j) = R(P_j - t)$), so that

$$\delta_{\text{PCS}, j} = \frac{c_{\text{PCS}}}{\|P_j\|} \int_{R, t} \frac{p(R, t)}{\|R(P_j - t)\|} (R(P_j - t))'(R(P_j - t)) dR dt. \quad (7)$$

Different metal ions $M_k$ may be substituted in the same binding site belonging to the $N$ terminal without influencing the fold of the protein [1]. We suppose that each set of measurements relative to metal $M_k$ is obtained by averaging values relative to conformers, using the same probability distribution $d \in D$.

We now briefly report some properties of RDC and PCS. In the supporting material we expand this section for the readers which are not acquainted with the subject. Independent PCS and RDC measurements may be obtained from at most 5 different metal ions $M_k$ [26]. The problem of determining $d$ from the measurements of the moving terminal is underdetermined. The target distribution $d$ is a function of six variables, those defining the Euler transformation.
On the other hand we only have a finite number of measurements. Moreover, it is well known that the maximal number of independent RDC measurements from atoms of the C terminal is 25, see for instance [22, 23]. Also, the information content of PCS is weak [2]. Hence, no matter how many measurements are available, the distribution \( d \) cannot be recovered except in some trivial cases.

3. The simplex algorithm

3.1. Geometrical setting

Suppose we have \( n \leq 5 \) metal ions, and that \( \chi^k \) are already given or determined via the RDC and PCS of the metal domain. Take any \( d \in D \), we can calculate the mean RDC and PCS with the general formula

\[
\delta_j = c_j \int_{R_{1}} \frac{p(R,t)}{||E(P)||} E(P) f^k E(P) dRdt,
\]

where \( p(R,t) \) is the probability density of \( d \) at \( (R,t) \), and \( E(P) \) is defined by (3). The values \( P_j, c_j \) and \( k \leq n \) depend on the choice of atoms and the type of measurement. The term \( ||E(P)|| \) is constant for RDC. In the case of PCS, for physical reasons the distance between the metal ion and any other atom is anyway bounded away from 0. Hence we may suppose that the measurements \( \delta_j \) are uniformly bounded. We can obtain a certain number of RDC and PCS measurements for each of the \( n \) metals, not necessarily referring to the same atoms. Let \( n_{\text{RDC}} \) be the total number of mean RDC, \( n_{\text{PCS}} \) be the total number of mean PCS, and let \( n_{\text{meas}} = n_{\text{RDC}} + n_{\text{PCS}} \).

We can collect the measurements calculated from (12) in a vector, so that each \( d \in D \) defines a point \( \delta \in \mathbb{R}^{n_{\text{meas}}} \). The key point of the geometrical approach is the projection from the space of finite distributions to the space of the measurements. Let \( \Pi \) be such a projection, we may also decompose \( \Pi \) into the RDC and PCS components:

\[
\Pi(d) \equiv \begin{pmatrix} \delta_{\text{RDC},1} \\ \vdots \\ \delta_{\text{RDC},n_{\text{RDC}}} \\ \delta_{\text{PCS},1} \\ \vdots \\ \delta_{\text{PCS},n_{\text{PCS}}} \end{pmatrix}.
\]

Let

\[
V = \{ v \in \mathbb{R}^{n_{\text{meas}}} : v = \Pi(d), d \in D \}.
\]

The set \( V \) is compact because the measurements are uniformly bounded. The set \( V \) is convex because if \( v_1, v_2 \in V \), \( v_i = \Pi(d_i), i = 1, 2 \), then \( \lambda d_1 + (1 - \lambda) d_2 \in D, \forall \lambda \in [0, 1] \), so that

\[
\lambda v_1 + (1 - \lambda) v_2 = \Pi(\lambda d_1 + (1 - \lambda) d_2) \in V.
\] (11)

Each convex set is the convex hull of its extreme points (also called vertices), i.e. the points that cannot be reconstructed using a convex combination of different points of the set. Let \( \Delta \subset D \) the set of finite probability distributions, and let \( \tilde{\Delta} \subset \Delta \) the set of probability distributions made by a single point. Because of the non-linearity, in general it is not true that each \( \Pi(d) \) is a vertex of \( V \), though we suspect this is the case in our setting. On the other hand, the set of vertices is a subset of \( \Pi(d) \), since \( V \) is the set of convex combinations of these points. We do not need the property that each \( \Pi(d) \) may be uniquely reconstructed, so we can nevertheless identify the set of vertices with \( \Pi(d) \).
Proposition 3.1. For each \( d \in D \) there exists a \( \hat{d} \in \Delta \) such that \( \Pi(\hat{d}) = \Pi(d) \).

