J Struct Funct Genomics (2009) 10:67–81
DOI 10.1007/s10969-009-9062-2

Distance matrix-based approach to protein structure prediction

Andrzej Kloczkowski · Robert L. Jernigan · Zhijun Wu · Guang Song · Lei Yang · Andrzej Kolinski · Piotr Pokarowski

Received: 23 September 2008 / Accepted: 1 February 2009 / Published online: 18 February 2009

Abstract Much structural information is encoded in the internal distances; a distance matrix-based approach can be used to predict protein structure and dynamics, and for structural refinement. Our approach is based on the square distance matrix $\mathbf{D} = [r_{ij}^2]$ containing all square distances between residues in proteins. This distance matrix contains more information than the contact matrix **C**, that has elements of either 0 or 1 depending on whether the distance r_{ij} is greater or less than a cutoff value r_{cutoff}. We have performed spectral decomposition of the distance matrices $\mathbf{D} = \sum \lambda_k \mathbf{v}_k \mathbf{v}_k^T$, in terms of eigenvalues λ_k and the corresponding eigenvectors $\mathbf{v}_k$ and found that it contains at most five nonzero terms. A dominant eigenvector is proportional to r^2—the square distance of points from the center of mass, with the next three being the principal components of the system of points. By predicting r^2 from the sequence we can approximate a distance matrix of a protein with an expected RMSD value of about 7.3 Å, and by combining it with the prediction of the first principal component we can improve this approximation to 4.0 Å. We can also explain the role of hydrophobic interactions for the protein structure, because r is highly correlated with the hydrophobic profile of the sequence. Moreover, r is highly correlated with several sequence profiles which are useful in protein structure prediction, such as contact number, the residue-wise contact order (RWCO) or mean square fluctuations (i.e. crystallographic temperature factors). We have also shown that the next three components are related to spatial directionality of the secondary structure elements, and they may be also predicted from the sequence, improving overall structure prediction. We have also shown that the large number of available HIV-1 protease structures provides a remarkable sampling of conformations, which can be viewed as direct structural information about the dynamics. After structure matching, we apply principal component analysis (PCA) to obtain the important apparent motions for both bound and unbound structures. There are significant similarities between the first few key motions and the first few low-frequency normal modes calculated from a static representative structure with an elastic network model (ENM) that is based on the contact matrix **C** (related to **D**), strongly suggesting that the variations among the observed structures and the corresponding conformational changes are facilitated by the low-frequency, global motions intrinsic to the structure. Similarities are also found when the approach is applied to an NMR ensemble, as well as to atomic molecular dynamics (MD) trajectories. Thus, a sufficiently large number of experimental structures can directly provide

A. Kloczkowski (✉) · R. L. Jernigan · L. Yang
Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, 112 Office and Lab Bldg, Ames, IA 50011-3020, USA
e-mail: kloczkow@iastate.edu

A. Kloczkowski · R. L. Jernigan · L. Yang
Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA

Z. Wu
Department of Mathematics, Iowa State University, Ames, IA 50011, USA

G. Song
Department of Computer Science, Iowa State University, Ames, IA 50011, USA

A. Kolinski
Laboratory of Theory of Biopolymers, Department of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland

P. Pokarowski
Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland

important information about protein dynamics, but ENM can also provide a similar sampling of conformations. Finally, we use distance constraints from databases of known protein structures for structure refinement. We use the distributions of distances of various types in known protein structures to obtain the most probable ranges or the mean-force potentials for the distances. We then impose these constraints on structures to be refined or include the mean-force potentials directly in the energy minimization so that more plausible structural models can be built. This approach has been successfully used by us in 2006 in the CASPR structure refinement (http://predictioncenter.org/caspR).

Keywords Distance matrix · Spectral analysis · Protein structure prediction · Protein structure refinement · Elastic networks · Distance geometry

Introduction

Mathematical approach to studies of various protein properties by the analysis of the corresponding matrices has been quite popular in bioinformatics. In our earlier work we tried to approximate 20 × 20 dimensional matrices corresponding to contact potentials by 20-dimensional vectors of various physical properties of amino acids by using a simple linear $c_0 + x_i + x_j$ and quadratic functions $c_0 + x_i x_j + y_i y_j$ of two amino acid properties x and y [1]. We analyzed 29 different matrices of contact potentials published in literature. We used AAindex database of over 500 amino acid indices collected by Kanehisa [2, 3] http://www.genome.jp/aaindex/ and have found that all matrices of contact potentials can be approximated with correlation 0.9 by hydrophobicities and isoelectric points of amino acids. A dominant role of hydrophobicity in interactions among residues in proteins has been already well known, and our study have shown that isoelectric points, that measure electric charges of various amino acids are also important for contact potentials. We have found two classes of contact potentials. The first class of contact potentials can be approximated by a linear combination of hydrophobicities. Major contribution comes from the one body transfer energy of amino acids from water to protein environment. The second class of contact potentials can be approximated by a quadratic function of hydrophobicities and isoelectric points of amino acids. Potentials of this class represent energies of contact of amino acid pairs within an average protein environment.

In our later work we have extended our method to substitution matrices [4]. We have analyzed 29 different substitution matrices known in the literature, plus five statistical contact potentials. We found that substitution matrices can be approximated with correlation 0.9 by a quadratic expression $c_0 + x_i x_j + y_i y_j + z_i z_j$ with vectors x, y and z corresponding to hydrophobicity, molecular volume, and coil preferences of amino acids. We also found that some substitution matrices correlate well with contact potentials.

In our present work we apply a similar approach to matrices containing structural information for proteins. We try to express these original matrices in terms of their eigenvectors, try to connect these eigenvectors with physical properties of amino acids and predict them from the amino acid sequence. Our work was motivated by a recent work of Vendruscolo [5] who found that the eigenvector corresponding to the dominant eigenvalue of the contact matrix in proteins correlates well with the vector of hydrophobicities of the amino acid sequence. We show that the structural matrices relate to experimental B-factors (temperature factors) that measure thermal fluctuations of atoms around their mean positions in crystals (for X-ray determined structures) or in solution (for NMR-determined structures). We discuss elastic network models of proteins that mathematically relate contact matrices to mean square fluctuations of residues. We show that motions of amino acids in proteins computed from elastic network models better fit NMR-determined. Finally we discuss methods of refinement of protein structures based on libraries of interatomic distances in proteins, and propose a new optimization method of solving a generalized distance geometry problem for determination of NMR structures by using B-factors.

Methods

Matrices containing structural information

There are several different matrices that contain structural information for proteins. The most common is the distance matrix

$$\mathbf{d} = (d_{ij}) \tag{1}$$

where the ij-the element of the matrix is the distance d_{ij} between residues i and j. Usually the distance is measured between the C^{α} atoms of the residues, although other definitions of distances, such as: the distance between the centers of side chains of the amino acids, or the distance between the closest heavy atoms of the two residues, are also popular.

Distance matrices have been used for a long time in structural bioinformatics mostly for protein structure comparison and alignment and for inferring protein–protein interactions [6–24]. In particular Snyder and Montelione used this approach for identification of core atom sets and for the assessment of the precision in NMR-derived protein structure ensembles [20].

From the mathematical point of view it is easier to deal instead of distances with their squares. Because of this we define the matrix of square distances

$$\mathbf{D} = (\mathrm{d}_{\mathrm{ij}}^2) \tag{2}$$

containing information about square distances d_{ij} between amino acids i and j. All diagonal elements of the distance matrix **d** and the square distance matrix **D** are zeros.

Another matrix that is very popular in computational biology and contains significantly less information than the distance matrix is the contact matrix

$$\mathbf{C} = (\mathrm{c}_{\mathrm{ij}}) \tag{3}$$

with elements c_{ij} defined as:

$$\begin{aligned} \mathrm{c}_{\mathrm{ij}} &= 1 \quad \text{if } \mathrm{d}_{\mathrm{ij}} \leq \mathrm{d}_{\mathrm{cutoff}} \\ \mathrm{c}_{\mathrm{ij}} &= 0 \quad \text{otherwise} \end{aligned} \tag{4}$$

Here $\mathrm{d}_{\mathrm{cutoff}}$ is the cutoff distance defining residues being in contact.

Laplacian of **C** that is frequently called the Kirchhoff matrix is defined

$$\mathbf{L}_{\mathrm{c}} = \mathbf{diag}\left(\Sigma \mathrm{c}_{\mathrm{ij}}\right) - \mathbf{C} \tag{5}$$

The diagonal elements of $\mathbf{L}_{\mathrm{c}}$ are the sums of its off-diagonal elements taken with the negative sign. Because of such definition the sums of all elements in each row or column are zero and the determinant of $\mathbf{L}_{\mathrm{c}}$ is zero, i.e. the matrix is singular and has no inverse.

We may however, define a generalized inverse (pseudoinverse that might be right or left) $\mathbf{L}_{\mathrm{c}}^{-1}$ of the Laplacian matrix $\mathbf{L}_{\mathrm{c}}$ Such generalized inverse $\mathbf{L}_{\mathrm{c}}^{-1}$ of the Laplacian of the contact matrix is introduced in elastic network models of proteins (described in detail in the next section) and its elements represent covariances between instantaneous fluctuations of residues i and j.

Similarly we can define Laplacian of the matrix of square distances **D**:

$$\mathbf{L}_{\mathrm{D}} = \mathbf{diag}\left(\Sigma \mathrm{d}_{\mathrm{ij}}^2\right) - \mathbf{D} \tag{6}$$

and its generalized inverse $\mathbf{L}_{\mathrm{D}}^{-1}$.

