Markov Methods for Hierarchical Coarse-Graining of Large Protein Dynamics

Chakra Chennubhotla and Ivet Bahar

Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15261 {chakra,bahar}@ccbb.pitt.edu

Abstract. Elastic network models (ENMs), and in particular the Gaussian Network Model (GNM), have been widely used in recent years to gain insights into the machinery of proteins. The extension of ENMs to supramolecular assemblies/complexes presents computational challenges, however, due to the difficulty of retaining atomic details in mode decomposition of large systems dynamics. Here, we present a novel approach to address this problem. Based on a Markovian description of communication/interaction stochastics, we map the full-atom GNM representation into a hierarchy of lower resolution networks, perform the analysis in the reduced space(s) and reconstruct the detailed models dynamics with minimal loss of data. The approach (hGNM) applied to chaperonin GroEL-GroES demonstrates that the shape and frequency dispersion of the dominant 25 modes of motion predicted by a full-residue (8015 nodes) GNM analysis are almost identically reproduced by reducing the complex into a network of 35 soft nodes.

1 Introduction

With advances in sequence and structure genomics, an emerging view is that to understand and control the mechanisms of biomolecular function, knowledge of sequence and structure is insufficient. Additional knowledge in the form of *dynamics* is needed. In fact, proteins do not function as static entities or in isolation; they are engaged in functional motions, and interactions, both within and between molecules. The resulting motions can range from single amino acid side chain reorientations (*local*) to concerted domain-domain motions (*global*). The motions on a local scale can be explored to a good approximation by conventional molecular dynamics (MD) simulations, but the motions at a global scale are usually beyond the range of such simulations. Elastic network models (ENM), based on polymer mechanics, succeed in providing access to global motions [1–3].

A prime example of an EN is the Gaussian Network Model (GNM) [4, 5]. In graph-theoretic terms, each protein is modeled by an undirected graph $\mathcal{G}$, given by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with residues $\mathcal{V} = \{v_i | i = 1, \ldots, n\}$ defining the nodes of the network, and edges $\mathcal{E} = \{e_{ij}\}$ representing interactions between residues v_i and v_j. The set of all pairwise interactions is described by a non-negative, symmetric *affinity* matrix $\boldsymbol{A} = \{a_{ij}\}$, with elements $a_{ij} = a_{ji}$. GNM chooses a simple

interaction model, which is to set the affinity $a_{ij} = a_{ji} = 1$, for a pair of residues v_i and v_j whose C^α atoms are within a cut-off distance of r_c. The interactions represent both bonded and non-bonded contacts in the native configuration of the protein. The cutoff distance represents the radius of the first coordination shell around residues observed in Protein Data Bank (PDB) [6] structures and is set to be $7\mathring{A}$ [7, 8].

The motions accessible under native state conditions are obtained from the Kirchhoff matrix $\boldsymbol{\Gamma}$, defined in terms of the affinity and degree matrices as $\boldsymbol{\Gamma} = \boldsymbol{D} - \boldsymbol{A}$. Here $\boldsymbol{D}$ is a diagonal matrix: $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_n)$ and d_j represents the degree of a vertex v_j: $d_j = \sum_{i=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji}$. $\boldsymbol{\Gamma}$ is referred to as the *combinatorial Laplacian* in graph theory [9]. The Kirchhoff matrix multiplied by a force constant γ that is uniform over all springs defines the *stiffness* matrix of an equivalent mass-spring system. The eigenvalue decomposition of $\boldsymbol{\Gamma}$ yields the shape and frequency dispersion of equilibrium fluctuations. In most applications it is of interest to extract the contribution of the most cooperative modes, i.e. the low frequency modes that have been shown in several systems to be involved in functional mechanisms [1, 2]. Also, of interest is the inverse of $\boldsymbol{\Gamma}$, which specifies the covariance matrix for the Boltzmann distribution over equilibrium fluctuations.

GNM is a linear model, and as such it cannot describe the transition between configurations separated by an energy barrier (or any other non-linear effect), so it only applies to fluctuations in the neighborhood of a single energy minimum. The energy well is approximated by a harmonic potential, which limits the magnitude of the predicted motion. The topology of inter-residue contacts in the equilibrium structure is captured by the Kirchhoff matrix $\boldsymbol{\Gamma}$. Also, there is no information on the 'directions' of motions in different vibrational modes, but on their sizes only. The fluctuations are assumed to be isotropic and Gaussian, but for anisotropic extension of GNM called ANM see [10, 11] or equivalent EN-based normal mode analyses (NMA) [12, 13]. Despite this simplicity, many studies now demonstrate the utility of GNM and other EN models in deducing the machinery and conformational dynamics of large structures and assemblies (for a recent review see [2]).

The application and extension of residue-based ENMs to more complex processes, or larger systems, is computationally expensive, both in terms of memory and time, as the eigen decomposition scales on the order of $O(n^3)$, where n is the number of nodes in the graph. Given that the Kirchhoff matrix is sparse, there are a plethora of efficient sparse eigensolvers that one can use [14–17], including eigensolvers designed specifically for decomposing graph Laplacians [18].

Another way to reduce complexity is to adopt coarser-grained models. For example, in the hierarchical coarse-graining (HCG) approach, sequences of m consecutive amino acids are condensed into unified nodes - which reduces the computing time and memory by factors of m^3 and m^2, respectively [19]; or a mixed coarse-graining has been proposed in which the substructures of interest are modeled at single-residue-per-node level and the surrounding structural units at a lower resolution of m-residues-per node [20]; another common representation

of the structure is to adopt rigidly translating and rotating blocks (RTB) [21, 22], or the so-called block normal mode analysis (BNM) [23].

While these methods have been useful in tackling larger systems, the choice and implementation of optimal model parameters to retain physically significant interactions at the residue-, or even atomic level, has been a challenge. The level of HCG has been arbitrarily chosen in the former group of studies, requiring *ad-hoc* readjustments to spring constants or cutoff distances of interaction. In the case of RTB or BNM approaches, all atomic, or residue level information is lost, and substructures that may contain internal degrees of freedom – some of which being functional – are assumed to move as a rigid block. Overall, information is lost on local interactions as structures are coarse-grained. Clearly, the challenge is to map a high resolution model to a low resolution, with a minimal loss of information. In this paper, we present a novel approach to address this problem.