**Proof.** By Carathéodory’s theorem, each \( \Pi(d) \in V \) may be reconstructed with a convex combination of at most \( n_{\text{meas}} + 1 \) vertices of \( V \). Let \( \Pi(d) = \sum_{i=0}^{n_{\text{meas}}} p_i \Pi(\hat{d}_i) \), with \( \hat{d}_i \in \hat{\Delta} \). Then \( \hat{d} = \sum_{i=0}^{n_{\text{meas}}} p_i \hat{d}_i \in \Delta \) is the required distribution. \( \square \)

**Remark.** Proposition 3.1 entitles us to work with finite distributions of probability without loss of generality. If \( d \equiv (p_1, \hat{R}_1, t_1) \in \Delta, \) formula (8) may be rewritten as

\[
\delta_j = c_j \sum_i \frac{1}{\|E(P_i)\|} \int_R \left( \frac{R\hat{P}_j}{\hat{P}_j} \right)^{h} H(R) \, dR.
\]

(12)

Proposition 3.2. There exists a \( d \in \Delta \) such that \( \Pi(d) = 0 \).

**Proof.** The proposition is proven in [13] for the case of RDC, and a constructive example with a finite distribution is given in [30]. Fix the origin of the Cartesian system in the binding site of the metal. Let \( \hat{d} \in \hat{\Delta} \) such that the translation \( t \) is fixed and the rotational part coincides with the Haar measure \( H(R) \), see for instance [34]. Then \( \Pi_{\text{RDC}}(\hat{d}) = 0 \). With these choices we also have \( \Pi_{\text{PCS}}(\hat{d}) = 0 \). Fix a \( j \) relative to a PCS in formula (12). Let \( \hat{P}_j = P_j - t \), then \( \|E(P_j)\| = \|R\hat{P}_j\| = \|\hat{P}_j\| \) for every rotation \( R \) because the metal is in the origin. Hence

\[
\delta_j = c_j \frac{1}{\|\hat{P}_j\|^2} \int_R \left( \frac{R\hat{P}_j}{\hat{P}_j} \right)^{h} H(R) \, dR = c_j \frac{1}{\|\hat{P}_j\|^2} \int_R \left( R^t \chi^h(R) H(R) \right) \, dR = 0.
\]

(13)

This is due to the fact that the integrand in parenthesis is the mean paramagnetic tensor, and its integral is 0 for the Haar measure [13]. The existence of a \( d \in \Delta \) is then guaranteed by proposition 3.1. \( \square \)

The dimension \( N \leq n_{\text{meas}} \) of the set \( V \) is a key point which can be determined from the data. Using the results of the previous section, if we suppose that we have at least 5 independent RDC measurements for each of the \( n \) metal ions, then \( N = 5n + n_{\text{PCS}} \). However, since the PCS are only marginally linearly independent, it is to be expected that there are directions where the set \( V \) is very thin, so that the effective determination of \( N \) should involve also some numerical considerations.

3.2. **Definition of the MAP**

Let \( \hat{d} \) the true unknown distribution of probability. Then, given any Euler transformation \((R, t)\) we define

\[
p_{\max}(R, t) = \max_{d \in \Delta} \{ p : (p, R, t) \in d \text{ and } \Pi(d) = \Pi(\hat{d}) \}.
\]

(14)

In other words given any conformer, identified by the Euler transformation \((R, t)\), we look for the maximal coefficient \( p \) that we can apply to this conformer in a convex combination such that the projection in \( V \) is the same as that of \( \hat{d} \). Suppose \( \Pi(\hat{d}) \) belongs to the interior of \( V \). Let \( \Pi(\hat{d}) = \Pi(1, R, t) \) be the vertex corresponding to the position \((R, t)\). Consider the line passing through \( \Pi(\hat{d}) \) and \( \Pi(\hat{d}) \), the segment in between the two points belongs to \( V \) because of the convexity. Moreover, since \( \Pi(\hat{d}) \) is internal, there exists a point \( \Pi(q) \in V \) on the continuation
of the segment on the side of $\Pi(\bar{d})$. Then $\Pi(\bar{d})$ is the convex combination of $\Pi(q)$ and $\Pi(\bar{d})$, i.e. there exists a $p \in (0, 1)$ such that
\[ \Pi(\bar{d}) = p\Pi(\bar{d}) + (1 - p)\Pi(q). \]  
By definition we have $p_{\max}(R, t) \geq p$. The value $p$ can be explicitly determined using the distances (i.e. the $L^2$ norms) in $\mathbb{R}^N$, in fact
\[ \Pi(\bar{d}) = \frac{\|\Pi(\bar{d}) - \Pi(q)\|}{\|\Pi(\bar{d}) - \Pi(q)\|} \Pi(\bar{d}) + \frac{\|\Pi(q) - \Pi(\bar{d})\|}{\|\Pi(\bar{d}) - \Pi(q)\|} \Pi(q). \]  
The maximal $p$ which verifies (15) is then obtained from the $q$ with projection in $V$ having the maximal distance from $\Pi(\bar{d})$. Because of the convexity, $\Pi(q)$ is the point on the boundary of $V$ on the continuation of the segment connecting $\Pi(\bar{d})$ and $\Pi(\bar{d})$. Unfortunately, except in some trivial cases, there is no analytical procedure for determining if a point $\Pi(q)$ belongs to the boundary of $V$, so that we have to use an iterative procedure.