Elastic network models of proteins

Elastic network models treat proteins as elastic bodies. A coarse-grained representation of proteins with a single site per residue is usually used. Positions of these sites are generally identified with the coordinates of the C^{α} atoms in proteins. Residues separated by a distance less than or equal to a certain cutoff value R_{c} (including neighbors along the sequence) are assumed to be in contact, and are connected with identical mass-less harmonic springs. This leads to an elastic network representation of a protein structure in the folded state that resembles a random polymer network. Figure 1 illustrates the basic idea of this model.

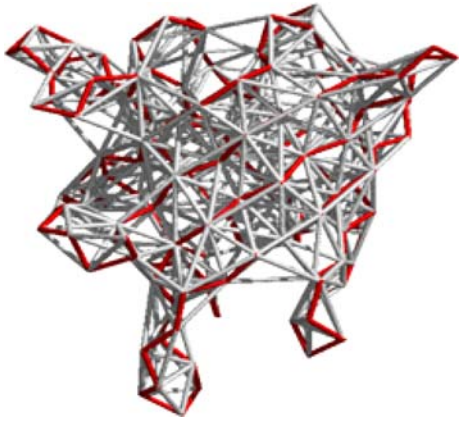

Fig. 1 Beads and springs representation of proteins in elastic network models. Protein backbone is shown in *red*

The simplest of the elastic network models is Gaussian Network Model (GNM). This model was originally developed for the theory of rubber-like elasticity of random polymer networks [25, 26] to calculate fluctuations of junctions and chains inside the network. The model was adapted to proteins by Bahar and Erman [27, 28] using an earlier result of Tirion [29] who used a single harmonic force parameter to analyze atomic motions in proteins.

The total potential energy for the network composed of N nodes is

$$V_{\mathrm{tot}} = \frac{1}{2}\gamma \sum_{i<j}^{N} \left(\Delta R_{ij}\right)^2 H(R_c - R_{ij}) \tag{7}$$

where γ is a uniform universal spring constant, and $H(x)$ is the Heaviside step function that equals 1 if $x > 0$, and is zero otherwise. Here $\Delta \boldsymbol{R}_{ij} = \boldsymbol{R}_{ij} - \boldsymbol{R}_{ij}^0$ is the instantaneous displacement of the distance vector $\boldsymbol{R}_{ij}$ between the ith and the jth sites from the mean value $\boldsymbol{R}_{ij}^0$ observed in the native structure. Equation 7 can be rewritten in the following form

$$\mathrm{V}_{\mathrm{tot}} = (\gamma/2)\{\Delta \boldsymbol{R}\}^T \boldsymbol{\Gamma} \{\Delta \boldsymbol{R}\} \tag{8}$$

where $\boldsymbol{\Gamma}$ is the Kirchhoff matrix of size $N \times N$, defined on the basis of the cutoff distance R_c, with off-diagonal elements ij being either -1 if nodes i and j are in contact or zero otherwise, and the diagonal elements are defined as the sum of the off-diagonal elements in the i-th row (or column) taken with a negative sign. Mathematical definition of the Kirchhoff matrix was given earlier by Eq. 5. Kirchhoff matrices were introduced first in physics to study electric currents in networks. Similarly as for electrical circuits where all currents at a given node sum up to zero, for a system of connected springs forming a network a sum of elastic

forces at each node is zero. The matrix is equivalent to the Laplacian $\mathbf{L}_c$ of the contact matrix in Eq. 5. Here $\{\Delta \boldsymbol{R}\}$ is the N-dimensional fluctuation vector $\Delta \boldsymbol{R} = \text{col}\,(\Delta \boldsymbol{R}_1, \Delta \boldsymbol{R}_2, \ldots, \Delta \boldsymbol{R}_N)$ of $\Delta \boldsymbol{R}_i$ for all N nodes, and the superscript T denotes the transpose. We should note that

$$\Delta \boldsymbol{R}_{ij} = \Delta \boldsymbol{R}_i - \Delta \boldsymbol{R}_j \tag{9}$$

Then the average changes in positions, given either as the correlation $\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle$ between the displacements of pairs of residues i and j or as the mean-square fluctuations $\langle (\Delta R_i)^2 \rangle = \langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_i \rangle$ for a single residue i, are

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle = \frac{\int (\Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j) \exp(-V_{\text{tot}}/k_B T) d\{\Delta \boldsymbol{R}\}}{\int \exp(-V_{\text{tot}}/k_B T) d\{\Delta \boldsymbol{R}\}} \tag{10}$$

for all i, j

This can be rewritten [25] in a simple form as

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle = \frac{3 k_B T}{2\gamma} (\boldsymbol{\Gamma}^{-1})_{ij} \tag{11}$$

where $(\boldsymbol{\Gamma}^{-1})_{ij}$ is the ij-th element of the inverse of the Kirchhoff matrix $\boldsymbol{\Gamma}$, k_B is the Boltzmann constant, T is the absolute temperature, and γ is a spring constant. Mean-square fluctuations $\langle (\Delta R_i)^2 \rangle$ of the i-th residue in a protein are given by the i-th diagonal element of $\boldsymbol{\Gamma}^{-1}$. Since the Laplacian matrix $\boldsymbol{\Gamma}$ is singular because $\det(\boldsymbol{\Gamma}) = 0$ only the pseudoinverse of $\boldsymbol{\Gamma}$ can be computed through the use of the singular value decomposition method. The pseudoinverse of $\boldsymbol{\Gamma}$ may be written as $\boldsymbol{\Gamma}^{-1} = \boldsymbol{U}(\boldsymbol{\Lambda}^{-1})\boldsymbol{U}^T$ where $\boldsymbol{U}$ is the matrix composed of eigenvectors $\boldsymbol{u}_i$ $(1 \le i \le N)$ of $\boldsymbol{\Gamma}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues of $\boldsymbol{\Gamma}$. Additionally, it can be proven that all eigenvalues λ_i of $\boldsymbol{\Gamma}$ are non-negative.

Mean-square fluctuations of the position of each C^α computed from Eq. 11 can be compared with the Debye-Waller thermal factors, which are measured by X-ray crystallography and deposited in the Protein Data Bank. The relationship between the B-factor and the mean square fluctuations for the i-th residue is given by

$$B_i = 8\pi^2 \langle (\Delta R_i)^2 \rangle / 3 \tag{12}$$

The B-factors computed by GNM usually are in excellent agreement with experimental data [30, 31].

The Gaussian Network Model is based on the assumption that all instantaneous fluctuations are isotropic. A more sophisticated elastic network model of proteins is the Anisotropic Network Model (ANM) [32]. Equation 8 is then replaced by

$$V = (1/2) \Delta \boldsymbol{R}^T \boldsymbol{H} \Delta \boldsymbol{R} \tag{13}$$

where $\Delta \boldsymbol{R}$ is the $3N$-dimensional vector of fluctuations, $\Delta \boldsymbol{R}^{\mathbf{T}}$ its transpose and $\mathbf{H}$ is the $(3N \times 3N)$ Hessian matrix, whose elements are the second derivatives of the total potential energy with respect to the Cartesian coordinates of the ith and jth nodes.

Spectral decomposition of structural matrices

Decomposition of matrices is a standard algebraic procedure to factorize them into a canonical form. There are various different methods of decomposition, such as for example **LU** decomposition where the original matrix is expressed a product of a lower triangular matrix **L** and an upper triangular matrix **U**. Decomposition based on eigenvalues of a square matrix **A** is called eigen decomposition or spectral decomposition. It allows us to express the original square matrix **A** of size in $N \times$ N terms of its eigenvalues λ_k and corresponding eigenvectors $\mathbf{v}_k$

$$\mathbf{A} = \sum_k \lambda_k \mathbf{v}_k \mathbf{v}_k^T \tag{14}$$

The inverse matrix $\mathbf{A}^{-1}$ is then expressed by the same Eq. 14 with eigenvalues λ_k replaced by their inverses λ_k^{-1}. In mathematical problems related to system dynamics the eigenvalues correspond to frequencies of motions that are called modes.

The matrix $\boldsymbol{\Gamma}^{-1}$ for the Gaussian Network Model can be written as the sum of contributions from individual modes:

$$\boldsymbol{\Gamma}^{-1} = \sum_k \lambda_k^{-1} \mathbf{u}_k \mathbf{u}_k^T \tag{15}$$

where zero eigenvalues of the Kirchhoff matrix $\boldsymbol{\Gamma}$ (related to rigid body motions of the center of mass of the system) are excluded from the summation. The i-th component of the eigenvector $\boldsymbol{u}_k$ specifies the magnitude of fluctuations of the i-th residue in the protein in the k-th mode. If we order the eigenvalues according to their ascending values starting from zero, then the most important contributions to $\boldsymbol{\Gamma}^{-1}$ in Eq. 15, and therefore (because of Eqs. 11–12) also to temperature factors are given by the smallest non-zero eigenvalues λ_k of $\boldsymbol{\Gamma}$ that correspond to the large-scale slow modes. The slowest modes play a dominant role in the fluctuational dynamics of protein structures, because their contributions to the mean-square fluctuations scale with λ_k^{-1}. It has been shown that the most important functional motions of proteins [33–35] or large biological structures (such as the ribosome [36–38]) correspond only a few of the slowest modes derived from the GNM.