Our approach is to model structures as networks of interacting residues and study the Markov propagation of "information" across the network. We rely on the premise that, the components (residues) of a protein machinery (network) communicate with each other and operate in a coordinated manner to perform their function successfully. Using the Markov chain perspective, we map the full atom network representation into a hierarchy of intermediate ENMs, while retaining the Markovian stochastic charactersitcs, i.e. transition probabilities and stationary distribution, of the original network. The communication properties at different levels of the hierarchy are intrinsically defined by the network topology. This new representation has several features, including: soft clustering of the protein structure into stochastically coherent regions thus providing a useful assessment of elements serving as hubs and/or transmitters in propagating information/interaction; automatic computation of the contact matrices for ENMs at each level of the hierarchy to facilitate computation of both Gaussian and anisotropic fluctuation dynamics; and a fast eigensolver for NMA. We illustrate the utility of the hierarchical decomposition by presenting its application to the bacterial chaperonin GroEL–GroES.

2 A Markov Model for Network Communication

We model each protein as a weighted, undirected graph $\mathcal{G}$ given by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with residues $\mathcal{V} = \{v_i | i = 1, \ldots, n\}$ defining the nodes of the network, and edges $\mathcal{E} = \{e_{ij}\}$ representing interactions between residues v_i and v_j. The set of all pairwise interactions is described by a non-negative, symmetric *affinity* matrix $\boldsymbol{A} = \{a_{ij}\}$, with elements $a_{ij} = a_{ji}$ and where a_{ij} is the total number of *atom–atom* contacts made within a cutoff distance of $r_c = 4.5\mathring{A}$ between residues v_i and v_j. The self-contact a_{ii} is similarly defined, but all bonded pairs are excluded. This representation takes into account the difference in the size of amino acids, and captures to a first approximation the strong (weak) interactions expected to arise between residue pairs with large (small) number of atom-atom contacts. The degree of a vertex v_j is defined as $d_j = \sum_{i=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji}$, which are organized in a diagonal matrix of the form $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_n)$.

A *discrete-time, discrete-state Markov* process of network communication is defined by setting the communication (or signalling) probability m_{ij} from residue v_j to residue v_i in *one time-step* to be proportional to the affinity between nodes, $a_{i,j}$. In matrix notation, this conditional probability matrix $\boldsymbol{M} = \{m_{ij}\}$, also called the Markov transition matrix, given by

$$\boldsymbol{M} = \boldsymbol{A}\boldsymbol{D}^{-1}. \tag{1}$$

defines the stochastics of a *random walk* on the protein graph $\mathcal{G}$. Note, $m_{ij} = d_j^{-1} a_{ij}$ where d_j gives a measure of local packing density near residue v_j and serves as a normalizing factor to ensure $\sum_{i=1}^{n} m_{ij} = 1$. Alternatively, m_{ij} can be viewed as the conditional probability of interacting with residue v_i, that is transmitting information to residue v_i, given that the signal (or perturbation) is initially positioned, or originates from, v_j. Suppose this initial probability is p_j^0. Then, the probability of reaching residue v_i using link e_{ij} is $m_{ij} p_j^0$. In matrix notation, the probability of ending up on any of the residues $\boldsymbol{v} = [v_1, v_2, \cdots, v_n]$ after one time step is given by the distribution $\boldsymbol{p}^1 = \boldsymbol{M}\boldsymbol{p}^0$, where $\boldsymbol{p}^k = \left[p_1^k, \ldots, p_n^k\right]$. Clearly this process can be iterated, so that after β steps we have

$$\boldsymbol{p}^\beta = \boldsymbol{M}^\beta \boldsymbol{p}^0. \tag{2}$$

Assume the graph is connected, i.e. there is a path connecting every pair of residues in the graph. Then, as $\beta \to \infty$ the Markov chain $\boldsymbol{p}^\beta$ approaches a unique *stationary* distribution $\boldsymbol{\pi}$, the elements of which are given by: $\pi_i = d_i / \sum_{k=1}^{n} d_k$. While the evolution of the random walk is a function of the starting distribution, the stationary distribution is invariant to the precise details of how the random walk is initiated.

The main goal in undertaking random walks is to reveal the communication patterns *inherent* to the network because of its architecture. However, a naive random walk on a large protein, as will be presented below for the GroEL-GroES complex, is computationally challenging. We address this problem by building a hierarchy of intermediate resolution network models, performing the analysis in the reduced space and mapping the results back to the high resolution representation as illustrated in Fig. 1.

3 Network Hierarchy to Reduce Communication Complexity

The objective in designing a network hierarchy is to map the Markov process operating at the highest resolution onto successively lower resolution network models, while maintaining its stochastic characteristics [24]. In particular, using the stationary distribution $\boldsymbol{\pi}$ and the Markov transition matrix $\boldsymbol{M}$, we build a coarse-scale Markov propagation matrix $\widetilde{\boldsymbol{M}}$ (size: $m \times m$, where $m \ll n$) and its stationary distribution $\boldsymbol{\delta}$. The random walk initiated on the coarse-scale network $\widetilde{\mathcal{G}}(m)$, and reaching distribution $\boldsymbol{\delta}$, is equivalent to the random walk on the full resolution network $\mathcal{G}(n)$ with stationary distribution $\boldsymbol{\pi}$. To