3.3. The simplex algorithm

Let $N \leq 5n + n_{PCS}$ be the dimension of $V$. By Carathéodory’s theorem there are $N + 1$ vertices of the convex $V$ such that
\[ \Pi(\bar{d}) = \sum_{j=0}^{N} p_j \Pi(\bar{d}_j^0), \]  
with $p_j \geq 0$ and $\sum_j p_j = 1$. Note again that we cannot suppose that $\bar{d} = \sum_j p_j \bar{d}_j^0$ because in general the solution is not unique, we can only recover the projection. Let $S_0 \subset V$ be the simplex formed by the convex combinations of the vertices $\Pi(\bar{d}_j^0)$. We may suppose $S_0$ is a simplex in $\mathbb{R}^N$, i.e. the vectors $\Pi(\bar{d}_j^0) - \Pi(\bar{d}_0^0)$ are linearly independent in $\mathbb{R}^N$. Since the set $\Pi(\bar{d})$ is connected we may choose $S_0$ so that $\Pi(\bar{d})$ is internal to $S_0$, i.e. $p_i > 0 \forall i$ [20].

Now take any position $(R, t)$ and let $\bar{d} = (1, R, t)$, look at figure 1 for reference. Take the line $r$ through $\Pi(\bar{d})$ and $\Pi(\bar{d})$. Since $\Pi(\bar{d})$ is internal to $S_0$ there is a point $P_0 \in \partial S_0$ on $r$ on the side opposite to $\Pi(\bar{d})$ with respect to $\Pi(\bar{d})$. The point $P_0$ identifies a face $F_{j_0} \subset S_0$ such that $P_0 \in F_{j_0}$. The fact $F_{j_0}$ is identified by removing the vertex $j_0$ from the set of vertices of $S_0$.

The point $P_0$ is either on the boundary or internal to $V$. In the first case we are finished because we have found the point needed by the definition of $p_{\max}$. In the second case, consider the hyperplane $H_{j_0}$ containing the face $F_{j_0}$. The hyperplane $H_{j_0}$ cannot be a support hyperplane since it contains an internal point, thus there will be at least a vertex $\bar{d}_{j_0}^1$ on the half-plane defined by $H_{j_0}$ and not containing $\Pi(\bar{d})$. Define $S_1$ to be the simplex with $\bar{d}_{j_0}^0$ replaced by $\bar{d}_{j_0}^1$.

We can iterate the algorithm, each time finding the two intersections of $r$ with the simplex $S_{\epsilon}$. The intersection point $P_1$ on $r$ farther from $\Pi(\bar{d})$ determines a face $F_{j_1}$ of the simplex $S_1$. If this intersection point is internal we can replace the vertex $j_1$ of $S_1$ not belonging to $F_{j_1}$ with a new one, lying in the half-space determined by the hyperplane $H_{j_1}$ and not containing $\Pi(\bar{d})$.

Thus we determine a monotonic sequence of points $P_\epsilon \in r$ converging to a point $P$. The point $P$ cannot be internal to $V$, otherwise the algorithm would have found a new replacement vertex. Then $P \in H$ is the point needed by the definition of $p_{\max}$. 
3.4. Determination of the projection matrix

In principle the algorithm may be carried out in the ambient space $\mathbb{R}^{n_{\text{meas}}}$ without any modifications. However, the dimension $N$ of $V$ is in general strictly smaller than the number of measurements $n_{\text{meas}}$. The simplex algorithm works in the linear subspace spanned by $V$, which has dimension $N$. Using the $n_{\text{meas}}$ ambient coordinates in this linear subspace is definitely a bad idea, because any numerical approximation in the calculations is likely to bring the points out of the linear subspace. Thus the first step is to determine the dimension $N$ of $V$, and the projection operator from $\mathbb{R}^{n_{\text{meas}}}$ into $\mathbb{R}^N$.

The dimension $N$ is the maximal number of linearly independent vectors of the form $\Pi(d_i) - \Pi(d_0)$, where $d_0$ is a fixed point in $\Delta$, and $d_i \in \Delta$. Because of proposition 3.2, we may take $d_0$ such that $\Pi(d_0) = 0$. Since each point in $V$ is a convex combination of vertices, $N$ is then the maximal number of linearly independent vectors $\Pi(d_i)$, where $d_i \in \hat{\Delta}$.

The singular value decomposition (SVD, see for instance [25]) may be used to determine $N$, as already done in [31] in a different context. Take points $\Pi(d_i) \in V$, $i = 1, \ldots, M$, with $M \gg n_{\text{meas}}$, and form the matrix

$$A = (\Pi(d_1), \ldots, \Pi(d_M)), \quad (18)$$

which has dimension $n_{\text{meas}} \times M$. The SVD is based on the singular values of $A$, which are the square roots of the eigenvalues of the symmetric and positive semi-definite matrix $A^T A$.

The SVD decomposes the matrix $A$ in the form

$$A = U \Lambda W,$$  

where $\Lambda$ is a $n_{\text{meas}} \times n_{\text{meas}}$ diagonal matrix containing the singular values of $A$ in decreasing order, $W$ is an orthogonal $n_{\text{meas}} \times n_{\text{meas}}$ matrix, and $U$ is a $M \times n_{\text{meas}}$ column-orthogonal matrix, i.e. $\sum_{i} u_{ij} u_{kj} = \delta_{ij}$. Since the rank of $A$ is by definition $N$, the matrix $A$ has exactly $N$ non-zero singular values.

