Predictions of Protein Flexibility: First-Order Measures

Julio A. Kovacs,[1]* Pablo Chacón,[2] and Ruben Abagyan[1]
[1] *Department of Molecular Biology, The Scripps Research Institute La Jolla, California*
[2] *Centro de Investigaciones Biológicas (CIB/CSIC) 28040, Madrid, Spain*

ABSTRACT **The normal modes of a molecule are utilized, in conjunction with classical conformal vector field theory, to define a function that measures the capability of the molecule to deform at each of its residues. An efficient algorithm is presented to calculate the local chain deformability from the set of normal modes of vibration. This is done by considering each mode as an off-grid sample of a deformation vector field. Predictions of deformability are compared with experimental data in the form of dihedral angle differences between two conformations of ten kinases by using a modified correlation function. Deformability calculations correlate well with experimental results and validate the applicability of this method to protein flexibility predictions. Proteins 2004;56:661–668.**

Key words: protein flexibility; deformation; hinge regions; normal mode analysis; elasticity; conformal mapping; conformal vector field

INTRODUCTION

Many cellular functions depend on the conformational changes of the molecules involved in such process. Following the basic principle *if you know how it moves, you can infer how it works,* the knowledge of structural flexibility offers a straight-line connection between structure and function. To date, several efforts have been made regarding the study of molecular flexibility. These endeavors can be divided into two major groups: (1) those based on the comparative analysis of two or more conformational states at atomic resolution; and (2) those that attempt to forecast the intrinsic flexibility from a single conformation. Within the first group, a number of approaches exist for characterizing the intrinsic deformability inside a protein from different crystallographic or NMR structures of the same protein.[1–3] In this case, one is limited by the availability of experimental conformational states. Due to their predictive power, the second group is more interesting from the biological point of view. In this context, there are several conventional methods that simulate the protein dynamics numerically. Among them, Molecular Dynamics (MD) and Monte Carlo (MC) simulation allow the direct study of the trajectories.[4–8] The disadvantage of these methods is that the computational time needed to reach large-scale conformational changes is beyond practical wide-range application. (There are, however, faster low-resolution models that achieve good results[9]). Other approaches try to identify flexible hinge joints or rigid domains directly from a single conformation of the molecule.[10–14] Very recently, an ingenious and elegant approach was developed for estimating the flexibility of proteins using graph theory.[14] Despite the excellent results obtained, the outcome is rather qualitative and mainly a distinction between rigid and nonrigid residues of the protein. The methods in this class are in general fast, but need to be carefully validated with experimental observations.

Midway between methods that simulate the protein motion with a detailed potential and those using a more drastic modeling approach, normal mode analysis (NMA) of simplified protein models is a very attractive tool that permits the study of dynamics of proteins and other macromolecules. Modal analysis has been applied successfully to predict large-scale motions from a single conformation. This analysis, based on strongly softened and reduced models and on the harmonic approximation of the motion, yields the vibrational modes of a given structure or model.[15–17] These modes constitute an orthonormal vector-space basis of the system displacements, i.e., any displacement can be described as a linear combination of such modes.

Several authors pointed out the excellent correlation between modal analysis results and the observed functional motions.[16,18,21] A good example of this correlation can be observed in the molecular movements database,[19] where it is possible to compare a collection of observed protein motions with those analyzed by NMA.[20] Recent efforts extend the applicability range of this technique to large systems by reducing the spatial detail and using a more simplified harmonic interaction potential.[21,22] Furthermore, even without atomic resolution and from low resolution structures, good correlation has been found between the motion characterized by the low frequency modes and the experimentally observed functional motions of large macromolecules.[23]

Here we present a method that combines NMA of an improved "contact" model of a protein with conformal

*Corresponding authors; email addresses: jkovacs@scripps.edu, abagyan@scripps.edu.

Grant sponsor: DOE; Grant number: DE-FG03-00ER6304. P.C. was supported by Fundación BBVA.

*Correspondence to: J.A. Kovacs, Dept. of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Rd., La Jolla, CA 92037. E-mail: jkovacs@scripps.edu

Received 24 November 2003; Accepted 4 February 2004

Published online 20 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20151

TABLE I. Testing Benchmark Consisting of Ten Pairs of Kinases†

Molecule	Conf. A	Conf. B	N	min r_c (Å)
Protein kinase spk1	1j4p (A)	1k3q (A)	151	10
Pyrophosphokinase	1hka (A)	1eqo (A)	158	7
hprk protein	1kkl (A)	1jb1	167	11
Adenosylcobinamide kinase	1cbu (A)	1c9k (B)	180	7
Guanylate kinase	1ex6 (A)	1ex7 (A)	186	9
Adenylate kinase	1dvr (A)	1aky	220	7
Cyclin-dependent kinase	1fin (A)	1hcl	298	8
Adenosine kinase	1lio	1lik	329	8
cAMP-dependent protein kinase	1cmk (E)	1jlu (E)	350	8
3-phosphoglycerate kinase	16pk	13pk (A)	415	8

†Shown are the PDB codes of both conformations, the number N of residues, and the minimum cutoff radius (r_c) such that a NMA performed with spring strengths that vanish outside r_c yields no "floppy modes" (i.e., those with eigenvalue 0, other than the six corresponding to rigid motions).

vector field theory. Considering the modes as vector fields over the molecule, we can define what we call *deformability function.* The proposed method represents the molecular structure as a network of point masses interconnected with springs, whose strengths are dependent on the distance of the corresponding point masses and on the residue contact areas. This mechanical model of the molecule is then subjected to a normal mode analysis, and the set of modes and frequencies obtained are subsequently merged together, in a precise mathematical way, to give a measure of the amount of deformation that the protein molecule can undergo at each of its residues. Here, we present the results and performance achieved using as input several protein kinases with two known conformational states. The results obtained demonstrate that the deformability function is well correlated with experimental results and validate its applicability to predict protein flexibility.