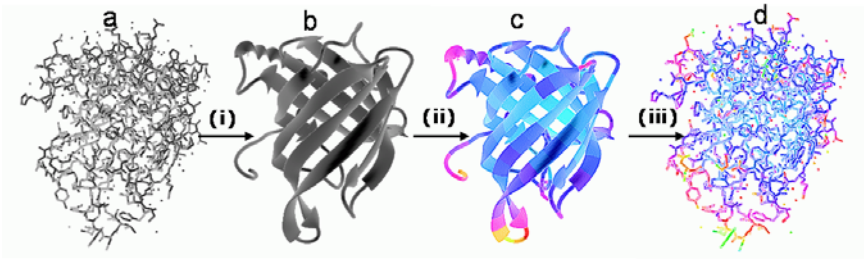


Fig. 1. Hierarchical Network Decomposition Overview: step (**i**) map the structure (**a**) to its optimal reduced level representation (illustrated here for retinol-binding protein mapped from full atomic scale to intermediate-chain representation). This step may involve several intermediate levels of resolution (**b**) (e.g. see Fig. 2); step (**ii**) perform structural analysis (e.g. GNM) at a coarse-grained scale (**c**); and step (**iii**) reconstruct the detailed structure-dynamics (**d**). The communication/coupling of residues at a given level are assumed to obey a Markov process controlled by atom-atom contact topology. The steps (**i**) and (**iii**) are achieved by two operators, $\boldsymbol{R}$ for model reduction, and $\boldsymbol{K}$ for model reconstruction. $\boldsymbol{R}$ and $\boldsymbol{K}$ ensure that similar stochastic characteristics (transition probabilities and stationary distributions) are retained between successive levels of the hierarchy.

build a hierarchy of intermediate resolution networks we devise two sets of new operators at each level of the hierarchy: $\boldsymbol{R}$ for model reduction, and $\boldsymbol{K}$ for model expansion/reconstruction.

3.1 Deriving Stationary Distribution in the Reduced Model

We begin by expressing the stationary distribution $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_n]$ as a probabilistic mixture of *latent* distributions,

$$\boldsymbol{\pi} = \boldsymbol{K}\boldsymbol{\delta}, \tag{3}$$

where $\boldsymbol{\delta} = [\delta_1, \delta_2, \cdots, \delta_m]$ is an unknown stationary distribution in a reduced ($m-$dimensional) representation of the structure; $\boldsymbol{K} = \{K_{ij}\}$ is an $n \times m$ non-negative *kernel* matrix with elements K_{ij} and columns K_j being latent probability distributions that each sum to 1, and $m \ll n$. The kernel matrix acts as an *expansion* operator, mapping the low-dimensional distribution $\boldsymbol{\delta}$ to a high-dimensional distribution $\boldsymbol{\pi}$.

We derive a maximum likelihood approximation for $\boldsymbol{\delta}$ using an expectation-maximization (EM) type algorithm [25]. To this aim we minimize the *Kullback-Liebler* distance measure [26, 27] between the two probability distributions $\boldsymbol{\pi}$ and $\boldsymbol{K}\boldsymbol{\delta}$, subject to the constraint that $\sum_{j=1}^{m} \delta_j = 1$ and ensured by the Lagrange multiplier λ in the equation below:

$$E = -\sum_{i=1}^{n} \pi_i \ln \sum_{j=1}^{m} K_{ij}\delta_j + \lambda\left(\sum_{j=1}^{m} \delta_j - 1\right). \tag{4}$$

Setting the derivative of E with respect to δ_j to be zero we obtain

$$\sum_{i=1}^{n} \frac{\pi_i K_{ij} \delta_j}{\sum_{k=1}^{m} K_{ik} \delta_k} = \lambda \delta_j. \tag{5}$$

The contribution made by kernel j to a node i (or its stationary probability π_i) is given by K_{ij} (or the product $K_{ij}\delta_j$), and hence we can define an *ownership* of node i in the high resolution representation by a node j in the low resolution representation as

$$R_{ij} = \frac{K_{ij}\delta_j}{\sum_{k=1}^{m} K_{ik}\delta_k}. \tag{6}$$

R_{ij} is also referred to as the responsibility of node j in the low resolution representation, for node i in the high resolution. We note that the mapping between the two resolutions is not deterministic, but probabilistic in the sense that $\sum_{j=1}^{m} R_{ij} = 1$.

Using this relation, and the equalities $\sum_{j=1}^{m} \delta_j = 1$ and $\sum_{i=1}^{n} \pi_i = 1$, summing over j in Eq. 5 gives $\lambda = 1$. This further leads to the stationary distribution $\boldsymbol{\delta}$ at the coarse scale

$$\delta_j = \textstyle\sum_{i=1}^{n} \pi_i R_{ij}. \tag{7}$$

The matrix $\boldsymbol{R}$ therefore maps the high dimensional distribution $\boldsymbol{\pi}$ to its low-dimensional counterpart $\boldsymbol{\delta}$ and hence the name *reduction* operator. Following Bayes theorem, K_{ij} can be related to the *updated* $\boldsymbol{\delta}$ values as

$$K_{ij} = \frac{R_{ij}\pi_i}{\delta_j}. \tag{8}$$

In summary, the operators $\boldsymbol{K}$ and $\boldsymbol{R}$ and stationary distribution $\boldsymbol{\delta}$ are computed using the following EM type procedure: (1) select an initial estimate for $\boldsymbol{K}$ and $\boldsymbol{\delta}$ (see § 3.2); (2) *E-step*: compute ownership maps $\boldsymbol{R}$ using Eq. 6; (3) *M-step*: estimate $\boldsymbol{\delta}$ and update $\boldsymbol{K}$ using Eqs. 7 and 8 respectively; and finally, (4) repeat *E-* and *M-* steps until convergence.

3.2 Kernel Selection Details

As an initial estimate for $\boldsymbol{\delta}$, a uniform distribution is adopted. The kernel matrix $\boldsymbol{K}$ is conveniently constructed by diffusing $\boldsymbol{M}$ to a small number of iterations β to give $\boldsymbol{M}^{\beta}$ and selecting a small number of columns. In picking the columns of $\boldsymbol{M}^{\beta}$, a greedy decision is made. In particular, column i in $\boldsymbol{M}^{\beta}$ corresponds to information diffusion from residue v_i. The first kernel K_i that is picked corresponds to the residue v_i with the highest stationary probability π_i. Following the selection of K_i, all other residues j (and the corresponding columns K_j in $\boldsymbol{M}^{\beta}$) that fall within the half-height of the peak value of the probability distribution in K_i are eliminated from further consideration. This approach generates kernels that are spatially disjoint. The selection of kernels continues until every

residue in the protein is within a half-height of the peak value of at least one kernel. While other kernel selection procedures are conceivable, we chose the greedy method for computational speed. In practice, we observed the EM algorithm generates results of biological interest that are insensitive to the initial estimates of $\boldsymbol{K}$ and $\boldsymbol{\delta}$.

