A methodology for determining mechanical properties of macromolecules from ensemble motion data

Gregory S. Chirikjian

This paper presents a methodology for directly determining macromolecular stiffness parameters from ensemble measurements of interatomic distances. These stiffness parameters can be used to derive empirical statistical potentials which, by definition, have the effects of solvent built in. A Gaussian network model is used together with methods from equilibrium statistical mechanics to formulate the problem. Ensemble distance measurements could come from a number of experimental modalities including FRET or NMR. The computational method presented here relies on the existence of a complete baseline structure (e.g., from crystallography), but no *a-priori* assumption of interatomic potentials is required.

Keywords: Elastic network model; Inverse problem; Macromolecules; Stiffness

Gregory S. Chirikjian*
Department of Mechanical Engineering, The Johns Hopkins University, Baltimore, MD 21218, USA

*Corresponding author;
E-mail: gregc@jhu.edu

1. Introduction

The shape fluctuations and conformational transitions exhibited by biological macromolecules provide insight into the relationship between structure and function. Having the ability to predict motion reliably from structure would therefore assist in our understanding of the machinery of life.

One approach to gaining insight into the structure-motion-function relationship is molecular dynamics (MD) simulation [1–3]. This is a method where Newtonian mechanics is used together with classical potentials to formulate a very large system of non-linear, second order, ordinary differential equations that describe the motion of the system. In principle, if the position and velocity of all atoms within the macromolecule and the surrounding solvent are known at an initial time, then these equations can be integrated to find a trajectory of the motion of all atoms within the system (provided quantum mechanical effects can be considered negligible). While the results of MD are promising, one well-recognized limitation is the high computational cost.

Incomplete information about the unstructured solvent environment and the sensitivity of these methods to the form and values of potential functions has led some to investigate coarse-grained models of macromolecular motion. In that school of thought, principles of equilibrium statistical mechanics and chemical physics are used to extract effective potentials from large numbers of crystal structures [4]. The basic idea is that given a sufficiently large collection of non-homologous structures in the Protein Data Bank (PDB) [5], one can obtain useful statistical potentials.

The goal of this paper is to introduce a new methodology for extracting stiffness parameters directly from ensemble motion data of biological macromolecules in solution rather than from crystal structures. These parameters define Gaussian contact potentials for macromolecules that fluctuate around an equilibrium in solution.

A Gaussian network model built from a baseline crystal structure is used together with methods from equilibrium statistical mechanics to formulate the problem. This

0165-9936/03/$ - see front matter © 2003 Published by Elsevier B.V. doi:10.1016/S0165-9936(03)00909-9

model can be made to include atomic detail (in which case, the stiffness matrix for the molecule is exactly the form of the Hessian matrix that would be obtained from Amber or Charmm), or it can be constructed using coarse-grained models as in [6–12]. Ensemble distance measurements serve as the input to this model. Such measurements could come from a number of experimental modalities including FRET or NMR.

2. Principles and models from mechanics

This section describes a discrete mechanical model for small conformational fluctuations in macromolecules (in particular, proteins) about an equilibrium conformation.

Consider a protein in which the mass of the i^{th} residue is labeled as m_i, and the interaction between residues i and j is modeled with a linear spring having stiffness $k_{i,j}$. Given the full set of masses, stiffnesses and C_α positions, one can derive the global mass matrix and the global stiffness matrix. Our goal is to illustrate how unknown parameters in the stiffness matrix can be estimated from experimental measurements of motion.

3. Classical equilibrium statistical mechanics

In classical statistical mechanics, the partition function is defined as

$$Z = \int_{\mathbf{q}}\int_{\mathbf{p}} \exp(-\beta\mathcal{H}(\mathbf{p},\mathbf{q}))\ d\mathbf{p}d\mathbf{q} \tag{1}$$

where $\beta = 1/k_BT$, $p_i=\mathbf{p}\cdot\mathbf{e}_i$ is the momentum conjugate to the i^{th} generalized coordinate $q_i=\mathbf{q}\cdot\mathbf{e}_i$, $\mathcal{H}$ is the Hamiltonian for the system, and $d\mathbf{p}d\mathbf{q}=dp_1\cdots dp_N dq_1\cdots dq_N$ for a system with N degrees of freedom. The range of integration is over all possible states of the system.

For any classical mechanical system the Hamiltonian is of the form

$$\mathcal{H}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\mathbf{p}^T\left\{M^{-1}(\mathbf{q})\right\}\mathbf{p} + V(\mathbf{q})$$

where $V(\mathbf{q})$ is the potential energy and $M(\mathbf{q})$ is the mass matrix.

