Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization

Erik Lindahl, Marc Delarue

Published online August 8, 2005

4496–4506 *Nucleic Acids Research, 2005, Vol. 33, No. 14*
doi:10.1093/nar/gki730

Refinement of docked protein–ligand and protein–DNA structures using low frequency normal mode amplitude optimization

Erik Lindahl[1,2] and Marc Delarue[1,*]

[1]Unité de Biochimie Structurale, URA 2185 du CNRS, Institut Pasteur, 25 Rue du Dr Roux, F-75015 Paris, France and [2]Stockholm Bioinformatics Center, Stockholm University, SE-17156 Stockholm, Sweden

Received May 24, 2005; Revised and Accepted July 6, 2005

ABSTRACT

Prediction of structural changes resulting from complex formation, both in ligands and receptors, is an important and unsolved problem in structural biology. In this work, we use all-atom normal modes calculated with the Elastic Network Model as a basis set to model structural flexibility during formation of macromolecular complexes and refine the non-bonded intermolecular energy between the two partners (protein–ligand or protein–DNA) along 5–10 of the lowest frequency normal mode directions. The method handles motions unrelated to the docking transparently by first applying the modes that improve non-bonded energy most and optionally restraining amplitudes; in addition, the method can correct small errors in the ligand position when the first six rigid-body modes are switched on. For a test set of six protein receptors that show an open-to-close transition when binding small ligands, our refinement scheme reduces the protein coordinate cRMS by 0.3–3.2 Å. For two test cases of DNA structures interacting with proteins, the program correctly refines the docked B-DNA starting form into the expected bent DNA, reducing the DNA cRMS from 8.4 to 4.8 Å and from 8.7 to 5.4 Å, respectively. A public web server implementation of the refinement method is available at http://lorentz.immstr.pasteur.fr.

INTRODUCTION

Structural flexibility is an important feature of biological macromolecules. The most obvious conception of flexibility is probably the ensemble of protein structures derived from NMR experiments in solution, but X-ray crystallography has also often revealed large backbone rearrangements for the same protein under different conditions, e.g. alternative packing arrangements, or in the presence of a bound ligand or inhibitor. A wide range of such experimentally observed molecular 'motions' have been classified in the MolMovDB database provided by the Gerstein lab (1,2).

Over the last decade, many efforts have focused on protein–protein interactions and docking (3). Significant progress has been made, but the community-wide CAPRI (Critical Assessment of PRedicted Interactions) experiment (4) has highlighted the limitations in handling conformational changes in protein–protein interactions (5), and it is only quite recently that receptor flexibility has been introduced explicitly (6–9). Protein flexibility is particularly important for understanding molecular interactions because many proteins exhibit significant structural changes when binding ligands, going from an open to a closed form. Historically, this prompted the development of the so-called induced fit theory (10). While the global cRMS (coordinate root-mean-square) difference between the open and closed structures can vary widely (11), a good model of the rearrangement around the active site is often crucial to correctly predict substrate binding (12).

Structural change due to protein interactions is intimately related to the refinement problems that currently receive a lot of attention in other applications. Homology modeling is a classical example, where fold recognition backbone templates are often displaced 3–6 Å relative to the target structure (13–15). Refining these models is an extremely hard but very important challenge. A lot of effort has been invested in improving models, often based on molecular dynamics (16–18) and using sampling-enhancing methods such as replica exchange (19). While many of these results are useful, the general conclusions at the 5th CASP (Critical Assessment of Structure Prediction) experiment (20) was still that they do not yet reliably improve blind prediction results (21).

Refinement of intial models using experimental data such as medium-resolution X-ray structure factors (22), electron densities obtained from electron cryo-microscopy (23) or small-angle X-ray scattering (SAXS) (24) intensities have been

*To whom correspondence should be addressed. Tel: +33 1 45 68 8605; Fax: +33 1 45 68 8604; Email: delarue@pasteur.fr

more successful, but the radius of convergence is small, even when using dihedral angles instead of Cartesian coordinates in combination with the *R*-free factor (25). Usually, this kind of 'docking into an envelope' problem is handled by rigid-body refinement of domains, but it is difficult to correctly define which domains to use. Also, some movements such as shear motion are not described adequately by rigid-body moves.

The reason why refinement is difficult is a combination of imperfect energy functions and the high-dimensional optimization space. Even with a perfect potential function it would be a tremendously hard task to find the global minimum for a protein structure, and when there are errors in the parameters any signal can easily drown in noise. A key challenge for refinement algorithms is thus to drastically reduce the dimensionality of the conformational space.

One way of introducing this type of restricted conformational space is to use the collective motions described by low-frequency normal modes as a basis. Normal mode analysis has long been used to extract characteristic motions of protein structures (26–28), since it provides a computationally unexpensive description of the motion close to an energy minimum. New simplified models pioneered by Tirion (29) have made it possible to rapidly determine normal modes for very large macromolecular assemblies such as a virus capsid (30) and even the ribosome (31) structure. With the availability of experimental structures for several conformations of many proteins, it has further become possible to correlate predicted transitions between conformations with experimentally observed ones (32,33). Individual mode frequencies are usually not accurate since real motions will be damped by the solvent (34), and with the Elastic Network Model the absolute frequency information is lost. However, the shape of motions associated with low-frequency modes is quite insensitive to the model, and these are the most important motions: Since the mode energy is proportional both to squared frequency and amplitude, thermal energy equipartition predicts that low-frequency modes will have the largest amplitudes. In most documented cases, a handful of the lowest frequency modes are sufficient to describe transitions between observed conformations (33).

The value of these features has recently been verified by successful application of normal modes to enforce large collective movements in X-ray refinement (35), molecular replacement experiments (36), and low-resolution electron microscopy density fitting (37,38). A related promising approach has been published by Qian *et al*., who utilized multiple homology modeling templates to define principal component variations along which an initial homology model could be refined (39). Similar principal component motions have also been used to extract essential motions from molecular dynamics simulations to improve ligand geometry prediction (8).

In this work, we show how simple Elastic Network normal modes provide an efficient basis for refinement entirely without additional experimental data, and design a robust optimization scheme. One of the key challenges has been to create a scheme that is not critically sensitive to the number of normal modes used, and equally important to never damage a structure. The latter is particularly vital in case the initial ligand placement is imperfect.

Zacharias *et al*. have previously used classical normal modes or principal components derived from simulations for docking small ligands into the minor groove of DNA (40) as well as predicting the FKBP–FK506 complex (8). Their primary aim was however to better discriminate between alternative ligand placements. Refined B-DNA structures were only compared to the same B-DNA minimized in Cartesian coordinates in presence of the ligand rather than with experimental structures (40). For FKBP, the refined receptor structure had a cRMS of 1.45 Å compared to the initial 1.5 Å, which was a simulation average (8)—the difference between bound and unbound experimental FKBP structures is only 0.6 Å. Our study instead focuses on the refinement of the experimental free form of the receptor, in particular mid- or low-resolution refinement of large motions that occur when the ligand binds.

MATERIALS AND METHODS

Refinement fundamentally consists of two parts: choosing the degrees of freedom to sample, and designing energy functions to discriminate between structures. For the refinement, it does not matter exactly how sampling moves are derived—the only important point is whether they form a useful basis for the motions.

