iMod: multipurpose normal mode analysis in internal coordinates

José Ramón López-Blanco1, José Ignacio Garzón2 and Pablo Chacón1,∗

1Department of Biological Chemical Physics, Rocasolano Physical Chemistry Institute, CSIC, Serrano 119, Madrid 28006, Spain and 2Department 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

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 structures (Chacon et al., 2007; Kim et al., 2009; Krebs et al., 2002; Tama and Sanjogand, 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...
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):

\[ HU = \lambda_T U \quad \text{where} \quad H = \frac{\partial^2 V}{\partial q^2} \quad \text{and} \quad U = (u_1, u_2, \ldots, u_N). \tag{1} \]

where \( \lambda_T \) is the eigenvalue associated with the \( k \)-th normal mode \( u_k \) and \( \beta \) are the IC indices, and \( H \) is the Hessian matrix. The eigenvalues are related to the frequencies, \( \nu_k \), as \( \lambda_T = (2\pi \nu_k)^2 \). The potential energy expressed in \( N \) ICs, \( q_i \), is given by

\[ V = \frac{1}{2} q^T \Phi q + \sum_{i=1}^{N} u_i \]

(2)

being \( q \) 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:

\[ T = \frac{1}{2} \dot{q}^T \Phi \dot{q} \]

(3)

The mass of the atom \( i \) is \( m_i \) and \( r_i \) the corresponding CC.

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

\[ q_i = \sum_{k=1}^{\infty} a_k \cos(2\pi \nu_k t + \phi_k) \tag{4} \]

Where \( a_k \) and \( \phi_k \) depend on the initial conditions and \( \nu_k \) is the angular frequency associated at each normal mode. The direct calculation of \( T \) scales to \( O(N^3) \), whereas \( H \) even reaches \( O(N^4) \). This computational burden can be significantly reduced in both cases to \( O(N^3) \) 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. \( \phi \) and \( \psi \) in proteins and \( \alpha, \beta, \gamma, \zeta, \chi \) in nucleic acids. By default, side chains, sugars and bases are considered to be rigid bodies but optionally the dihedral angle \( \chi \) can also be included. To avoid ring closure problems, \( \phi \) is fixed for prolines and \( \psi \) in nucleic acids. The first \( \phi \) angles and the last \( \psi \) 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.
- HC: each residue is represented by five pseudo-atoms—three for the backbone (NH, CO, and two for the side chain (C\( \beta \) and virtual mass located at the mass center of the remaining side chain atoms) (Cavasotto et al., 2005; Kovacs et al., 2005).
- C\( \alpha \): a single C\( \alpha \) atom per amino acid (Lu et al., 2006; Mendez and Bastolla, 2010). In this case, the backbone carbonic 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 coordinate 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:

\[ V = \sum_{i<j} F_{ij}(r_{ij} - r_{ij}^0)^2 + \sum_{a} (\theta_{a} - \theta_{a}^0)^2 \]

(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 (Tiezno, 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:

\[ F_{ij} = \frac{k}{1 + a (\Delta^2 / \sigma^2)} \quad \text{if} \quad \Delta^2 < r_{ij}^0 - \text{cut}, \quad \text{otherwise} \quad F_{ij} = 0. \]

(6)

In this equation, \( \Delta \) gives the maximum stiffness, \( r_{ij}^0 \) represents the inflexion point, \( a \) denotes the sigmoid shape and \( \sigma \) a convenient cut-off for removing ineffective very weak springs from calculations. The parameters \( k, r_{ij}^0, p \) and \( r_{ij} \) 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 extraatomic stiffness, \( \kappa \), which is related to each dihedral angle, \( \theta_i \). 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 \( \Delta E \) 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 stiff factor was adjusted to yield structures with average RMSD around the 60% of the motion amplitude between hole 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 structures are discarded. The remaining 90% are used to generate the next iteration. This process is repeated several times to generate a structural ensemble. In this study, iMorph was used for generating ensembles of 1000 conformations around a set of known apo structures at 300 K. In each case, a stiff factor was adjusted to yield structures with average RMSD around the 60% of the motion amplitude between hole and apo structures. This procedure also avoided large distortions of the initial structure.
The most straightforward application for our method is the Table S1. representative characteristic modes of this NMA are shown as an angles. This calculation required on a standard 4 GB RAM Linux box by fixing 75% of the dihedral large structures in commodity hardware. For example, the Cowpea resolution. The high versatility and efficiency permit NMA of very residues), takes 11 s to complete the analysis at maximal model fast. For example, the biggest protein test case (ATPase pump, 994 of the macromolecular system. Our NMA implementation is very variables can be freely removed, thereby allowing for freezing parts 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).