For a macromolecule fluctuating about one conformation that globally minimizes its potential energy, the potential energy function can be expressed as [13]

$$V(\mathbf{q}) \approx V_0 + \frac{1}{2}\mathbf{q}^T K\mathbf{q} \tag{2}$$

where the elements of K are

$$k_{ij} = \frac{\partial^2 V}{\partial q_i \partial q_j}\Big|_{\mathbf{q}=\mathbf{0}}$$

and $\mathbf{q}=\mathbf{0}$ is defined to be the value for which $V(\mathbf{0})=V_0$ is the minimum attainable potential energy. By appropriate choice of datum, one can take $V_0=0$. Since $\mathbf{q}(t)$ never strays far from $\mathbf{0}$, it follows that $M(\mathbf{q})$ is approximated well as the constant matrix $M=M(\mathbf{0})$.

Therefore,

$$\begin{aligned} Z &\approx \int_{\mathbf{q}}\int_{\mathbf{p}} \exp\left(-\beta\left\{\frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T K\mathbf{q}\right\}\right) d\mathbf{p}d\mathbf{q} \\ &= \frac{(2\pi/k_BT)^N}{\left|\det\left\{M^{-\frac{1}{2}}KM^{-\frac{1}{2}}\right\}\right|^{\frac{1}{2}}} \end{aligned} \tag{3}$$

where M and K are constant matrices. Equation 3 holds for systems with one global minimum that is deeper than, and well separated from, any local energy minima. This equation holds regardless of whether Cartesian or internal coordinates are used to describe the motion.

The following subsection presents an intuitive geometric method for constructing K, in which one or more stiffness parameters are left free. In a subsequent section, it is shown how these parameters can be fixed based on experimental measurements of molecular motion.

4. Elastic network model

Given a set of C_α crystal structure coordinates for a protein, $\{\boldsymbol{x}_i(0)\}$, the Cartesian displacement of the i^{th} α-carbon at time t can be written without loss of generality as

$$\boldsymbol{x}_i(t) = \boldsymbol{x}_i(0) + \boldsymbol{\delta}_i(t) \tag{4}$$

We define $\boldsymbol{\delta}_i$(t) to be a vector of small displacements.

The total kinetic energy in a network of n residues (each of which is treated as a point mass) then has the form

$$T = \frac{1}{2}\sum_{i=1}^{n} m_i\left\|\dot{\boldsymbol{x}}_i(t)\right\|^2 = \frac{1}{2}\dot{\boldsymbol{\delta}}^T M\dot{\boldsymbol{\delta}} \tag{5}$$

where the constant matrix M is the global mass matrix for the whole network and

$$\boldsymbol{\delta} = \left[\boldsymbol{\delta}_1^T, \cdots, \boldsymbol{\delta}_n^T\right]^T \in \mathbb{R}^{3n} \tag{6}$$

In the current context, M is diagonal.

The total potential energy in a network of connected springs has the form

$$V = \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{i,j}\left\{\left\|\boldsymbol{x}_i(t) - \boldsymbol{x}_j(t)\right\| - \left\|\boldsymbol{x}_i(0) - \boldsymbol{x}_j(0)\right\| + \epsilon_{i,j}\right\}^2 \tag{7}$$

and

$$k_{i,j} = \kappa_{[i],[j]} c_{ij} \tag{8}$$

$\epsilon_{i,j}$ is a measure of the residual strain in the contact between residues i and j in the equilibrium conformation. $c_{i,j}$ is the (i, j) element of an $n \times n$ matrix called the *linking* or *contact* matrix. $c_{i,j}$ is equal to 1 if residues i and j are in contact, and zero otherwise. In coarse-grained models, one often sets $\kappa_{[i],[j]} = \alpha$, a single parameter. However, it is possible to partition the interactions in the macromolecule into several different types of interactions (e.g., covalent backbone interactions, disulfide bonds, hydrophobic contacts, and solvent-mediated surface interactions). That is, if contacting residue pairs (i, j) and (i', j') are in the same class, then $\kappa_{[i],[j]} = \kappa_{[i'],[j']}$. The stiffnesses in each of these cases can be left as variables to be determined directly from experimental measurements, or some of them can be set using *a-priori* knowledge of contact potentials.

