BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 20 2011, pages 2843–2850
doi:10.1093/bioinformatics/btr497

Structural bioinformatics

Advance Access publication August 27, 2011

iMod: multipurpose normal mode analysis in internal coordinates

José Ramón Lopéz-Blanco[1], José Ignacio Garzón[2] and Pablo Chacón[1,*]

[1]Department of Biological Chemical Physics, Rocasolano Physical Chemistry Institute, CSIC, Serrano 119, Madrid 28006, Spain and [2]Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Howard Hughes Medical Institute, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Dynamic simulations of systems with biologically relevant sizes and time scales are critical for understanding macromolecular functioning. Coarse-grained representations combined with normal mode analysis (NMA) have been established as an alternative to atomistic simulations. The versatility and efficiency of current approaches normally based on Cartesian coordinates can be greatly enhanced with internal coordinates (IC).

Results: Here, we present a new versatile tool chest to explore conformational flexibility of both protein and nucleic acid structures using NMA in IC. Consideration of dihedral angles as variables reduces the computational cost and non-physical distortions of classical Cartesian NMA methods. Our proposed framework operates at different coarse-grained levels and offers an efficient framework to conduct NMA-based conformational studies, including standard vibrational analysis, Monte-Carlo simulations or pathway exploration. Examples of these approaches are shown to demonstrate its applicability, robustness and efficiency.

Contact: pablo@chaconlab.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 30, 2011; revised on July 30, 2011; accepted on August 22, 2011

1 INTRODUCTION

Dynamic simulations of large molecules long enough to observe functional changes are challenging. Normal mode analysis (NMA) merged with coarse-grained (CG) models has proven to be a powerful and popular alternative to simulate collective motions of macromolecular complexes at extended timescales (Bahar and Rader, 2005; Bahar *et al.*, 2010; Cui and Bahar, 2007; Ma, 2005; Skjaerven *et al.*, 2009; Tama and Brooks, 2006). The CG-NMA application range includes: the prediction of biologically relevant motions of proteins and nucleic acids (Hinsen *et al.*, 1999; Krebs *et al.*, 2002; Tama and Sanejouand, 2001), even with low-resolution structures (Chacon *et al.*, 2003); X-ray refinement (Delarue and Dumas, 2004; Kidera *et al.*, 1992; Lindahl *et al.*, 2006); protein and ligand docking (Cavasotto *et al.*, 2005; Zacharias, 2010); flexible fitting of atomic structures into electron microscopy density maps (Delarue and Dumas, 2004; Hinsen *et al.*, 2010; Tama *et al.*, 2004); efficient generation of conformational pathways (Franklin *et al.*, 2007; Kim *et al.*, 2002; Miyashita *et al.*, 2003); and identification of conserved dynamic patterns within protein families (Leo-Macias *et al.*, 2005).

NMA describes the relevant collective motions based on the harmonic approximation around a local minimum, which allows for solving the Lagrangian equations of motion by diagonalizing the Hessian and kinetic energy matrices. The resulting eigenvectors are a set of orthogonal displacements or normal modes. The high-frequency modes represent localized displacements, whereas low-energy modes correspond to collective conformational changes. These collective motions are closely related to functional motions (Krebs *et al.*, 2002; Yang *et al.*, 2007) and they have been correlated with essential motions extracted from molecular simulations (Ahmed *et al.*, 2010; Rueda *et al.*, 2007). These results and many others validate the use of NMA and CG modeling to describe molecular flexibility, serving as a powerful alternative to costly atomistic simulations.

For relatively large systems, the main bottleneck of NMA is the diagonalization step that can easily go beyond standard computers. The vast majority of current NMA approaches have adopted Cartesian coordinates (CC) as variables. However, internal coordinate (IC) method requires at least one-third less degrees of freedom (DoF) and hence reduces both computational time and memory usage. Moreover, in CC the covalent bonding geometry is not implicitly preserved, thus allowing potential non-physical geometrical distortions. The mathematical simplicity (e.g. kinetic energy matrix is reduced to the identity) and its straightforward implementation are behind the CC preference. Nevertheless, several exceptions have shown the IC potential rewards. Early work by Go and others (Go *et al.*, 1983; Levitt *et al.*, 1985; Noguti and Go, 1983a) led to the development of a complete mathematical framework using dihedral angles. This full atomic approximation has been extended to include even bond stretching and angle bending (Kamiya *et al.*, 2003). ProMode is a full-atom NMA repository where users can only explore pre-computed results of >3000 proteins (Wako *et al.*, 2004). The integration of CG approximations with IC also has been successfully explored. Pioneering work by Tirion with elastic network models in IC (Tirion, 1996) has been continued by others (Kovacs *et al.*, 2005; Lu *et al.*, 2006; Mendez and Bastolla, 2010). However, the scope of these approaches is mainly theoretical and none of them are accessible as functional tool.

Here, we present a new multipurpose tool chest, named iMod, to exploit the benefits of classical NMA formulations in IC while extending them to cover multiscale modeling. Ordinary torsion angles are maintained as variables, whereas different graining levels were incorporated to represent protein structure (e.g. with only α carbons). These atomic models can be easily combined

*To whom correspondence should be addressed.

with several elastic networks in a highly customizable framework, including user-defined potential or the immobilization of parts of the macromolecular system by removing their dihedral angle variables. iMod has been designed to be versatile and can handle also multiple chains, nucleic acids and rigid ligands. The versatility and efficiency of this new integrative framework expand the applicability range of NMA especially to very large systems. Here, we show its robustness describing several representative conformational transitions. We also illustrate its applicability in two simulation contexts: conformational sampling and pathway trajectory generation.