To calculate the normal modes for the Anisotropic Network Model, the Hessian matrix **H** is diagonalized to the canonical form $\mathbf{S}^{\mathbf{T}}\mathbf{H}\mathbf{S} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a $(3N \times 3N)$ diagonal matrix with diagonal elements corresponding to eigenvalues $(\lambda_1, \ldots, \lambda_{3N})$ and **S** is an orthogonal $(3N \times 3N)$ matrix (i.e. $\mathbf{S}^{\mathbf{T}}\mathbf{S} = \mathbf{I}$) built from eigenvectors.

The mean-square fluctuations of the residue i can be expressed as a sum over all normal modes (except the first

six zero modes that corresponds to translations and rotations of the system) as

$$\left\langle (\Delta \boldsymbol{R}_i)^2 \right\rangle = (k_B T/\gamma) \sum_{l=1}^{3N-6} \frac{S_{il}^2}{\lambda_l} \tag{16}$$

where $\left\langle (\Delta \boldsymbol{R}_i)^2 \right\rangle$ are the mean-square fluctuations of residue i.

Structure determination and refinement using distances

We consider a problem of the determination of a structure or an ensemble of structures for a protein with a given set of inter-atomic distances or their ranges. This problem arises in modeling proteins using NMR distance data. Mathematically, it requires the solution for a nonlinear system of equations or inequalities. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ be the coordinate vector of atom i; ($i = 1, \ldots, n$), with n being the total number of atoms in the protein. The problem can be formulated to find $\boldsymbol{x}_i$, $i = 1, \ldots, n$ such that

$$\begin{aligned} &||\mathbf{x}_i - \mathbf{x}_j|| = d_{i,j}, \\ &(i,j) \in S \subset D = \{(i,j) : i = 1, \ldots, n,\ j > i\} \end{aligned} \tag{17}$$

where $d_{i,j}$ are the given distances between atoms i and j or

$$\begin{aligned} &l_{i,j} \leq ||\mathbf{x}_i - \mathbf{x}_j|| \leq u_{i,j}, \\ &(i,j) \in S \subset D = \{(i,j) : i = 1, \ldots, n,\ j > i\} \end{aligned} \tag{18}$$

Here $l_{i,j}$ and $u_{i,j}$ are the given lower and upper bounds on $d_{i,j}$, respectively.

The problem formulated in Eq. 17 has been studied in several fields and has many applications. It is called the distance geometry problem in mathematics, the multidimensional scaling problem in statistics, and the graph embedding problem in computer science. Distance geometry methodology for proteins has been developed 30 years ago by Havel and Crippen. [39–43]

The problem can be solved in polynomial time by using for example the well-known singular value decomposition algorithm if the distances for all the pairs of atoms in the protein are given, but it is NP-hard for an arbitrarily given subset of all the distances. The problem defined in Eq. 18 has a particular application in NMR protein modeling, where only a lower and upper bound can be estimated for a distance. A set of solutions can be obtained for this problem, which corresponds to an ensemble of structures, all satisfying the given distance constraints. It is of great practical interest to obtain the whole ensemble of structures, since it shows how a structure may change dynamically given the possible ranges of their distances. However, the problem to obtain the whole solution set, even for a linear system of inequalities, is NP-hard.

Heuristic methods have been developed for the solution of the first problem (Eq. 17), and been extended to the solution of a generalized problem (Eq. 18). A common approach to the later (Eq. 18) is to generate repeatedly a set of distances within the given distance ranges, and solve Eq. 17 with the generated distances. In the end, a set of solutions is obtained that represents the whole solution set for the problem defined by Eq. 18. The obtained solutions form an ensemble of structures. They can be put together to show how they deviate from each other at different times. A long-standing issue with this approach is that the solution set for Eq. 18 is often underdetermined or not well represented by the obtained solutions, and therefore, the ensemble of structures cannot fully reflect the dynamic behavior of structures. Besides, solving Eq. 17 for each generated set of distances can be very costly.

Results

Spectral decomposition of a square distance matrix

The eigenvalue spectrum of contact matrices or Laplacian (Kirchhoff) matrices is rather complex, with only one eigenvalue out of N being zero for GNM, and six out of 3 N being zero for ANM. In the case of the square distance matrix **D** (Eq. 2) the eigenspectrum is much simpler. Spectral decomposition of a square distance matrix is a complete and simple description of a system of points and has at most five nonzero, interpretable terms:

A dominant eigenvector associated with the dominant eigenvalue is proportional to r^2—the square distance of points to the center of the mass, and the next three are principal components of the system of points. It can be shown that these principal components are related to the directionality of the secondary structure elements. This means that the square distance matrix **D** that contains almost complete information about protein structure (except impossibility to distinguish protein from its mirror image) can be completely reconstructed from the dominant r^2-related eigenvector and three eigenvectors corresponding to the principal components.

To illustrate relationships with the square distance of residues from the center of mass and the secondary structure let us consider protein G. Figure 2a shows the plots of experimental B-factors of C^{α} atoms measured by the X-ray crystallography (shown in black), mean-square fluctuations computed from the Gaussian Network Model, and the values of the square distance of C^{α} atoms from the protein center of mass plotted versus the residue index. We see that r^2 correlates with B-factors better than predictions provided by elastic network model. Figure 2b shows the plot of the first principal component versus the residue index for protein G. The relation with directionality of the secondary structure elements is obvious if we compare Fig. 2 with

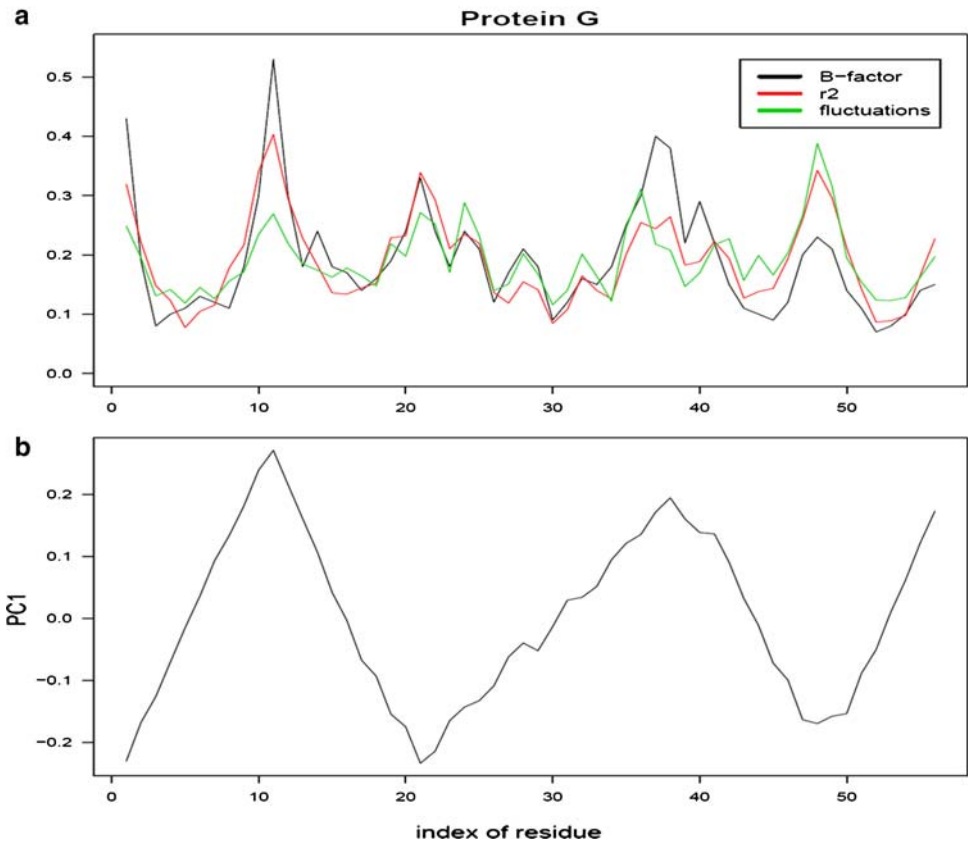

Fig. 2 **a** The plot experimental B-factors (*black*), fluctuations computed from Gaussian Network Model (*green*), and the square of the distance from the center of mass (*red*) versus the residue index for protein G. **b** The plot of the first principal component versus residue index for protein G

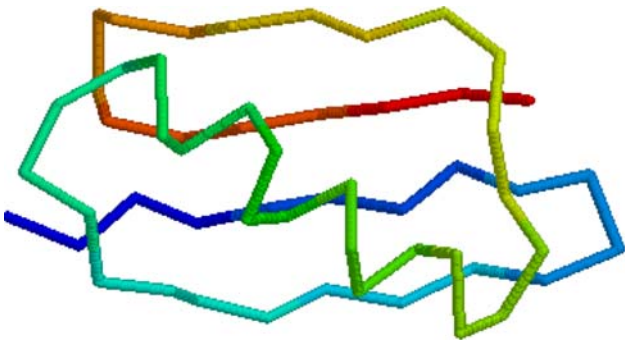

Fig. 3 Structure of protein G in the direction of the first principal component

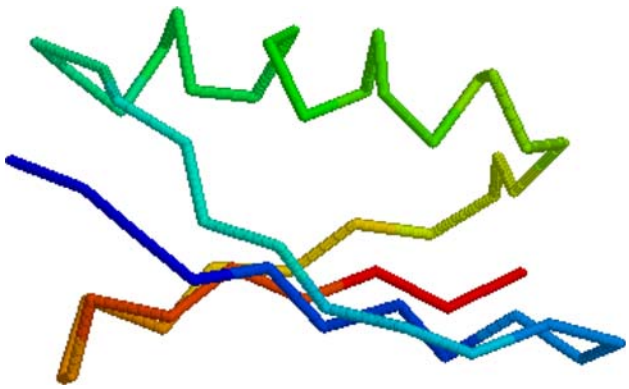

Fig. 4 Structure of protein G in the direction of the second principal component

Fig. 3 that shows protein G oriented in the direction of the first principal component.