Equation 7 is a non-quadratic function of the deformations even though the springs are linear. However, when we assume that the deformations are small, V becomes a classical quadratic potential energy function. In particular, using the Taylor series approximation

$$\|\boldsymbol{x}_i + \boldsymbol{\delta}_i\| \approx \|\boldsymbol{x}_i\| + \frac{\boldsymbol{x}_i \cdot \boldsymbol{\delta}_i}{\|\boldsymbol{x}_i\|}, \tag{9}$$

then for small deflections, the total potential energy (7) can be written in the form

$$V = V_0 + \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{i,j}\left[\boldsymbol{\delta}_i(t) - \boldsymbol{\delta}_j(t)\right]^T G_{i,j}\left[\boldsymbol{\delta}_i(t) - \boldsymbol{\delta}_j(t)\right]. \tag{10}$$

where $G_{i,j} \in \mathbb{R}^{3\times 3}$ is defined as

$$G_{i,j} = \frac{\left[\boldsymbol{x}_i(0) - \boldsymbol{x}_j(0)\right]\left[\boldsymbol{x}_i(0) - \boldsymbol{x}_j(0)\right]^T}{\left\|\boldsymbol{x}_i(0) - \boldsymbol{x}_j(0)\right\|^2} \tag{11}$$

Note that the effects of residual strain do not appear in (10) in any material way (they do however appear in V_0). Therefore, any residual strain effects are completely removed by setting a datum such that V_0=0.

The stiffness matrix for the whole network is then the matrix K such that

$$V = \frac{1}{2}\boldsymbol{\delta}^T K \boldsymbol{\delta}, \tag{12}$$

where $\boldsymbol{\delta}$ is defined in Equation (6). In the elastic network model, which is purely mechanical, Equation 12 replaces the Hessian matrix in (2) and $\boldsymbol{\delta}$ replaces $\boldsymbol{q}$ as the generalized coordinate. For more details regarding these models see [11,12].

5. Solving the inverse problem

Given an elastic network model, which determines a specific form for the stiffness matrix K=$K(\{\kappa_{[i],[j]}\})$, one can then determine the unknown parameters $\{\kappa_{[i],[j]}\}$ from ensemble measurements of inter-residue distances, assuming that (3) holds. The following subsections respectively formulate and solve the problem of determining these stiffness parameters from ensemble motion measurements.

5.1. Problem formulation

Suppose a pair of FRET probes are attached to residues i and j, and a probability distribution (histogram) of inter-residue distances, $r = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|$, is recorded. Let us assume one of the following: (1) the inter-residue distances recorded by the probes are exact, and that fluctuations in the position of a probe itself relative to its attachment point are negligible; or, (2) the elastic network model of the original macromolecule is modified to include the probes, and the mechanical properties of the probes are known in advance. In other words, let us either assume that $\rho_{ij}(r)$ is either a true measure of the ensemble of inter-residue distances or, if it is not, we have enough information to deconvolve errors in the experimental data due to probe flexibility.

This measurement can be matched to the network model as follows. First, define the probability density

$$f(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n) = \frac{\exp\left(\frac{-\boldsymbol{\delta}^T K \boldsymbol{\delta}}{2k_B T}\right)}{Z_c} \tag{13}$$

where

$$Z_c = \frac{(k_B T/2\pi)^{\frac{3n}{2}}}{|\det K|^{\frac{1}{2}}}$$

is the conformational partition function.

Next, integrate $f(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n)$ over all displacements except for $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_j$. Then the resulting probability distribution $f(\boldsymbol{\delta}_i,\boldsymbol{\delta}_j)$ should be related in some way to the observed $\rho_{ij}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$. In particular, if we compute the variance of ρ_{ij}, this can be matched to the mechanical model as follows:

$$\sigma^2_{\rho ij} = \int_{\boldsymbol{\delta}_i}\int_{\boldsymbol{\delta}_j}\Big\{\|\boldsymbol{x}_i(0) + \boldsymbol{\delta}_i - \boldsymbol{x}_j(0) - \boldsymbol{\delta}_j\| - \|\boldsymbol{x}_i(0) - \boldsymbol{x}_j(0)\|\Big\}^2 f(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j)d\boldsymbol{\delta}_j d\boldsymbol{\delta}_i. \tag{14}$$

Under the assumption that f ($\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_j$) vanish rapidly as $\|\boldsymbol{\delta}_i\|,\|\boldsymbol{\delta}_j\|$ deviate from zero, the right-hand side of this equation can be computed as a closed-form expression that depends on the stiffness parameter(s) $\{\kappa_{[i],[j]}\}$. If there is only one such parameter, one experimentally measured $\sigma^2_{\rho ij}$ is sufficient to fix this parameter. If multiple different stiffness parameters are to be determined, then a system of simultaneous equations of the form in (14) must be solved.