Low-frequency normal modes for sampling

The basic idea of normal modes is to provide a simplified model of the potential energy landscape and motions of an N-particle system close to a local minimum. In this state, all first derivatives disappear by definition, so the simplest approximation of the energy variation is $\mathbf{x}^{\mathbf{T}}\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is the $(3N)^2$ Hessian matrix obtained from the second derivatives of energy with respect to coordinates $\mathbf{x}$. For biomolecules, the Hessian is normally expressed in mass weighted coordinates. With a matrix $\mathbf{M}$ containing atomic masses on the diagonal, we obtain $\mathbf{H}' = \mathbf{M}^{-1/2}\,\mathbf{H}\mathbf{M}^{-1/2}$. By restricting the Taylor expansion of the energy close to the minimum to $\mathbf{x}^{\mathbf{T}}\mathbf{H}\mathbf{x}$, an analytical solution to the equations of motion can be found in terms of the superposition of normal modes, whose directions are given by the eigenvectors of $\mathbf{H}$ and the normal mode frequencies will be proportional to the square root of the corresponding eigenvalues. This type of normal mode calculation has been applied to biomolecules for over two decades (26–28), but there are significant limitations. First, memory and CPU requirements increase as N^2 and N^3, respectively, which limits the method to fairly small systems. Some degrees of freedom can be removed by using, e.g. torsional instead of Cartesian coordinates (28), but it does not address the scaling problem. Another complication is that structures need to be energy minimized prior to the normal mode calculation. Not only is this computationally expensive, but the minimization often distorts the cRMS of a protein by several Ångstroms due to approximate potential functions and lack of solvent (28,41).

Interestingly, the appearance of low-frequency modes depends more on overall geometric properties such as the number of surrounding neighbors rather than force field details. Tirion was the first to note this and introduce the so-called Elastic Network Model (ENM) where the energy is described by a network of simple pair potentials (29),

$$U = \sum_{i,j;r_{ij}<R_c} U_{ij}(\mathbf{R}_i - \mathbf{R}_j), \qquad \mathbf{1}$$

where R_c is a cut-off distance for the interactions. By defining the pair potential as a harmonic function of interatomic distance with the equilibrium length equal to the distance in the input structure

$$U_{ij}(\mathbf{r}) = k_{ij}(|\mathbf{r}| - |\mathbf{R}_{ij}^0|)^2, \quad \mathbf{2}$$

the initial structure will by definition be at the global energy minimum, and no minimization necessary. The method was simplified by Hinsen (42) who limited it to alpha carbons, and also by Bahar *et al*. (43). In the original model, the force coefficient k_{ij} was constant, and the only variable parameter was the cut-off distance R_c (typically 10 Å). Interactions that decay with distance are more realistic and much less sensitive to the cut-off distance: For compatibility reasons, we use exponential weighting (42) with a screening length r_0 (normally 3 Å for Cα networks).

Due to the arbitrary force constant, mode frequencies are not physically meaningful, but the low frequency normal mode motions are in excellent agreement both with modes derived from classical force fields (34,42) and experiments (44,45). The method additionally handles missing atoms transparently.

In the present work, we keep all atoms when building the elastic network model, but to avoid artifacts in the light hydrogen motions the screening length r_0 was increased to 5 Å. Theoretically, the resulting Hessian could be extremely large, but with the 10 Å cutoff employed most elements vanish, and it can be represented efficiently through sparse-matrix storage. The lowest m eigenvectors of the sparse Hessian are determined with a computational cost proportional to $O(mN)$ by using the implicitly restarted Lanczos algorithm provided in the ARPACK library (46). Calculating the lowest 50 modes of a 3000-atom protein takes ~40 s on a 2.8 GHz Pentium IV workstation, and the code has been tested successfully with hundreds of thousands of atoms. A web server implementing this algorithm is freely available at http://lorentz.immstr.pasteur.fr. As an alternative, Sanejouand and co-workers (47,48) have developed an interesting general approach, the so-called rotation-translation block (RTB) method, where groups of atoms or even multiple residues are merged into a single block that has just translational and rotational degrees of freedom. This can be used for arbitrary reduction of the conformational space, and the motions of the removed atomic coordinates are easily reconstructed from the normal modes of the rotation-translation blocks. The differences between the two normal mode algorithms are fairly small for the low-frequency modes, and normal modes derived from the RTB approach provided essentially identical refinement results when used for double-checking in this study.

Energy-based refinement and discrimination

The reason why energy minimization along modes derived from a minimum works is partly that the elastic network model is based on receptor geometry instead of the real force field, but also that the energy optimization will be limited to the intermolecular interaction between the receptor and ligand. The entire normal mode calculation step can be considered a black box from which a set of search directions are obtained that are used to conduct the actual refinement.

The optimization energy function consists of all intermolecular electrostatic and Lennard–Jones interactions, with parameters taken from the CHARMM19 force field (49). Polar hydrogen atoms were added to the structures prior to determining normal modes. To avoid diverging energies when atoms get too close, both electrostatic and Lennard–Jones interactions were modified with a Levitt soft-core scaling factor (41)

$$U'(r) = U(r)\frac{1}{C_{12}r^{-12}(1 + br^2)/h + 1} \quad \mathbf{3}$$

where C_{12} is the repulsive Lennard–Jones parameter. Default values (41) are used for the two model parameters, $b = 0.1$ Å^{-2} and $h = 10$ kcal/mol. The soft-core interaction asymptotically approaches the standard form for long distances, but goes to a constant value (h) instead of diverging when atoms overlap.

A number of geometric optimization functions such as surface solvation (50,51) or maximizing the interaction area were attempted but not used since they are hard to balance with the repulsion term. Refinement using only Lennard–Jones interaction works well, but since electrostatics improves a couple of receptors it was kept for all cases. To avoid introducing arbitrary parameters, the dielectric coefficient was fixed to the vacuum value.

The energy refinement is performed in normal mode space by using mode amplitudes c_k as degrees of freedom, and the Cartesian coordinates reconstructed from m excited normal mode eigenvectors $\mathbf{a}_k$ as

$$\mathbf{x} = \mathbf{x}_0 + \sum_{k=1}^{m} c_k \mathbf{a}_k. \quad \mathbf{4}$$

The lowest six modes are rigid body motions, and normally not included in the refinement. Energy minimization was performed with a quasi-Newtonian algorithm (52). If vectors are orthogonal in the Cartesian norm, the partial derivatives with respect to c_k can be calculated analytically.

Even with fewer degrees of freedom and soft-core potentials, it is common to get stuck in local minima. If 1–2 modes provide enough flexibility, it is possible to discretize the problem by introducing a two-dimensional grid of mode amplitudes and sample this subspace exhaustively. Since some structures are dimers or have other flexible parts, it is not trivial to a priori select which modes to employ. As a compromise, a semi-exhaustive sampling was introduced by iteratively scanning each mode for a minimum while the amplitudes of previously added modes were kept constant. The main complication is again low-frequency modes that do not contribute to the refinement: if a false minimum is found at large amplitude, the remaining refinement is ruined when this amplitude is frozen. This was addressed by pre-scanning of each normal mode with all other set to zero, and sorting in order of largest reduction in energy. For the actual refinement, this means more promising modes are used first, and their amplitude frozen before the next degree of freedom is added.

To avoid too large excitations, the concept of applying artificial restraints (53) on the amplitudes was borrowed from quantum chemistry charge fitting. Our model is similar to those commonly used for NMR distance restraints, in that mode amplitudes are allowed to vary freely in the range $c_k \in [-100, 100]$, and a harmonic restraint $U = 0.5k'_0(|c_k| - 100)^2$ applied outside this region. For the present work, a value of

$k_0' = 0.01$ kcal/mol was used. The restraints contribute to more robust refinement on average, but also deteriorate the very best unrestrained results slightly.