The first principal component (Fig. 2b) increases as the residue index follows the direction of the secondary structure in proteins (Fig. 3), when the secondary structure reverses its direction the principal components starts decreasing, etc. In the case of the second (or the third) principal component the relationship between the values of these components and orientation of the secondary structure in the direction of the principal component (Fig. 4) is much more difficult to visualize.

We used a nonredundant database of 680 structures derived from the ASTRAL database and computed average correlations between experimental B-factors and various theoretically computed quantities, as well as correlations among them. We analyzed the square distance of each residue from the center of mass (r^2), principal eigenvector of the contact matrix (PECM), contact number (the number of residues being in contact) for each residue (CN), and mean-square fluctuations computed from the Gaussian Network Model (GNM). We tried also to predict B-factors from the sequence alone using Support Vector Regression (SVR) that is a variant of Support Vector Machines for continuous variables. The results of our computations are

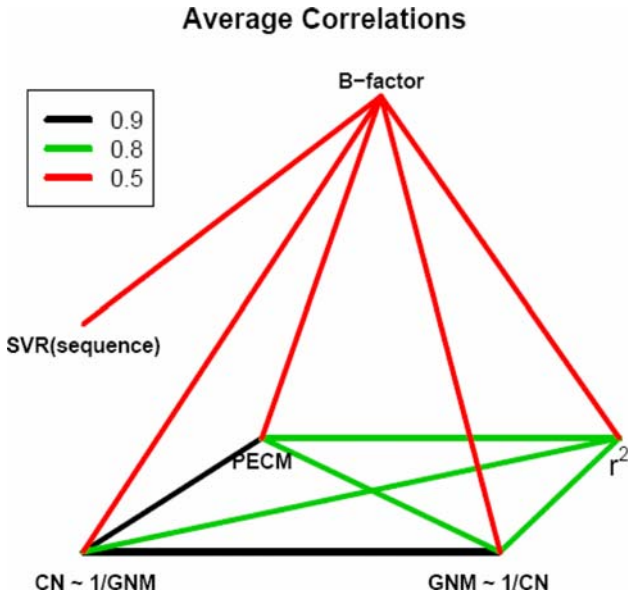

Fig. 5 Correlations among experimental B-factors, contact numbers (CN), principal eigenvectors of the contact matrix PECM, square distances from the center of mass r^2, and predictions of GNM. Support Vector Regression (SVR) predictions of B-factors from the sequence alone are also shown. The extent of correlations is illustrated by *colors*

shown in Fig. 5. Highest correlations of the order 0.9 are shown in black, correlations 0.8 are shown in green, and correlations of order 0.5 are shown in red. We see that all four quantities (r^2, PECM, CN and GNM) are very well correlated with each other. Especially the correlations between the fluctuations predicted from GNM and the inverse of the contact number CN, or PECM are surprisingly high (0.9). Accuracy of predictions of experimental B-factors from the sequence alone using SVR is almost the same ($\sim$0.5) as for predictions based on structural information contained in the contact matrix (for GNM, CN and PECM), or in the square distance matrix (for r^2).

Some of these observations have been already reported in literature. In 1980 Petsko found that crystallographic B-factors correlate with the distances of residues from the center of mass r^2 [44]. Correlations between fluctuations of residues and the inverse of their contact numbers have been pointed out by Halle [45]. Prediction of B-factors from the sequence using SVM was recently reported [46].

Approximation of distance matrices

We tried to reconstruct the original structure described by the square distance matrix by using eigenvalue decomposition (Eq. 14). The inclusion of all four terms in the summation in Eq. 14 gives to the original square distance matrix. By using only the first term related to the dominant eigenvector, or the first two terms (the dominant eigenvector and the first principal component) we can assess the contribution of these terms to the reconstruction of the original square distance matrix from the eigenvalue decomposition.

The computations were performed on our nonredundant database of 680 structures derived from the ASTRAL database. We found that the dominant eigenvector r^2 alone approximates protein structures with average RMSD 7.3 Å. However if we used two terms in Eq. 14 by combining r^2 with the first principal component the original structures were approximated with much better RMSD 4.0 Å. Addition of the second principal component would of course additionally improve these approximations. Since both r^2 and the first principal component can be predicted from the sequence alone, that allows us to predict the tertiary structure of proteins with RMSD better than 4.0 Å from the sequence. Such predictions can be based only on the predicted distances of residues from the center of mass, and prediction of the secondary structure elements and their orientation in space. We are currently working on this problem by using Support Vector Regression.

Principal component analysis of multiple HIV-1 protease structures

We used 164 X-ray-determined and 28 NMR-determined structures of HIV-1 protease deposited in PDB structures [47]. Figure 6 shows the structure of HIV-1 protease. We used also 10,000 structures (snapshots) obtained from the Molecular Dynamics simulations of HIV-1 protease. We performed the Principal Component Analysis of the structural matrices for all these three different datasets. Then we compared the results of Principal Component Analysis with normal modes computed from the Anisotropic Network Model. We computed the overlap (measured as the dot products of vectors) between directions of motions computed from ANM and principal components for X-ray determined structures and NMR-determined structures for first few slowest modes. The results are shown in Table 1 with the best cases marked in bold.

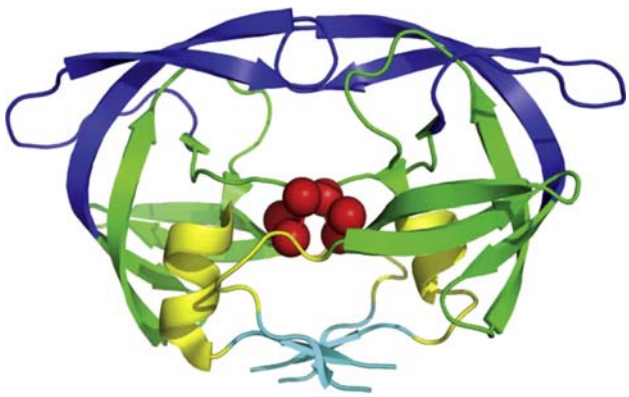

Fig. 6 The α-carbon trace of the HIV-1 structure

Table 1 Overlap between ANM directions of motions and principal components

	X-ray			NMR		
	Mode 1	Mode 2	Mode 3	Mode 1	Mode 2	Mode 3
PC1	0.06	0.06	0.24	0.25	**0.91**	0.02
PC2	0.07	0.04	**0.64**	0.88	0.28	0.04
PC3	0.46	0.53	0.13	0.02	0.05	0.30

Table 2 Cumulative overlap between directions of motions and principal components

	X-ray			NMR		
	PC1	PC2	PC3	PC1	PC2	PC3
3 Modes	0.25	0.65	0.71	0.94	0.92	0.31
6 Modes	0.25	0.65	0.74	0.95	0.94	0.35
20 Modes	0.32	0.69	**0.84**	**0.96**	0.95	0.46

Table 1 suggests that NMR-determined structures fit predictions of elastic network models better than X-ray-determined structures. This idea was further evidenced after the computation of the cumulative overlap (a sum of overlaps for the first k-modes), shown in Table 2.

NMR-derived structures fit prediction of Anisotropic Network Model much better than X-ray-derived structures. A possible explanation is that NMR experiments enable us to study single isolated molecules in solution, and elastic networks are basically also single molecule models, whereas in X-ray crystallography motions of protein residues are affected by interactions with the rest of the crystal lattice.

An optimization approach for structure determination and refinement using distances

We propose a new model for the solution of the problem defined by Eq. 17 by making a similar assumption as in X-ray crystallography that a protein has an equilibrium structure and the atoms fluctuate around their equilibrium positions. These thermal fluctuations are represented by the B-factors in the X-ray crystal structure. With this model, we can then reformulate the problem for determining an ensemble of structures for a given set of distance ranges as an optimization problem, i.e., to find the equilibrium positions and maximal possible fluctuation radii for the atoms in the protein, subject to the condition that the fluctuations should be within the given distance ranges (see Fig. 7). Let r_i be the fluctuation radius of atom i.

