

Structural bioinformatics

Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics

Lee-Wei Yang^{1,*}, Eran Eyal², Ivet Bahar² and Akio Kitao^{1,3}

¹University of Tokyo, Institute of Molecular and Cellular Bioscience, Tokyo, Japan, ²University of Pittsburgh, School of Medicine, Department of Computational Biology, PA, USA and ³Japan Science and Technology Agency, Core Research for Evolutional Science and Technology, Japan

Received on September 08, 2008; revised on December 11, 2008; accepted on January 08, 2009

Advance Access publication January 15, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: To efficiently analyze the 'native ensemble of conformations' accessible to proteins near their folded state and to extract essential information from observed distributions of conformations, reliable mathematical methods and computational tools are needed.

Result: Examination of 24 pairs of structures determined by both NMR and X-ray reveals that the differences in the dynamics of the same protein resolved by the two techniques can be tracked to the most robust low frequency modes elucidated by principal component analysis (PCA) of NMR models. The active sites of enzymes are found to be highly constrained in these PCA modes. Furthermore, the residues predicted to be highly immobile are shown to be evolutionarily conserved, lending support to a PCA-based identification of potential functional sites. An online tool, PCA_NEST, is designed to derive the principal modes of conformational changes from structural ensembles resolved by experiments or generated by computations.

Availability: http://ignm.cccb.pitt.edu/oPCA_Online.htm

Contact: lwy1@iam.u-tokyo.ac.jp

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

1 INTRODUCTION

Principal Component Analysis (PCA) is a widely used technique to retrieve dominant patterns and representative distributions from noisy data (Jolliffe, 2002). The idea is to map the investigated complex system from a multidimensional space to a reduced space spanned by a few principal components (PCs), thus elucidating the principal/dominant features underlying the observed data. In structural biology, the ensembles of structures accessible to a given biomolecule form such complex data, which upon suitable analyses may provide information on conformational motions (Amadei *et al.*, 1993; Garcia 1992; Hayward *et al.*, 1993; Kitao *et al.*, 1991). Studies such as the quasi-harmonic analysis of molecular dynamics (MD) (Hayward *et al.*, 1993; Kitao *et al.*, 1991; Kitao *et al.*, 1998) or Monte Carlo (MC) trajectories (Kenzaki and Kikuchi, 2008),

the essential dynamics analysis of MD runs (Amadei *et al.*, 1993), the analyses of the ensembles generated by non-Newtonian approaches (Barrett *et al.*, 2004), the definition of human Major Histocompatibility Complex supertypes (Doytchinova *et al.*, 2004) and the clustering of NMR ensembles of structural models determined by Nuclear Magnetic Resonance (NMR) (Howe, 2001) are just a few examples of PCA applications. Despite its widely recognized utility, a user-friendly, high-throughput online PCA tool that takes as input experimentally determined (e.g. X-ray, NMR) or theoretically generated (e.g. by MD, MC or comparative modeling) structural ensembles to yield functionally relevant information is lacking.

The PCA of the native ensemble of structures accessible to a given biomolecule is shown in the present study to provide information on functional dynamics and biomolecular design features. Calculations performed for NMR ensembles of structures show that, in 20 out of 24 examined proteins, the conformational differences $\Delta\{\Phi\}$ observed between the NMR structures and their X-ray counterparts are consistent with the principal modes of motion identified by PCA and supported by the physics-based anisotropic network model (ANM; Atilgan *et al.*, 2001; Eyal *et al.*, 2006) (Section 3.1). These 'robust modes', insensitive to model parameters and methods, point to directions of conformational changes that are energetically favored by the particular topology of native contacts. We argue that the molecules are designed to so move for functional reasons (Alexandrov *et al.*, 2005; Bahar and Rader, 2005; Ma, 2005; Tama and Brooks, 2006; Yang and Chng, 2008). Evidence is given here that catalytic residues of 12 examined enzymes are shown to occupy highly constrained, 'immobile', positions along the dominant PC profiles (Section 3.2). Notably, the residues identified to be highly constrained in these robust PC modes are verified to be evolutionarily conserved, further supporting the utility of the PCA of structural ensembles for identifying *potential functional sites* (PFSs).

Finally, we will present in Section 3.3 a new web resource, PCA_NEST (PCA of Native Ensembles of Structures), which uses as input ensembles of protein and/or polynucleotide structures accessible under native state conditions, and releases the corresponding principal modes of structural changes and PFSs via a user-friendly interface.