TEST STRUCTURES

Protein–small ligand systems

There are plenty of structures known to be involved in docking where one or both free and complexed conformations have been determined, and a number of protein interaction benchmarks have been published (54–57). To test the efficiency of our refinement, additional conditions were applied that resulted in a smaller set of benchmark structures. First, since the methods are coarse the structural change should be significant, and both the open and closed receptor structures available. Individual amino acid mutations and missing atoms do not pose any problems, but there cannot be insertions or deletions between the alternative receptor structures. With these constraints, six suitable protein–ligand pairs were identified in the PDB: maltodextrin binding protein, chicken citrate synthase, glutamine binding protein, HIV-1 protease, phosphoglycerate kinase and lactoferrin. All systems are summarized in Table 1 (58,59). It is noteworthy that chicken citrate synthase is a rather large dimer structure. Both monomers need to be included in the refinement, and since their relative motions are the largest in the system it is an interesting challenge to normal mode-based refinement. The specific docking interactions in the Citrate synthase have recently been studied in detail (60,61).

All-atom Elastic Network normal mode eigenvectors were determined from the unbound conformations, since the whole point of non-benchmark refinement will be to predict the complex structure without having any experimental information about it.

Each open receptor conformation was superimposed on the closed experimental (keeping the ligand coordinates) using the PROFIT package by A.C.R. Martin (http://www.bioinf.org.uk), which implements the McLachlan algorithm (62). Therefore, the ligand positioning is almost perfect in these tests, although we did study cases where the ligand was intentionnally misplaced (see below).

To investigate the extent to which low-frequency normal modes can describe the transition between the pairs of conformations, overlap coefficients (63) were calculated as scalar products of the coordinate displacement vector with each normal mode eigenvector,

$$c_k = \frac{\mathbf{x}_{\text{closed}} - \mathbf{x}_{\text{open}}}{|\mathbf{x}_{\text{closed}} - \mathbf{x}_{\text{open}}|} \cdot \mathbf{a}_k. \qquad \mathbf{5}$$

Due to mutations between the two structures and/or missing atoms, the displacement vectors could only be calculated for Cα coordinates. The scalar product must then be taken with Cα-only elastic network normal modes, since the Cα subset of all-atom modes would not be orthogonal (the modes used in refinement are always all-atom, though).

As Figure 1 shows, the transitions for the maltodextrin- and glutamine-binding proteins as well as citrate synthase are described almost entirely with 1–2 modes. The other three structures are also heavily biased against low modes, with the possible exception of HIV-1 Protease for which modes with index up to 15–20 might play a role. The ability of low-frequency normal mode eigenvectors to predict docking flexibility is thus in very good agreement with the findings of Tama (32) and Krebs (33) for general transitions.

The lowest cRMS for a given set of modes is obtained when each amplitude is set to the overlap coefficient c_k. The striking efficiency is evident from Figure 2; the cRMS of maltodextrin-binding protein could theoretically be reduced from almost 4 Å to just >1 Å using only two degrees of freedom, and all receptors except the HIV-1 protease exhibit quite dramatic improvements. The cRMS improvement can also be estimated directly from the overlap coefficients as

$$\text{cRMS} = \text{cRMS}_0 \left[1 - \sum_{k=1}^{m} c_k^2\right]^{1/2}, \qquad \mathbf{6}$$

where cRMS_0 is the value for the starting conformation and m the number of normal modes (64).

DNA–Protein systems

To test the applicability of the final refinement protocol on larger structures and nucleic acids, the DNA–protein complexes 1BER (Catabolite gene protein, CAP), 1LWS (Intein homing endonuclease), and 1A1H (A variant of Zif 268 zinc finger) were selected. For the CAP protein, we also compared the results with the free crystal structure 1G6N (no free structures are available for the other cases). Straight B-DNA structures were created using the program NAHELIX with the same sequence as the bound motifs in the PDB structures and superimposed on the bound structure using PROFIT. All protein structures were assumed to be rigid and the normal mode calculation only applied to the DNA ligands. The initial cRMS value between the straight B-DNA and the structure in the PDB file was 8.42 Å for 1BER, 8.72 Å for 1LWS and 2.26 Å for 1A1H when all DNA heavy atoms are included in the calculation.

Table 1. Summary of protein–ligand structure pairs

Receptor	Open	Closed	Ligand	cRMS (Å)
Maltodextrin binding protein	1OMP	1ANF	Glucose	3.77
Chicken citrate synthase	5CSC	6CSC	Citrate, trifluoroacetonyl-coenzymeA	2.84
Glutamine binding protein	1GGG	1WDN	Glutamine	5.33
HIV-1 protease	1HHP	1AJX	Cyclic sulfamide	1.91
Phosphoglycerate kinase	1V6S	1VPE	Mg, phosphoglyceric acid, 5′-adenyly-imido-triphosphate	3.45
Lactoferrin	1CB6	1LCF	Copper carbonate, copper oxalate	6.44

All cRMS values are calculated from Cα atoms only.
Non-protein parameters were obtained through the hic-up (58) and prodrg (59) web servers.

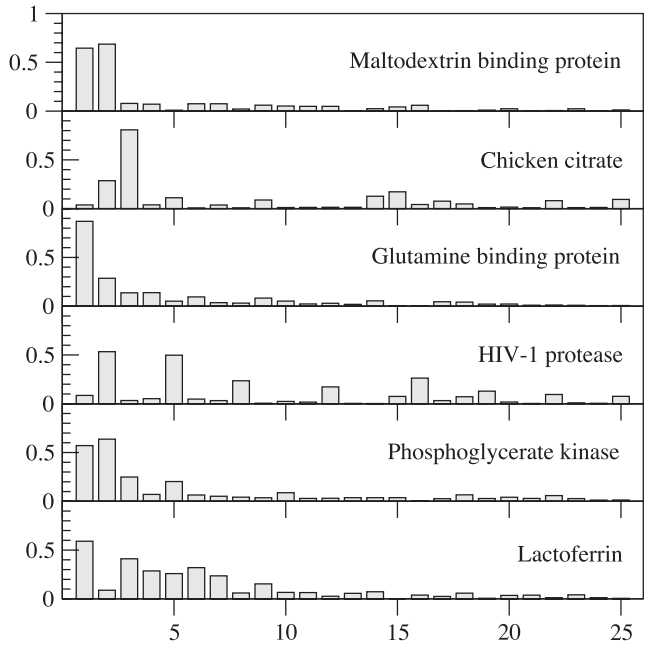

Figure 1. Overlap between structural change and low frequency normal modes, measured as the scalar product between the Cα displacement vectors from open to closed conformation and the first 25 non-trivial normal mode eigenvectors. From top to bottom: maltodextrin binding protein, citrate synthase, glutamine binding protein, HIV protease, phosphoglycerate kinase and lactoferrin.