3.3 Transition and Affinity Matrices in the Reduced Model

The Markov chain propagation at the reduced representation obeys the equation $\boldsymbol{q}^{k+1} = \widetilde{\boldsymbol{M}}\boldsymbol{q}^k$, where $\boldsymbol{q}^k$ is the coarse scale m-dimensional probability distribution after k steps of the random walk. We *expand* $\boldsymbol{q}^k$ into the fine scale using $\boldsymbol{p}^k = \boldsymbol{K}\boldsymbol{q}^k$, and *reduce* $\boldsymbol{p}^k$ back to the coarse scale by using the ownership value $R_{i,j}$ as in $q_j^{k+1} = \sum_{i=1}^{n} p_i^k R_{i,j}$. Substituting Eq. 6 for ownerships, followed by the expression for $\boldsymbol{p}^k$, in the equation for q_j^{k+1}, we obtain

$$\widetilde{\boldsymbol{M}} = \mathrm{diag}(\,\boldsymbol{\delta}\,)\,\boldsymbol{K}^{\mathsf{T}}\,\mathrm{diag}\,(\boldsymbol{K}\boldsymbol{\delta}\,)^{-1}\;\boldsymbol{K}. \tag{9}$$

Using the definition of $\widetilde{\boldsymbol{M}}$, and the corresponding stationary distribution $\boldsymbol{\delta}$, we generate a *symmetric* affinity matrix $\widetilde{\boldsymbol{A}}$ that describes the node-node interaction strength in the low resolution network

$$\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{M}}\mathrm{diag}(\boldsymbol{\delta}). \tag{10}$$

To summarize, we use the stationary distribution $\boldsymbol{\pi}$ and Markov transition matrix $\boldsymbol{M}$ at the fine-scale to derive the operator $\boldsymbol{K}$ and associated reduced stationary distribution $\boldsymbol{\delta}$, using the EM algorithm described in the previous section. $\boldsymbol{K}$ and $\boldsymbol{\delta}$ are then used in Eq. 9 and 10 to derive the respective transition $\widetilde{\boldsymbol{M}}$ and affinity $\widetilde{\boldsymbol{A}}$ matrices in the coarse-grained representation. Clearly, this procedure can be repeated recursively to build a hierarchy of lower resolution network models.

4 Hierarchical Decomposition of the Chaperonin GroEL-GroES

We examine the structure and dynamics of the bacterial chaperonin complex GroEL-GroES-$(ADP)_7$ [28], from the perspective of a Markov propagation of information/interactions. GroEL is a cylindrical structure, 150$\mathring{A}$ long and 140$\mathring{A}$ wide, consisting of 14 identical chains organized in two back-to-back stacked rings (*cis* and *trans*) of seven subunits each. The GroES co-chaperonin, also heptameric, binds to the apical domain of GroEL and closes off one end of the cylinder. During the allosteric cycle that mediates protein folding, the *cis* and *trans* rings alternate between open (upon binding of ATP and GroES) and closed (unliganded) forms, providing access to, or release from, the central cylindrical cavity, where the folding of an encapsulated (partially folded or misfolded) protein/peptide is assisted.

First, the inter-residue affinity matrix $\boldsymbol{A}$ based on all atom-atom contacts is constructed (Fig. 2**a**), from which the fine-scale Markov transition matrix $\boldsymbol{M}$ is derived using Eq. 1. The kernel selection algorithm applied to $\boldsymbol{M}^{\beta}$ ($\beta = 4$) yields 1316 (reduced level 1) kernels. Using these kernels as an initialization, a recursive application of the EM procedure derives stationary distributions $\boldsymbol{\delta}$ (Eq. 7), updated expansion matrices $\boldsymbol{K}$ (Eq. 8), reduced level probability transition matrices $\widetilde{\boldsymbol{M}}$ (Eq. 9) and the corresponding residue interaction matrices $\widetilde{\boldsymbol{A}}$ (Eq. 10). The respective dimensions of $\widetilde{\boldsymbol{A}}$ turn out to be 483 (reduced level 2), 133 (reduced level 3), 35 (reduced level 4, Fig. 2**c**) and 21 (reduced level 5, Fig. 2**d**). We note that the individual subunits of the GroEL/GroES are distinguished by their strong intra-subunit interactions, and a number of inter-subunit contacts are maintained at all levels, which presumably establish the communication across the protein at all levels. The dimension m of the reduced model is automatically defined during the kernel selection at each level of the hierarchy. The method thus avoids the arbitrary choices of sampling density and interaction cutoff distances at different hierarchical levels.

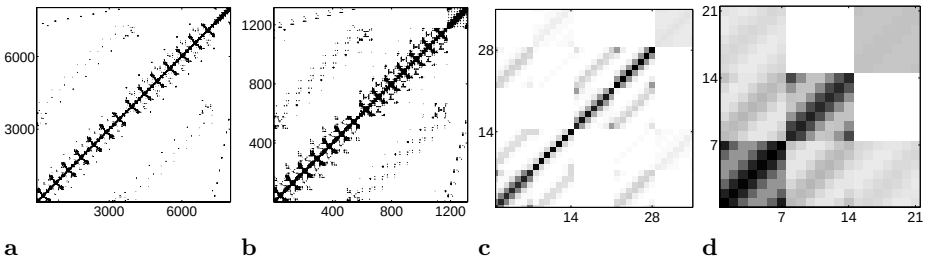


Fig. 2. Affinity matrix hierarchy for the protein GroEL/GroES (PDB: **1AON**). The respective sizes of the reduced models, and the associated affinity matrices, across the hierarchy are $n = 8015$ (fine-scale, panel **a**) and $m = 1316$ (coarse-scale 1, panel **b**), 483 (coarse-scale 2), 133 (coarse-scale 3), 35 (coarse-scale 4, panel **c**) and 21 (coarse-scale 5, panel **d**). The affinity matrices are real-valued but are shown here as *dot* plots (panels **a-b**), to highlight the similarity in the matrix structure across the hierarchy. The affinity matrices for the two lowest resolution models (panels **c-d**) are shown as images, where the affinity value is *inversely* proportional to the brightness of a pixel.