MATERIALS AND METHODS

Protein Vibrational Analysis

Normal mode analysis (NMA) furnishes a way to study the atomic motions of a molecule by decomposing them into their different vibrational modes and frequencies.[24,25] It has been originally applied to small molecules, but as computer power has been growing, it is now possible to study proteins with more than 200 amino acids using an all-atom model. However, since we intend our method to be fast, we use a reduced C_α normal mode analysis approach, which, according to observations of many researchers,[22] reproduces quite accurately the large-scale molecular motions as predicted by the all-atom model. Also, since vibrations are computed based on a single minimum of the harmonic energy landscape, the low-frequency modes should correspond to the directions of more shallow energy increase, i.e., to those of more deformability of the structure.

Thus, given a protein molecule, its C_α atoms are interconnected with "springs" of certain strengths (described below). This harmonic model of the molecule gives rise to a potential energy function E. If $\mathbf{H}$ denotes the Hessian of E, we have the standard secular equation:

$$\det(\mathbf{H} - \lambda\mathbf{M}) = 0, \tag{1}$$

where $\mathbf{M}$ is the mass matrix, which in our case is diagonal. In order to get a symmetric eigenvalue problem, the secular equation is rewritten as:

$$\det(\tilde{\mathbf{H}} - \lambda I) = 0, \tag{2}$$

where $\tilde{\mathbf{H}} = \mathbf{M}^{-1/2}\mathbf{H}\mathbf{M}^{-1/2}$. The eigenvalues $\lambda_n = \omega_n^2$ (where the ω_n are the vibrational frequencies) and the corresponding eigenvectors $\tilde{\mathbf{u}}^n$ of $\tilde{\mathbf{H}}$ (the normal modes of vibration) are determined by a diagonalization procedure. The "modified modes" $\mathbf{u}^n$ of the system are easily obtained from the $\tilde{\mathbf{u}}^n$:

$$\mathbf{u}^n = \mathbf{M}^{-1/2}\tilde{\mathbf{u}}^n. \tag{3}$$

Each of these modified modes can be visualized as the velocity vectors that atoms have when, while vibrating according to that mode, they pass through their initial positions.

We perform the normal mode analysis (NMA) on the C_α atoms, setting the mass m_i of each as the total mass of the corresponding ith residue. The spring strengths are set in the following way:

$$C_{ij} = \left(\frac{r_0}{r_{ij}}\right)^6 + a\ s_{ij}, \tag{4}$$

where i, j denote residue numbers, r_{ij} is the distance between the α carbons of residues i and j, and the s_{ij} are normalized residue contact areas, which are computed by the ICM program,[26] using the algorithm presented in Shrake and Ruptey.[27] The parameter r_0 was set to 3.8Å, which is approximately the mean distance between consecutive α carbons. The sixth power for the contact strengths was determined empirically as the lowest power that produces predictions close to those obtained by using constant strength within a cutoff radius and zero outside. The parameter a is determined so as to optimize the correlation of the predictions with the experimental data contained in our benchmark (Table I).

Using the above strengths and masses, and for a fixed value of a, an NMA is performed by means of a highly efficient subroutine in the LAPACK linear algebra library,[28] yielding eigenvalues λ_n and eigenvectors $\tilde{\mathbf{u}}^n$ (1 ≤

$n \le 3N$). Assuming that the λ_n are sorted in increasing order, we have $\lambda_1 = \ldots = \lambda_6 = 0$ (modes corresponding to rigid motions). The frequencies of vibrations are $\omega_n = \sqrt{\lambda_n}$. Each normal mode $\tilde{\mathbf{u}}^n$ is normalized so that $\sum_{i=1}^{N} \|\tilde{\mathbf{u}}_i^n\|^2 = 1$. Then the "modified" (velocity) modes $\mathbf{u}^n$ are obtained from the $\tilde{\mathbf{u}}^n$ through Eq. 3.

Conformal Vector Field Theory

Each normal mode $\mathbf{u}^n$ obtained in the vibrational analysis can be viewed as a vector field over the molecule. Thus, we can harness the vector field theory framework to process the normal mode results. Using this viewpoint allows us to define a function that measures the capability of the molecule to deform at each of its residues. To this end, we briefly review a few concepts from vector field theory.