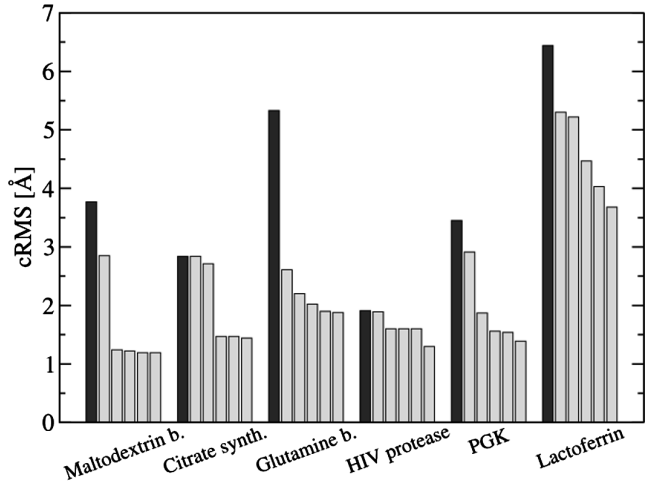

Figure 2. Theoretical limit of refinement efficiency. This is derived by projecting Cα displacement vectors between open and closed conformations on normal mode eigenvectors. The first bar for each target represents the initial cRMS value and the remaining bars are projection results using 1–5 normal mode eigenvectors.

RESULTS

Test 1: Minimization in Cartesian coordinates

To assess the difficulty, the free receptor structures were docked with the ligands and subjected to classical molecular mechanics energy minimization using the GROMACS package (65). Parameters were taken from the OPLS-AA/L force field (66), and a limited-memory quasi-Newtonian minimizer used (52). Interactions were cut off at 10 Å and smoothly switched off between 8 and 10 Å. The systems were minimized to double precision machine accuracy, which in all cases completed in <10 000 steps. The results are summarized in Table 2, with coordinate displacements relative to the initial state of 1.5–2.1 Å. Only the maltodextrin binding protein shows any significant improvement, and 3 of the 6 receptors are deteriorated. On average, Cartesian minimization results in 0.1 Å worse cRMS. There is consequently very little, if any, predictive power from this type of free refinement. Unfortunately, this agrees quite well with CASP observations that unconstrained energy minimization does not improve homology models (21).

Table 2. Test Protocol 1—naive unconstrained Cartesian energy minimization of docked receptor structures using the OPLS-AA force field in GROMACS

Receptor	cRMS move (Å)	Δ cRMS target (Å)
Maltodextrin binding protein	2.12	−1.23
Chicken citrate synthase	1.89	−0.01
Glutamine binding protein	1.83	+0.67
HIV-1 protease	1.45	−0.13
Phosphoglycerate kinase	1.47	+0.77
Lactoferrin	1.66	+0.56

The first column displays the cRMS displacement relative to the starting configuration, and the second the cRMS change during minimization with respect to the closed target structure.

Test 2: Minimization in normal mode space

Minimization along normal modes has proven very successful when the refinement is guided by experimental data (35,36), but purely theoretical energy functions is a harder challenge. The detailed interactions result in a more rugged energy landscape that is harder to optimize, and for large mode amplitudes the approximations used give rise to false minima with extremely low energies. Figure 3 displays the energy landscape of the intermolecular interaction between ligand and receptor, and cRMS variation along the two lowest non-trivial modes for the maltodextrin and glutamine binding proteins. Even with soft-core interactions (28), the extremely reduced two-dimensional landscape is non-trivial to search, while the cRMS variation is approximately harmonic, as expected.

For both these cases, minimization works fine when only the first two modes are used, but the structures are easily ruined when additional modes are employed. The maltodextrin binding receptor ends up at 5.7 Å (see Figure 4), while the glutamine binding protein diverges to >200 Å. For HIV-1 Protease, the cRMS drops to 1.6 Å when one or two modes are used, but increases again for 3–5 modes. Lactoferrin does not exhibit any significant improvement before it starts to diverge with four modes. On the other hand, both citrate synthase and phosphoglycerate kinase improve up to at least five modes.

The results of this first test are thus somewhat mixed: the best normal mode minimization results are excellent, but because of stability issues, the blind refinement predictive power is limited.

Test 3: unconstrained scanning along modes

Semi-exhaustive scanning was attempted mainly to avoid the divergence problems encountered in normal mode space

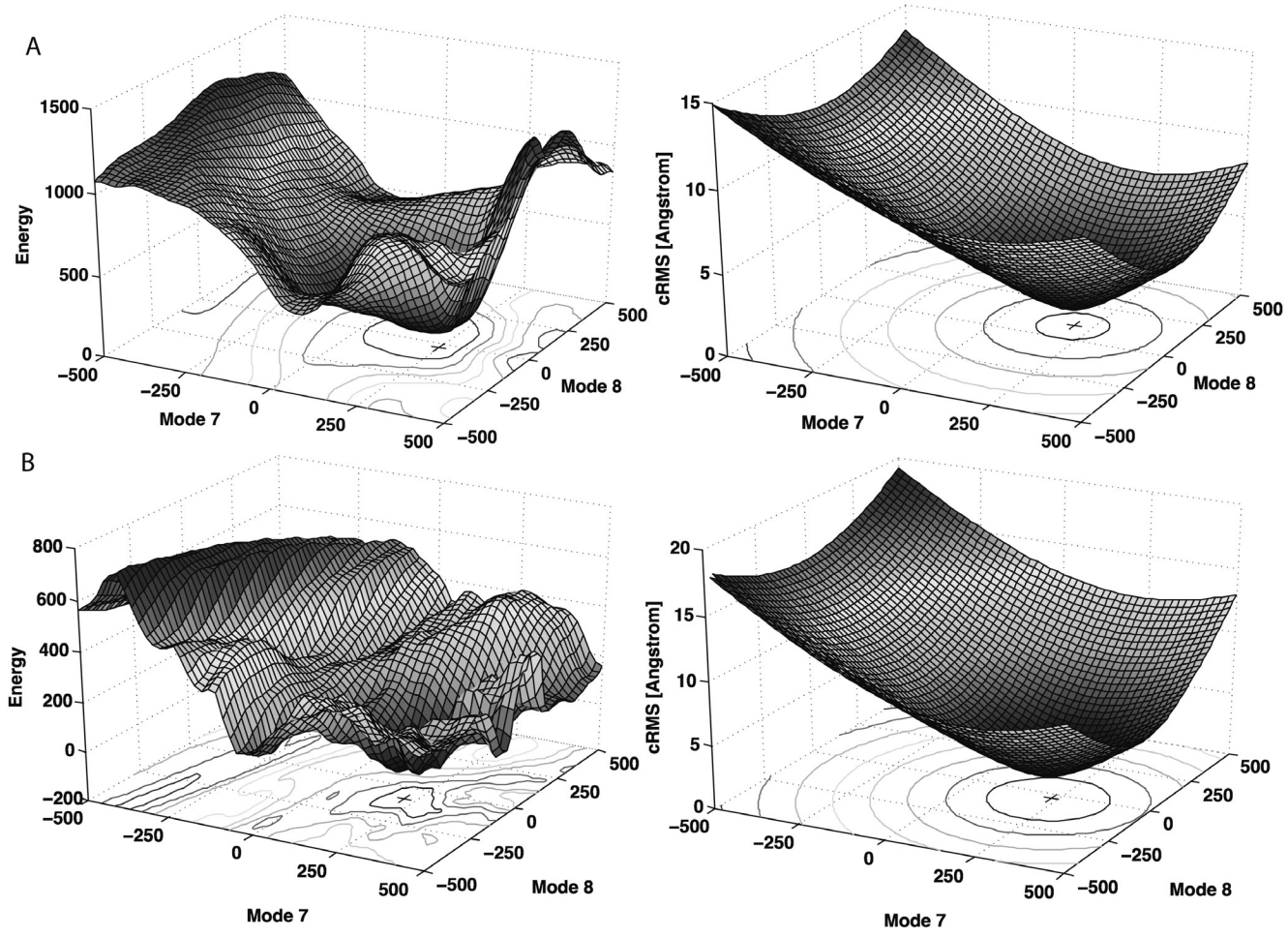

Figure 3. (**A**) Non-bonded intermolecular energy (kcal/mol) for the maltodextrin binding protein, as a function of the two lowest non-trivial normal modes (left) and cRMS variation relative to the closed target state (right). The minimum is slightly offset from both axes, indicating that both modes contribute to refinement. The global cRMS minimum is 1.21 Å, and the value at the energy minimum 1.26 Å. (**B**) Similar plots for the glutamine binding protein. In this case, the lowest attainable cRMS is 2.20 Å, and at the energy minimum it is 2.28 Å.