Then, the problem can be written as to find $\boldsymbol{x}_i$ and r_i, $i = 1, \ldots, n$ such that we maximize the total volume of

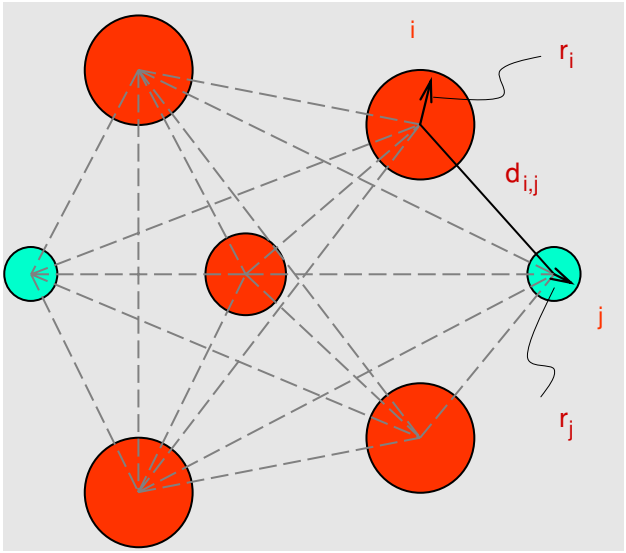

Fig. 7 Protein model. A protein is assumed to have an equilibrium structure and the atoms fluctuate around their equilibrium positions (as represented by the B-factors in an X-ray crystal structure)

spheres corresponding to fluctuations of atoms, subject to the lower and upper distance constraints imposed on interatomic distances:

$$\begin{array}{ll} \max\limits_{x_i, r_i} & \sum_{i=1}^{n} r_i^3 \\ \text{subj.} & ||\mathbf{x}_i - \mathbf{x}_j|| + r_i + r_j \le u_{i,j} \\ & ||\mathbf{x}_i - \mathbf{x}_j|| - r_i - r_j \ge l_{i,j}, \quad (i,j) \in S \\ & r_i \ge 0, \quad i = 1, \ldots, n. \end{array} \tag{19}$$

We call this problem a generalized distance geometry problem. This problem is not exactly equivalent to Eq. 18, but the solution of the problem can provide a meaningful description for the structure to be determined and its dynamic behavior. Moreover, the formulation given by Eq. 19 has many advantages over Eq. 18. First, it is a much better defined problem, because it requires only a single solution rather than a solution set. Second, it is computationally more tractable because there are well-developed methods for solving optimization problems. Third, the solution of the problem can deliver an NMR structure in a similar form as an X-ray crystal structure, with a single structural file containing the coordinates and fluctuation radii (or B-factors) for the atoms. These advantages make it possible for us to develop an efficient algorithm for the determination of a structure using a set of distance data and improve the way to represent a structural ensemble in NMR modeling.

A buildup algorithm

In practice, there can be more than tens of thousands of variables and constraints for the problem in Eq. 19. A constrained optimization problem of such complexity can

still be very difficult to solve. We therefore propose a novel so-called buildup algorithm for the solution of the problem. The idea of this algorithm is to determine the coordinate vectors and fluctuation radii of the atoms, one at a time, using the distance constraints from the determined atoms to the undetermined ones. Let $\boldsymbol{x}_j$ and r_j be the coordinate vector and fluctuation radius of an atom to be determined. Suppose that there are l determined atoms $\boldsymbol{x}_i$, $i = 1, \ldots, l$ from which the lower and upper bounds on the distances to atom j are given. Then, a subproblem for determining atom j can be formulated as follows:

$$\begin{array}{ll} \max\limits_{x_j, r_j} & r_j^3 \\ \text{subj.} & \|\mathbf{x}_i - \mathbf{x}_j\| + r_i + r_j \le u_{i,j} \\ & \|\mathbf{x}_i - \mathbf{x}_j\| - r_i - r_j \ge l_{i,j} \\ & r_i \ge 0, \quad i = 1, \ldots, l. \end{array} \tag{20}$$

This subproblem has only four variables and $2l$ constraints, and can be solved easily. By repeatedly solving such subproblems for undetermined atoms, the coordinate vectors and fluctuation radii of the all atoms in the protein can all be determined eventually. We have implemented such a buildup algorithm in Matlab and applied it to a set of test problems. We demonstrate how the algorithm works in the following.

Let us consider the structure of protein 1AX8 as an example. In order to test the algorithm, we first used the PDB data for 1AX8 to compute all the distances less than or equal to 5 Å. We then computed the root-mean-square fluctuations for all the atoms based on their B-factors. Let $\boldsymbol{y}_i$ and b_i be the coordinate vectors and B-factors for atom i, respectively, $i = 1, \ldots, n$. We then set a fluctuation radius for atom i to be

$$f_i = C\,\mathrm{sqrt}(b_i/D) \tag{21}$$

where constants C and D are the scaling factors that are evaluated by solving later the optimization problem defined by Eq. 19. Let $d_{i,j}$ be the distance between atoms i and j. We then set

$$\begin{aligned} & l_{i,j} = d_{i,j} - f_i - f_j, \\ & u_{i,j} = d_{i,j} + f_i + f_j, \\ & (i,j) \in S = \{(i,j) : d_{i,j} = \|\mathbf{y}_i - \mathbf{y}_j\| \le 5\text{Å}\}. \end{aligned} \tag{22}$$

With such a set of distance intervals, we then solve an optimization problem (Eq. 19) by using a buildup procedure. Figure 8 shows the X-ray crystal structure for 1AX8 and the equilibrium structure determined after solving Eq. 19 using the distance data given in Eq. 22. Let $\boldsymbol{Y} = \{\boldsymbol{y}_i, i = 1, \ldots, n\}$ and $\boldsymbol{X} = \{\boldsymbol{x}_i, i = 1, \ldots, n\}$ be two $n \times 3$ coordinate matrices for the two structures, respectively. Then, RMSD $(\boldsymbol{X}, \boldsymbol{Y}) = 2.0\text{e}{-}04$ Å, showing that the two structures are almost the same. After solving Eq. 19, we have also obtained the fluctuation radii for the atoms. Figure 9 shows the computed radii r_i and the radii f_i derived from the B-factors of the crystal structure, $i = 1, \ldots, n$. Clearly, the two sets of radii correlate very well.

Structure refinement using statistical distances

We propose a computational approach to refining an NMR structure (and possibly other types of structures as well) by statistically deriving additional distance data from a large set of known protein structures. General idea of our approach is based on earlier work of Sippl [48–53], Melo and Feytmans [54, 55], Garbuzynskiy [56] among others.

By statistically deriving additional distance data, we mean that we can search for the distances between certain pairs of atoms, especially for those missing in the experimental data, in a database of known protein structures such as PDB, and then obtain a statistical distribution of each distance type, say the distance between the two C_β atoms in two neighboring residues, alanine (ALA) and tryptophan (TRP). Using these distributions, a probable range or a

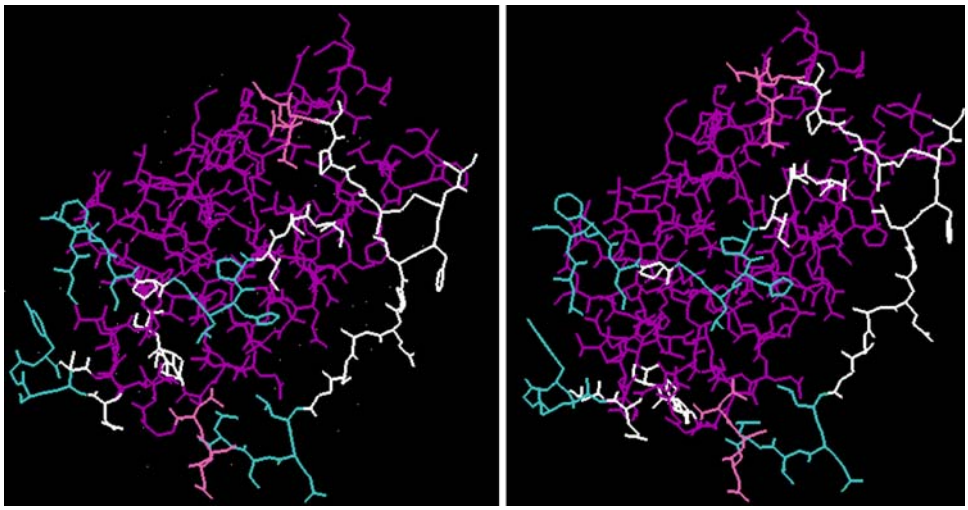

Fig. 8 Computational versus experimental structures. The structure on the left is the crystal structure for protein 1AX8. The one on the right is the structure for 1AX8 obtained by solving a generalized distance geometry problem using a set of distance bounds

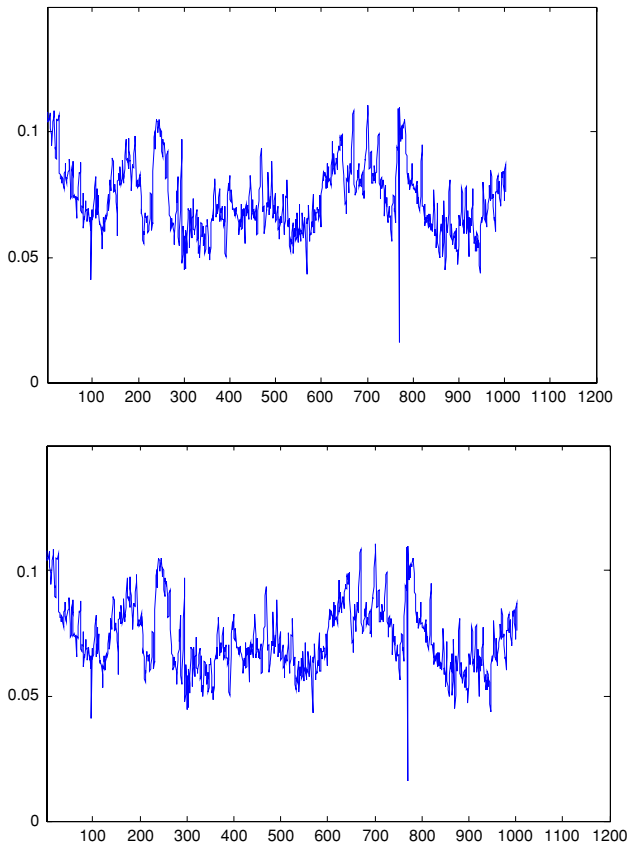

Fig. 9 Atomic fluctuation radii. The upper graph shows the fluctuation radii of the atoms extracted from the B-factors of the original crystal structure of protein 1AX8. The lower graph shows the fluctuation radii obtained by solving a generalized distance geometry problem

mean-force potential of each distance type can be defined, and applied to refining a structure.