### 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 iMode 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 Cα 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 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™ i7 8 Quad Q6600 processor with 4 GB of RAM. The CC-NMA was performed 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://chalconlab.org/mod/index.html.

<table>
<thead>
<tr>
<th>α1</th>
<th>α2</th>
<th>α3</th>
<th>δ5</th>
<th>δ8</th>
<th>δ15</th>
<th>δ36</th>
<th>δ54</th>
<th>δ72</th>
<th>δ90</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.23</td>
<td>2.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 1. Protein conformational transitions overlaps

<table>
<thead>
<tr>
<th>α1</th>
<th>α2</th>
<th>α3</th>
<th>δ5</th>
<th>δ8</th>
<th>δ15</th>
<th>δ36</th>
<th>δ54</th>
<th>δ72</th>
<th>δ90</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>0.34</td>
<td>0.25</td>
<td>0.78</td>
<td>0.84</td>
<td>0.89</td>
<td>1.7</td>
<td>107</td>
<td>0.98</td>
<td>0.93</td>
</tr>
</tbody>
</table>

### 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 Cα 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, δ3, 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, α1, dropped from 0.77 to 0.63, and the overlap δ10 was reduced ~7% when the closed conformation was used. It is well known that NMA performs better from open forms (Tama and Sanjoseanud, 2001). Nevertheless, the 8 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 low-frequency modes correspond mainly to dihedral angle...
The variance ratio. These results suggest that the conformational space CC approaches, especially when molecular size is big enough. For transitions also correlated very well with the low modal subspace. affected the overlaps, e.g. C5 and HA yielded gradually lower values. Inclusion of more atoms in the representation also eigenvectors, showing a clear divergence as more modes were taken into account. In contrast, our approach took only 11 s for this case (1942 DoF). In the biggest protein of our validation test (ATPase pump, 994 residues, ∼23 000 DoF in CC) cannot be computed with HA representation and just 4.7 s for iMode (1412 DoF). Furthermore, the CC memory requirements become a bottleneck with relatively large proteins. The process that only uses the 10% lowest frequency modes; the initial α model is the most favorable for 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 overlaps of 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 (α = 0.63), but the deformational space clearly diverged (γ = 0.66) from the complete Cα 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 almost no extra Ramachandran outliers were observed in the original crystallographic quality with scores expected for resolutions close to 3 Å from an initial value of 2.7 (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

---

**Table 2.** Protein and RNA conformational transitions overlaps

<table>
<thead>
<tr>
<th>Model</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>Ntot</th>
<th>NCG</th>
<th>γ1</th>
<th>γ2</th>
<th>γ3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cα</td>
<td>0.70</td>
<td>0.34</td>
<td>0.23</td>
<td>0.77</td>
<td>0.83</td>
<td>0.88</td>
<td>1.7</td>
<td>118</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>C5</td>
<td>0.68</td>
<td>0.33</td>
<td>0.23</td>
<td>0.75</td>
<td>0.81</td>
<td>0.86</td>
<td>1.7</td>
<td>226</td>
<td>0.94</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>HA</td>
<td>0.70</td>
<td>0.33</td>
<td>0.22</td>
<td>0.76</td>
<td>0.82</td>
<td>0.87</td>
<td>1.8</td>
<td>368</td>
<td>0.91</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>Cα-50%</td>
<td>0.69</td>
<td>0.37</td>
<td>0.24</td>
<td>0.78</td>
<td>0.84</td>
<td>0.89</td>
<td>1.6</td>
<td>85</td>
<td>0.98</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Cα-90%</td>
<td>0.63</td>
<td>0.40</td>
<td>0.25</td>
<td>0.74</td>
<td>0.80</td>
<td>0.88</td>
<td>2.0</td>
<td>30</td>
<td>0.75</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>RNA</td>
<td>0.66</td>
<td>0.38</td>
<td>0.25</td>
<td>0.77</td>
<td>0.82</td>
<td>0.86</td>
<td>1.6</td>
<td>45</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