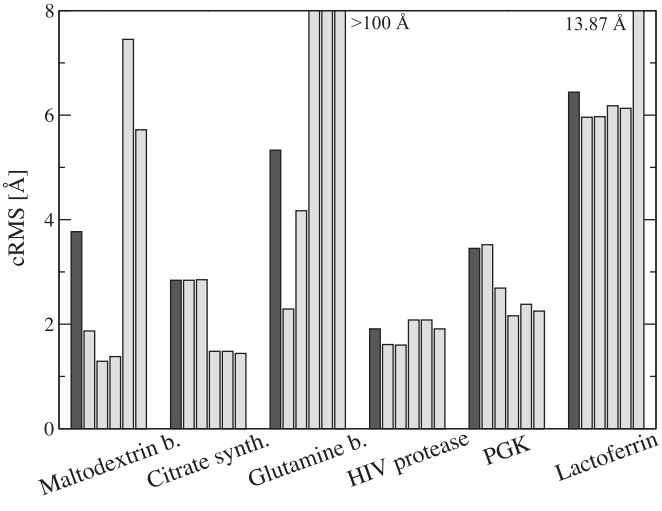

Figure 4. Test protocol 2—L-BFGS minimization of non-bonded energy along 1–5 normal modes. Shaded bars indicate the initial cRMS values before refinement. This is substantially more efficient than unconstrained Cartesian minimization, but not entirely stable—structures sometimes end up in worse states when additional normal modes are used.

minimization, but also because it is insensitive to energy barriers. Normal mode amplitudes in the range −500 to +500 were sampled with a spacing of 10, and the amplitude corresponding to lowest energy frozen before the next higher mode in frequency order was added. Figure 5 illustrates the results for all the test receptors when between 1 and 5 non-trivial normal modes are used in the scanning. Hypothetically, a simple scanning refinement approach using the lowest three normal modes of each structure would reduce the cRMS of the maltodextrin binding protein, citrate synthase and glutamine binding protein to their target conformations by >50%, it would improve phosphoglycerate kinase and lactoferrin slightly, and finally leave HIV-1 protease close to its initial state. The only problem is that some structures still deteriorate when additional modes are included. We choose to address this in two ways, first by restricting scanning to the normal modes that are most likely to improve the structure, and second by constraining the receptor structure distortion.

Final protocol: mode sorting and restrained scanning

The refinement would probably be improved if it was known a priori which degrees of freedom to apply first. This was

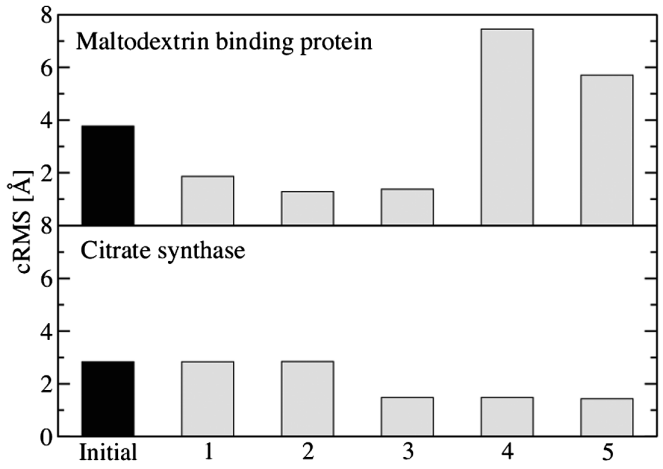

Figure 5. Test protocol 3—energy optimization through discrete amplitude scanning along low frequency normal modes. Bars in each group indicate the initial cRMS value (shaded) followed by refinement with 1–5 normal modes. Scanning provides a better alternative than minimization along the normal mode eigenvectors, but is still somewhat unstable.

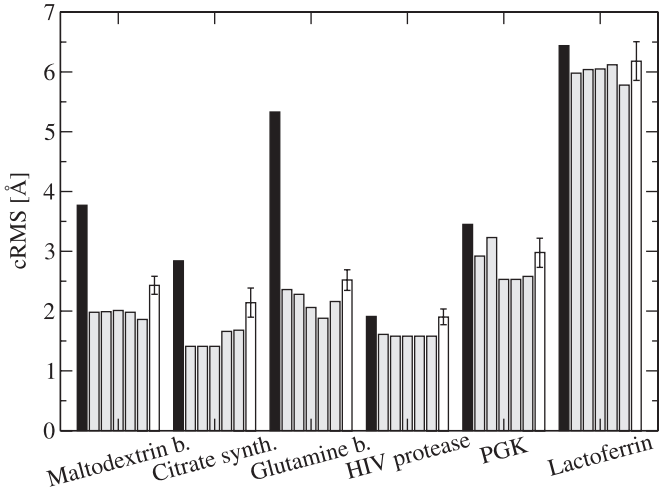

Figure 6. Final refinement protocol—restrained discrete amplitude scanning along normal modes first sorted by pre-scanning. The first dark shaded bar for each target is the cRMS value prior to refinement and the next five the results using 1–5 normal modes. The final open bar in each group is the average result of refinement with five modes after random translation/rotation. Restraints combined with pre-sorting makes the scheme quite robust, at the cost of slightly higher final cRMS values.

estimated through pre-scanning of individually excited modes and subsequent sorting in order of largest energy reduction. The receptor distortion was simultaneously controlled by adding restraints on the mode amplitudes, as discussed in the Methods section. The combination of these methods leads to a dramatic improvement in stability, as evident from Figure 6. When using five sorted normal modes, the cRMS value of maltodextrin binding protein is refined from 3.77 to 1.86 Å, chicken citrate synthase from 2.84 to 1.68 Å, glutamine binding protein from 5.33 to 2.16 Å, HIV-1 protease from 1.91 to 1.58 Å, phosphoglycerate kinase from 3.45 to 2.58 Å, and the lactoferrin cRMS drops from 6.44 to 5.78 Å.

This refinement scheme reduces the cRMS value of all structures compared to the respective initial states, and is much less sensitive to the number of extra modes used.