5.2. *Solution method*

The integral of (13) over all displacements except $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_j$ can be computed in closed form since it is a Gaussian probability density [14,15]. This follows from the fact that if we partition any symmetric matrix A as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix}$$

then

$$\frac{1}{2}\mathbf{y}^T A\mathbf{y} = \frac{1}{2}\mathbf{y}_1^T A_{11}\mathbf{y}_1 + \mathbf{y}_1^T A_{12}\mathbf{y}_2 + \frac{1}{2}\mathbf{y}_2^T A_{22}\mathbf{y}_2$$

where $\mathbf{y} = \left[\mathbf{y}_1^T, \mathbf{y}_2^T\right]^T$. Hence, for $A=\beta K$,

$$\begin{aligned}&\int_{\boldsymbol{\delta}_1}\cdots\int_{\boldsymbol{\delta}_{i-1}}\int_{\boldsymbol{\delta}_{i+1}}\cdots\int_{\boldsymbol{\delta}_{j-1}}\int_{\boldsymbol{\delta}_{j+1}}\cdots\int_{\boldsymbol{\delta}_n} f(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n)d\boldsymbol{\delta}_1\cdots d\boldsymbol{\delta}_{i-1}d\boldsymbol{\delta}_{i+1}\cdots d\boldsymbol{\delta}_{j-1}d\boldsymbol{\delta}_{j+1}\cdots\boldsymbol{\delta}_n \\ &= \frac{1}{Z_c}\int_{\mathbf{y}_2}\exp\left(-\frac{1}{2}\mathbf{y}^T A\mathbf{y}\right)d\mathbf{y}_2 \\ &= \frac{(2\pi)^{\frac{n_2}{2}}}{Z_c|\det A_{22}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\mathbf{y}_1^T A'_{11}\mathbf{y}_1\right)\end{aligned}$$

where n_2 is the dimension of the vector $\mathbf{y}_2$ and

$$A'_{11} = A_{11} - A_{12}A_{22}^{-1}A_{12}^T \tag{15}$$

In the current context, $\mathbf{y}_1 = \left[\boldsymbol{\delta}_i^T, \boldsymbol{\delta}_j^T\right]^T \in \mathbb{R}^6, \mathbf{y}_2 \in \mathbb{R}^{3n-6}$, and A is partitioned accordingly. Performing the integration yields $f(\boldsymbol{\delta}_i,\boldsymbol{\delta}_j)$. We may now approximate (14) in closed form by observing that when $\boldsymbol{y}_1$ is of dimension six,

$$\int_{\mathbf{y}_1}\mathbf{y}_1^T G\mathbf{y}_1\exp\left(-\frac{1}{2}\mathbf{y}_1^T A'_{11}\mathbf{y}_1\right)d\mathbf{y}_1 = \frac{(2\pi)^3\mathrm{tr}\left(GA'^{-1}_{11}\right)}{\left|\det A'_{11}\right|^{\frac{1}{2}}}$$

By comparing (7) when $\epsilon_{i,j}=0$ with (10), we observe that, for the above expression to approximate (14), it must be the case that

$$G = \begin{pmatrix} G_{ij} & -G_{ij} \\ -G_{ij} & G_{ij} \end{pmatrix} \in \mathbb{R}^{6\times 6}$$

Combining the results of this section, we see that

$$\begin{aligned}\frac{\sigma^2_{\rho ij}}{k_B T} &= \frac{(2\pi/k_B T)^{\frac{3n-6}{2}}}{Z_c|\det K_{22}|^{\frac{1}{2}}}\frac{(k_B T)^3\mathrm{tr}\left(GK'^{-1}_{11}\right)}{(2\pi)^3\left|\det K'_{11}\right|^{\frac{1}{2}}} \\ &= \frac{(2\pi/k_B T)^{\frac{3n}{2}}}{Z_c|\det K|^{\frac{1}{2}}}\mathrm{tr}\left(GK'^{-1}_{11}\right) = \mathrm{tr}\left(GK'^{-1}_{11}\right)\end{aligned} \tag{16}$$

(The factor of $k_B T$ in the denominator on the left-hand side of the equation comes from the fact that $A'_{11} = \beta K'_{11}$).

If one assumes that a single stiffness parameter describes all pair-wise contacts, $\kappa_{[i],[j]}=\alpha$, then we can solve for the best estimate of α in terms of the single experimentally measured quantity $\sigma^2_{\rho ij}$ by substituting into (16) to find

$$\alpha = \frac{k_B T}{\sigma^2_{\rho ij}}\mathrm{tr}\left(GK''^{-1}_{11}\right) \tag{17}$$

where $K''_{11} = K'_{11}/\alpha$ is what the stiffness matrix would be if all contacts had unit stiffness.