2 METHODS

Here, we outline the basic mathematical framework for performing NMA. Briefly, macromolecular motion can be described as a combination of normal modes determined by solving the following generalized eigenvalue problem (Noguti and Go, 1983a):

$$\boldsymbol{HU}=\lambda_k\boldsymbol{TU} \text{ where } \boldsymbol{H}=\frac{\partial^2 V}{\partial q_\alpha \partial q_\beta} \text{ and } \boldsymbol{U}=(\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_N), \quad (1)$$

where λ_k is the eigenvalue associated with the k-th normal mode $\mathbf{u}_k$; α and β are the IC indices; and $\boldsymbol{H}$ is the Hessian matrix. The eigenvalues are related to the frequencies, ν_k, as $\lambda_k=(2\pi\nu_k)^2$. The potential energy expressed in N ICs, q, is given by

$$V=\frac{1}{2}\boldsymbol{qHq}^T \text{ with } \boldsymbol{q}=(q_\alpha-q_\alpha^0) \text{ and } \alpha=1,\ldots,N \quad (2)$$

being $\boldsymbol{q}$ is the coordinate's displacement from the equilibrium conformation at a given energy minimum, q^0. In a similar way, the kinetic energy can be expressed as follows:

$$T=\frac{1}{2}\dot{\boldsymbol{q}}\boldsymbol{T}\dot{\boldsymbol{q}}^T \text{ where } \boldsymbol{T}_{\alpha\beta}=\sum_i m_i\cdot\frac{\partial \boldsymbol{r}_i}{\partial q_\alpha}\cdot\frac{\partial r_i}{\partial q_\beta} \quad (3)$$

The mass of the atom i is m_i and $\boldsymbol{r}_i$ the corresponding CC.

The diagonalization of Lagrange's Equation (1) yields solutions of the form:

$$\boldsymbol{q}_=\boldsymbol{q}^0+\sum_{k=1}^{N}a_k\mathbf{u}_k\cos(2\pi\nu_k\cdot t+\delta_k) \quad (4)$$

Where a_n and δ_n depend on the initial conditions and ν_n is the angular frequency associated at each normal mode. The direct calculation of $\boldsymbol{T}$ scales to $O(N^3)$, whereas $\boldsymbol{H}$ even reaches $O(N^4)$. This computational burden can be significantly reduced in both cases to $O(N^2)$ by employing recursion relationships (Braun *et al.*, 1984; Noguti and Go, 1983b). Then, the $O(N^3)$ diagonalization step performed with LAPACK subroutines (Anderson *et al.*, 1999) becomes the main computational bottleneck.

The ICs are defined by the canonical backbone dihedral angles, i.e. ϕ and ψ in proteins and α, β, γ, ε and ζ in nucleic acids. By default, side chains, sugars and bases are considered to be rigid bodies but optionally the dihedral angle χ can also be included. To avoid ring closure problems, ϕ is fixed for prolines and δ in nucleic acids. The first ϕ angles and the last ψ of the chains are also not considered. The remaining dihedral angles and all covalent bond lengths and angles are assumed to be fixed, thereby preserving the underlying covalent structure. To account for multiple chains, the corresponding six rigid body variables are added to describe their relative motion. Moreover, any subset of the above-described internal variables can be fixed to allow arbitrary definition of the rigid parts of the system. This technique is very useful to reduce the computational cost of large systems or to prevent flexibility in known rigid regions. Although this is the most direct way to reduce the number of variables, CG can be done at many other levels. Three different representations can be selected:

- HA: considers all heavy atoms.
- C5: each residue is represented by five pseudo-atoms—three for the backbone (NH, C_α, CO) and two for the side chain (C_β and virtual mass located at the mass center of the remaining side chain atoms) (Cavasotto *et al.*, 2005; Kovacs *et al.*, 2005).
- C_α: a single C_α atom per amino acid (Lu *et al.*, 2006; Mendez and Bastolla, 2010). In this case, the backbone carbonylic carbon and nitrogen atoms are only considered to define the dihedral angles.

Only the heavy atom representation is currently available for nucleic acids. Note that in all cases the backbone covalent structure is always maintained.

Independently of the atomic model used, the non-bonded atoms (or pseudo-atoms) are interconnected by harmonic springs. The potential energy can be formulated as follows:

$$V=\sum_{i<j}F_{ij}(r_{ij}-r_{ij}^0)^2+s\sum_\alpha(\theta_\alpha-\theta_\alpha^0)^2 \quad (5)$$

The first term describes the atom pairwise part of the harmonic potential, where r_{ij} is the distance between atoms i and j, the super-index 0 indicates the initial equilibrium conformation and F_{ij} represents the spring stiffness matrix whose elements describe the force constant associated with each atom pair. This generic function is also a customizable element in our implementation. Users can choose between a basic cut-off function (Tirion, 1996), exponential-like functions (Hinsen *et al.*, 1999; Rueda *et al.*, 2007) and an essential dynamics (ED) refined potential (Orellana *et al.*, 2010) or even define their own stiffness matrix. By default, we used the following sigmoid function:

$$F_{ij}=\frac{k}{\left(1+\left(\frac{r_{ij}^0}{r_0}\right)^p\right)}, \text{ if } r_{ij}^0<r_{cut}, \text{ otherwise } F_{ij}=0. \quad (6)$$

In this equation, k gives the maximum stiffness, r_0 represents the inflexion point, p denotes the sigmoid shape and r_{cut} is a convenient cut-off for removing ineffective very weak springs from calculations. The parameters k, r_0, p and r_{cut} were set to 1, 3.8 Å, 6 and 10 Å, respectively, to obtain the same behavior of the exponential function used in Rueda *et al.* (2007). We found this parameterization quite robust with all models used.

The second term of (5) is an extratorsional stiffness, s, which is related to each dihedral angle, θ_α. This term prevents the so-called tip effect, i.e. the presence of irrational low frequencies typically caused by floppy small regions [for more details, see Lu and coworkers (Lu *et al.*, 2006)].