*To whom correspondence should be addressed.

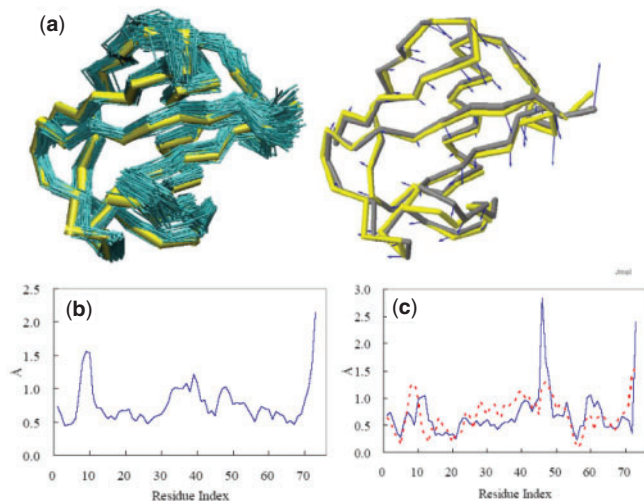


Fig. 1. An ensemble of conformations and corresponding variations in residue positions. (a) An ensemble of NMR models (teal) for ubiquitin (1xqj; Lindorff-Larsen *et al.*, 2005) and corresponding X-ray structure (1ubj; yellow), are optimally superimposed as described in Method (left). The mean structure of the NMR ensemble (gray) moves towards its X-ray counterpart (yellow) along the first PC mode, $\mathbf{v}_i^{(1)}$, indicated by the blue arrows (right). (b) Per residue RMSD, $\langle (\Delta \mathbf{q}_i)^2 \rangle^{1/2}$ [based on α -carbons; see Equation (2)] as a function of residue index i , derived from the NMR ensemble shown in (a). (c) Difference, $\Delta\{\Phi\}_i$ (blue, solid curve) in the position of residue i , between the average NMR and the X-ray structures and comparison with the first PC mode $|\mathbf{v}_i^{(1)}|$ (red, dotted curve).

2 THEORY AND METHOD

2.1 Ensembles of structures

In the present study, the term ‘ensemble’ refers to a collection of structures sampled under native state conditions by a given biomolecule, namely a protein, a polynucleotide, or a complex (Fig. 1a). The members of the ensemble are referred to as conformations, structural frames or simply ‘models’. The flow diagram in Figure 2 describes how PCA results are generated for a given set of input data. The input can be an ensemble of NMR models (>7000 such ensembles have been deposited in the PDB as of February, 2008) or a collection of X-ray structures for the same protein (or those having high sequence identity). The PDB presently contains >5987 such non-single-member families (Qi *et al.*, 2005) that share >90% sequence identity. Alternatively, MD or MC snapshots may be used as structural models. Finally, models predicted from sequence data using *ab initio* or comparative modeling methods may serve as input.

2.2 Optimal superimposition of structural frames

Our main interest is to analyze the internal motions. We remove the contributions from rigid-body translations and rotations by optimal alignment of the models using Kabsch (1976) algorithm. This algorithm gives a unique solution that minimizes the root-mean-square deviation (RMSD) between the examined models. Each model f is represented by a $3N$ -dimensional vector

$$\{\mathbf{q}^f\} = [\mathbf{q}_1^f \ \mathbf{q}_2^f \ \mathbf{q}_3^f \dots \mathbf{q}_N^f]^T \quad (1)$$