A given vector field $\mathbf{u} : G \to \mathbb{R}^3$ is called *conformal vector field* if, for every t, $\phi_t : G \to G$ is a conformal transformation[29] of G. Here $\{\phi_t\}$ is the 1-parameter group of transformations[30] defined by the vector field, and $G \subset \mathbb{R}^3$ is an open set containing the points that represent our molecule (in our case, the C_α atoms). Thus, conformality of $\mathbf{u}$ means that following the integral curves of the field produces transformations that preserve shape (locally), i.e., angles are not changed. The following theorem regarding conformal vector fields is fundamental for our method:

Theorem. *The vector field* $\mathbf{u}$ *is conformal if and only if the tensor field* $S = S_\mathbf{u}$ *with components*

$$S_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right) - \frac{1}{3}\,\mathrm{div}\,\mathbf{u}\;\delta_{kl} \quad (1 \le k, l \le 3) \qquad (5)$$

vanishes identically.

For a proof of this theorem, see Weber and Goldberg.[29] Here δ_{kl} denotes the *Kronecker delta:* $\delta_{kl} = 1$ if $k = l$ and 0 otherwise. Also, div $\mathbf{u}$ denotes the *divergence* of $\mathbf{u}$: div $\mathbf{u} = \sum_{k=1}^{3} \partial u_k/\partial x_k$.

Note: In this section, u_1, u_2, u_3 stand for the three components of the vector field $\mathbf{u}$ as functions of the spatial coordinates x_1, x_2, x_3. This should not be confused with an expression such as $\mathbf{u}_i$, which means "$\mathbf{u}$ at the ith residue" (a 3D vector).

We note that the "main part" of the tensor S is nothing but the *strain tensor* of linear elasticity theory[31]:

$$E_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right). \qquad (6)$$

This tensor has the property that its vanishing is equivalent to $\mathbf{u}$ representing, through its flow, rigid motions of G. (In this case, each ϕ_t is an isometry, and $\mathbf{u}$ is called a *Killing vector field*).

The correction term 1/3 div $\mathbf{u}$, substracted from the diagonal entries of E_{kl}, measures the change in volume as one follows the flow of $\mathbf{u}$, thereby taking care of any change of scale. Hence, the tensor S is insensitive to changes in scale; it only detects changes of shape. In other words, it gives a measure of the change of shape produced when following the flow of $\mathbf{u}$. By the way, this correction term is important in order to compensate for the necessary "unphysical softness" of our protein model.

If we denote with $\nabla\mathbf{u}$ the gradient of $\mathbf{u}$, that is, the matrix whose entries are

$$(\nabla\mathbf{u})_{kl} = u_{k,l} = \frac{\partial u_k}{\partial x_l} \quad (1 \le k, l \le 3), \qquad (7)$$

and with sym the "symmetrization operator":

$$\mathrm{sym}\,A = \frac{1}{2}(A + A^T), \quad \text{where } A^T = \text{transpose of } A, \qquad (8)$$

then the tensor $S_\mathbf{u}$ can be written as:

$$S_\mathbf{u} = \mathrm{sym}(\nabla\mathbf{u}) - \frac{1}{3}\,\mathrm{div}\,\mathbf{u}\,I, \qquad (9)$$

where I is the 3×3 identity matrix.

Deformability

According to the theorem and remarks in the previous section, for a given vector field $\mathbf{u}$ we can define the *deformation* function $d_\mathbf{u}$: $G \to \mathbb{R}$ as:

$$d_\mathbf{u} = \|S_\mathbf{u}\|. \qquad (10)$$

Here $\|\cdot\|$ denotes the *norm* of a tensor; see Appendix A for details. The function $d_\mathbf{u}$ quantifies the deformation (change of shape) produced by the vector field $\mathbf{u}$ on the molecule. Since every normal mode $\mathbf{u}^n$ is considered as a vector field, it is possible to characterize the deformation of the molecule by taking the norm of the "conformal tensor" $S_{\mathbf{u}^n}$ associated to each mode. Thus, in analogy to the classical formula for the atomic fluctuations[32]:

$$\langle(\Delta r_i)^2\rangle = k_B T \sum_{n=7}^{3N} \left(\frac{\|\mathbf{u}_i^n\|}{\omega_n}\right)^2, \qquad (11)$$

where i denotes residue number, k_B is Boltzmann's constant, and T is temperature, we call *deformability* of the molecule M to the function $d_M : M \to \mathbb{R}$ defined by:

$$d_M^2 = \sum_{n=7}^{3N} \left(\frac{d_{\mathbf{u}^n}}{\omega_n}\right)^2, \qquad (12)$$

or, by making the residue number i explicit:

$$d_M(i) = \left(\sum_{n=7}^{3N} \left(\frac{d_{\mathbf{u}^n}(i)}{\omega_n}\right)^2\right)^{1/2}. \qquad (13)$$

Thus, by means of the normal modes we capture all possible ways in which the molecule can deform, and then combine the deformation measures $d_{\mathbf{u}^n}$ corresponding to each mode by using their statistical thermal amplitudes ω_n^{-1}, in order to obtain the function d_M describing how much, in average, the molecule can deform at each point (residue). Although at first sight it might seem that the concept of deformability is a local one, it is actually global, i.e., its value at a particular point depends on the global structure of the molecule. Also, it can be easily checked