Simulation applications: low-frequency IC modal space was effectively used in two applications: iMorph and iMC (see flowcharts in Supplementary Fig. S1). The iMC tool performs a Monte-Carlo (MC) sampling to get a trajectory using the 5–10 lowest frequency modes. In each step, a new trial displacement is obtained by randomly selecting a mode and its amplitude. Such displacement is accepted according Metropolis criteria, with a probability defined by the minimum of $(1, e^{-\Delta E/kT})$, where ΔE is the difference between the harmonic energy of the new and old conformations. To prevent low acceptance rates and improve the sampling efficiency, the mode amplitude was balanced following the scaled collective variable MC method (Yamashita *et al.*, 2001). After 1000 MC steps, a new conformation is generated by applying the resulting modal displacement to the initial structure. The whole MC protocol is repeated several times to generate a structural ensemble. In this study, iMC was used for generating ensembles of 1000 conformations around a set of known apo structures at 300 K. In each case, a stiffness factor was adjusted to yield structures with average RMSD around the 60% of the motion amplitude between holo and apo structures. This procedure also avoided large distortions of the initial structure.

In iMorph, the collective deformation directions of the modes are used for simulating feasible transitions between two known conformations. This iterative process starts by calculating only the 10% lowest frequency part from the initial structure. After collectively scaling as before, 10% of these modes are selected and merged. The modes selection is inversely proportional to their eigenvalues and the merging is done with random amplitudes. Thus, the resulting displacement includes deformational directions of the first

modes but with random relative contributions. Then a new conformation is generated by moving the structure along the directions encoded into the merged displacement. Such conformation is accepted only if its RMSD is closer to the target structure; otherwise, new modes are selected and combined. If the new structure diverges 0.1 Å from the last structure used for the NMA, then the modes are computed again. This procedure is repeated until the flexed structure converges to the target. Alternatively, during the pathway generation part of dihedral angles can be fixed. To this end, every time the NMA is computed new subsets of dihedral angles variables are randomly removed from the calculations.

Test datasets: we selected 23 different open/closed protein pairs with displacements >2 Å RMSD from the molecular motions database (Flores *et al.*, 2006). These examples have different sizes (186–994 amino acids) and correspond to a wide variety of macromolecular motions. The structures are relatively large and conform to the quality scores of the standard structure validation program Molprobity (Chen *et al.*, 2010). Hydrogen atoms were added to each conformation before checking its structural integrity. Although hinge motions are predominant, the test set also includes shear and other complex protein motions. The average displacement was 7.57±4.39 Å. To complement our test with nucleic acid motions, we screened the PDB for RNA collective conformational changes. We selected 11 conformational pairs with motion displacements >3 Å that passed the Molprobity criteria. The average displacement was 7.23±2.37 Å. The size of the RNAs is small (40–126 nt) except for two large rRNAs formed by 721 and 1529 nt. The complete list of protein and RNA test cases is detailed in Supplementary Table S1.

Technical details: all calculations were performed using a Linux box with an Intel® Core™2 Quad Q6600 processor with 4 GB of RAM. The CC-NMA was preformed with an updated version of DFprot (Garzon *et al.*, 2007). All the tools and databases presented here, including full documentation and tutorials, are available at http://chaconlab.org/imod/index.html,

3 RESULTS

In this section, we illustrate the use of low-frequency modal spectra by our IC CG-NMA approximations to describe protein flexibility.

3.1 Vibrational analysis

The most straightforward application for our method is the computation and exploration of normal modes to identify potential functional motions. The collective character of such motions can be captured by a few low-frequency normal modes. Within our new tool chest, iMode can compute the vibrational modes from multiple chains of proteins or nucleic acids, even supporting rigid ligands. To animate the resulting soft modes, the iMove tool generates a trajectory file that can be visualized with standard molecular viewers. In addition to visual inspection, other parameters such as B-factors and deformabilities (Garzon *et al.*, 2007) can also be obtained. The users can easily adapt the CG level to their needs by selecting an atomic model resolution from a heavy atom representation to a simple C_α model. Moreover, the dihedral variables can be freely removed, thereby allowing for freezing parts of the macromolecular system. Our NMA implementation is very fast. For example, the biggest protein test case (ATPase pump, 994 residues), takes 11 s to complete the analysis at maximal model resolution. The high versatility and efficiency permit NMA of very large structures in commodity hardware. For example, the Cowpea Chlorotic Mottle Virus (CCMV) NMA can be computed in just 1 h on a standard 4 GB RAM Linux box by fixing 75% of the dihedral angles. This calculation required ∼15 000 internal variables. Two representative characteristic modes of this NMA are shown as an animation in the Supplementary Material. The first mode has been proposed to explain the maturation pathway from the native to the swollen state of the CCMV virus (Tama and Brooks, 2002). The second animation corresponds to the lowest energy mode of the icosahedral symmetry group (van Vlijmen and Karplus, 2005).

Table 1. Protein conformational transitions overlaps

	α_1^a	α_2	α_3	δ_3^b	δ_5	δ_{10}	$N\alpha_1^c$	$N\sigma_{90\%}^d$	γ_3^e	γ_{10}	γ_{50}
Average IC[f]	0.70	0.34	0.25	0.78	0.84	0.89	1.7	107	0.98	0.93	0.90
Open to closed	0.77	0.30	0.23	0.86	0.89	0.92	1.3	90	0.98	0.94	0.90
Closed to open	0.63	0.38	0.28	0.71	0.80	0.86	2.2	125	0.97	0.93	0.89
Average CC	0.70	0.34	0.24	0.77	0.83	0.89	1.8	155	1.00	1.00	1.00

[a]Overlap between the transition vectors Δr and the first, second and third most overlapping modes (v_n). Calculated as: $\alpha_{1,2,3} = |\Delta r \cdot v_n| / [\|\Delta r\| \|v_n\|]$.
[b]The cumulative overlap contribution of the 3, 5 and 10 lowest energy modes was computed from: $\delta_n = (\sum_{k=1}^{n} \alpha_k^2)^{1/2}$.
[c]Rank of the best overlapping normal mode.
[d]Number of modes required to cover the 90% of the modal variance.
[e]Overlaps between deformation spaces **u** and **v** were computed using (Noy *et al.*, 2006): $\gamma_n = {}^1/_n \sum_{i,j}^{n} (u_i \cdot v_j)^2$ where n is the number of eigenvectors considered.
[f]NMA in IC and CC was performed with iMode and DFprot (Garzon *et al.*, 2007), respectively. In both cases, C_α model was used with the EDs refined potential (Orellana *et al.*, 2010). The averages include all protein transitions detailed in Supplementary Table S1.