Consider the distances between two atoms in two residues separated by some residues in sequence. Let A_1 and A_2 be the two atoms, R_1 and R_2 the two residues, and S_1, …, S_N the residues between R_1 and R_2. Let the distances between A_1 and A_2 in R_1 and R_2 separated by S_1, …, S_N be collected from a database of know protein structures and grouped into a set of uniformly divided distance intervals $[D_i, D_{i+1}]$, where $D_i = 0.1 * i$ Å, $i = 0, 1, \ldots, n-1$. Then, the distribution of this particular type of distances can be defined by a function $P[A_1, A_2, R_1, R_2, S_1, \ldots, S_N](D)$ for any distance D, and

$$P[A_1, A_2, R_1, R_2 S_1, \ldots, S_N](D) = \frac{\#\text{distances in } [D_i, D_{i+1}]}{\#\text{ distances in } [D_0, D_n]}, \quad D \in [D_i, D_{i+1}]. \tag{23}$$

The distribution graphs for most distance types should have non-uniform patterns if the two residues are not too far apart. This is primarily due to the fact that large portions of protein segments form regular secondary

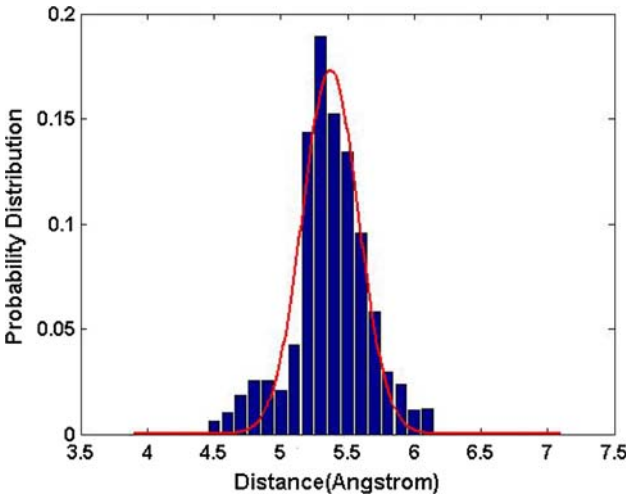

Fig. 10 Statistical distances. The distances of a specific type have been collected from known protein structures and used to form a probability distribution function of the distances

structures, i.e., α-helices or β-sheets, where short-range distances always have certain ranges (see Fig. 10). Based on the distribution of the distances of a given type, we can extract a probable range for the distances by using the mean minus and plus a few standard deviations of the distances. Let l and u be the lower and upper bounds of the distances between A_1 and A_2 in R_1 and R_2 separated by S_1, …, S_N. We can define $l = \mu - k\sigma$ and $u = \mu + k\sigma$, where μ and σ are the mean and the standard deviation of $P[A_1, A_2, R_1, R_2, S_1, \ldots, S_N]$ and k is a constant. Alternatively, we can also use the distribution of the distances to define a mean-force potential.

For example, for the distances between A_1 and A_2 in R_1 and R_2 separated by S_1, …, S_N, we can define a potential function E such that for any distance D of this type

$$E[A_1, A_2, R_1, R_2, S_1, \ldots, S_N](D) = -kT \log P[A_1, A_2, R_1, R_2, S_1, \ldots, S_N](D) \tag{24}$$

where k_B is the Boltzmann constant and T the temperature.

Once a set of distance bounds or mean-force potentials are obtained, we can impose the bounds on a structure to be refined or include the mean-force potentials in energy minimization so that a more plausible structural model may be built.

Results of structure refinement

We have downloaded around 2000 X-ray crystal structures with resolution of ≤2.0 Å and sequence similarity of ≤90% from PDB, and calculated a set of short-range distances and their distributions [57]. The types of the distances calculated can be specified in terms of five parameters $[A_1, A_2, R_1, R_2, S]$, where A_1 and A_2 are the atoms, R_1 and R_2 the residues, and S the residue separating R_1 and R_2. Also, only five different

types of atoms were considered: the amide N, C_α, and the carbonyl C and O along the backbone and the carbon C_β in the side-chain. The residue types included all twenty different amino acid types. For convenience, we call them cross-residue distances. For each set of A_1, A_2, R_1, R_2, and S, all corresponding distances in the downloaded crystal structures were computed and collected into a set of uniformly divided distance intervals $[D_i, D_{i+1}]$, where $D_i = 0.1\ i$ Å, $i = 0, 1, \ldots, 200$. The distribution function $P[A_1, A_2, R_1, R_2, S](D)$ for any D in $[D_i, D_{i+1}]$ was defined as the number of distances in $[D_i, D_{i+1}]$ normalized by the total occurrences of distances in all intervals.

The distribution functions for a subset of cross residue distances were used to generate a set of bound constraints for the corresponding distance types, with the lower and upper bounds equal to the mean values of the distances minus and plus twice the standard deviations, respectively. The generated distance bounds were then taken as additional distance constraints to refine a set of NMR structures, including five structures for 1EPH, 1GB1, 1IGL, 2IGG, 2SOB and five for 1CEY, 1CRP, 1E8L, 1ITL, 1PFL. The last five were selected because they have X-ray structures available. The original NMR experimental constraints for the structures were downloaded from NMR structure database BioMagResBank [58]. The structures were refined using the default torsion angle dynamic simulated annealing protocol implemented in CNS [59, 60]. The results obtained with and without additional database distance constraints were examined on the deviations of all simple cross-residue distances from their average distributions, and compared and assessed in terms of several criteria used in NMR modeling, including the acceptance rates of the structures, the RMSD values of the ensembles of structures, and the RMSD values of the structures compared with their X-ray structures (for available ones).

The distribution functions for a set of cross residue distances were also used to define a set of mean force potentials [61]. Let P be the distribution function for any distance of interest between two atoms. Then, the mean-force potential E for the distance was computed from Eq. 24. The potentials for all the cross residue distances were then summed up and inserted into the energy function in CNS software. The extended energy function was applied to refining a set of selected NMR structures. Again, the original NMR experimental constraints for the structures were downloaded from NMR structure database BioMagResBank. The embedding and energy minimization routines in CNS were used for the refinement. The results obtained with and without using the mean-force potentials were compared and assessed in terms of several standard measures, including the potential energy of the structures in various categories, the RMSD values of the ensembles of structures, and the RMSD values of the structures compared with their X-ray reference structures (for available ones), and the Ramachandran plots.

Table 3 RMSD of the ensembles of refined NMR structures

Protein	#Res	Data	Means ± standard deviations[a]	
			Backbone[b]	Non-H[c]
1EPH	53	NMR	2.04 ± 0.61	2.94 ± 0.70
		NMR + DB	1.78 ± 0.40	2.76 ± 0.54
1GB1	56	NMR	0.45 ± 0.12	1.04 ± 0.18
		NMR + DB	0.38 ± 0.09	0.91 ± 0.16
1IGL	67	NMR	4.50 ± 1.52	5.49 ± 1.55
		NMR + DB	3.81 ± 1.24	4.70 ± 1.43
2IGG	64	NMR	2.62 ± 0.85	3.29 ± 0.83
		NMR + DB	2.16 ± 0.90	2.87 ± 0.85
2SOB	103	NMR	7.25 ± 1.60	8.06 ± 1.67
		NMR + DB	5.54 ± 1.77	6.41 ± 1.77

[a] The means and standard deviations of the RMSD values of the structure ensembles refined with and without database distance constraints

[b] RMSD values in terms of backbone atoms

[c] RMSD values in terms of all non-hydrogen atoms

As shown in Table 3, the means and standard deviations of the RMSD values for the listed ensembles of NMR structures all became smaller after the structures were refined with the statistically derived distance constraints. Note that the RMSD values were calculated in terms of either just backbone atoms or all non-hydrogen atoms. The results were consistent in both calculations.

The refined NMR structures for five proteins (1CEY, 1CRP, 1E8L, 1ITL, and 1PFL) were compared with their corresponding X-ray structures for the RMSD values of the pairs of NMR and X-ray structures. As shown in Table 4, both means and standard deviations of the RMSD values for the ensembles of structures refined with the derived distance constraints were clearly smaller than those refined without them, indicating strongly that the structures agreed

Table 4 Refined NMR Structures Compared to X-ray Structures

NMR ID	X-ray ID	#Res	Means ± standard deviations[a]	
			NMR[b]	NMR + DB[c]
1CEY	3CHY	128	1.85 ± 0.19	1.80 ± 0.17
1CRP	1IAQ_A	166	1.77 ± 0.29	1.60 ± 0.26
1E8L	193L	129	2.05 ± 0.22	2.02 ± 0.19
1ITL	1RCB	129	2.88 ± 0.76	2.79 ± 0.21
1PFL	1FIK	139	1.66 ± 0.07	1.65 ± 0.07

[a] The means and standard deviations of the RMSD values for the ensembles of NMR structures compared with their X-ray structures

[b] Refined with only NMR distance constraints

[c] Refined with NMR and database distance constraints

Table 5 Energy of refined NMR structures (KJ/MOL)