In contrast to the *deterministic* assignment of one-node-per-residue in the original ENM, the Markov-chain-based representation adopts a *stochastic* description in the sense that each node *probabilistically* 'owns', or 'is responsible for' a subset of residues. To see this, consider the ownership matrix $\boldsymbol{R}^{(l,l+1)} = \{R_{ij}^{(l,l+1)}\}$ that relates information between two adjacent levels l and $l+1$ of the hierarchy. Likewise, the matrix $\boldsymbol{R}^{(0,L)} = \prod_{l=0}^{L-1} \boldsymbol{R}^{(l,l+1)}$ ensures the passage from the original high resolution representation 0 to the top level L of the hierarchy. In particular, the ij^{th} element $\boldsymbol{R}_{ij}^{(0,L)}$ describes the probabilistic participation of residue v_i (at level 0) in the cluster j (at level L), and $\sum_j \boldsymbol{R}_{ij}^{(0,L)} = 1$. Hence,

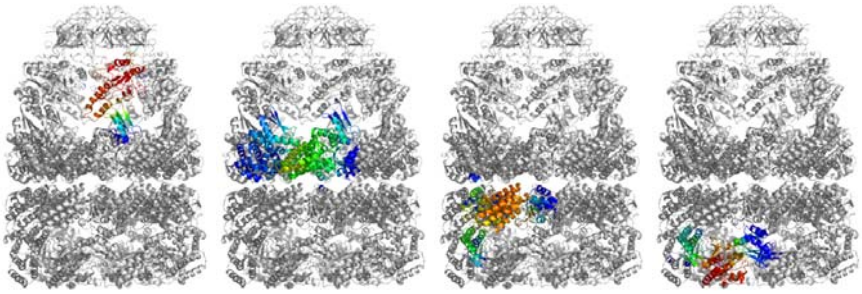


Fig. 3. Four different *soft* clusters located on GroEL.

the nodes at level L perform a *soft partitioning* of the structure. *This type of soft distribution of residues among the m nodes, or their partial/probabilistic participation in neighboring clusters, establishes the communication between the clusters, and is one of the key outcomes of the present analysis.* Of interest is to examine the ownership of clusters at a reduced representation. We select the coarse-scale 4, for example, which maps the structure into a graph of 35 clusters (Fig. 2**c**). Fig. 3 demonstrates the ownership of the individual clusters at this level. Essentially there are five sets of seven clusters each, centered near the apical and equatorial domains of the *cis* and *trans* rings, and at the individual GroES chains. The intermediate domains are being shared between the clusters at the apical and equatorial domains. As such, they play a key role in establishing intra-subunit communication. The color-coded ribbon diagrams in Fig. 3 display the loci of representative clusters from each of these four distinct types (excluding the GroES clusters). The color code from red-to-blue refers to the higher-to-lower involvement (or responsibility) of the individual residues in the indicated clusters. Evidently, the regions colored red serve as hubs for broadcasting the information within clusters, and those colored blue play the key role of establishing the communication, or transferring information between clusters. Detailed examination of the ownership of these clusters reveal several interesting features, correlating with the experiments and summarized in §6.

Next, we benchmark the utility and robustness of the presently introduced methodology in so far as the equilibrium dynamics of the examined structure is concerned. Mainly, we compare the collective modes of motion predicted for the GroEL-GroES complex using a full-residue (8015 nodes) ENM [29], with those captured by the hierarchy of reduced models. The newly introduced representation hierarchy will be shown below to successfully map structure-dynamics information between successive levels with minimal loss in accuracy[1].

[1] The ownership matrix can also be used to propagate the location information of the residues from one level of the hierarchy to another. This in turns help perform anisotropic fluctuation modeling, but for lack of space this procedure will not be elaborated any further.

5 Hierarchical Gaussian Network Model (*h*GNM)

Here we present a methodology for generating GNM modes at different levels of coarse-graining the information on contact topology inherent in $\mathcal{G}$, and reconstructing the detailed mode behavior by projecting the eigenvectors and eigenvalues generated at low levels of resolution back to their fine scale counterparts using the Markov chain propagation formalism, a method shortly referred to as hierarchical GNM (*h*GNM).

For *h*GNM, assume that the dimensions of the Kirchhoff matrices at the coarse, intermediate and fine scales are e, m and n respectively, where $e \leq m \ll n$. The affinity and Kirchhoff matrices at the coarsest level are not likely to be sparse, however a full eigen decomposition of the coarsest Kirchhoff matrix (size: $e \times e$) will be computationally the least expensive step.

To reconstruct the eigen information at the fine-scale, assume we have access to the leading eigenvectors $\widehat{\boldsymbol{U}}$ (size: $m \times e$) for $\widehat{\boldsymbol{\Gamma}}$ (size: $m \times m$). Using this we generate the leading eigenvectors $\widetilde{\boldsymbol{U}}$ (size: $n \times e$), and the leading eigenvalues $\widetilde{\boldsymbol{\Lambda}} = [\lambda_1, \lambda_2 \cdots \lambda_e]$ (size: $e \times 1$) of the fine-scale Kirchhoff matrix $\boldsymbol{\Gamma}$ (size: $n \times n$). Let $\{\boldsymbol{U}, \boldsymbol{\Lambda}\}$ denote the eigenvectors and eigenvalues obtained from a direct decomposition of $\boldsymbol{\Gamma}$. There are several steps to the eigen reconstruction process. (**i**) The coarse-scale eigenvectors $\widehat{\boldsymbol{U}}$ can be transformed using the kernel matrix $\boldsymbol{K}$ as $\widetilde{\boldsymbol{U}} = \boldsymbol{K}\widehat{\boldsymbol{U}}$ to generate $\widetilde{\boldsymbol{U}}$ as an approximation to $\boldsymbol{U}$. (**ii**) This transformation alone is unlikely to set the directions of $\widetilde{\boldsymbol{U}}$ exactly aligned with $\boldsymbol{U}$. So, we update the directions in $\widetilde{\boldsymbol{U}}$ by repeated application of the following iteration (called *power* iterations [30]): $\widetilde{\boldsymbol{U}} \Leftarrow \boldsymbol{\Gamma_g}\widetilde{\boldsymbol{U}}$ Note, here instead of using $\boldsymbol{\Gamma}$ we use an adjusted matrix $\boldsymbol{\Gamma_g}$ given by $\boldsymbol{\Gamma}_g = \nu\boldsymbol{I} - \boldsymbol{\Gamma}$, where ν is a constant and $\boldsymbol{I}$ is an identity matrix. The power iterations will direct the eigenvectors to directions with large eigenvalues. But for fluctuation dynamics, we are interested in the *slow* eigen modes with *small* eigenvalues, hence the adjustment $\boldsymbol{\Gamma}_g$ is made. In particular, because of Gerschgorin disk theorem [30] the eigenvalues of $\boldsymbol{\Gamma}$ are bound to lie in a disk centered around the origin with a radius ν that is no more than twice the largest element on the diagonal of $\boldsymbol{\Gamma}$. (**iii**) Steps **i** and **ii** need not preserve orthogonality of the eigenvectors in $\boldsymbol{U}$. We fix this by a Gram-Schmidt orthogonalization procedure [30]. Finally, the eigenvalues are obtained from $\widetilde{\boldsymbol{\Lambda}} = \mathrm{diag}(\widetilde{\boldsymbol{U}}^{\mathsf{T}}\,\boldsymbol{\Gamma}\,\widetilde{\boldsymbol{U}})$. In [24] we present more details of this coarse to fine eigen mapping procedure, including a discussion on the number of power iterations to use; setting the thresholds for convergence and a comparison of the speed ups obtained over a standard sparse eigensolver for large matrices.