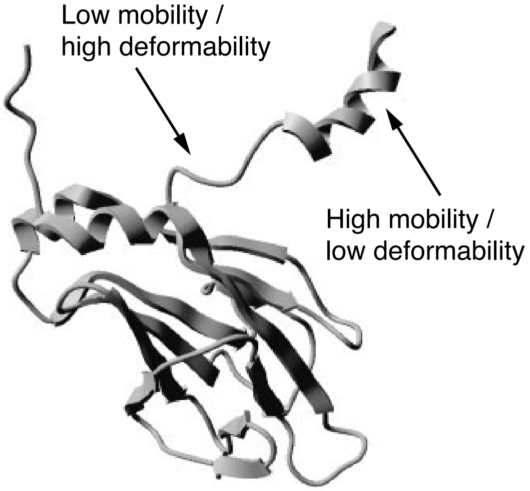

Fig. 1. This protein (PDB code 1j4p) illustrates the contrast between mobility [given by the classical fluctuation formula (Eq. 11), and reflected in B-factors] and deformability (defined by Eq. 12). The helix is relatively rigid (low deformability values), but points near its tip are quite mobile (high mobility values), whereas the loop will have low mobility but high deformability values, acting as a hinge region around which the helix can rotate. See also Figures 2 and 3.

that our definition of deformability, Eq. 12, is invariant under scaling of the masses m_i, and also under scaling of the contact strengths C_{ij}

Note that deformability (Eq. 12) gives us a measure of the flexibility of the protein, whereas the classical fluctuation formula (Eq. 11) is related to protein mobility (typically reflected in B-factors). Both measures are complementary but distinct. For example, Figure 1 shows one of the molecules in our set (1j4p), which has a protruding helix connected to the rest of the protein by a loop. Points in this loop have lower mobility values (B-factors) than points closer to the tip of the helix, but have higher deformability values than points in the helix, since the loop acts as an elbow or hinge region around which the helix can rotate. In this way, our deformability measure can be utilized to detect hinge points. Another example is 1kkl (see Figs. 2 and 3), although in this case the loop region (*f*) is not as flexible due to its particular ("helix-like") geometry, which makes it somewhat rigid (but less rigid than the helix).

Calculation of the Spatial Derivatives

In order to compute the deformability, we have to compute the tensors $S_{\mathbf{u}^n}$. For this, in turn, we need a numerical way to evaluate derivatives of the normal mode vector fields, whose values are known only at the N points (residues) of the molecule. Several methods exist for performing interpolation and gradient estimation of scattered data.[33, 34] Gradient estimation is usually done by differentiating an interpolant fitted to the data.[34] We chose *Hardy's hyperbolic multiquadric* interpolant,[35] which, according to tests,[33, 34] performs extremely well. The only reported drawback of this method is that it is more time consuming than other methods, but this begins to be an issue only for 10^4–10^5 points. Indeed, even for the largest of our test cases (415 residues), it takes a small fraction of a second to compute the gradient of a function on the whole molecule. Another advantage of this method is that *the same interpolant can be used for the whole set of points.*

The outline of Hardy's hyperbolic multiquadric method is as follows. Suppose the molecule M consists of points $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N$ (in our case, α carbons). Then the interpolant has the form:

$$g(\mathbf{p}) = \sum_{i=1}^{N} c_i \sqrt{\mathrm{dist}(\mathbf{p},\mathbf{p}_i)^2 + b}, \qquad (14)$$

where "dist" denotes the distance between two points, and b is an adjustable parameter. Its value has negligible effect on the quality of the gradient estimation; we set it to $\frac{1}{10}$ of the squared diameter of the molecule.[34]

The coefficients c_i are determined by imposing the condition that the interpolant and the given function agree on all points $\mathbf{p}_j$:

$$g(\mathbf{p}_j) = f(\mathbf{p}_j) \ \ (1 \leq j \leq N). \qquad (15)$$

Then, analytical derivatives of the interpolant g furnish approximate values for the derivatives of f.

RESULTS AND DISCUSSION

We applied our method to the cases contained in our testing benchmark (Table I). The benchmark consists of ten kinases of various shapes and sizes available in two distinct atomic conformations (conformation A and B in Table I). Specifically, we computed the deformation functions d_M from the conformation A of each protein. Examples of the results are shown in Figure 2, which are color- and size-coded according to the deformability values at each residue: blue/small means more rigid, red/large more deformable. As can be seen, the results are consistent with the expected protein structural flexibility. In all cases, the terminal regions, external loops, or hinge regions are coincident with high deformability values. In fact, a good qualitative agreement can be noticed between the flexibility observed in the relative motion of the two conformations available for each protein and the deformability measures. Nevertheless, an accurate manner to assess the correspondence level of deformability and protein flexibility is required. To this end, the deformability predictions obtained were compared with the dihedral angle difference (DAD) between both conformations of each kinase (see Appendix B for DAD definition). The DADs corresponding to the examples in Figure 2 are plotted versus residue number in Figure 3. There, it can be seen that there is a qualitative agreement between the predictions and the DADs, but the location of the peaks of the predictions appear somewhat shifted, some to the left, some to the right. This is presumably due to the "linking effect" that the interpolation of the vectors has over nearby regions of the molecule.