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

bThe γ overlaps were restricted to Cα atoms. Compatible eigenvectors for C5 and HA cases have been obtained by diagonalizing the corresponding Cα covariance matrices.
Table 3. Conformational pathways results

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSD [Å]</th>
<th>Clash(^b)</th>
<th>%C(_{\text{all}})</th>
<th>Molprobity(^a)</th>
<th>τ(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>F</td>
<td>I</td>
<td>A</td>
<td>F</td>
</tr>
<tr>
<td>C(_5)</td>
<td>7.57 0.79</td>
<td>21 36 48 0.6 0.6 0.7 2.7 2.9 3.1</td>
<td>5.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C(_5)-50%</td>
<td>7.57 0.86</td>
<td>21 34 45 0.6 0.6 0.6 2.7 2.9 3.1</td>
<td>1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C(_5)-90%</td>
<td>7.57 1.06</td>
<td>21 37 53 0.6 0.6 0.6 2.7 2.9 3.1</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C(_5)</td>
<td>7.57 0.75</td>
<td>21 29 37 0.6 0.6 0.8 2.7 2.9 3.0</td>
<td>6.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td>7.57 0.77</td>
<td>21 27 35 0.6 0.7 0.8 2.7 2.8 2.9</td>
<td>6.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA-50%</td>
<td>7.57 0.80</td>
<td>21 27 35 0.6 0.6 0.7 2.7 2.8 2.9</td>
<td>2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA-90%</td>
<td>7.57 1.06</td>
<td>21 33 50 0.6 0.6 0.6 2.7 2.9 3.1</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA</td>
<td>7.23 1.34</td>
<td>32 33 37</td>
<td>– – – – – – – – – –</td>
<td>13.08</td>
<td></td>
</tr>
<tr>
<td>RNA-50%</td>
<td>7.23 1.39</td>
<td>32 33 37</td>
<td>– – – – – – – – – –</td>
<td>5.45</td>
<td></td>
</tr>
<tr>
<td>RNA-90%</td>
<td>7.23 1.58</td>
<td>32 41 54</td>
<td>– – – – – – – – – –</td>
<td>3.96</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)RMSD computed with C\(_{\text{α}}\) or P atoms.
\(^b\)Scores computed with Molprobity (Chen et al., 2010).
\(^c\)Log-weighted combination of scores which reflects the crystallographic resolution at which these values would be expected.
\(^d\)Mean running time in minutes.
\(^e\)Average values conformational pathway test cases obtained with δMorph using a sigmoid potential. Values shown for initial (i), final (f) and average conformations (A). CG representations as indicated in Table 2.

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.

Table S5. RNA

| RNA-90% | 7.23 1.58 | 32 41 54 | – – – – – – – – | 3.96 |
| RNA-50% | 7.23 1.39 | 32 33 37 | – – – – – – – – | 5.45 |
| RNA     | 7.23 1.34 | 32 33 37 | – – – – – – – – | 13.08 |

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 (δ\(_5\) > 0.90). The closest
Table 4. IC based Monte-Carlo conformational sampling

<table>
<thead>
<tr>
<th></th>
<th>Unbiased</th>
<th>Biased Rg ±1 Å</th>
<th>RgΔ RMSD</th>
<th>Clashes</th>
<th>Vmol</th>
<th>Molprobity %</th>
<th>αα</th>
<th>ββ</th>
<th>γγ</th>
<th>δδ</th>
<th>εε</th>
<th>ζζ</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSD</td>
<td>4.04</td>
<td>0.94</td>
<td>1.28</td>
<td>1.6</td>
<td>6.6</td>
<td>1.08</td>
<td>16.2</td>
<td>33.7</td>
<td>12.40</td>
<td>45.0</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 4 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 (Fieg and Burton, 2010; Fulle and Gehlke, 2010; Orozco et al., 2008; Skjærven 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 iMode 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 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.

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 1wv5) and holo (yellow, PDB 1wv2) structures is shown (B, left). On the right, the holo structure was superimposed with the closest conformation (cyan) found in the biased ensemble.
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, L., et al., 2010). Finally, the progress of this type of methods with nucleic acids (Fullé and Göhlke, 2010; Orozco et al., 2008) is also a promising research area.

ACKNOWLEDGEMENTS

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

REFERENCES


J.R.López-Blanco et al.