5.1 Collective Dynamics in the Reduced Space: Benchmarking against GNM

As discussed earlier, the eigenvalue decomposition of $\boldsymbol{\Gamma}$ yields the shape and frequency dispersion of equilibrium fluctuations. The shape of mode k refers to the normalized distribution of residue displacements along the principal axis k, given by the elements $\boldsymbol{u}_i^{(k)}$ $(1 \leq i \leq n)$ of the k^{th} eigenvector $\boldsymbol{u}^{(k)}$, and the

associated eigenvalue λ_k scales with the frequency of the k^{th} mode. In most applications, it is of interest to extract the contribution of the most cooperative modes, i.e. the low frequency modes that have been shown in several systems to be involved in functional mechanisms. To this end, we used the Markov-chain based hierarchy to build reduced Kirchhoff matrices $\widetilde{\boldsymbol{\Gamma}}$ at increasingly lower levels of resolution. We then performed their mode decompositions and propagated the information back over successive levels of the hierarchy, so as to generate the eigenvectors and eigenvalues for the fine-scale Kirchhoff matrix $\boldsymbol{\Gamma}$. We now show that *h*GNM maps the structure-dynamics information between successive levels of the hierarchy with minimal loss in accuracy.

First, our previous study identified ten slowest modes of interest, including the counter-rotation of the two rings around the cylindrical axis (non-zero mode 1) and other collective deformations proposed to be involved in chaperonin function [29]. Results presented in Fig. 4 show the mechanism of the dominant mode, mainly a global twisting of the structure where the *cis* and *trans* undergo counter rotation about the cylindrical axis (mode 1). The most important point is that these results corroborate previous findings [29, 1] and are reproduced here by adopting a reduced representation down to $m = 21$ nodes and mapped back to full-residue level.

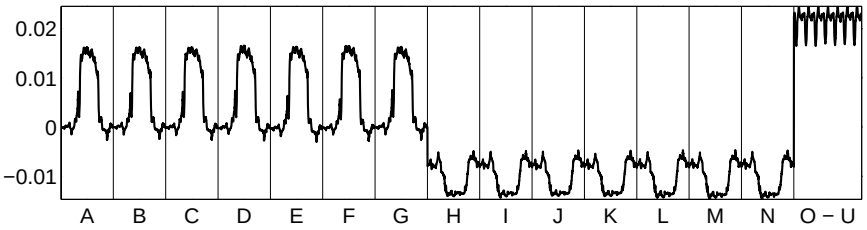
Second, Figure 5**a** compares the frequencies obtained by the full-residue-level representation, with those obtained by *h*GNM, upon propagation of the topology information from reduced level 4 (Fig. 2**c**). An excellent agreement is observed between the reconstructed eigenvalues $\widetilde{\lambda}$ (red curve) and their original values λ (open circles). In Fig. 5**b**, we display the correlation cosine between the eigenvectors $\boldsymbol{u}^{(k)}$ and $\widetilde{\boldsymbol{u}^{(k)}}$ obtained by the full-residue representation and the reconstruction from reduced level 4 respectively. Notably, the reduced representation contains only 35 nodes. Yet, the correlation cosine with the detailed representation containing 8015 nodes is almost unity throughout all the leading 25 modes, and above 0.8 for all modes, except the terminal four modes. The contribution of the latter to the overall dynamics is negligibly small compared to the large group of slow modes.

Finally, in order to assess the effect of coarse-graining on fluctuation dynamics, we compared in Fig. 5**c** the mean-square fluctuations obtained from different levels of the hierarchy with the experimental B-factor values. The theoretical B-factor for each residue v_i is computed using [31]

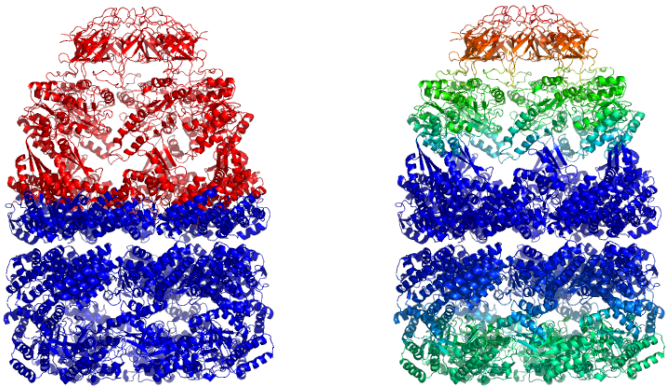
$$B_i = \frac{8\pi^2 k_B T}{\gamma} \sum_{k=2}^{n} \lambda_k^{-1} \left(\boldsymbol{u}_i^{(k)} \right)^2 , \tag{11}$$