3.2 Conformational transitions

As a validation test, we contrasted the overlaps between the 23 motions observed in our transition dataset with the modes obtained from each conformer. To compare with the CC approach, we restricted this study to the C_α model using the potential optimized against a set of representative MD trajectories (Orellana *et al.*, 2010). The overlap has been measured with a normalized dot product between the vector calculated from the two crystallographic conformers and the modal displacements (Table 1). On average, the best overlapping mode yielded a value of 0.70, indicating excellent agreement with the experimental transition vectors. Moreover, these best modes usually corresponded to the first or second low-energy modes. The cumulative overlap of the first three modes, δ_3, was 0.77. If we included up to the 5th or 10th modes, the scores increased to 0.83 and 0.90, respectively. In other words, only 10 modes were needed to account for 90% of the observed change. These results are in agreement with the fact that the low-frequency modes correlate very well with the biologically relevant motions. Similar overlaps within the macromolecular motion database have been found (Krebs *et al.*, 2002). Remarkably, lower values have been observed when the NMA is performed from the closed conformation. The best mode overlap, α_1, dropped from 0.77 to 0.63, and the overlap δ_{10} was reduced ∼7% when the closed conformation was used. It is well known that NMA performs better from open forms (Tama and Sanejouand, 2001). Nevertheless, the δ overlap values from close conformations were still high, and 5 modes accounted for 80% of the motion. Overlaps for individual cases are found in the Supplementary Table S2.

No major differences occurred between computing the NMA in internal or in CC. The overlaps and hence the first modes are almost identical (Table 1). The deformation spaces determined by the two methods overlapped considerably with γ scores >0.9, indicating that CC low-frequency modes correspond mainly to dihedral angle

Table 2. Protein and RNA conformational transitions overlaps

Model[a]	α_1	α_2	α_3	δ_3	δ_5	δ_{10}	Nα_1	N$\sigma_{90\%}$	γ_3^b	γ_5	γ_{10}
C_α	0.70	0.34	0.23	0.77	0.83	0.88	1.7	118	1.00	1.00	1.00
C5	0.68	0.33	0.23	0.75	0.81	0.86	1.7	226	0.94	0.89	0.87
HA	0.70	0.33	0.22	0.76	0.82	0.87	1.8	368	0.91	0.87	0.85
C_α-50%	0.69	0.37	0.24	0.78	0.84	0.89	1.6	85	0.98	0.95	0.94
C_α-90%	0.63	0.40	0.25	0.74	0.80	0.88	2.0	30	0.75	0.69	0.66
RNA	0.66	0.38	0.25	0.77	0.82	0.86	1.6	45	–	–	–

[a]All the calculations were performed as in Table 1 but using the default sigmoid potential (see Section 2). CG protein models: C_α, a C_α atom per residue; C5, 3 atoms for backbone and 2 for the side chain; HA, considering all heavy atoms; C_α-50%, as C_α but fixing randomly 50% of the dihedral angles; C_α-90%, fixing 90%; and for RNA only the heavy atoms model was used.

[b]The γ overlaps were restricted to C_α atoms. Compatible eigenvectors for C5 and HA cases have been obtained by diagonalizing the corresponding C_α covariance matrices.

motions (Kitao *et al.*, 1994). However, we detected differences in the number of modes needed to account for 90% of the variance. In this case, the IC required ∼30% fewer modes to express the same variance ratio. These results suggest that the conformational space described by modes computed in IC is more compact.

iMode is faster and consumes much less memory than classical CC approaches, especially when molecular size is big enough. For example, the NMA using the C_α model for acetyl-CoA synthetase (1oao, 728 residues) with CC takes twice the time of IC (6.6 s versus 3.7 s). The comparison with C_α model is the most favorable for the CC approach because three DoF per C_α are needed, whereas IC approximation required only two dihedrals (i.e. only one-third less). The profit becomes more apparent as more pseudo-atoms are considered in the atomic model. A heavy atom representation for the same synthetase demanded 43 min for CC (5739 atoms, 17217 DoF) and just 4.7 s for iMode (1412 DoF). Furthermore, the CC memory requirements become a bottleneck with relatively large proteins. The biggest protein of our validation test (ATPase pump, 994 residues, ∼23 000 DoF in CC) cannot be computed with HA representation without exceeding the 2 GB memory allocation limit of 32-bit PCs. In contrast, our approach took only 11 s for this case (1942 DoF). In principle, iMode can handle proteins ∼15 000 DoF (∼7500 residues) in a 32-bit machine.

Table 2 shows the average results obtained at different CG levels. For comparative reasons, we employed the default sigmoid potential described in Section 2. No major differences were observed, as the motion was already captured by the first two or three modes ($N\alpha_1 < 2$, $\delta_3 > 0.75$). These modes depend mainly on the molecular shape, which is the same for all of the CG levels used. Although the first modes are equal, the deformation spaces differ from each other specially at higher frequencies. Taking C_α as a reference, the γ scores become smaller as a larger number of eigenvectors are considered or as more detailed representations are used (Table 2). For example, C_α and HA had γ overlaps of 0.91, 0.87 and 0.85 for deformation spaces comprising the 3, 5 and 10 lowest energy eigenvectors, showing a clear divergence as more modes were taken into account. Inclusion of more atoms in the representation also affected the overlaps, e.g. C5 and HA yielded gradually lower values.