The matrix $W$ can be used as a projection matrix. In fact, for any $v \in V$ we have that $Wv$ is the decomposition of $v$ with respect to the base of eigenvectors of $A^T A$. Moreover, since the matrix $\Lambda$ has only $N$ non-zero eigenvalues, we are only interested in the matrix $W_N \subset W$, including only the first $N$ rows of $W$.

While the matrix $W_N W_N$ is not the identity, it works as such on the points of $V$. In fact the points of the linear space spanned by $V$ can be uniquely identified either by a subset of the $n_{\text{meas}}$ coordinates satisfying $n_{\text{meas}} - N$ linear conditions, or by the $N$ intrinsic coordinates obtained by applying $W_N$.

The following diagram pictures the situation:
4. Implementation and self-consistency tests

In this section we describe the implementation of the simplex algorithm and report the results of simulations run with exact measurements. It is a ‘proof of concept’ that the method works, and a necessary step to understand the interactions of the measurements before considering noisy data.

4.1. Experimental setting

As a model for calmodulin we use the pdb fold as determined by [7], shown in figure 2. In physiological conditions the long $\alpha$-helix breaks in the zone between residuals 77 and 81 [3, 4], resulting in some conformational freedom of the C terminal. Any position such that the linker length is between 6 Å and 12 Å is considered to be attainable. Any position such that there exists two $C_\alpha$ atoms with a distance smaller than 3 Å is considered to be a physical violation. A conformer satisfying both conditions is an allowable state for the C terminal. The positions of the C terminal shown are marked $C_1$ through $C_3$, with $C_1$ being the position where the linker remains folded into the $\alpha$-helix. In position $C_2$ the C terminal is close to the N terminal, but far from the metal. In position $C_3$ the C terminal is close both to the N terminal and to the metal. We have highlighted in red the first two residuals of the C terminal for positions $C_2$ and $C_3$ to show the connection point to the linker.

The Euler transformations are mapped using the following coordinate system. We set the origin of a Cartesian system in the last atom of the backbone of the N terminal, namely the nitrogen atom of residual 77. We kept the lab frame orientation of the system. The translation is measured as the distance (calculated in angstroms) between the origin of the system and the first atom of the backbone of the C terminal, namely the $C_\alpha$ atom of residual 82. For the rotation, the roll, pitch, yaw system of Euler angles (see for instance [34]) is used, with the identity rotation corresponding to the orientation $C_1$.

The measurements are generated using a continuous probability distribution $d_l \in D$ centered in conformers $C_l \equiv (R_l, t_l)$, $l = 1, 2, 3$. The measurements are calculated using formula (12) as the arithmetic average of a large number $M$ of allowable states drawn according to the distribution $d_l$. More precisely, given two positive numbers $\sigma_t$ and $\sigma_R$ we draw conformers in the following way. The translation is drawn according to a Gaussian distribution with average $t_l$ and standard deviation of the module $\sigma_t$. The rotation is drawn according to a von Mises–Fisher distribution (see for instance [34]) with average $R_l$ and standard deviation $\sigma_R$ of the rotational distance, calculated using quaternions. We only retain allowable states. The number $M$ is large enough to stabilize the measurements, thus simulating a continuous probability distribution. Loosely speaking the C terminal moves around the center position $C_1$ in such a way that the average deviation from the central position is $\sigma_t$ angstroms for the translation, and $\sigma_R$ degrees for the rotation. In this paper we use the numerical values $\sigma_t = 3$ Å and $\sigma_R = 20^\circ$.

We generated mean measurements with respect to three different paramagnetic tensors, corresponding to Tb, Tm and Dy lanthanide ions substituted for Ca in the second binding site of the N-terminal. We simulated a total of 112 mean RDC using N-H dipoles from residuals of the C terminals, and a total of 132 mean PCS using HN atoms from the C terminal.
In principle the distribution is symmetric around the center. However the constraint on the physical violations may introduce asymmetries in the distribution. This happens in cases \( C_2 \) and \( C_3 \), when the center position of the \( C \) terminal is close to the \( N \) terminal. As a consequence there is a small shift in the most probable position of the distribution. In the supplementary information (stacks.iop.org/IP/33/025011/mmedia) we discuss in detail this issue.

4.2. Steps to determine a central tendency

We used the following steps for the central tendency determination (CTD) of the measurement.

1. The orientation \( R_0 \) of the conformer with the largest \( p_{\text{max}} \) is determined using RDC alone.
2. The translation \( t_0 \) of the conformer with the largest \( p_{\text{max}} \) is determined using PCS alone.

   The rotation is kept fixed at \( R = R_0 \).
3. The orientation \( R_0 \) of the conformer with the largest \( p_{\text{max}} \) is refined using RDC and PCS together. The translation is kept fixed at \( t = t_0 \) determined in step 2.