where the summation is performed over all $n-1$ modes in the GNM, or over all the $m-1$ reduced eigenvectors and eigenvalues reconstructed from different levels of the hierarchy in *h*GNM. Because experimental B-factors correspond to each atom and our representation at the fine-scale is a summary of atom-atom contact information for each residue, we average the experimental B-factors over all atoms for each residue. As shown in Fig. 5**c**, a correlation coefficient value of 0.86 is achieved between the experimental and theoretical B-factors after mapping

the structure of 8015 residues into a representative network of 21 nodes. Thus, the fluctuation behavior of individual residues is accurately maintained despite a drastic reduction in the complexity of the examined network. Interestingly, a maximum in correlation coefficient is obtained at an intermediate level of resolution, $m = 133$, which may be attributed to an optimal elimination of noise in line with the level of accuracy of experimental data at this level of representation.



a. Shape of the dominant mode $\boldsymbol{u}_i$ over elements i

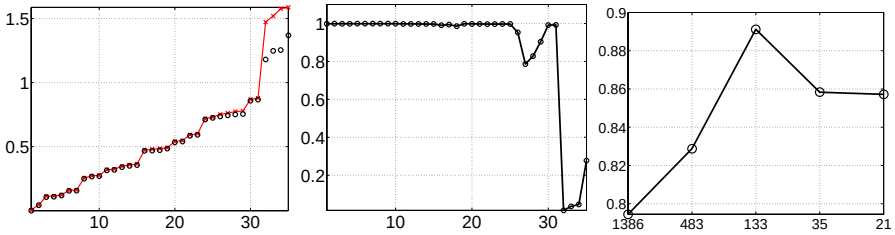


b. Displacement Polarity **c.** Mobility

Fig. 4. Dominant mode shape and mobility **a.** The labels on the abscissa indicate the chain identities, A-G belong to the *cis* ring, H-N come from the *trans* ring and O-U are from the GroES cap. The black curve gives the shape of the *slowest* eigen mode. The ordinate value is the normalized distribution of residue displacements along the dominant mode coordinate. **b.** Ribbon diagram illustrating the polarity of the displacement, color coded to be red for positive and blue for negative, indicating the anticorrelated motions of the two halves of the complex. **c.** Ribbon diagram color-coded after residue mobilities in mode 1. The mobility of residue v_i given by the squared displacement: $\left(\boldsymbol{u}_i^{(1)}\right)^2$, with a color code that is red for high and blue for low.

6 Discussion

A new method is introduced in the present study, which permits us to use structural information at atomic level in building network representations of different



a. $\widetilde{\lambda}_k$ (red), λ_k **b.** $\cos(\widetilde{\boldsymbol{u}}^{(k)}, \boldsymbol{u}^{(k)})$ **c.** Correlation Coefficient

Fig. 5. hGNM results (**a**) comparing eigenvalues λ (circles) from a direct decomposition of the $\boldsymbol{\Gamma}$ with multi-scale eigensolver spectrum $\widetilde{\lambda}$ (red line). For the direct eigen decomposition, we use the Matlab program `svds.m` which invokes the compiled ARPACKC routine [14], with a default convergence tolerance of `1e-10`. (**b**) Mode shape correlation: diag $\left(|\widetilde{\boldsymbol{U}}^{\mathsf{T}}\boldsymbol{U}|\right)$, between the matrix of eigenvectors $\widetilde{\boldsymbol{U}}$ derived by hGNM and $\boldsymbol{U}$ from direct decomposition. (**c**) Correlation coefficient between the theoretical B-factors (derived at each level of the hierarchy) *vs* experiment. The abscissa labels indicate the size m of the network at successive levels of the hierarchy.

complexity, which lend themselves to efficient analysis of collective dynamics and information propagation stochastics. The approach is particularly useful for analyzing large structures and assemblies, or cooperative/allosteric processes that are usually beyond the range of conventional molecular simulations.

We illustrated the utility of the methodology by way of application to the chaperonin GroEL-GroES, a widely studied structure composed of $n = 8015$ residues. Notably, we start with the full-atomic representation of the complex, which involves a total of $\approx 10^6$ atom-atom contacts (based on an interaction range of $4.5\mathring{A}$). Interatomic contacts define the affinities of pairs of residues, which are, in turn, used to define the weights of the connectors between residues (nodes) in the graph/network representation of the structure. The affinities also define the conditional probabilities of information transfer across residues following a Markovian process. The original network of n nodes is mapped into lower dimensional representations, down to $m = 21$ nodes, by an EM algorithm that maintains two basic properties of the original stochastic process: its Markovian conditional probabilities and stationary distribution (i.e. communication probability/potential) of individual residues. Two sets of operators, ensuring model reduction and reconstruction at different hierarchical levels permit us to perform the analysis at reduced scales but reconstructing the behavior.

Acknowledgments: Partial support by the NSF-ITR grant # EIA-0225636 and the NIH grant #1 R01 LM007994-01A1 is gratefully acknowledged.