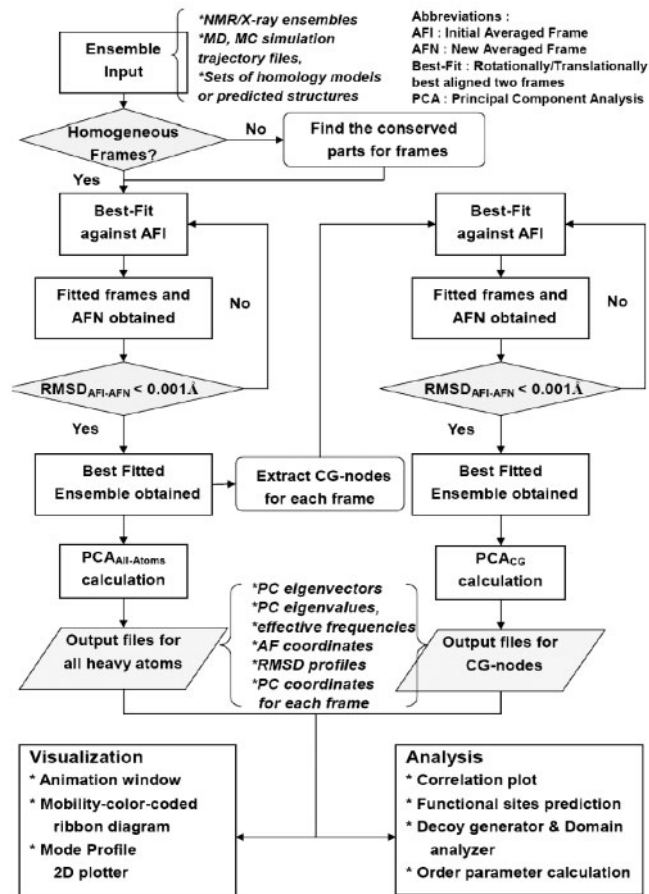


Fig. 2. Flow diagram of PCA_NEST. The ensemble of native structures is subjected to PCA at two levels, atomic (left column) and CG (right column). See the text for more details.

where $\mathbf{q}_i^f = [x_i^f \ y_i^f \ z_i^f]$ is the position vector of atom i in model f for a protein of N atoms (or residues in the coarse-grained (CG) representation where C^α atom coordinates are used for each residue position). The RMSD, $\langle (\Delta \mathbf{q}_i)^2 \rangle^{1/2}$, in the position of atom i is found from

$$\langle (\Delta \mathbf{q}_i)^2 \rangle^{1/2} = \sqrt{\frac{1}{M} \sum_{f=1}^M \|\mathbf{q}_i^f - \bar{\mathbf{q}}_i\|^2} \quad (2)$$

where $\bar{\mathbf{q}}_i$ is the position vector of the i -th atom averaged over the ensemble of M models. Figure 1b displays the $\langle (\Delta \mathbf{q}_i)^2 \rangle^{1/2}$ profile for ubiquitin, deduced from the NMR models shown in Figure 1a, for example. The RMSD averaged over all atoms is written as

$$\overline{\text{RMSD}} = \sqrt{\frac{1}{N} \sum_{i=1}^N \langle (\Delta \mathbf{q}_i)^2 \rangle} \quad (3)$$

A best-fitting procedure is iterated until the RMSD between successive average frames is converged (see flow diagram in Fig. 2).

2.3 Principal Component Analysis (PCA)

For an ensemble containing M frames ($1 \leq f \leq M$) and N heavy atoms (or CG-nodes) ($1 \leq i \leq N$) per frame, we build a *covariance matrix*

$$\mathbf{C} = \mathbf{Q}\mathbf{Q}^T \quad (4)$$

Here \mathbf{Q} is a matrix of M columns consisting each of $3N$ -dimensional vectors of N super-elements (3D vectors). The corresponding i -th super-element

$$\Delta \mathbf{q}_i^f = \frac{\mathbf{q}_i^f - \bar{\mathbf{q}}_i}{\sqrt{M-1}} = \frac{1}{\sqrt{M-1}} (\Delta x_i^f, \Delta y_i^f, \Delta z_i^f)^T \quad (5)$$

describes the deviation of atom i from its mean position $\bar{\mathbf{q}}_i$. \mathbf{C} can be decomposed as

$$\mathbf{C} = \mathbf{Q}\mathbf{Q}^T = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T = (\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{V}^T)^T (\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{V}^T) \quad (6)$$