To quantify the agreement between deformability (variable X) and DAD (variable Y), one could use the usual correlation function, defined by:

$$\mathrm{corr}_0 = \frac{1}{N-1} \sum_{j=1}^{N} \frac{X_j - \bar{X}}{\sigma_X} \cdot \frac{Y_j - \bar{Y}}{\sigma_Y}, \qquad (16)$$

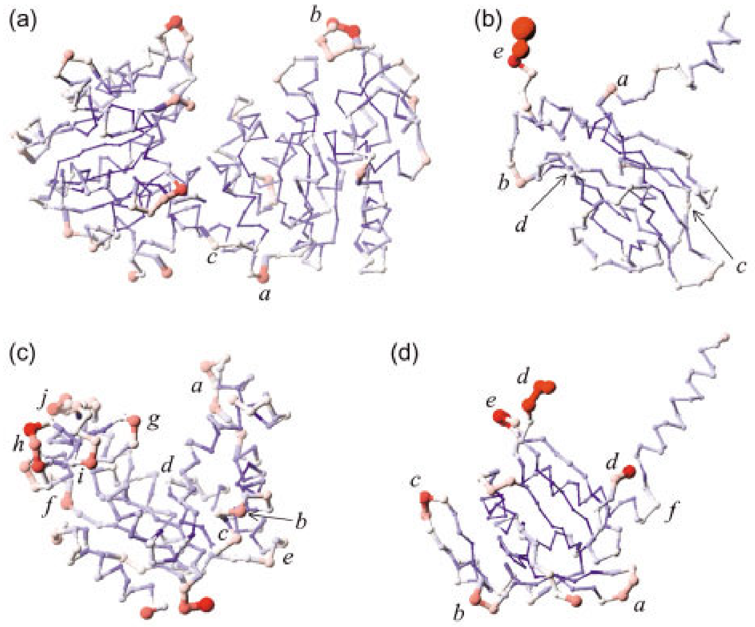

Fig. 2. Color- and size-coded deformability maps for four of the cases tested: **(a)** 16pk; **(b)** 1j4p; **(c)** 1dvr; (**d**) 1kkl. Red and large features indicate more flexible residues. Small italic letters refer to features indicated in Figure 3. The value of the parameter *a* is 0.25. Besides the double cue of color and size, the "fog" effect helps in distinguishing the depth of different parts of the molecules.

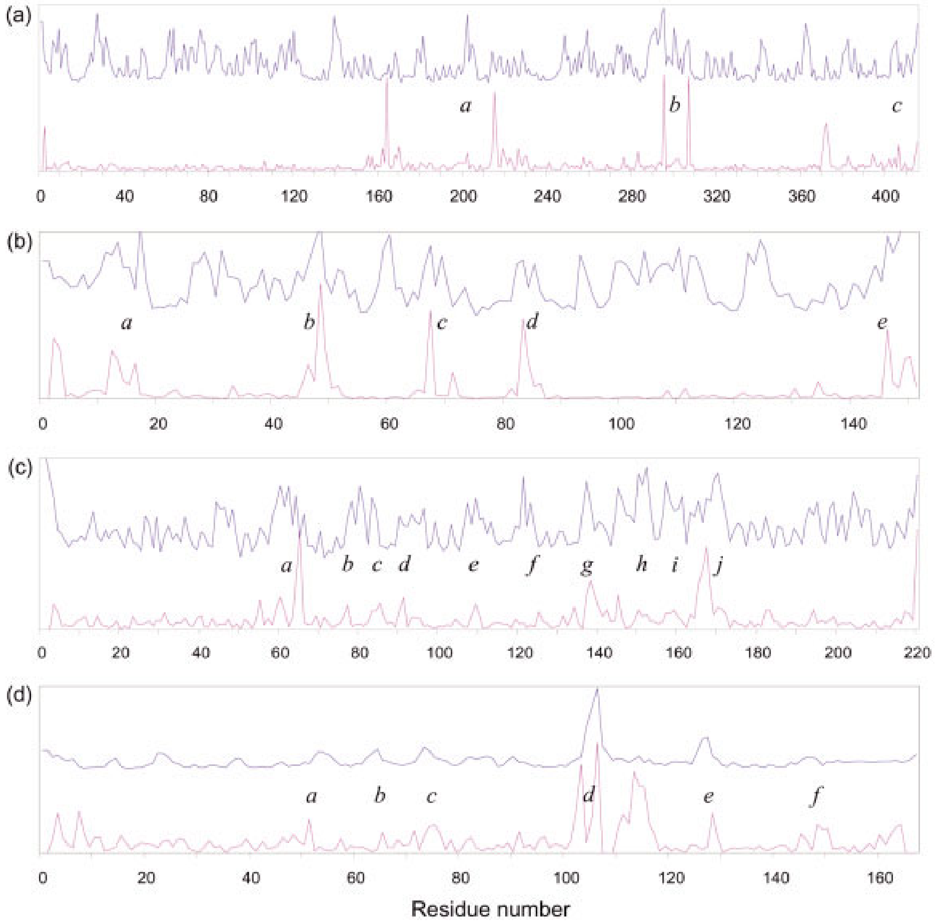

Fig. 3. Deformability (blue) and dihedral angle difference (magenta) between conformations A and B (Table I), plotted vs. residue number, for the four cases shown in Figure 2 (italic letters indicate corresponding features), using the optimal *a* (0.25). The deformability curves have been raised to avoid clutter.