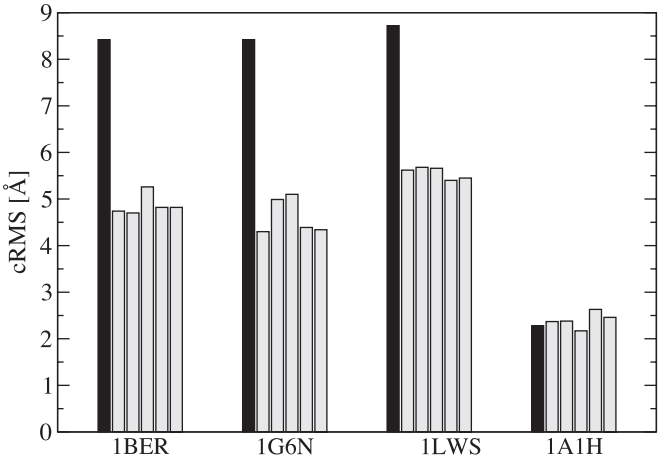

Figure 7. Refinement of DNA structures complexed with proteins using the final protocol. Dark shaded bars indicate initial cRMS values for all heavy atoms, and the remaining ones are the results from refinement using 1–5 normal mode degrees of freedom. Even a single degree of freedom reduces the cRMS of DNA bound to 1BER, 1G6N and 1LWS by almost a factor 2. The starting B-DNA was generated with NAHELIX (68).

In almost all cases, the most significant reduction in cRMS is seen when the first sorted normal mode is applied. The maltodextrin binding protein result is slightly worse than with unrestrained scanning (1.86 Å instead of 1.26 Å), but we believe it is a reasonable sacrifice for the massive gain in robustness.

Refinement of DNA–protein complexes

The final protocol was also applied to the DNA–protein complexes previously described. Both for the 1BER and 1LWS complexes, the cRMS is reduced significantly even when only a single normal mode degree of freedom is used. For 1BER, the cRMS value drops from 8.42 to 4.74 Å, i.e. slightly better than the 4.82 Å achieved when the first five modes are used simultaneously. The results when 1BER is replaced with the unbound protein structure 1G6N are similar; DNA refinement with five modes results in a cRMS of 4.34 Å.

For 1LWS, the single-mode result is 5.62 Å (starting from 8.72 Å), which improves to 5.45 Å with the first five modes. In both these cases, the DNA wraps around the protein. The DNA bound to the protein of the 1A1H structure (a variant of Zif 268 zinc finger) does not refine further from the initial 2.28 Å cRMS, but the structure remains fairly stable—with five normal modes, the final structure has a cRMS of 2.46 Å compared to the target. The DNA–protein refinement results are summarized in Figure 7.

In general, iterating the normal mode calculation and refinement did not enhance the results, except for DNA bound to the Catabolite Activator Protein (1BER)—in this case, a second iteration improved the cRMS further from 4.8 to 4.1 Å.

Refinement of relative rotation/translation

To test the effect of errors in ligand placement, we performed an additional 50 restrained scans for each target, in cases where the ligands were randomly translated 1.0 Å and rotated 30°, which resulted in an average ligand cRMS displacement of 2.04 Å. The extra rotational/translational flexibility was

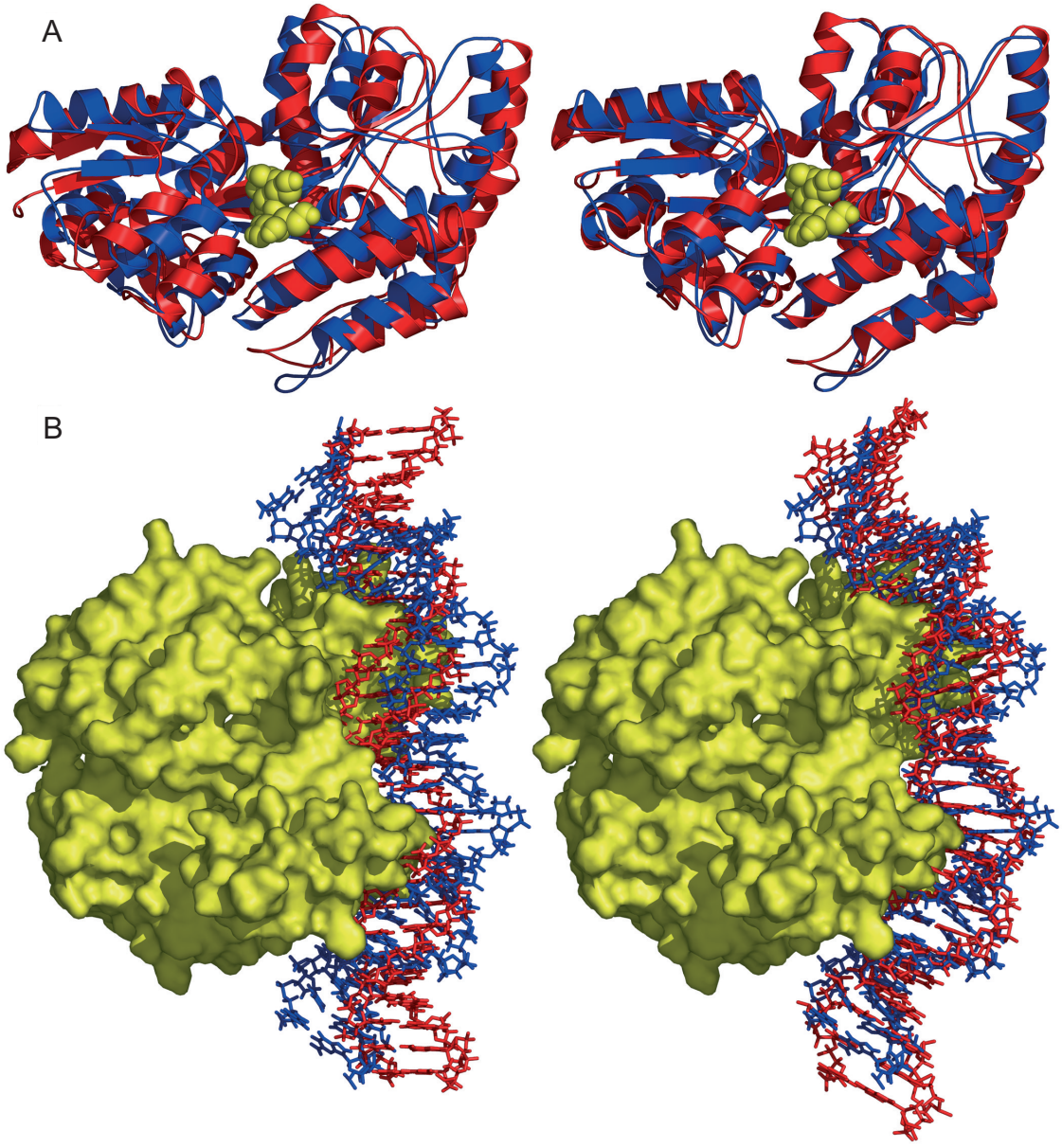

Figure 8. Refined structures of maltodextrin binding protein (**A**) and DNA bound to Catabolite Gene Protein (**B**). The left panels show the initial unbound (red) compared with the bound target state (blue), while the right panels display the structure after normal mode energy optimization (red) compared with the same target state. Fixed protein/ligand parts are shown in yellow. The maltodextrin binding protein cRMS reached 1.86 Å (starting from 3.77 Å) and the DNA structure was refined from 8.42 to 4.82 Å. Secondary structure elements move significantly, but the positions of individual helices and in particular beta sheets agree very well with the target state after refinement. Figures drawn with PyMol (W. L. DeLano, http://www.pymol.org).

handled transparently by including the six lowest normal modes in the scanning, thus using 11 modes in total. There are obviously better choices for rigid body moves since all normal mode deformations are linear, and 30° of rotation will introduce some structural distortion. Still, it is quite entertaining as a proof-of-concept and the general conclusions remain valid (Figure 6). On average, the cRMS value of the displaced maltodextrin binding protein was refined to 2.4 Å (compared to 1.86 Å without the random translation), citrate synthase achieved a cRMS of 2.1 Å (1.68 Å when correctly superpositioned), glutamine binding protein reached cRMS 2.5 Å (instead of 2.16 Å), HIV-1 protease remained at the initial cRMS 1.9 Å (refined to 1.58 Å above), phosphoglycerate kinase refined to a cRMS of 3.0 Å (2.58 Å without translation), and lactoferrin a cRMS of 6.2 Å (was 5.78 Å). These results are shown as open bars in Figure 6. Compared to the initial unrefined cRMS values, all structures except the HIV protease are still improved. This was not unexpected, since the random translation is five times larger than the best cRMS improvement reported for the ideal ligand placement for HIV protease.