where \mathbf{V} is the matrix of the $3N$ -dimensional eigenvectors $\mathbf{v}^{(k)}$ ($1 \leq k \leq M$) associated with the M non-zero PC modes, and $\mathbf{\Sigma}^{1/2}$ is the diagonal matrix of the square root $\xi_k^{1/2}$ of the corresponding eigenvalues, obtained from the singular value decomposition (SVD) of \mathbf{Q} . The $3N$ -elements of $\mathbf{v}^{(k)}$ describe the variations in the positions of the N nodes associated with PC mode k , each given by a 3D vector $\mathbf{v}_i^{(k)}$ ($1 \leq i \leq N$) (see the arrows in Figs 1a (right) and 3a) and the $\xi_k^{1/2}$ represents the weight of the mode k , the modes being rank-ordered as $\xi_1 \geq \xi_2 \geq \dots \geq \xi_M$. The largest contributions to conformational variations come from the top-ranking PC modes. For a system of $M < 3N$ frames, the decomposition of \mathbf{C} yields M non-zero modes. \mathbf{U} is the $M \times M$ PC coordinates matrix ($\mathbf{U}\mathbf{U}^T = \mathbf{I}$) that maps the frames in the PC space back to their original coordinate system, as may be inferred from Equation (6). The SVD of \mathbf{Q}^T is computationally faster than the eigenvalue decomposition of \mathbf{C} when $M < 3N$, which is the case for NMR models; hence the use of SVD in our web implementation.

2.4 Correlation between structural changes and PC modes

The conformational change between two conformations for a given protein is given by a $3N$ -dimensional vector $\Delta\{\Phi\}$. The elements of $\Delta\{\Phi\}$ are the x -, y - and z -components of the changes in the position vectors of the N atoms (or nodes in the CG model). The correlation, $\alpha_{P,k}$, between $\Delta\{\Phi\}$ and the k -th PC mode $\mathbf{v}^{(k)}$ is defined as

$$\alpha_{P,k} = \frac{\Delta\{\Phi\} \bullet \mathbf{v}^{(k)}}{\|\Delta\{\Phi\}\| \|\mathbf{v}^{(k)}\|} \quad (7)$$

where \bullet designates the dot product, and the bars denote the magnitudes. The cumulative contribution of M PC modes, δ_{PC} is evaluated from $\delta_{PC} = (\sum \alpha_{P,i})^{1/2}$ ($1 \leq i \leq M$). Alternatively, the summation over M' ($\leq M$) modes gives the cumulative contribution $\delta_{PC,M'}$ of M' modes (see the footnote of Table 1). Similar expressions apply to the correlations with ANM modes.