Of course, if multiple experimental measurements are taken with different values of i and j, a more robust estimate of α would be obtained. In the case when multiple stiffness parameters are present in the model, multiple equations can be solved simultaneously to find the values that best match the ensemble motion measurements captured in $\sigma^2_{\rho ij}$. In such a case, multiple values of i and j would have to be chosen, each pair potentially corresponding to a different experiment. Numerical methods for solving simultaneous non-linear equations can be employed to obtain optimal estimates of stiffness parameters for any such experimental data set. The number of contact classes to use can be determined rationally by starting with a single class and increasing the number until it is no longer feasible to do more experiments, or if there is diminishing returns in the

reduction of model error as the number of classes increases for a fixed set of experimental data.

6. Conclusions

A methodology has been presented in which ensemble measurements of distances between labeled points within a macromolecule can be used to determine stiffness parameters. These stiffness parameters describe the mechanical characteristics of the macromolecule, and are closely related to contact potentials. Knowing such parameters would be useful when defining coarse-grained mechanical models of large biomolecular structures, and provide a way to associate frequencies of motion with the shapes of normal modes.

Acknowledgement

The author would like to thank Moonki Kim for carefully checking the equations in this article.

References

[1] P.K. Weiner, P.A. Kollman, J. Comput. Chem. 2 (1981) 287.
[2] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M.J. Karplus, Comput. Chem. 4 (1983) 187.
[3] A.R. Leach, Molecular Modelling: Principles and Applications, 2nd Ed., Longman, Harlow, England, 2001.
[4] S. Miyazawa, R.L. Jernigan, Macromolecules 18 (1985) 534.
[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235.
[6] I. Bahar, R.L. Jernigan, J. Mol. Biol. 281 (1998) 871.
[7] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, I. Bahar, Biophys. J. 80 (2001) 505.
[8] D. Ming, Y.F. Kong, M.A. Lambert, Z. Huang, J.P. Ma, Proc. Natl. Acad. Sci. 99 (2002) 8620.
[9] D. Ming, Y.F. Kong, Y.H. Wu, J.P. Ma, Proc. Natl. Acad. Sci. 100 (2003) 104.
[10] F. Tama, W. Wriggers, C.L. Brooks, J. Mol. Biol. 321 (2002) 297.
[11] M. Kim, R.L. Jernigan, G.S. Chirikjian, Biophys. J. 83 (2002) 1620.
[12] M. Kim, G.S. Chirikjian, R.L. Jernigan, J. Mol. Graphics Modelling 21 (2002) 151.
[13] E.B. WilsonJr., J.C. Decius, P.C. Cross, Molecular vibrations: the theory of infrared and Raman vibrational spectra, Dover Publications, New York, USA, 1980.
[14] Z.W. Birnbaum, Introduction to probability and mathematical statistics, Harper, New York, USA, 1962.
[15] G.S. Chirikjian, A.B. Kyatkin, Engineering Applications of Noncommutative Harmonic Analysis, CRC Press, Boca Raton, FL, USA, 2001.

ERRATUM: A Methodology for Determining Mechanical Properties of Macromolecules from Ensemble Motion Data

GREGORY S. CHIRIKJIAN

Department of Mechanical Engineering

The Johns Hopkins University

Baltimore, Maryland 21218

email: gregc@jhu.edu

Due to minor typesetting errors, in several places in the text, $2\pi/k_BT$ should be $2\pi k_BT$. In particular:

Eq. 3 should be:

$$Z = \frac{(2\pi k_BT)^N}{|det\{M^{-\frac{1}{2}}KM^{-\frac{1}{2}}\}|^{\frac{1}{2}}}$$

The equation at the bottom of p 551 should be:

$$Z_c = \frac{(2\pi k_BT)^{\frac{3n}{2}}}{|detK|^{\frac{1}{2}}}$$

and Eq. 16 should read:

$$\frac{\sigma^2_{\rho_{ij}}}{k_BT} = \frac{(2\pi k_BT)^{\frac{3n-6}{2}}}{Z_c|\det K_{22}|^{\frac{1}{2}}}\frac{(2\pi k_BT)^3\mathrm{tr}(GK'^{-1}_{11})}{|\det K'_{11}|^{\frac{1}{2}}} = \frac{(2\pi k_BT)^{\frac{3n}{2}}}{Z_c|\det K|^{\frac{1}{2}}}\mathrm{tr}(GK'^{-1}_{11}) = \mathrm{tr}(GK'^{-1}_{11}).$$