PDB	Method	Overall	Bond	Angle	Improper	Van der Waals	NOE	DIH
1AFI	CNS	160.9 ± 72.0	6.2 ± 3.3	63.6 ± 18.8	8.4 ± 7.2	54.2 ± 21.7	27.6 ± 20.1	0.9 ± 0.9
	CNS + PMF	122.1 ± 56.5	4.2 ± 2.3	53.9 ± 15.8	6.2 ± 4.7	37.8 ± 17.3	19.0 ± 15.4	1.0 ± 1.1
1BA4	CNS	93 ± 60.8	4.0 ± 3.0	34.3 ± 21.8	4.4 ± 5.9	26.0 ± 14.3	24.3 ± 15.9	NA
	CNS + PMF	57.8 ± 14.7	2.1 ± 0.7	24.1 ± 3.7	2.1 ± 1.2	17.1 ± 4.0	12.4 ± 5.2	NA
1DKC	CNS	155.7 ± 90.1	7.4 ± 4.1	40.1 ± 10.6	4.7 ± 2.5	48.9 ± 48.6	54.6 ± 24.3	NA
	CNS + PMF	118.6 ± 40.4	5.2 ± 2.0	31.4 ± 8.1	3.2 ± 2.1	34.6 ± 12.4	44.3 ± 15.8	NA
1DVV	CNS	85.6 ± 19.6	3.1 ± 0.9	40.7 ± 5.8	4.0 ± 1.1	23.7 ± 7.8	14 ± 5.2	0.05 ± 0.06
	CNS + PMF	73.3 ± 15.8	2.5 ± 0.9	37.5 ± 3.7	3.5 ± 0.9	18.4 ± 4.7	11.2 ± 5.5	0.03 ± 0.02
1I6F	CNS	190.0 ± 73.2	1.4 ± 2.1	24.4 ± 8.8	1.3 ± 1.9	113.8 ± 47.3	48.9 ± 12.9	0.16 ± 0.47
	CNS + PMF	173.8 ± 8.3	0.9 ± 0.3	22.6 ± 1.8	0.9 ± 0.5	103.4 ± 3.3	45.9 ± 2.4	0.06 ± 0.09

Listed are means and standard deviations of the energies of the structural ensembles in various categories: Overall, total energy; Bond, bond-length energy; Improper, improper angle energy; Van der Waals, Van der Waals interaction energy, NOE, energy for NOE distance constraint satisfaction; DIH, energy for dihedral angle constraints. CNS, refined with original NMR data and CNS built-in energy function. CNS + PMF, refined with original NMR data, CNS built-in energy function, and database derived mean-force potentials

more closely with their reference structures after being refined with the derived distance constraints.

As a case study, we have also applied the derived distance constraints to refining the NMR structure of the human PrP^c E200 K variant of the prion protein. Two biologically critical but under-determined loop regions (residues 167–171 and 195–199) were targeted particularly for improvement. The Ramachandran plots of the average and energy-minimized structure and the lowest energy structure of the refined structural ensemble showed a significantly higher percentage (89.6%) of residues in the most favorable regions of the plots than the 85.4% of such residues found in the regularly refined structures, which was a clear indication on the improvement of the structures due to the use of the statistically derived distance constraints. Table 5 shows the energy values for a list of refined structures in various categories and in particular, the means and standard deviations of the energy values in each structural ensemble. Note that for a fair comparison, the calculation of the overall energy did not count the contribution from the mean-force potentials although the latter were used in the CNS + PMF refinement. Note also that the energy due to electrostatic interactions was not listed because the corresponding potentials were not included in the default CNS refinement protocol. Table 3 shows that the means and standard deviations of the energy values of the ensembles of structures became smaller in almost all categories after the structures were refined with the addition of the mean force potentials. The results suggested that the refined structures, when using the mean-force potentials, were clearly more favorable energetically. Surprisingly, they also satisfied the experimental constraints better as the NOE and DIH energies were decreased in many cases as well. Overall, in terms of the means and standard deviations of the energy values in the structural ensembles, of the 70 selected NMR structures, about 80% had the overall energy significantly reduced, in average by 7.5%, and about 65% had the NOE energy decreased, in average by 5%, after refined with mean-force potentials. Here we have not calculated the statistics for the DIH energy because some structures did not have the DIH data and energy available.

Refining comparative models

We have also participated in the CASPR 2006 structure refinement experiments. In these experiments eight structural models (predicted with comparative modeling) were provided for further refinement. The RMSD values of the models compared with PDB structures ranged from 2.0 to 5.0 Å. To illustrate our methodology we focus on a model of a protein with 70 residues and a 2.19 Å RMSD from its crystal structure (1WHZ, see Fig. 11). We have used the following procedure to refine the structure. First, 16 different structures were generated by randomly perturbing the residues of the target structure. Energy minimization was then carried out using CHARMM [62] with the generated structures as starting points. Of the 16 obtained minima, four were selected randomly, and each was used to generate 16 more structures for further energy minimization. The process was repeated until the maximum number of structures was generated.

In the end, total 100 minimum energy structures were selected from the structures obtained in the energy minimization stage. Based on the energy values and the Ramachandran plots of the structures, a small set of structures were selected and the one with both low energy and good residual distribution in the Ramachandran plot was used as an initial model. The RMSD value of the initial model against the experimental structure was 1.92 Å. From

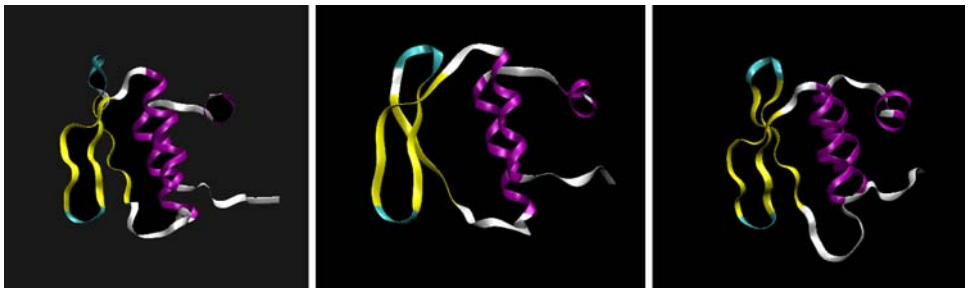

Fig. 11 Refining comparative models. From left to right are the crystal structure, the given target, and the refined model. The RMSD of the given target is 2.19 Å against the crystal structure, while the refined model is 1.80 Å. The differences can be observed in both the helical (*purple*) and extended (*yellow*) regions

this initial model, a large set of distances between atoms contact distances was computed. A set of lower and upper bounds for the distances was then generated by subtracting 20% from or adding 20% to the distances. Then, the CNS NMR refinement protocols were used to further refine the model with the generated distance constraints.

The distributions of distances between certain pairs of atoms, especially the distances between heavy atoms in different residues separated by several residues in the primary sequence, were also computed. A set of mean-force potentials for the distances was constructed using the distribution functions, and was added to the CNS energy function. The initial model was refined with the modified energy function. Total 50 structures were generated by CNS as an ensemble of models for the protein. The structures were analyzed based on their total energies and residual distributions in the Ramachandran plots. The one with both low energy and good Ramachandran plot was selected as the final model. This model had a 1.80 Å RMSD from the experimental 1WHZ structure. The improvement in this sense was significant compared to the RMSD value (2.19 Å) of the original model.

Discussion

Conclusions

We show that mathematical approach based on distance matrices is very powerful and enable us to predict protein structure from the sequence. The information contained in the square distance of residues from the center of mass, and the first principal component allows us to reconstruct protein structure with RMSD 4.5 Å. We demonstrate that crystallographic B-factors can be predicted from the sequence using Support Vector Regression. We also prove that protein structures can be refined by using statistical interatomic distances, and that generalized distance geometry problem for solving NMR structures based on distances between atoms subject to upper and lower bounds can be reduced to an optimization problem that involves maximization of the volume of spheres with the radii equal to the range of corresponding thermal fluctuations of atoms. All methods presented are still being improved and may lead to a significant progress in prediction of protein structure and dynamics and to substantial refinement of protein models.

Summary

We have applied distance matrices and the related contact matrices to several different, although interconnected problems relevant to structural bioinformatics. We have performed eigenvalue decomposition of square distance matrices, and we have shown that a dominant eigenvector is proportional to r^2—the square distance of points from the center of mass, while the next three eigenvectors are the principal components of the system of points. We have shown that both the dominant eigenvector and the first principal component can be predicted from the sequence alone that allows us to predict the tertiary structure of proteins from sequence with RMSD around 4.0 Å.

We have performed elastic network analysis (based on contact matrices) of the large number of available HIV-1 protease structures, and have shown that they provide a remarkable sampling of conformations, which can be viewed as direct structural information about the dynamics. Finally, we have used distance constraints from databases of known protein structures for structure refinement.

Acknowledgements It is a pleasure to acknowledge the financial support provided by the National Institutes of Health through grants 1R01GM081680, 1R01GM072014, and 1R01GM073095.