We remark that the CTD does not depend on any initial conformation. The space of the conformers is simply sampled in order to determine a tabulated MAP function. If necessary, the conformer maximizing the MAP can be refined using a local maximization technique. Also, the only numerical threshold which should be set for the simplex algorithm is the stopping criterion for the sequence \( P_k \) of section 3.3. In principle the sequence can be stopped when no further point \( P_{k+1} \) can be determined, which implies that the MAP value has been correctly determined. In practice a small tolerance is anyway set on \( \|P_{k+1} - P_k\| \) to avoid useless iterations which would only refine the \( p_{\text{max}} \) by a negligible quantity. Thus the MAP technique is resilient to the numerical choices which may influence the results of different algorithms.

During each step of the CTD the relevant space (rotational for steps 1 and 3, translational for step 2) is sampled. For most tests we used a sample of \( 35^3 = 42\,875 \) conformers, uniform in the Haar metric for the rotations (see for instance [16]), uniform in the Cartesian metric for the translations. Note that the MAP function is rather smooth. The density of points in figures 3–7...
below depends on the sample distribution and not on the shape of the MAP function itself. Conformers producing physical violations are removed, and the MAP of the remaining states is determined. In the first step the orientation $R_0$ with the largest MAP is determined using the RDC alone. This is equivalent to the application of the MAP algorithm as reported in [20], and each conformer involved in the simplex algorithm only depends on three rotational variables.

In the second step, the translation with the largest MAP is determined from the PCS, while the orientation of the sample is kept fixed at $R_0$. Finally in the third step RDC and PCS are used together to refine the orientation with the largest MAP. The use of PCS during steps 2 and 3 implies that each conformer involved in the algorithm depend on the full set of 6 variables of the Euler motion. Thus steps 2 and 3 are much more demanding in terms of computer time. On our standard PC the determination of a single MAP value takes no more than a few seconds during the first step. During steps 2 and 3 the time needed to determine a single MAP value can reach a couple of minutes, depending on the number of iterations of the simplex algorithm needed to stabilize the $p_{\text{max}}$ value.

In the supplementary information we prove that the information of PCS and RDC is largely linearly independent. As a consequence, the final output of the CTD is very similar to the direct sampling of the full set of states with the combined use of PCS and RDC.

4.3. Observability of a central tendency

Suppose there is a central tendency in the data. The MAP estimate is able to detect whether this tendency exists. To show this fact, we considered a probability distribution $d_0$ where all the allowable states are equally probable in the correct metric. If there were no constraints on the conformers the simulated measurements would be 0 by proposition 3.2. The large conformational freedom is however detectable from the RDC measurements. The standard deviation of the simulated RDC is in fact 0.36 for $d_0$, while is larger than 5 for the $d_l$ cases, $l \geq 1$. The situation is different for PCS. In this case, small values are obtained both when there is a large conformational freedom and when the C terminal is far from the binding site of the metal. The standard deviation of the simulated PCS is 0.10 for $d_0$, 0.14 for $d_1$, 0.08 for $d_2$ and finally 0.94 for $d_3$, the case where the C terminal is closer to the metal binding site.

The MAP estimate detects this difference in the data. In the $d_0$ case using RDC we found $0.31 \leq p_{\text{max}}(R) \leq 0.34$ for all the orientations of the C terminal. Typical values for the other $d_l$ cases are $0.10 \leq p_{\text{max}}(R) \leq 0.70$, thus having a much larger span. The MAP estimate is then in principle able to detect an asymmetry in the data due to a restricted conformational freedom.

4.4. Determination of the central tendency

We applied the CTD of section 4.2 to our test cases. In the following figures we present the results of the tests.

In figure 3 we show the results of step 1 of the CTD in cases $C_1$ (circles), $C_2$ (squares), $C_3$ (triangles). The points represent the orientations of the sample with MAP larger than a certain threshold. The larger green dots, here and in the subsequent figures, mark the positions of the centers. In cases $C_2$ and $C_3$ there are two different zones with large MAP, due to the so-called phenomenon of ghost cones [20], which derives from the symmetries of the RDC formula.

In figure 4 we show the results of step 2 of the CTD in cases $C_1$ and $C_3$, case $C_2$ being very similar to case $C_1$. Note that the MAP of a large number of translations is close to 1, as a consequence of the poor resolving power of the PCS. A slightly better reconstruction is obtained for $C_3$. In this case the center position is close to the metal, so that there are some PCS values
which are rather large, see formula (7). To obtain these values the C terminal must remain close to the C$_3$ position for a not negligible fraction of time. As a consequence, the $p_{\text{max}}$ values for positions far from the metal is slightly reduced.

In figure 5 we show the joint PCS + RDC case for determining an orientation. While the PCS do not resolve well the translation, they are useful to eliminate the ghost cones, thus determining the correct region of space for the orientations. The RDC and PCS formulas have the same type of symmetries, however the $E(P_j)$ vectors from (3) are different, so that in general the symmetries do not coincide.

5. Tests with experimental error

5.1. Determination of the central tendency

We added an uncorrelated Gaussian error to the mean measurements to take into account the experimental error. The error level was kept to ±0.01 ppm ±10% for PCS and ±1 Hz ±10% for RDC. We applied the CTD of section 4.2 and report the results in the following figures.