The overlap results of the 22 RNA test cases were in agreement with our results from proteins (Table 2). The collective RNA transitions also correlated very well with the low modal subspace. The best overlapping mode scores were similar to those of the proteins using the same potential and HA model. The high overlap values (e.g. $\delta_{10} > 0.85$) reflect the good description for RNA motion provided by the low-frequency modes. The size of the RNA cases was smaller than the protein dataset, so the number of modes that were needed to account for 90% variance was reduced to 45. As for the proteins, the memory limitations and computational cost were strongly reduced using IC. For example, the 7S RNA case took 1 s with IC (1mfq, 126 nt, 628 DoF), whereas CC (2708 atoms, 8124 DoF) took 5 min. The two biggest rRNA cases were also beyond the scope of CC using the HA model with a 32-bit machine. In contrast, the biggest case in IC (3e1a, 1529 nt, 32811 atoms, 7651 DoF) took ∼9 min. On a case-by-case basis, the conformational RNA transitions results are shown in Supplementary Table S3.

Notably, we observed that the space described by the low-frequency modes was quite robust when we randomly removed different percentages of dihedral variables. The first modes in proteins were essentially the same (γ close to 1 and very similar α and δ values) for fixing percentages <50%. The equivalence and the overlaps with the motions decayed more drastically as more dihedral angles are fixed. By removing 90% of the DoF, part of the conformational transition was kept in the most overlapping mode ($\alpha_1 = 0.63$), but the deformational space clearly diverged ($\gamma_{10} = 0.66$) from the complete C_α model case. Similar behavior was observed for more detailed models and for RNA (data not shown). As discussed below, removing a fraction of the dihedral angles from the NMA calculations is an effective CG approach.

3.3 Conformational pathways

Simulating the structural transition between two known conformations of a macromolecule has been successfully performed by CG-NMA (Franklin *et al.*, 2007; Krebs and Gerstein, 2000; Lindahl *et al.*, 2006; Miyashita *et al.*, 2003). Here, we tested a simple approach, named iMorph, to generate plausible trajectories using the space encoded into low-energy IC modes. In an iterative process that only uses the 10% lowest frequency modes; the initial structure is flexed gradually toward the target by minimization of their relative RMSD. In all of the 46 protein test cases, the initial structure converged fast and smoothly to the target as illustrated in Figure 1 with a representative case. On average, the target and closest conformations were <1 Å from an initial deviation of 7.57 Å (Table 3). As before, a very small difference with the morphing direction has been observed. In the case of HA (Supplementary Table S4), the average deviation was 0.74 Å from open to closed and 0.80 Å from closed to open.

The quality of the pathway structures was checked with Molprobity (Chen *et al.*, 2010). Final conformations preserved the original crystallographic quality with scores expected for resolutions close to 3 Å from an initial value of 2.7 (Table 3). Few more clashes and almost no extra Ramachandran outliers were observed compared with initial structures. This structural quality was also preserved along the simulation trajectory (6–9 more clashes) and only the final part models usually had more collisions (14–27). See a typical variation of the clashes on Figure 2A. Considering that we did not use any minimization strategy, the absolute number of collisions was still very small. These few clashes are mainly caused by the added hydrogens of the lateral chains. In contrast, other CC CGNMA approaches have incorporated regularization steps to avoid

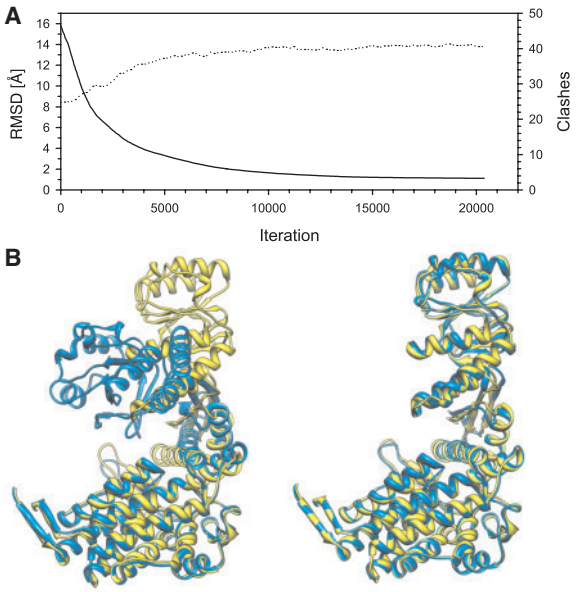

Fig. 1. GroEL conformational pathway. (**A**) Variation of the RMSD (solid line) and Molprobity clashes (dotted line) along pathway from closed (1oel, cyan) to open (1sx4, yellow) conformation. (**B**) Structural superposition of the initial (left) and final (right) conformers with the open target structure. The corresponding pathway animation is available in Supplementary Material. The final deviation is only 1.12 Å with 41 clashes using the HA model.

Table 3. Conformational pathways results

	RMSD [Å][a]		Clash[b]			%r$^c_{out}$			Molprobity[d]			t[e]
Model[f]	*I*	*F*	*I*	*A*	*F*	*I*	*A*	*F*	*I*	*A*	*F*	
C_α	7.57	0.79	21	36	48	0.6	0.6	0.7	2.7	2.9	3.1	5.1
C_α-50%	7.57	0.86	21	34	45	0.6	0.6	0.6	2.7	2.9	3.1	1.2
C_α-90%	7.57	1.06	21	37	53	0.6	0.6	0.6	2.7	2.9	3.1	0.6
C5	7.57	0.75	21	29	37	0.6	0.6	0.8	2.7	2.9	3.0	6.1
HA	7.57	0.77	21	27	35	0.6	0.7	0.8	2.7	2.8	2.9	6.9
HA-50%	7.57	0.80	21	27	35	0.6	0.6	0.7	2.7	2.8	2.9	2.4
HA-90%	7.57	1.06	21	33	50	0.6	0.6	0.6	2.7	2.9	3.1	2.1
RNA	7.23	1.34	32	33	37	–	–	–	–	–	–	13.0[g]
RNA-50%	7.23	1.39	32	33	37	–	–	–	–	–	–	5.1[g]
RNA-90%	7.23	1.58	32	41	54	–	–	–	–	–	–	3.9[g]