DISCUSSION

Many of the optimization results are quite significant, and even more so when the refined structures are studied in detail.

Figure 8 displays the effect of restrained scanning refinement using five sorted modes for the maltodextrin binding protein and DNA bound to the Catabolite Activator Protein (1G6N), and the other structures are illustrated in the Supplementary Material. Helices overlap very well for the maltodextrin binding protein, and the structure around the active site agrees almost perfectly with the closed form. There is also room for future improvements, since the unrestrained scanning or minimization reached cRMS values of 1.26 Å, at the cost of stability. The DNA structure bound to 1G6N moves significantly to wrap around the protein—in particular, the middle section agrees very well with the target state. Incidentally, starting from this refined state it is actually possible to also improve the protein (1G6N) structure from 2.3 to 1.82 Å. While encouraging, this is not yet a general result; we are currently working on methods to allow simultaneous flexibility in both ligand and receptor structures.

It is tempting but a bit risky to make predictions about the actual *in vivo* docking process from these results. Induced fit has sometimes been described alternatively as a selection of a pre-existing state (67), where multiple receptor structures exist in equilibrium and the introduction of the ligand displaces the equilibrium towards the closed form. Normal mode refinement can be viewed as a computational simulation of this process; the variable normal mode amplitudes are simply introduced as a way to generate a pre-existing ensemble of conformations from a simple basis set, and non-bonded energy used to select the most advantageous of these.

Only a single test case (1BER) was improved by re-determining normal modes as the structure is deformed. This agrees well with our previous results that the normal mode eigenvectors of the open form usually is an excellent basis for the closed and intermediate states normal modes (35).

One nice feature of normal mode refinement is that it always produces realistic protein structures, containing very few steric clashes. This is an expected result, because the elastic network Hamiltonian is expressed as constraints in pairwise distance space r_{ij}, and the low frequency normal modes tend to preserve typical bonded interactions as well as secondary structure elements by construction, while emphasizing collective motions. This also sets the limits of the model—this kind of refinement will not be good at reproducing structural transitions involving only local movements, such as loop rearrangements.

While the refinement protocol described here does not solve the problem of ligand placement, which is crucial for success of the method, it is fast enough to be incorporated in conjunction with a docking algorithm, as suggested by Zacharias *et al*. (8,40).

CONCLUSIONS

The efficiency of normal mode docking refinement without experimental data is considerably better than what we first expected, considering the absence of fitted parameters and very simple energy functions. All six protein receptor systems show cRMS improvements ranging from 0.3 to 3.2 Å (12–65%) when the ligand was superimposed based on the experimental coordinates. With errors in the ligand position, 5 of the 6 receptors are still improved by the refinement, and the last one did not deteriorate—it was left at its initial conformation. Complications such as the dimer structure of the citrate synthase turned out to be a non-issue after pre-scanning and sorting of modes. For DNA bound to proteins, 2 of the 3 test systems exhibit significant reduction of cRMS (from 8.4 to 4.8 Å, and 8.7 to 5.4 Å) while a third one stays close to the initial cRMS value of 2.3 Å.

Two of the protein targets (maltodextrin binding protein and citrate synthase) studied here have recently been used for refinement based on X-ray structure factors (35), and another two (Glutamine binding protein and HIV protease) to test molecular replacement refinement (36). With the present work, we achieve almost the same accuracy entirely without experimental guidance.

The refinement appears to be most successful when the transition between the states is well described by 2–3 low frequency normal modes, not only in the obvious sense of larger absolute cRMS improvements, but the optimization also appears to attain a large fraction of the best possible improvement for the modes used. This is not a severe limitation, since Gerstein and coworkers (33) have shown that transitions between alternate structures in the PDB usually can be described with a handful of low-frequency elastic network modes. Given the primitive energy function, the current implementation is primarily useful for low or medium resolution refinement, but with more efficient discrimination that can handle 25–30 modes it should be possible to approach high-resolution refinement.

In conclusion, normal modes seem to provide a strikingly simple but quite promising approach to refinement in general, with or without experimental data. Source code for both refinement and our newly developed fast all-atom normal mode calculation is freely available by contacting either of the authors, and a public web server implementation of the algorithms is provided at http://lorentz.immstr.pasteur.fr/.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Grégoire Lenssens for performing most of the initial calculations on DNA refinement. E.L. also wants to acknowledge École Normale Supérieure and Région Ile-de-France for kind support in the form of a Blaise Pascal fellowship. We thank Y.H. Sanejouand for kindly providing the RTB code. Funding to pay the Open Access publication charges for this article was provided by Institut Pasteur.

Conflict of interest statement. None declared.

REFERENCES

1. Gerstein,M. and Krebs,W. (1998) A database of macromolecular motions. *Nucleic Acids Res.*, **26**, 4280–4290.
2. Echols,N., Milburn,D. and Gerstein,M. (2003) MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res.*, **31**, 478–482.
3. Janin,J. (1997) The kinetics of protein–protein recognition. *Proteins*, **28**, 153–161.
4. Janin,J., Henrick,K., Moult,J., Eyck,L.T., Sternberg,M.J.E., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: a critical assessment of predicted interactions. *Proteins*, **52**, 2–9.