References

1. Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A (2005) Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins: Struct Funct Bioinform 59:49–57. doi:10.1002/prot.20380
2. Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28:374. doi:10.1093/nar/28.1.374
3. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205. doi:10.1093/nar/gkm998
4. Pokarowski P, Kloczkowski A, Nowakowski S, Pokarowska M, Jernigan RL, Kolinski A (2007) Ideal amino acid exchange forms

for approximating substitution matrices. Proteins: Struct Funct Bioinform 69:379–393. doi:10.1002/prot.21509
5. Bastolla U, Porto M, Roman HE, Vendruscolo M (2005) Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins: Struct Funct Bioinform 58:22–30. doi:10.1002/prot.20240
6. Choi IG, Kwon J, Kim SH (2004) Local feature frequency profile: a method to measure structural similarity in proteins. Proc Natl Acad Sci USA 101:3797–3802. doi:10.1073/pnas.0308656100
7. Domingues FS, Rahnenfuhrer J, Lengauer T (2007) Conformational analysis of alternative protein structures. Bioinformatics 23:3131–3138. doi:10.1093/bioinformatics/btm499
8. Godzik A, Skolnick J, Kolinski A (1993) Regularities in interaction patterns of globular-proteins. Protein Eng 6:801–810. doi:10.1093/protein/6.8.801
9. Heger A, Lappe M, Holm L (2004) Accurate detection of very sparse sequence motifs. J Comput Biol 11:843–857. doi:10.1089/cmb.2004.11.843
10. Holm L, Park J (2000) DaliLite workbench for protein structure comparison. Bioinformatics 16:566–567. doi:10.1093/bioinformatics/16.6.566
11. Huang YM, Bystroff C (2006) Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. Bioinformatics 22:413–422. doi:10.1093/bioinformatics/bti828
12. Jaroszewski L, Li WZ, Godzik A (2002) In search for more accurate alignments in the twilight zone. Protein Sci 11:1702–1713. doi:10.1110/ps.4820102
13. Kolodny R, Linial N (2004) Approximate protein structural alignment in polynomial time. Proc Natl Acad Sci USA 101:12201–12206. doi:10.1073/pnas.0404383101
14. Mooney SD, Liang MHP, DeConde R, Altman RB (2005) Structural characterization of proteins using residue environments. Proteins: Struct Funct Bioinform 61:741–747. doi:10.1002/prot.20661
15. Pazos F, Valencia A (2008) Protein co-evolution, co-adaptation and interactions. EMBO J 27:2648–2655. doi:10.1038/emboj.2008.189
16. Rodionov MA, Galaktionov SG (1992) Analysis of the 3-dimensional structure of proteins in terms of residue contact matrices. 1. The contact criterion. Mol Biol 26:773–776
17. Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21:3482–3489. doi:10.1093/bioinformatics/bti564
18. Sato T, Yamanishi Y, Horimoto K, Kanehisa M, Toh H (2006) Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. Bioinformatics 22:2488–2492. doi:10.1093/bioinformatics/btl419
19. Schneider TR (2000) Objective comparison of protein structures: error-scaled difference distance matrices. Acta Crystallogr D Biol Crystallogr 56:714–721. doi:10.1107/S0907444900003723
20. Snyder DA, Montelione GT (2005) Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. Proteins: Struct Funct Bioinform 59:673–686. doi:10.1002/prot.20402
21. Snyder DA, Bhattacharya A, Huang YPJ, Montelione GT (2005) Assessing precision and accuracy of protein structures derived from NMR data. Proteins: Struct Funct Bioinform 59:655–661. doi:10.1002/prot.20499
22. Szustakowski JD, Weng ZP (2000) Protein structure alignment using a genetic algorithm. Proteins-Structure Funct Genet 38:428–440. doi:10.1002/(SICI)1097-0134(20000301)38:4<428::AID-PROT8>3.0.CO;2-N
23. Ye JP, Janardan R (2004) Approximate multiple protein structure alignment using the sum-of-pairs distance. J Comput Biol 11:986–1000. doi:10.1089/cmb.2004.11.986
24. Zhou XB, Chou J, Wong STC (2006) Protein structure similarity from principle component correlation analysis. BMC Bioinformatics 7:40 (10pp)
25. Flory PJ (1976) Statistical thermodynamics of random networks. Proc R Soc Lond A: Math Phys Eng Sci 351:351–380
26. Kloczkowski A, Mark JE, Erman B (1989) Chain dimensions and fluctuations in random elastomeric networks 1 phantom Gaussian networks in the undeformed state. Macromolecules 22:1423–1432. doi:10.1021/ma00193a070
27. Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2:173–181. doi:10.1016/S1359-0278(97)00024-2
28. Haliloglu T, Bahar I, Erman B (1997) Gaussian dynamics of folded proteins. Phys Rev Lett 79:3090–3093. doi:10.1103/PhysRevLett.79.3090
29. Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 77:1905–1908. doi:10.1103/PhysRevLett.77.1905
30. Kundu S, Melton JS, Sorensen DC, Phillips GN (2002) Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J 83:723–732. doi:10.1016/S0006-3495(02)75203-X
31. Sen TZ, Feng YP, Garcia JV, Kloczkowski A, Jernigan RL (2006) The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. J Chem Theory Comput 2:696–704. doi:10.1021/ct600060d
32. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80:505–515. doi:10.1016/S0006-3495(01)76033-X
33. Keskin O, Bahar I, Flatow D, Covell DG, Jernigan RL (2002) Molecular mechanisms of chaperonin GroEL-GroES function. Biochemistry 41:491–501. doi:10.1021/bi011393x
34. Keskin O, Durell SR, Bahar I, Jernigan RL, Covell DG (2002) Relating molecular flexibility to function: a case study of tubulin. Biophys J 83:663–680. doi:10.1016/S0006-3495(02)75199-0
35. Navizet I, Lavery R, Jernigan RL (2004) Myosin flexibility: structural domains and collective vibrations. Proteins-Structure Funct Genet 54:384–393. doi:10.1002/prot.10476
36. Wang YM, Rader AJ, Bahar I, Jernigan RL (2004) Global ribosome motions revealed with elastic network model. J Struct Biol 147:302–314. doi:10.1016/j.jsb.2004.01.005
37. Wang YM, Jernigan RL (2005) Comparison of tRNA motions in the free and ribosomal bound structures. Biophys J 89:3399–3409. doi:10.1529/biophysj.105.064840
38. Yan A, Wang Y, Kloczkowski A, Jernigan RL (2008) Effects of protein subunits removal on the computed motions of partial 30S structures of the ribosome. J Chem Theory Comput 4:1757–1767
39. Crippen GM, Havel TF (1978) Stable calculation of coordinates from distance information. Acta Crystallogr A 34:282–284. doi:10.1107/S0567739478000522
40. Havel TF, Crippen GM, Kuntz ID (1979) Effects of distance constraints on macromolecular conformation. 2. Simulation of experimental results and theoretical predictions. Biopolymers 18:73–81. doi:10.1002/bip.1979.360180108
41. Havel TF, Kuntz ID, Crippen GM (1983) The combinatorial distance geometry method for the calculation of molecular-conformation. 1. A new approach to an old problem. J Theor Biol 104:359–381. doi:10.1016/0022-5193(83)90112-1
42. Havel TF, Crippen GM, Kuntz ID, Blaney JM (1983) The combinatorial distance geometry method for the calculation of molecular-conformation. 2. Sample problems and computational statistics. J Theor Biol 104:383–400. doi:10.1016/0022-5193(83)90113-3

43. Havel TF, Kuntz ID, Crippen GM (1983) The theory and practice of distance geometry. Bull Math Biol 45:665–720
44. Petsko GA, Frauenfelder H (1980) Crystallographic approaches to the dynamics of ligand-binding to myoglobin. Fed Proc 39:1648
45. Halle B (2002) Flexibility and packing in proteins. Proc Natl Acad Sci USA 99:1274–1279. doi:10.1073/pnas.032522499
46. Chen P, Wang B, Wong HS, Huang DS (2007) Prediction of protein B-factors using multi-class bounded SVM. Protein Pept Lett 14:185–190. doi:10.2174/092986607779816078
47. Yang L, Song G, Carriquiry A, Jernigan RL (2008) Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. Structure 16:321–330. doi:10.1016/j.str.2007.12.011
48. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force—an approach to the knowledge-based prediction of local structures in globular-proteins. J Mol Biol 213:859–883
49. Sippl MJ (1992) Detection of native-like models for amino-acid-sequences of unknown 3-dimensional structure in a data-base of known protein conformations. Proteins 13:258–271
50. Sippl MJ (1993) Recognition of errors in 3-dimensional structures of proteins. Proteins-Structure Funct Genet 17:355–362. doi:10.1002/prot.340170404
51. Sippl MJ (1995) Knowledge-based potentials for proteins. Curr Opin Struct Biol 5:229–235
52. Sippl MJ, Scheraga HA (1986) Cayley-Menger coordinates. Proc Natl Acad Sci USA 83:2283–2287. doi:10.1073/pnas.83.8.2283
53. Sippl MJ, Scheraga HA (1985) Solution of the embedding problem and decomposition of symmetric-matrices. Proc Natl Acad Sci USA 82:2197–2201. doi:10.1073/pnas.82.8.2197
54. Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. J Mol Biol 267:207–222. doi:10.1006/jmbi.1996.0868
55. Melo F, Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 277:1141–1152. doi:10.1006/jmbi.1998.1665
56. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV (2005) Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray and NMR-resolved protein structures? Proteins: Struct Funct Bioinform 60:139–147. doi:10.1002/prot.20491
57. Wu D, Cui F, Jernigan R, Wu ZJ (2007) PIDD: database for protein inter-atomic distance distributions. Nucleic Acids Res 35:D202–D207. doi:10.1093/nar/gkl802
58. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao HY, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408. doi:10.1093/nar/gkm957
59. Brunger AT, Adams PD, Clore GM, Delano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921. doi:10.1107/S0907444998003254
60. Brunger AT (2007) Version 1.2 of the crystallography and NMR system. Nat Protoc 2:2728–2733. doi:10.1038/nprot.2007.406
61. Wu D, Jernigan R, Wu ZJ (2007) Refinement of NMR-determined protein structures with database derived mean-force potentials. Proteins: Struct Funct Bioinform 68:232–242. doi:10.1002/prot.21358
62. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) Charmm—a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217. doi:10.1002/jcc.540040211