[a]RMSD computed with C_α or P atoms.
[b–d]Scores computed with Molprobity (Chen *et al.*, 2010).
[b]Number of serious clashes per 1000 atoms.
[c]Percentage of Ramachandran backbone ϕ and ψ angles outside the allowed region.
[d]Log-weighted combination of scores which reflects the crystallographic resolution at which those values would be expected.
[e]Mean running time in minutes.
[f]Average values conformational pathway test cases obtained with iMorph using a sigmoid potential. Values shown for initial (*I*), final (*F*) and average conformations (*A*). CG representations as indicated in Table2.
[g]For the two biggest RNA cases, the divergence limit to recompute NMA was changed from 0.1 Å to 1 Å to speed up the process. For detailed timings, see Supplementary Table S5.

improper structures (Yang and Sharp, 2009). Similar results were observed with RNA including excellent geometry maintenance. The final mean deviation was higher than in proteins, 1.34 Å, but such difference is irrelevant considering the different nature of data. In any case, the results validate the use of low-frequency modal space to generate feasible pathways also for the RNA transitions (see Supplementary Table S5 for detailed results).

The test was repeated with different CG models obtaining similar pathways and scores. We only detected fewer clashes with finer grained models, which could be a positive effect of considering more detailed representations. Surprisingly, when randomly removing the 50% of the dihedral angles in each NMA calculation, the final conformations and scores are the same as considering all the Dofs. Even more, when freezing 90% we still had reasonable results with RMSDs close to 1 Å but with more clashes. Analogous results have been obtained with other CG models and with RNA (Table 3). Note that this removal is not permanent, the subset of dihedrals to be fixed is randomly selected every NMA calculation (∼100 times during the trajectory). Thus, all dihedrals had a chance to move preventing critical levels of stiffness. Nonetheless, clear differences were observed for calculation times. The mean time to obtain a trajectory using the C_α model was only 5.1 min, whereas employing C5 or HA required 20 and 35% more time, respectively. Although computation of the plausible transitions was quick, 4- and 8-fold speed ups have been obtained by fixing 50 and 90% of dihedral angles, respectively. In addition to the efficiency gains of this CG procedure, the squared reduction of memory cost with the fixed DoFs greatly extends the applicability to larger systems.

iMorph is a proof of concept to test the sampling power of IC-NMA and to check the structure maintenance in an iterative modeling process. Even at this point of development, it already has advantages over C_α CG-NMA approximations: extended efficiency, larger size coverage and covalent structure preservation. Nevertheless, there are more sophisticated and powerful alternatives for fast motion planning including coarse-graining dynamics (Weiss and Levitt, 2009), Monte Carlo (Borrelli *et al.*, 2005) or path planning methods (Barbe *et al.*, 2011). Despite these methods being more realistic, there is room for improving their ability to describe large collective changes by incorporating IC-NMA.

3.4 Conformational sampling

We include a tool (iMC) to explore the low-frequency essential space of a given structure by activating its first modes according to Metropolis criterion. This procedure generates variable modal displacements that can be applied to the structure for producing a pseudo-trajectory. A sampling exercise for generating conformational ensembles around known structures is presented in Table 4 (and Supplementary Table S6). In this case, we employed a dataset of 10 proteins that undergo domain closure upon ligand binding. This dataset has been used recently for testing a protocol to predict holo structures from apo conformations (Seeliger and de Groot, 2010). Our intention was not to reproduce this sophisticated protocol, which includes biased conformational sampling, docking and molecular dynamics but rather to illustrate the sampling power of our iMC approximation. For each test case, an ensemble of 1000 conformations was generated based on the apo structure using only the first 5 low-frequency modes. Note that few modes encoded the majority of the conformational change ($\delta_5 > 0.90$). The closest

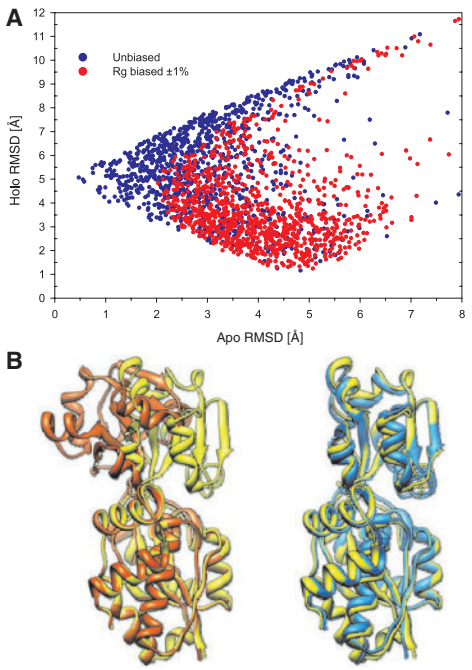

Fig. 2. Conformational sampling comparison of a biased (red circles) and an unbiased (blue circles) ensemble generation performed with iMC from the apo structure of the osmo-protection protein (**A**). The overlay of apo (orange, PDB 1sw5) and holo (yellow, PDB 1sw2) structures is shown (**B**, left). On the right, the holo structure was superimposed with the closest conformation (cyan) found in the biased ensemble.

Table 4. IC based Monte-Carlo conformational sampling

		Unbiased[a]			Biased Rg ±1%[b]			Clashes[c]			%r_{out}			Molprobity		
RMSD[d]	δ_5^e	Min[f]	<1.5 Å[g]	<2 Å[h]	Min	<1.5 Å	<2 Å	*I*	*A*	*B*	*I*	*A*	*B*	*I*	*A*	*B*
4.04	0.94	1.28	1.6	6.6	1.08	16.2	33.7	12	40	45	0.3	0.3	0.3	2.0	2.5	2.5

[a]Unbiased ensembles were generated using the default parameters of iMC from apo conformations with C_α model and ED refined potential.
[b]Constraining the sampling to obtain conformers deviated ±1% from the holo Rg.
[c]Molprobity scores of the unbiased sampling for: (*I*) initial structure; (*A*) average of 1000 conformers and (B) average of conformers <2 Å C_α-RMSD from the holo structure. Similar values have been obtained for the biased sampling (data not shown).
[d]Motion amplitude (C_α-RMSD) between apo and holo conformations.
[e]Cumulative overlap of the five lowest energy modes as defined Table 1.
[f]Minimum C_α-RMSD to the target holo structure found in the ensemble.
[g–h]Percentage of models with a C_α-RMSD from holo structure <1.5 or 2 Å.