5. Wodak,S.J. and Mendez,R. (2004) Predictions of protein–protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.*, **14**, 242–249.
6. Claussen,H., Buning,C., Rarey,M. and Lengauer,T. (2001) FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, **308**, 377–395.
7. Cavasotto,C.N. and Abagyan,R.A. (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.*, **337**, 209–225.
8. Zacharias,M. (2004) Rapid protein–ligand docking using soft modes from Molecular Dynamics simulations to account for protein deformability: binding of FK505 to FKBP. *Proteins*, **54**, 759–767.
9. Halperin,I., Ma,B., Wolfson,H.J. and Nussinov,R. (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–443.
10. Koshland,D. (1963) Correlation of structure and function in enzyme action. *Science*, **142**, 1533–1541.
11. Gutteridge,A. and Thornton,J. (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. *J. Mol. Biol.*, **346**, 21–28.
12. Teague,S.J. (2003) Implications of protein flexibility for drug discovery. *Nature Rev. Drug Discov.*, **2**, 527–541.
13. Bradley,P., Chivian,D., Meiler,J., Misura,K.M.S., Rohl,C.A., Schief,W.R., Wedemeyer,W.J., Scheuler-Furman,O., Murphy,P., Schonbrun,J., Strauss,C.E.M. and Baker,D. (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, **53**(Suppl. 6), 457–468.
14. Rohl,C.A., Strauss,C.E.M., Misura,K.M.S. and Baker,D. (2004) Protein structure prediction using Rosetta. *Methods Enzymol.*, **383**, 66–93.
15. Skolnick,J., Zhang,Y., Arakaki,A.K., Kolinski,A., Boniecki,M., Szilagyi,A. and Kihara,D. (2003) Touchstone: a unified approach to protein structure prediction. *Proteins*, **53**(Suppl. 6), 469–479.
16. Lee,M.R., Tsai,J., Baker,D. and Kollman,P.A. (2001) Molecular dynamics in the endgame of protein structure prediction. *J. Mol. Biol.*, **313**, 417–430.
17. Lu,H. and Skolnick,J. (2003) Application of statistical potentials to protein structure refinement from low resolution *ab initio* models. *Biopolymers*, **70**, 575–584.
18. Fan,H. and Mark,A. (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci.*, **13**, 211–220.
19. Chen,J., Won,H.S., Im,W., Dyson,H.J. and Brooks,C.L.,III. (2005) Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling. *J. Biomol. NMR*, **31**, 59–64.
20. Venclovas,C., Zemla,A., Fidelis,K. and Moult,J. (2003) Assessment of progress over the CASP experiments. *Proteins*, **53**(Suppl. 6), 585–595.
21. Tramontano,A. and Morea,V. (2003) Assessment of homology-based predictions in CASP5. *Proteins*, **53**(Suppl. 6), 352–368.
22. Chen,B., Vogan,E.M., Gong,H., Skehel,J.J., Wiley,D.C. and Harrisson,S.C. (2005) Determining the structure of an unliganded and fully glycosylated SIV gp120 envelope glycoprotein. *Structure*, **13**, 197–211.
23. Chen,J.Z., Furst,J., Chapman,M.S. and Grigorieff,N. (2003) Low-resolution structure refinement in electron microscopy. *J. Struct. Biol.*, **144**, 144–151.
24. Davies,J.M., Tsuruta,H., May,A.P. and Weis,W.I. (2005) Conformational changes of p97 during nucleotide hydrolysis determined by small-angle X-Ray scattering. *Structure (Camb.)*, **13**, 183–195.
25. Brünger,A.T. (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472–475.
26. Go,N., Noguti,T. and Nishikawa,T. (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl Acad. Sci. USA*, **80**, 3696–3700.
27. Brooks,B.R. and Karplus,M. (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl Acad. Sci. USA*, **80**, 6571–6575.
28. Levitt,M., Sander,C. and Stern,P.S. (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, **181**, 423–447.
29. Tirion,M.M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.
30. 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.
31. Tama,F., Valle,M., Frank,J. and Brooks,C.L.,III (2003) Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl Acad. Sci. USA*, **100**, 9319–9323.
32. Tama,F. and Sanejouand,Y.-H. (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, **14**, 1–6.
33. Krebs,W.G., Alexandrov,V., Wilson,C.A., Echols,N., Yu,H. and Gerstein,M. (2002) Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, **48**, 682–695.
34. Hayward,S., Kitao,A. and Berendsen,H.J.C. (1997) Model-free methods of analyzing domain motions in proteins from simulations: a comparison of normal mode analysis and molecular dynamics simulation. *Proteins*, **27**, 425–437.
35. 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.
36. Suhre,K. and Sanejouand,Y.-H. (2004) On the potential of normal-mode analysis for solving difficult molecular-replacement problems. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 796–799.
37. Tama,F., Miyashita,O. and Brooks,C.L.,III (2004) Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.*, **147**, 315–326.
38. Hinsen,K., Navaza,J., Stokes,D.L. and Lacapere,J.J. (2005) Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys. J.*, **88**, 818–827.
39. Qian,B., Ortiz,A.R. and Baker,D. (2004) Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc. Natl Acad. Sci. USA*, **101**, 15346–15351.
40. Zacharias,M. and Sklenar,H. (1999) Harmonic modes as variables to approximately account for receptor flexibility in ligand-receptor docking simulations: application to DNA minor groove ligand complex. *J. Comput. Chem.*, **20**, 287–300.
41. Levitt,M. (1983) Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, **170**, 723–764.
42. Hinsen,K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–429.
43. Bahar,I., Atilgan,A.R. and Erman,B. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**, 505–515.
44. Bahar,I., Atligan,A.R. and Erman,B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, **2**, 173–181.
45. Delarue,M. and Sanejouand,Y.-H. (2002) Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J. Mol. Biol.*, **320**, 1011–1024.
46. Lehoucq,R.B., Sorensen,D.C. and Yang,C. (1998) *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, Philadelphia.
47. Durand,P., Trinquier,G. and Sanejouand,Y.-H. (1994) A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, **34**, 759–771.
48. Tama,F., Gadea,F.X., Marques,O. and Sanejouand,Y.-H. (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, **31**, 1–7.
49. Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculation. *J. Comput. Chem.*, **4**, 187–217.
50. Eisenberg,D. and McLachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
51. Koehl,P. and Delarue,M. (1994) Polar and nonpolar atomic environments in the protein core: implications for folding and binding. *Proteins*, **20**, 264–278.
52. Byrd,R.H., Lu,P. and Nocedal,J. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Stat. Comput.*, **16**, 1190–1208.
53. Bayly,C.I., Cieplak,P., Cornell,W.D. and Kollman,P.A. (1993) A well-behaved electrostatic potential based method using charge

restraints for determining atom-centered charges: the RESP model. *J. Phys. Chem.*, **97**, 10269–10280.
54. Smith,G.R. and Sternberg,M.J. (2002) Predictions of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, **12**, 28–35.
55. Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 368–373.
56. Mendez,R., Leplae,R., Maria,L. and Wodak,S.J. (2003) Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins*, **52**, 51–67.
57. Chen,R., Mintseris,J., Janin,J. and Weng,Z. (2003) A protein–protein docking benchmark. *Proteins*, **52**, 88–91.
58. Kleywegt,G.J. and Jones,T.A. (1998) Databases in protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 1119–1131.
59. Schuettelkopf,A.W. and van Aalten,D.M.F. (2004) ProDRG—a tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 1355–1363.
60. Hayward,S. (2004) Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. *J. Mol. Biol.*, **339**, 1001–1021.
61. Daidone,I., Roccatano,D. and Hayward,S. (2004) Investigating the accessibility of the closed domain conformation of citrate synthase using essential dynamics sampling. *J. Mol. Biol.*, **339**, 515–525.
62. McLachlan,A.D. (1982) Rapid comparison of protein structures. *Acta Crystallogr. A*, **38**, 871–873.
63. Marques,O. and Sanejouand,Y.-H. (1995) Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins*, **23**, 557–560.
64. Cui,Q., Li,G., Ma,J. and Karplus,M. (2004) A normal mode analysis of structural plasticity in the biomolecular motor F(1)-ATPase. *J. Mol. Biol.*, **340**, 345–372.
65. Lindahl,E., Hess,B. and van der Spoel,D. (2001) Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, **7**, 306–317.
66. Kaminski,G.A., Friesner,R.A., Tirado-Rives,J. and Jorgensen,W.L. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, **105**, 6474–6487.
67. Goh,C.-S., Milburn,D. and Gerstein,M. (2004) Conformational changes associated with protein–protein interactions. *Curr. Opin. Struct. Biol.*, **14**, 104–109.
68. Westhof,E., Dumas,P. and Moras,D. (1985) Crystallographic refinement of yeast tRNA-Asp. *J. Mol. Biol.*, **184**, 119–28.