In figure 6 we show the results of step 1 of the CTD. Note that there are very few changes with respect to figure 3. This is due to the fact that the mean RDC have only 15 degrees of freedom, while we have 112 measurements. The SVD algorithm implicitly fits the 15 degrees
of freedom of the mean RDC. Since the error is assumed to be Gaussian, a large number of measurements reduces the standard error of the fitted quantities. Hence the information of the RDC is well recovered even when the experimental error is considered. As explained in the previous section, coupling PCS and RDC helps removing ghost cones.

In figure 7 we show the results of step 2 of the CTD, in case C_1 (left panel) and C_3 (right panel). Here the introduction of the experimental error worsens the results. This should not be a surprise, since most of the mean PCS are close to 0, so that their numerical value is destroyed by the error. As already explained, the case C_3 is better because the C terminal is close to the metal.

Figure 8 shows the final reconstruction. Here the dark colored conformers show the positions with the largest probability, while the lightly colored conformers show the reconstructed positions $\hat{C}_k$. 

Figure 5. Tests with RDC and PCS. Cases C_1 (circles), C_2 (squares), C_3 (triangles). The green dots mark the positions of the centers. Only orientations with a MAP larger than a threshold are shown.

Figure 6. Tests with RDC alone. Cases C_1 (circles), C_2 (squares), C_3 (triangles). The green dots mark the positions of the centers. Only orientations with a MAP larger than a threshold are shown.
We already noted that the CTD does not include a step where PCS and RDC are analysed together sampling the full space of Euler motions. The joint use of PCS and RDC is limited to refining the orientation, a step very useful to remove the ghost cones. We did attempt the full joined RDC + PCS analysis, using a local maximization technique. The results of this step are on average only marginally better than the results of the algorithm. If great care is not applied in the optimization, the error on the translation may even increase. Even in the joined RDC + PCS case, the translation is determined only by the PCS values, since there is no dependence on the RDC. However, the dependence on the translation is very weak, as it can be seen from figure 7, so that the variations in the MAP are largely due to the orientation of the conformer. More information on this additional attempt may be found in the supplementary material.

5.2. Estimate of the amplitude of the distribution

To show that the amplitude of the distribution can be estimated from the MAP information, we repeated the tests with noisy data in case $C_1$ using different $\sigma_\ell$ and $\sigma_r$, see section 4.1. In order to quantify with a single parameter the distance of the conformers $K_1$ and $K_2$ we introduce

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig7}
\caption{Tests with PCS alone. Cases $C_1$ (left panel) and $C_3$ (right panel). The green dots mark the positions of the centers. Only translations with a MAP larger than a threshold are shown.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig8}
\caption{Tests with RDC and PCS. Metal (pink), N terminal (blue). Final reconstruction of the $C_1$, $C_2$ and $C_3$ cases. Dark colors show the positions with the largest probability, light colors show the reconstruction.}
\end{figure}
the following distance, see also the supplementary material for a discussion on this topic. Let $a_i, \ldots, a_p$ be the atoms of the backbone of the $C$ terminal, and let $E_i(a_j) \equiv R_i(P_j - i)$ be the position of the atom $j$ in the $K_i$ conformer. We then define

$$d(K_1, K_2) = \frac{1}{\sqrt{P}} \sum_{j=1}^{P} ||E_1(a_j) - E_2(a_j)||^2.$$  \hspace{1cm} (20)

In other words $d(K_1, K_2)$ is the RMS of the distances between the positions of the backbone atoms assumed in the states $K_1$ and $K_2$.

In table 1 we report the results of the CTD for this set of tests. The third column contains the average distance $\bar{d}$ of the conformers used to construct the simulated data, while the fourth column contains the maximal value of the MAP obtained from step 3 of the CTD. We point out that there is an anticorrelation between these two values, both in the RDC and in the RDC + PCS cases. The closer the $\max_K p_{\max}(K)$ is to 1, the more the distribution tends to a single delta function.

In figure 9 the horizontal axis represents the distance $d$ of (20) with respect to the distribution center. The points $d_i$ on the x axis are chosen in such a way that $\int_{d_{i-1}}^{d_i} dV$ is constant. With this choice we correct for the different volume of the spherical annular sectors and obtain decreasing functions for the probability distributions. We set $d_0 = 0$. In the figure the probability $p_i$ that a conformer in the generating distribution has a distance from the center between $d_{i-1}$ and $d_i$ is reported as the point $(d_i, p_i)$ of the solid lines. The average MAP of the conformers of the sample with distance from the center between $d_{i-1}$ and $d_i$ is reported with symbols. We see that the shape of the MAP function is able to detect the sharp peak of the case $\sigma_t = 1$ Å and $\sigma_R = 10^\circ$. Moreover, the low values of the MAP in the case $\sigma_t = 4$ Å and $\sigma_R = 40^\circ$ indicate a broad shallow peak. Given the fact that the problem is under-determined we feel that establishing a more accurate quantitative relationship between the shape of the distribution and the MAP is not really appropriate.