conformation within the ensemble was 1.28 Å apart on average to the bound conformation. In addition, 1.6 and 6.6% of the conformations were <1.5 and 2.0 Å, respectively. These results are interesting, especially taking into account that the initial apo structures were deviated 4.04 Å from the holo structures. Imposing constraints can enrich the sampling. For example, if the sampling is biased toward structures within ±1% of the holo radius of gyration (Rg), 34% of the 1000 conformations are <2 Å and close to 16% are <1.5 Å. These sampling differences are depicted in Figure 2 for an illustrative case. Figure 2A shows how the Rg constraint biases the RMSD toward the holo structure. The biased ensemble (red) samples a subset of the unbiased conformational space (blue) that was much closer to the holo structure. As before, the geometric quality is reasonably maintained with only 40 clashes on average. Also, few seconds were needed to generate an unbiased ensemble and <4 min for the biased case. This small cost and the quality of the ensemble conformers already suggest their use as starting points for other simulation protocols.

4 DISCUSSION

We presented an efficient NMA tool for both protein and nucleic acid structures that considers the canonical dihedral angles as variables. The implicit maintenance of the covalent structure preserves the model geometry and minimizes the potential distortions of CC-NMA approaches. The robustness of the proposed approaches for modeling flexibility has been tested in diverse contexts. We showed how the low-frequency modes computed with iMode are well correlated with the protein collective transitions observed between different known conformers. Notably, similar correlations have been obtained with RNA transitions, corroborating the usefulness of NMA for estimating collective dynamics. Although there are several reports of CG-NMA and nucleic acids (Feig and Burton, 2010; Fulle and Gohlke, 2010; Orozco *et al.*, 2008; Skjaerven *et al.*, 2011; Yang *et al.*, 2006), to the best of our knowledge, this is the first time that a validation of IC-NMA with a representative set of RNA transitions has been performed. We have proven the utility of our approximations for generating plausible conformational pathways between protein and RNA transitions or for producing ensembles from protein apo structures using only the first low-frequency modes. Our results point out the sampling power of NMA to provide reasonable and rather inexpensive direct view of the relevant conformational space even at different CG levels. Simplified models will be especially useful to expand the conformational search capabilities to larger macromolecular systems in commodity hardware. Proteins up to 7500 residues (or nucleic acids ~3000 nt) can be analyzed with iMod in a 32-bit PC. Furthermore, in 64-bit machines the size of the biological system is only limited by the available RAM. For example, the NMA of the 3.2 mega-Dalton CCMV capsid (28 620 residues) will require ~25 GB of RAM. The size of any of these systems is out of the scope of CC-based NMA methods that can only approach them with much more aggressive simplifications. An effective CG approximation has been revealed by randomly removing a large fraction of the dihedral angles from NMA calculations. This simple procedure extends even more our application range to huge size systems as CCMV in standard PCs (see Supplementary Material). The overall efficiency gain is also reflected in the computation times. IC approaches are always faster than CC, and the gain grows cubically with the number of pseudo-atoms considered in the representation. Even in the simplest C_α level, the covalent backbone is maintained, which naturally reduces the potential distortions produced when the structures are displaced along the modes with CC-NMA. Structural quality is preserved even in the large iterative processes of generating pathways or ensembles. In addition, maintaining the backbone covalent structure greatly simplifies the process of transferring reduced models to full atomic coordinates, which is a common requirement in modeling and structural refinements.

As any other NMA-based approach, conformational changes far away from native structure or other non-linear dynamics behavior cannot be properly described. Since the major sources of anharmonicity are related to high-frequency side chain dynamics, limited coverage to local motions is expected by the approximations presented here. In these cases, detailed atomics simulations are preferred. Nevertheless, being able to predict the collective intrinsic motions at reduced costs is valuable for both understanding the functional conformational changes and introducing flexibility into the molecular modeling applications, especially for large systems. We are currently working in these directions, and we have just successfully introduced the iMod procedures for the flexible fitting of large macromolecular conformational changes into electron microscopy 3D reconstructions (López-Blanco,J.R. *et al.*, manuscript in preparation). Additional methodological work is being performed to extend the applicability and efficiency of our current approaches using parallelization techniques. Future efforts will be focused on the elastic network optimization of our CG models using atomistic MD simulations (Orellana *et al.*, 2010). Finally, the progress of this type of methods with nucleic acids (Fulle and Gohlke, 2010; Orozco *et al.*, 2008) is also a promising research area.

ACKNOWLEDGEMENTS

We thank J.Kovacs for useful comments and fruitful discussions.

Funding: MICNN BFU2009-09552; Human Frontier Science Program RGP0039/2008.

Conflict of Interest: none declared.

REFERENCES

Ahmed,A. *et al.* (2010) Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. *Proteins Struct. Funct. Bioinformatics*, **78**, 3341–3352.

Anderson,E. *et al.* (1999) *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics. Philadelphia, PA.

Bahar,I. and Rader,A.J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 586–592.

Bahar,I. *et al.* (2010) Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem. Rev.*, **110**, 1463–1497.

Barbe,S. *et al.* (2011) A mixed molecular modeling-robotics approach to investigate lipase large molecular motions. *Proteins*, **79**, 2517–2529.

Borrelli,K.W. *et al.* (2005) PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *J. Chem. Theory Comput.*, **1**, 1304–1311.

Braun,W. *et al.* (1984) Formulation of static and dynamic conformational energy analysis of biopolymer systems consisting of two or more molecules. *J. Phys. Soc. Jpn.*, **53**, 3269–3275.

Cavasotto,C.N. *et al.* (2005) Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.*, **127**, 9632–9640.

Chacon,P. *et al.* (2003) Mega-dalton biomolecular motion captured from electron microscopy reconstructions. *J. Mol. Biol.*, **326**, 485–492.