TABLE II. Correlation Values for *a* = 0.25†

Molecule	PDB (conf. A)	Correlation No window	 3-residue window
Protein kinase spk1	1j4p	0.370	0.496
Pyrophosphokinase	1hka	0.374	0.538
hprk protein	1kkl	0.412	0.796
Adenosylcobinamide kinase	1cbu	0.088	0.487
Guanylate kinase	1ex6	0.191	0.610
Adenylate kinase	1dvr	0.294	0.638
Cyclin-dependent kinase	1fin	0.384	0.691
Adenosine kinase	1lio	0.264	0.623
cAMP-dependent protein kinase	1cmk	0.196	0.578
3-phosphoglycerate kinase	16pk	0.119	0.308

†Shown are the correlations between the deformability values (computed using Eq. 12) and the dihedral angle difference between conformations A and B of each molecule. (These are plotted in Fig. 3). The "no window" column shows "raw" correlations, while the "3-residue window" column shows the correlations obtained by allowing an offset of 3 residues (see text for details).

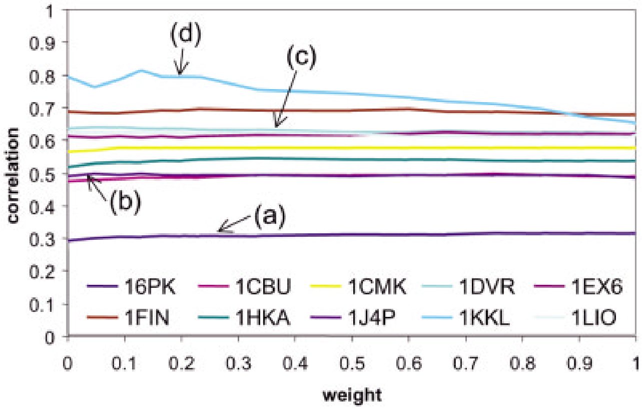

Fig. 4. Correlation (3-residue window) as a function of the weight *w* of the contact area term. This weight is related to the parameter *a* in Eq. 4 by *w* = *a*/(1 + *a*). The letters (a–d) refer to the four cases shown in Figures 2 and 3 (a–d). The correlation averaged over the ten cases has a maximum at *w* = 0.2 or *a* = 0.25. Table II lists the correlation values for this *a*.

where $\bar{X}, \bar{Y}$ are the averages of the samples, and σ_X, σ_Y are their standard deviations. However, due to the slight shifting of the peaks, the corr_0 values (shown in the "no window" column in Table II) turn out to be relatively low.

In order to properly measure the agreement, and to estimate the parameter a in Eq. 4, we need a robust objective function to be maximized against this parameter. Since corr_0 is too sensitive to small displacements of the peaks, we consider a "windowed correlation." For an integer $h \geq 0$, it is defined as follows:

$$\text{corr}_h = \frac{1}{N-1}\sum_{j=1}^{N}\frac{X_j - \bar{X}}{\sigma_X}\cdot\frac{Y_{k_j} - \bar{Y}}{\sigma_Y}, \qquad (17)$$

where k_j is an index between $j - h$ and $j + h$ satisfying:

$$\left|\frac{Y_{k_j} - \bar{Y}}{\sigma_Y} - \frac{X_j - \bar{X}}{\sigma_X}\right| \leq \left|\frac{Y_l - \bar{Y}}{\sigma_Y} - \frac{X_j - \bar{X}}{\sigma_X}\right| \qquad (18)$$

for all l between $j - h$ and $j + h$.

A window $h = 3$ was seen to be enough to capture the small shifts of the peaks across the benchmark. Therefore, we computed corr_3 for each of the cases in the benchmark and for a range of values of a. Results are shown in Figure 4. The average of the corr_3's over the benchmark is maximum for $a = 0.25$.

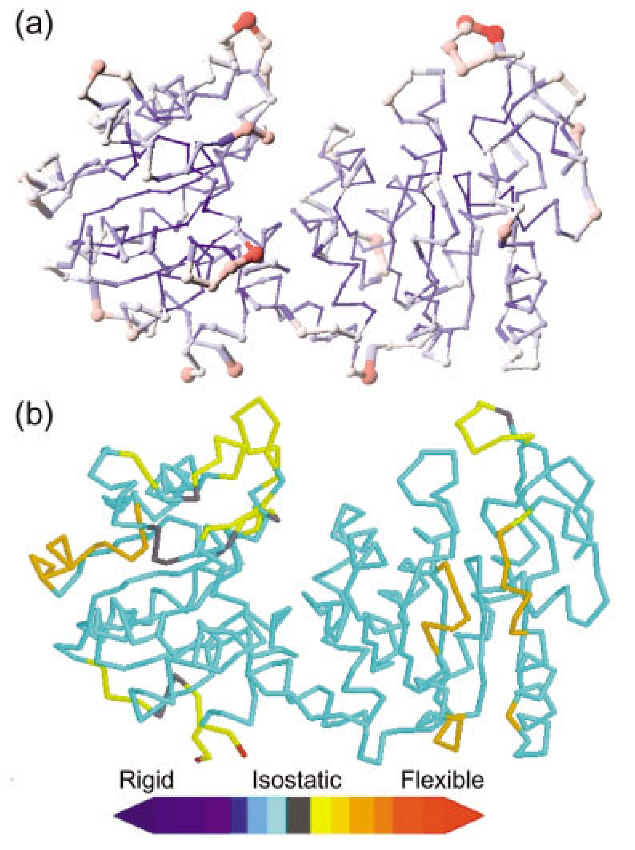

Fig. 5. **(a)** Deformability for the 16pk kinase [same as Fig. 2(a)]. **(b)** Flexibility index from the program FIRST. The picture was captured directly from the screen output of the FirstWeb website.[14]