References

1. Ma J.: Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular systems, *Structure*, **13**, 373–380, (2005).

2. Bahar, I. and Rader, A. J.: Coarse-grained normal mode analysis in structural biology, *Curr. Opi. Struct. Bio.* **15**, 1–7, 2005.
3. Rader, A. J., Chennubhotla, C., Yang, L.-W. & Bahar, I: The Gaussian Network Model: Theory and Applications, in *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, eds. Cui, Q. & Bahar, I., CRC Press, (2005).
4. Bahar, I., Atilgan, A. R. & Erman, B.: Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential, *Folding & Design* **2**, 173–181, 1997.
5. Haliloglu, T., Bahar, I. & Erman, B.: Gaussian dynamics of folded proteins, *Phys. Rev. Lett.* **79**, 3090, 1997.
6. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E.: The Protein Data Bank, *Nucleic Acids Research*, **28**, 235–242, 2000.
7. Miyazawa, S. & Jernigan, R. L.: Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation, *Macromolecules*, **18**, 534, (1985).
8. Bahar, I. & Jernigan, R. L.: Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation, *J. Mol. Biol.*, **266**, 195, (1997).
9. Chung, F. R. K. *Spectral Graph Theory* CBMS Lectures, AMS, (1997).
10. Doruker, P., Atilgan, A. R. & Bahar, I.: Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor, *Proteins* **40**, 512–524, (2000).
11. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O. & Bahar, I.: Anisotropy of fluctuation dynamics of proteins with an elastic network model, *Biophys. J.* **80**, 505, (2001).
12. Hinsen K.: Analysis of domain motions by approximate normal mode calculations, *Proteins* **33**, 417, (1998).
13. Tama, F. & Sanejouand, Y. H.: Conformational change of proteins arising from normal mode calculations, *Protein Eng.* **14**, 1–6, (2001).
14. Lehoucq, R. B., Sorensen, D. C. & Yang, C. *ARPACK User Guide: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods*, TR, Dept. of CAM, Rice University, (1996).
15. Simon, H. and Zha, H.: Low-rank matrix approximation using the Lanczos bidiagonalization process with applications, *SIAM J. of Sci. Comp.* **21**, 2257-2274, (2000).
16. Barnard, S. and Simon H.: A fast multi-level implementation of recursive spectral bisection for partitioning usntructured grid, *Concurrency: Practice and Experience*, **6**, 101–117, (1994).
17. Fowlkes, C., Belongie, S., Chung, F. & Malik, J.: Spectral Grouping Using the Nystrm Method, *I*EEE PAMI, **26**, 2, (2004).
18. Koren, Y., Carmel, L. & Harel, D.: Drawing Huge Graphs by Algebraic Multigrid Optimization, Multiscale Modeling and Simulation, **1**:4, 645–673, SIAM, (2003).
19. Doruker, P., Jernigan, R.L. & Bahar, I.: Dynamics of large proteins through hierarchical levels of coarse-grained structures, *J. Comp. Chem.*, **23**, 119, (2002).
20. Kurkcuoglu, O., Jernigan, R. L. & Doruker, P.: Mixed levels of coarse-graining of large proteins using elastic network model methods in extracting the slowest motions, *Polymers*, **45**, 649–657, (2004).
21. Marques O. *BLZPACK: Description and User's Guide*, TR/PA/95/30, CERFACS, Toulouse, France, (1995).

22. Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H.: Building-block approach for determining low-frequency normal modes of macromolecules, *Proteins*, **41**, 1–7, (2000).
23. Li, G. H. & Cui, Q.: A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca^{2+}-ATPase, *Bipohys. J.*, **83**, 2457, (2002).
24. Chennubhotla, C. & Jepson, A.: Hierarchical Eigensolver for Transition Matrices in Spectral Methods, *NIPS* **17**, 273–280, (2005).
25. McLachlan, G. J. & Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, N.Y.
26. Kullback, S.: *Information Theory and Statistics* Dover Publications, New York, (1959).
27. Kullback, S. & Leibler, R. A.: On Information and Sufficiency, *Ann. of Math. Stat.* **22**, 79–86, (1951).
28. Xu, Z. H., Horwich, A. L. & Sigler, P. B.: The crystal structure of the asymmetric GroEL-GroES$(ADP)_7$ chaperonin complex, *Nature* **388**, 741–750, (1997).
29. Keskin, O., Bahar, I., Flatow, D. Covell, D. G. & Jernigan, R. L.: Molecular Mechanisms of Chaperonin GroEL-GroES Function, *Biochemistry* **41**, 491–501, (2002).
30. Watkins, D. S. *Fundamentals of Matrix Computations*, Wiley-Interscience, (2002).
31. Kundu, S., Melton, J. S., Sorensen, D. C. & Phillips, G. N.: Dynamics of proteins in crystals: comparison of experiment with imple models, *Biophys. J.*, **83**, 723–732, (2002).
32. Landry, S. J., Zeilstra-Ryalls, J., Fayet, O., Georgopoulos, C., and Gierasch, L. M.: Characterization of a functionally important mobile domain of GroES, *Nature* **364**, 255–258, (1993).
33. Hohfeld, J., and Hartl, F. U.: Role of the chaperonin cofactor Hsp10 in protein folding and sorting in yeast mitochondria, *J. Cell Biol.* **126**, 305–315, (1994) .
34. Kovalenko, O., Yifrach, O., and Horovitz, A.: Residue lysine-34 in GroES modulates allosteric transitions in GroEL, *Biochemistry* **33**, 14974–14978, (1994).
35. Richardson, A., van der Vies, S. M., Keppel, F., Taher, A., Landry, S. J., and Georgopoulos, C.: Compensatory changes in GroEL/Gp31 affinity as a mechanism for allele-specific genetic interaction, *J. Biol. Chem.* **274**, 52–58, (1999).
36. Richardson, A., and Georgopoulos, C.: Genetic analysis of the bacteriophage T4-encoded cochaperonin Gp31, *Genetics* **152**, 1449–1457, (1999).
37. Richardson, A., Schwager, F., Landry, S. J., and Georgopoulos, C.: The importance of a mobile loop in regulating chaperonin/ co-chaperonin interaction: humans versus Escherichia coli, *J. Biol. Chem.* **276**, 4981–4987, (2001).
38. Shewmaker, F., Maskos, K., Simmerling, C., and Landry, S.J.: A mobile loop order-disorder transition modulates the speed of chaperoin cycling, *J. Biol. Chem.* **276**: 31257–31264, (2001).
39. Ma, J., Sigler, P. B., Xu, Z. H. & Karplus, M.: A Dynamic Model for the Allosteric Mechanism of GroEL, *J. Mol. Biol.* **302**, 303–313, (2000).
40. Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D.C., Joachimiak, A., Horwich, A.L., and Sigler, P.B.: The crystal structure of the bacterial chaperonin GroEL at 2.8$\mathring{A}$, *Nature* **371**, 578–586, (1994).
41. Yifrach, O. and Horovitz, A.: Nested cooperativity in the ATPase activity of the oligomeric chaperonin GroEL, *Biochemistry* **34**, 5303–5308, (1995).
42. Saibil, H. R., and Ranson, N. R.: The chaperonin folding machine, *Trends in Biochem. Sci.* **27**, 627–632, (2002).