Chen,V.B. *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 12–21.

Cui,Q. and Bahar,I. (2007) Normal mode analysis theoretical and applications to biological and chemical systems. *Mathematical & Computational Biology*. Chapman and Hall/CRC, Boca Raton.

Delarue,M. and Dumas,P. (2004) On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc. Natl Acad. Sci. USA*, **101**, 6957–6962.

Feig,M. and Burton,Z.F. (2010) RNA polymerase II flexibility during translocation from normal mode analysis. *Proteins Struct. Funct. Bioinformatics*, **78**, 434–446.

Flores,S. *et al.* (2006) The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic Acids Res.*, **34**, D296–301.

Franklin,J. *et al.* (2007) MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acids Res.*, **35**, 477–482.

Fulle,S. and Gohlke,H. (2010) Molecular recognition of RNA: Challenges for modelling interactions and plasticity. *J. Mol. Recogn.*, **23**, 220–231.

Garzon,J.I. *et al.* (2007) DFprot: a webtool for predicting local chain deformability. *Bioinformatics*, **23**, 901–902.

Go,N. *et al.* (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl Acad. Sci. USA*, **80**, 3696–3700.

Hinsen,K. *et al.* (1999) Analysis of domain motions in large proteins. *Proteins*, **34**, 369–382.

Hinsen,K. *et al.* (2010) From electron microscopy maps to atomic structures using normal mode-based fitting. *Methods Mol. Biol.*, **654**, 237–258.

Kamiya,K. *et al.* (2003) Algorithm for normal mode analysis with general internal coordinates. *J. Comput. Chem.*, **24**, 826–841.

Kidera,A. *et al.* (1992) Normal mode refinement: crystallographic refinement of protein dynamic structure. II. Application to human lysozyme. *J. Mol. Biol.*, **225**, 477–486.

Kim,M.K. *et al.* (2002) Efficient generation of feasible pathways for protein conformational transitions. *Biophys. J.*, **83**, 1620–1630.

Kitao,A. *et al.* (1994) Comparison of normal mode analyses on a small globular protein in dihedral angle space and Cartesian coordinate space. *Biophys. Chem.*, **52**, 107–114.

Kovacs,J.A. *et al.* (2005) Conformational sampling of protein flexibility in generalized coordinates: application to ligand docking. *J. Comput. Theor. Nanosci.*, **2**, 354–361.

Krebs,W.G. *et al.* (2002) Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, **48**, 682–695.

Krebs,W.G. and Gerstein,M. (2000) The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res.*, **28**, 1665–1675.

Leo-Macias,A. *et al.* (2005) An analysis of core deformations in protein superfamilies. *Biophys. J.*, **88**, 1291–1299.

Levitt,M. *et al.* (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, **181**, 423–447.

Lindahl,E. *et al.* (2006) NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Res.*, **34**, W52–W56.

Lu,M. *et al.* (2006) A new method for coarse-grained elastic normal-mode analysis. *J. Chem. Theory Comput.*, **2**, 464–471.

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

Mendez,R. and Bastolla,U. (2010) Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. *Phys. Rev. Lett.*, **104**, 228103–228107.

Miyashita,O. *et al.* (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl Acad. Sci. USA*, **100**, 12570–12575.

Noguti,T. and Go,N. (1983a) Dynamics of native globular proteins in terms of dihedral angles. *J. Phys. Soc. Jpn.* , **52**, 3283–3288.

Noguti,T. and Go,N. (1983b) A method of rapid calculation of a second derivative matrix of conformational energy for large molecules. *J. Phys. Soc. Jpn.* , **52**, 3685–3690.

Noy,A. *et al.* (2006) Data mining of molecular dynamics trajectories of nucleic acids. *J. Biomol. Struct. Dyn.*, **23**, 447–456.

Orellana,L. *et al.* (2010) Approaching elastic network models to molecular dynamics flexibility. *J. Chem. Theory Comput.*, **6**, 2910–2923.

Orozco,M. *et al.* (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–193.

Rueda,M. *et al.* (2007) Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**, 565–575.

Seeliger,D. and de Groot,B.L. (2010) Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Comput. Biol.*, **6**, e1000634.

Skjaerven,L. *et al.* (2009) Normal mode analysis for proteins. *J. Mol. Struct.*, **898**, 42–48.

Skjaerven,L. *et al.* (2011) Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins Struct. Funct. Bioinformatics*, **79**, 232–243.

Tama,F. and Brooks,C.L. III (2002) The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.*, **318**, 733–747.

Tama,F. and Brooks,C.L. (2006) Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 115–133.

Tama,F. and Sanejouand,Y.H. (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, **14**, 1–6.

Tama,F. *et al.* (2004) Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.*, **337**, 985–999.

Tirion,M.M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.

van Vlijmen,H.W. and Karplus,M. (2005) Normal mode calculations of icosahedral viruses with full dihedral flexibility by use of molecular symmetry. *J. Mol. Biol.*, **350**, 528–542.

Wako,H. *et al.* (2004) ProMode: a database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics*, **20**, 2035–2043.

Weiss,D.R. and Levitt,M. (2009) Can morphing methods predict intermediate structures? *J. Mol. Biol.*, **385**, 665–674.

Yamashita,H. *et al.* (2001) Sampling efficiency of molecular dynamics and Monte Carlo method in protein simulation. *Chem. Phys. Lett.*, **342**, 382–386.

Yang,Q. and Sharp,K.A. (2009) Building alternate protein structures using the elastic network model. *Proteins Struct. Funct. Bioinformatics*, **74**, 682–700.

Yang,L.W. *et al.* (2006) oGNM: online computation of structural dynamics using the Gaussian Network Model. *Nucleic Acids Res.*, **34**, 24–31.

Yang,L. *et al.* (2007) How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys. J.*, **93**, 920–929.

Zacharias,M. (2010) Accounting for conformational changes during protein-protein docking. *Curr. Opin. Struct. Biol.*, **20**, 180–186.