Values of corr_3 are listed in the last column of Table II for all the cases of our benchmark. According to Figures 3 and 4, we can see that a value of $\text{corr}_3 = 0.5$ or more indicates a reasonably good agreement with experimental data. Of all ten cases in the benchmark, there is only one outlier (case (a) in Figures 2–4), whose correlation is 0.3, while the rest have correlations between 0.5 and 0.8. We applied the FIRST program[14] to this case, obtaining predictions similar to ours (Fig. 5). This suggests that the low correlation value for this particular case is due to an inappropriate second conformation of the molecule (in which the molecule, being able to deform relative to the first conformation, does not).

We have compared the deformability values with accessibility and with B-factors, and observed that all three are essentially different measures. Naturally, in some structures (like 16pk) there are similarities in, e.g., flexible protruding loops. But in general there is low correlation between these measures (like in the protruding helices in Fig. 2).

CONCLUSIONS

We introduced the concept of *deformability,* which is a measure of the capability of a given molecule to deform at each of its residues. This was done in three steps: (1) obtaining the normal modes of vibration of the molecule; (2) measuring, by means of the norm of the "conformal tensor" associated to each mode, the amount of deformation that each mode produces on the molecular structure; (3) combining all these measures into a single deformability measure using the statistical thermal amplitudes ω_n^{-1} of the modes.

In order to perform the NMA, we utilized a mechanical model of the molecule in which the α carbons are connected with one another with springs whose strengths are expressed as a combination of two terms: an inverse sixth power of the distance, and a term proportional to the area of contact between the corresponding residues. This definition gives results similar to those using the classical cutoff radius, and is mathematically more convenient, since it avoids the discontinuity at the boundary of the cutoff sphere.

Deformability predictions performed by this method were compared with experimental data obtained by measuring the DAD of the two atomic conformations available for each kinase (Table I). To quantify this agreement, and at the same time to have an objective function to maximize against the coefficient a of the contact area term, we considered the "windowed correlations" corr_h for integers $h \geq 0$. The usual correlation, corr_0, is too sensitive to small shifts in the location of the peaks of the predictions relative to those of the DADs. We noted that taking $h = 3$ residues covers amply the observed shifts (which, presumably, are due to the "linking effect" produced by the interpolation of the vectors implicit in the numerical calculation of the spatial derivatives). Therefore, corr_3 was taken as the objective function, resulting in $a = 0.25$ being the optimal value across our benchmark.

For all test cases except one, correlations between 0.5 and 0.8 have been obtained. The 3-phosphoglycerate kinase was the only outlier with a correlation of 0.3. In this case, the low correlation value can be attributed to an inappropriate second conformation, since (1) the deformability observed in (a) in Figure 4 seems reasonable (high values in loop and hinge regions), and (2) very compatible results were obtained using the qualitative flexibility prediction method based on graph theory[14] (Fig. 5). Therefore, the good agreement found between the deformability function and the atomic experimental data validates our method as a quantitative way for estimating flexibility. This method inherits the low computational cost and wide applicability range of traditional modal analysis and extends its prediction power to a quantitative level.

In this work we demonstrate, with a small set of kinases, the potential application range of our method. However, the general applicability of this method—and the validity of the current parametrization—must be addressed with a more comprehensive data set. We are currently working on the improvement of the spring strengths by using the actual free energy of the contacts, in particular to distinguish strong hydrophobic contacts from weaker ones. We will present this in a future publication, applied to a larger family of protein kinases for which different conformations are available. This will be a crucial point to reflect more realistically the interaction between residues, thereby enhancing the deformability prediction ability of this approach.

APPENDIX A: OPERATOR NORMS

Let A be a symmetric 3×3 matrix. It naturally defines an operator (denoted with the same name) $A\colon \mathbb{R}^3 \to \mathbb{R}^3$ by multiplication: $A(x) = Ax$. We can define a number of measures to quantify the "magnitude" of this operator, e.g.:

$$\mathrm{m}_1 = \max\{|\mu_1|, |\mu_2|, |\mu_3|\},$$

$$\text{where the } \mu_i \text{ are the eigenvalues of } A, \quad (19)$$

$$\mathrm{m}_2 = \max_{\|x\|=1} |(Ax,x)|, \quad (20)$$

$$\mathrm{m}_3 = \max_{\substack{\|x\|=1\\ \|y\|=1}} |(Ax, y)|, \quad (21)$$

$$\mathrm{m}_4 = \max_{\|x\|=1} \|Ax\|. \quad (22)$$

It is straightforward to check that $\mathrm{m}_1 \leq \mathrm{m}_2 \leq \mathrm{m}_3 \leq \mathrm{m}_4$. On the other hand, the spectral theory of self-adjoint operators[36] shows that $\mathrm{m}_1 = \mathrm{m}_4$. Therefore, all the above measures are actually the same:

$$\mathrm{m}_1 = \mathrm{m}_2 = \mathrm{m}_3 = \mathrm{m}_4. \quad (23)$$

This common value is called the *norm* of A, and is denoted by $\|A\|$. Incidentally, m_1 gives us a concrete way to compute the norm.