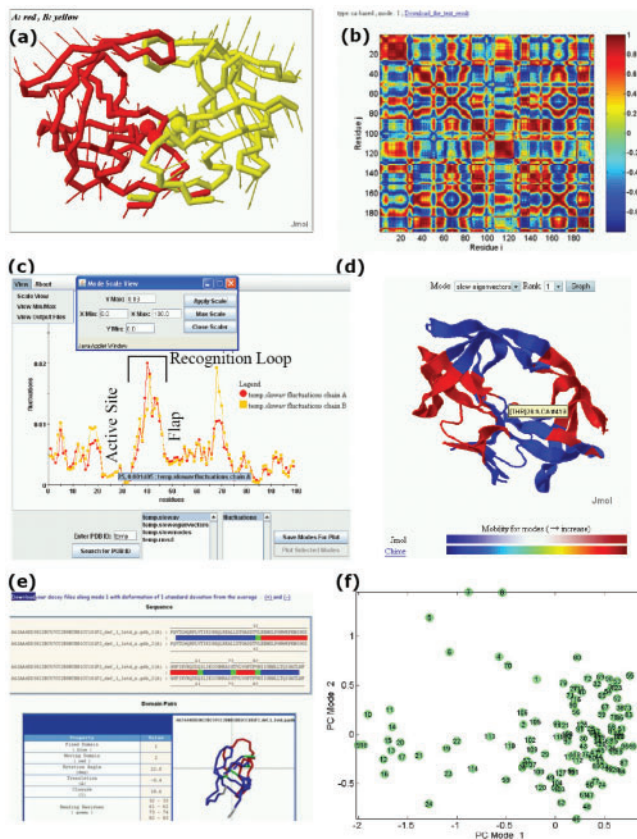


Fig. 3. PCA_NEST outputs and interfaces, illustrated for HIV-1 protease (PDB code: 1bve). (a) Conformational changes along the first PC mode. The diagram represents a snapshot from a Jmol animation describing the conformational fluctuations favored by this particular mode. The arrows refer to the elements of the first eigenvector $\mathbf{v}^{(1)}$; they describe the directions of the movements of each CG-node, and their lengths are proportional to the magnitude of the fluctuations. Active sites Asp25 and Thr26 are shown in spheres. (b) Residue-residue correlation map for the first PC mode (see Supplementary Material for details) (c) $M_{12,i}$ profiles [see Equation (8)] for monomers A and B of HIV-1 protease. The minima in the profile are further analyzed (with regard to the solvent exposure and spatial clustering of the corresponding residues) to identify PFSs. (d) Ribbon diagram illustrating the PCA_NEST-defined dynamic domains. Different colors indicate substructures subject to opposite direction (anticorrelated) movements in the first PC mode. We note that active sites are located near the interface between the dynamic domains in the center of the molecule. (e) Decoys along the PC mode 1. Decoys are created in pairs (along the positive and negative directions of the first PC), and they are automatically submitted to DynDom (Hayward and Berendsen, 1998) for dynamic domain analysis. (f) Distribution of the 128 NMR frames deposited for ubiquitin (1xqq) on the conformational subspace spanned by PC modes 1 and 2. Such projection onto a reduced space allows for clustering structures based on their dominant conformational features.

2.5 Potential Functional Sites (PFSs)

In accord with our previous study (Yang and Bahar, 2005), PFSs are found from minima in the mobility profiles

$$M_{12,i} = \sum_{k=1}^2 \xi_k \left(\mathbf{v}_i^{(k)} \bullet \mathbf{v}_i^{(k)} \right) \quad (8)$$

Table 1. Identification of the PC and ANM modes that dominate experimentally observed conformational changes^a

PDB structures				PCA results			Correlation between PCA and ANM			ANM results		
Protein	<i>M</i>	<i>N</i>	NMR/X-ray Δ	$\alpha_{P_{\max}}$	$m_{P_{\max}}$	$\delta_{PC}/\delta_{PC10}$	m_{PA1}, m_{PA2}	$\alpha_{PA1}, \alpha_{PA2}$	$\alpha_{A_{\max}}$	$m_{A_{\max}}$	$\delta_{ANM}/\delta_{ANM10}$	
Chicken cystatin	16	108	1a67/1cew	4.52	0.64	2	0.81/0.79	(2,2)(1,1)	0.84, 0.64	0.65	2	0.82/0.81
Serine protease	18	269	1ah2/1c9j	1.28	0.16	14	0.34/0.16	(3,3)(2,2)	0.60, 0.51	0.22	18	0.42/0.24
DHFR	21	163	1ao8/3dfr	0.86	0.29	1	0.57/0.47	(1,1)(2,3)	0.66, 0.56	0.34	1	0.56/0.54
Bov. phospholipase	20	123	1bvm/1bp2	1.16	0.31	2	0.70/0.59	(7,1)(1,1)	0.51, 0.34	0.43	1	0.71/0.63
Cyclophilin	12	166	1clh/1j2a	2.18	0.33	2	0.46/0.46	(1,2)(2,6)	0.86, 0.62	0.27	2	0.40/0.39
Hen lysozyme	50	129	1e8l/132l	1.61	0.33	2	0.83/0.57	(2,2)(6,1)	0.47, 0.44	0.32	1	0.71/0.50
Cytochrome C	32	129	1eky/1cpq	2.66	0.40	1	0.79/0.65	(1,2)(5,1)	0.47, 0.44	0.36	13	0.67/0.49
Pyrophosphokinase	20	158	1eq0/1dy3	4.22	0.38	3	0.77/0.73	(4,2)(1,1)	0.55, 0.50	0.54	4	0.78/0.72
Maltodextrin binding	10	370	1ezo/1anf	4.87	0.34	1	0.61/0.61	(1,2)(3,3)	0.93, 0.78	0.52	1	0.75/0.75
Ferricytochrome C	17	108	1fhh/1chh	1.03	0.32	5	0.65/0.58	(1,1)(4,5)	0.71, 0.52	0.36	5	0.67/0.62
Bovine angiogenin	10	125	1gio/1agi	1.80	0.43	1	0.59/0.59	(1,1)(1,3)	0.52, 0.46	0.37	7	0.68/0.68
Pleckstrin domain	50	106	1mpl/1btn	1.55	0.30	2	0.77/0.50	(2,1)(1,2)	0.61, 0.58	0.45	1	0.79/0.56
Fatty acid binding	10	125	1mvg/1tvq	1.38	0.26	1	0.47/0.47	(5,2)(2,1)	0.44, 0.39	0.53	2	0.60/0.60
Ompx	20	148	1orm/1qj8	4.44	0.28	1	0.57/0.49	(2,1)(1,1)	0.56, 0.52	0.44	8	0.76/0.68
Sterol carrier	20	123	1qnd/1c44	5.76	0.32	1	0.60/0.52	(2,1)(1,7)	0.75, 0.42	0.41	7	0.72/0.58
Ribonuclease F1	42	106	1rck/1fus	1.82	0.33	4	0.77/0.58	(5,4)(2,1)	0.55, 0.54	0.43	1	0.73/0.53
Frataxin orthologue	20	106	1soy/1ew4	1.28	0.33	4	0.59/0.50	(20,16)(1,3)	0.53, 0.48	0.31	3	0.48/0.42
Hemoglobin	29	147	1vre/1jff	1.43	0.25	1	0.64/0.56	(1,2)(2,1)	0.61, 0.49	0.38	1	0.69/0.63
Profilin I	19	125	2prf/1acf	2.00	0.29	6	0.58/0.49	(1,3)(2,3)	0.46, 0.43	0.18	9	0.43/0.36
Che Y protein	20	129	1eye/1e6k	1.56	0.39	1	0.71/0.67	(2,4)(6,2)	0.41, 0.41	0.29	2	0.60/0.53
H.C. variable domain	20	116	1g9e/1hcv	1.41	0.25	3	0.50/0.38	(7,11)(3,24)	0.46, 0.38	0.31	36	0.46/0.32
Ferrocyclochrome	20	111	1qn0/1wad	1.18	0.26	1	0.48/0.35	(1,1)(4,2)	0.59, 0.43	0.34	1	0.56/0.51
Fibronectin ³ F1 ¹ F1	15	89	2cku/2cg6	6.00	0.62	3	0.82/0.77	(1,1)(2,2)	0.76, 0.75	0.55	3	0.83/0.82
Ubiquitin	128	76	1xqq/1ubq	1.65	0.67	1	1.00/0.94	(1,1)(2,2)	0.75, 0.71	0.51	1	0.98/0.77
Average	26.6	140		2.40	0.35		0.65/0.56		0.61, 0.51	0.40		0.66/0.57