5.3. Stability of the algorithm

In order to explore the possibility to apply the MAP technique to smaller molecules, we analyzed the stability of the algorithm with respect to the number of measurements available. We applied the CTD procedure, using only measurements from a limited number of residuals of the $C$ terminal, and compare the results with the standard (full) case. Figure 10 shows the results in the case $C_1$ for the RDC measurements (left panel), and RDC + PCS (right panel). We note from the left panel that the identification of the center orientation is preserved even
when the minimal number of measurement needed is employed. However the curve tends to become flat, which is a sign that the discriminating power of the MAP is reduced. In the combined RDC + PCS case we have also plotted the standard deviation of the bins in the full and 5 residuals case. The plot shows that the standard deviation is larger in the 5 residuals case up to 10 Å from the center. It follows that the peak of the MAP function is shallower and wider, so that the identification of a central tendency is less precise.

6. Conclusions

In this paper we demonstrated the ability of PCS and RDC measurements to recover a central tendency of an unknown probability distribution representing the conformational freedom of a protein made by two rigid domains connected by a flexible linker.

We made use of the MAP algorithm, extended to include PCS (and indeed any other class of measurements). The MAP algorithm determines the largest probability of a conformer in a distribution satisfying the measurements. Taken globally the MAP function is not a probability distribution but a sharp bound from above. For each position there is however an
explicitly determined finite probability distribution with the MAP value as a weight for that conformer.

The RDC measurements are well able to determine any central tendency in the orientation of the conformer. Adding the PCS helps removing symmetric orientations, since the symmetries of the PCS are different from those of the RDC.

The identification of the translation is more difficult. The RDC do not depend on the translation, so we can only use PCS. However the information content of the PCS is very weak, and is further destroyed by the experimental noise. With exact data we can only approximatively determine the central tendency of the translation, see figure 4. The situation worsens when the experimental error is added, as shown by figure 7. Values of MAP larger than 0.9 are obtained for a large fraction of the sample, and especially for positions relatively far from the metal. In other words, the conformer can sample any positions far from the metal for the 90% of the time, resulting in very small partial values of the PCS. Adjusting the remaining 10% of the distribution is enough to obtain the correct values of the PCS. As a consequence, the determination of the translation is less accurate. Still no translation could by determined without the PCS, so their contribution is fundamental.

Let us remark here the difference in the approach with respect to the already mentioned EA procedure. The EA is surely very informative in presenting the motion as a combination of states (check figure 4 of [28]). The MAP approach focuses on determining a sharp bound for the probability of a single conformer, so it is probably less ambitious. The MAP technique however does not depend on any initial value or numerical threshold. All the exploitable information is fully recovered, the rest being filtered out by the SVD. The solution is found by adding a certain number of transitional states (the vertices of the simplex) with low probability, but the main attention is given to the central tendency. Information about the amplitude of the distribution can still be gathered from the MAP function, we pointed out two correlated estimators. We also tested the stability of the method with respect to the number of measurements, in order to assess the applicability of the MAP technique to smaller molecules.

We again stress that the MAP algorithm points out the extremal cases which should anyway be considered by any method trying to determine a solution. Or course there might be reasons to exclude some probability distributions, for instance by using some threshold on the number of conformers or on the spread of the distribution. These are additional hypotheses which reduce the set of compatible solutions. However, since the problem is underdetermined and the real solution is not known, the reasons for removing these particular solutions should be soundly justified.

Other methods might not able to detect the MAP solutions, for instance if a predetermined sample is used for the conformers. Even if the sample is large, there might be a correlation between the rotational and translational part of the Euler transformations of the sample. In this case the identification of the orientation might bias the translation towards the correct value.

Further developments will include the case when there is more than a single central tendency in the data. Giving the results of this study, we foresee that additional information might be extracted only for the rotational part of the distribution. A possible way of overcoming this difficulty might be the inclusion of different measurements such as SANS (small angle x-rays scattering) [32]. The NMR-SAXS integration has already been studied using the MO (maximum occurrence) method in [8]. Since the SAXS measurements depends on the global shape of the molecule this information might help in better determining the translational part of the Euler transformation.
Acknowledgment

As usual we wish to acknowledge the fruitful and long-established collaboration with the Center for Magnetic Resonance of the University of Florence. Discussions with applied scientists is always fruitful and helps gearing mathematics towards realistic problems.