APPENDIX B: DIHEDRAL ANGLE DIFFERENCE CALCULATION

For each pair of conformations, the dihedral angle difference (DAD) was computed in the following way:

$$\mathrm{DAD}_i = |\varphi_i^A - \varphi_i^B| + |\psi_i^A - \psi_i^B|, \quad (24)$$

where each angle difference was reduced to a value between $-180°$ and $+180°$ prior to taking its absolute value. The meaning of these angles is as follows: ϕ_i is the torsion angle around the bond connecting the N and C_α atoms of residue i, and ψ_i is the torsion angle around the bond connecting the C_α and C atoms of residue i.

REFERENCES

1. Nichols WL, Rose G, Eyck LFT, Zimm BH. Rigid domains in proteins: An algorithmic approach to their identification. Proteins Struct Funct Genet 1995;23:38–48.
2. Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definition. Protein Sci 1995;4:872–884.
3. Boutonnet N, Rooman M, Wodak S. Automatic analysis of protein conformational changes by multiple linkage clustering. J Mol Biol 1995;253:633–647.
4. McCammon JA, Harvey SC. Dynamics of proteins and nucleic acids. Cambridge: Cambridge University Press; 1987.
5. Leach AR. Molecular modelling: principles and applications. Essex, UK: Addison Wesley Longman; 1996.
6. Ma J, Karplus M. The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. Proc Natl Acad Sci USA 1998;95:8502–8507.
7. Case DA. Molecular dynamics and normal mode analysis of biomolecular rigidity. In: Doorpe M, Duxbury P, editors. Rigidity theory and applications. New York: Kluwer Academic/Plenum Publishers; 1999;p 329–344.
8. O Keskin, RL Jernigan, IB. Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys J 2000;78: 2093–2106.
9. Bahar I, Erman B, Haliloglu T, Jernigan RL. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. Biochemistry 1997:36: 13512–13523.
10. Holm L, Sander C. Parser for protein folding units. Proteins 1994;19:256–268.
11. Karplus PA, Schulz GE. Solvation: a molecular dynamics study of a dipeptide in water. Naturwissenschaften 1985;72:212–213.
12. Janin J, Wodak S. Structural domains in proteins and their role in the dynamics of protein function. Prog Biophys Mol Biol. 1983;42: 21–78.
13. Maiorov V, Abagyan R. A new method for modeling large-scale rearrangements of protein domains. Proteins 1997;27:410–424.
14. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins Struct Funct Genet 2001;44:150–165.
15. Goldstein H, Poole CP, Safko JL. Classical mechanics, 3rd ed., San Francisco: Addison Wesley; 2002.
16. Case DA. Normal mode analysis of biomolecular dynamics. In: Gunsteren WF, Weiner PK, Wilkinson AJ, editors. Computer simulation of biomolecular systems, Vol. 3. Dordrecht: Kluwer Academic Publishers; 1997, p 284–301.
17. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements in proteins. Biochemistry 1994;33:6739–6749.
18. Tama F, Wriggers W, Brooks CL. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. J Mol Biol 2002;321:297–305,
19. Echols N, Milburn D, Gerstein M. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res. 2003;31:478–482.
20. Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu, H, and Gerstein, M. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins 2002;48:682–695.
21. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold. Des. 1997;2:173–181.
22. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter atomic analysis. Phys Rev Lett 1996;77:1905–1908.
23. Chacón P, Tama F, Wriggers W. Mega-Dalton biomolecular motion captured from electron microscopy reconstructions. J Mol Biol 2003;326:485–492.
24. Case DA. Normal mode analysis of protein dynamics. Curr. Opin Struct Biol 1994;4:285–290.
25. Noguti T, Go N. Collective variable description of small-amplitude conformational fluctuations in a globular protein. Nature 1982;296: 776–778.
26. MolSoft. ICM 3.0 program manual. San Diego, CA: MolSolft LLC; 2002.
27. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and Insulin. J Mol Biol 1973;79:351–371.
28. LAPACK Users' Guide, 1999. http://www.netlib.org/lapack.
29. Weber WC, Goldberg SI. Conformal deformations of Riemannian manifolds. Queen's papers in pure and applied mathematics, no. 16. Queen's University, Kingston, Ontario, 1969.
30. Arnol'd V. Ordinary differential equations. New York: Springer-Verlag; 1992.
31. Love AEH. A treatise on the mathematical theory of elasticity, 4th ed. New York: Dover; 1944.
32. Brooks BR, Janežič, D, Karplus M. Harmonic analysis of large systems. I. Methodology. J Comp Chem 1995;16:1522–1542.
33. Franke R. Scattered data interpolation: tests of some methods. Math Comput 1982;38:181–200.
34. Stead SE. Estimation of gradients from scattered data. Rocky Mountain J Math 1984;14:265–279.
35. Hardy RL. Multiquadric equations of topography and other irregular surfaces. J Geophys Res 1971;76:1905–1915.
36. Reed M, Simon B. Functional analysis. San Diego: Academic Press; 1980.