References

[1] Allegrozzi M, Bertini I, Janik M B L, Lee Y-M, Liu G and Luchinat C 2000 Lanthanide induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 Å from the metal ion J. Am. Chem. Soc. 122 4154–61
[2] Andrałojc W, Berlin K, Fushman D, Luchinat C, Parigi P, Ravera E and Sgheri L 2015 Information content of long-range NMR data for the characterization of conformational heterogeneity J. Biomol. NMR 62 353–71
[3] Baber J, Szabo A and Tjandra N 2001 Analysis of slow interdomain motion of macromolecules using NMR relaxation data J. Am. Chem. Soc. 123 3953–9
[4] Barbato G, Ikura M, Kay L E, Pastor R W and Bax A 1992 Backbone dynamics of calmodulin studied by 15N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible Biochemistry 31 5269–78
[5] Berlin K, Castaneda C A, Schneidman-Duhovny D, Sali A, Nava-Tudela A and Fushman D 2010 Recovering a representative conformational ensemble from underdetermined macromolecular structural data J. Am. Chem. Soc. 122 4154–61
[6] Bertini I, Del Bianco C, Gelis I, Katsaros N, Luchinat C, Parigi G, Peana M, Provenzani A and Zoroddu M A 2004 Experimentally exploring the conformational space sampled by domain reorientation in calmodulin Proc. Natl Acad. Sci. USA 101 6841–6
[7] Bertini I, Gelis I, Katsaros N, Luchinat C and Provenzani A 2003 Tuning the affinity for lanthanides of calcium binding proteins Biochemistry 42 8011–21
[8] Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov M V, Pierattelli R, Ravera E and Svergun D I 2010 Conformational space of flexible biological macromolecules from average data J. Am. Chem. Soc. 132 13553–8
[9] Bertini I, Gupta K J, Luchinat C, Parigi G, Peana M, Sgheri L and Yuan J 2007 Paramagnetism-based NMR restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains J. Am. Chem. Soc. 129 12786–94
[10] Bonvin A M and Brünger A T 1996 Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? J. Biomol. NMR 7 72–6
[11] Cavalli A, Camilloni C and Vendruscolo M 2013 Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle J. Chem. Phys. 138 094112
[12] Frauenfelder H, Sligar S and Wolynes P 1991 The energy landscapes and motions of proteins Science 254 1598–603
[13] Gardner R, Longinetti M and Sgheri L 2005 Reconstruction of orientations of a moving protein domain from paramagnetic data Inverse Problems 21 879–98
[14] Huang Y J and Montelione G T 2005 Structural biology: proteins flex to function Nature 438 36–7
[15] Jaynes E 1979 Where do we stand on maximum entropy? The Maximum Entropy Formalism ed R Levine and M Tribus (Cambridge, MA: MIT Press) pp 1–104
[16] Kuffner J J 2004 Effective sampling and distance metrics for 3D rigid body path planning Proc. 2004 IEEE Int. Conf. on Robotics and Automation
[17] Kullback S and Leibler R A 1951 On information and sufficiency Ann. Math. Stat. 22 79–86
[18] Kurland J R and McGarvey B R 1970 Isotropic NMR shifts in transition metal complexes: the calculation of the Fermi contact and pseudocontact terms J. Magn. Reson. 2 286–301
[19] Lange O F, Lakomek N A, Farès C, Schröder G F, Walter K F, Becker S, Meiler J, Grubmüller H, Griesinger C and De Groot B L 2008 Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution Science 320 1471–5
[20] Longinetti M, Luchinat C, Parigi G and Sgheri L 2006 Efficient determination of the mostfavoured orientations of protein domains from paramagnetic NMR data Inverse Problems 22 1485–502
[21] Longinetti M, Parigi G and Sgheri L 2002 Uniqueness and degeneracy in the localization of rigid structural elements in paramagnetic proteins J. Phys. A: Math. Gen. 35 8153–69
[22] Longinetti M, Sgheri L and Sottile F 2010 Convex hulls of orbits and orientations of a moving protein domain Discrete Comput. Geom. 43 54–68
[23] Meiler J, Prompers J J, Peti W, Griesinger C and Brüschweiler R 2001 Model-free approach to the dynamic interpolation of residual dipolar coupling in globular proteins J. Am. Chem. Soc. 123 6098–107
[24] Palmer A G and Massi F 2006 Characterization of the dynamics of biomacromolecules using rotating-frame spin relaxation NMR spectroscopy Chem. Rev. 106 1700–19
[25] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 Numerical Recipes in C: the Art of Scientific Computing 2nd edn (New York: Cambridge University Press)
[26] Ramirez B E and Bax A 1998 Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium J. Am. Chem. Soc. 120 9106–7
[27] Ravera E, Sgheri L, Parigi G and Luchinat C 2016 A critical assessment of methods to recover information from averaged data Phys. Chem. Chem. Phys. 18 5686–701
[28] Russo L, Maestre-Martinez M, Wolff S, Becker S and Griesinger C 2013 Interdomain dynamics explored by paramagnetic NMR J. Am. Chem. Soc. 135 17111–20
[29] Salmon L and Blackledge M 2015 Investigating protein conformational energy landscapes and atomic resolution dynamics from NMR dipolar couplings: a review Rep. Prog. Phys. 78 126601
[30] Sgheri L 2010 Joining RDC data from flexible protein domains Inverse Problems 26 115021
[31] Sgheri L 2010 Conformational freedom of proteins and the maximal probability of sets of orientations Inverse Problems 26 035003
[32] Svergun D I, Barberato C and Koch M H J 1995 CRYSOL—a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates J. Appl. Cryst. 28 768–73
[33] Tolman J R, Flanagan J M, Kennedy M A and Prestegard J H 1995 Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution Proc. Natl Acad. Sci. USA 92 9279–83
[34] Weisstein E W 1998 CRC Concise Encyclopedia of Mathematics (Boca Raton: Chapman and Hall)