^aPDB codes for 24 pairs of homologous NMR and X-ray structures are listed in the column ‘NMR/X-ray’. *M* and *N* are the available number of NMR models and the number of residues for a given protein respectively. Δ is the RMSD (Å) between the average NMR structure and the corresponding X-ray structure. The $m_{P_{\max}}$ is the index of the PC mode that shows maximal correlation with $\Delta\{\Phi\}$, and $\alpha_{P_{\max}}$ is the corresponding correlation (Equation 7 with $k = m_{P_{\max}}$). δ_{PC} is the cumulative contribution of all PC modes, $\delta_{PC} = (\sum \alpha_{P_i})^{1/2} (1 \leq i \leq M)$. The δ_{PC10} is the counterpart of δ_{PC} based on the first 10 modes ($1 \leq i \leq 10$). The $m_{A_{\max}}$ is the index of the ANM mode that shows the highest correlation with $\Delta\{\Phi\}$. The corresponding correlation is $\alpha_{A_{\max}}$. Bold-faced $m_{P_{\max}}/m_{A_{\max}}$ values correspond to the 20 cases where *robust modes* describe most of observed conformational changes (see main text). δ_{ANM} is the ANM counterpart of δ_{PC} . The highest and second highest correlations observed between pairs of PC- and ANM-modes are α_{PA1} and α_{PA2} respectively. The corresponding pairs of mode numbers are m_{PA1} and m_{PA2} , respectively.

reflecting the weighted sum of the top-ranking two PC modes ($1 \leq i \leq N$). Minima of interest are based on groups of five sequential nodes, with the center being lower than the first two neighbors on each side. We also require that the $M_{12,i}$ value of the PFSs to be $< 10\%$ of that exhibited by the most mobile node in the structure. Figure 3c illustrates the $M_{12,i}$ profile for HIV-1 protease.

2.6 Minima-screening algorithm to rank-score the PFSs

Not all the PFSs (or minima) have equal functional importance. We further score those PFSs based on their relative solvent accessibility and spatial clustering properties (see Supplementary Material for details). The algorithm, referred to as minima-screening algorithm, depends merely on inter-residue contact topology. The top-scoring eight (or less) PFSs for the examined proteins are listed at the bottom of the ‘Potential Functional Sites’ page of PCA_NEST.

2.7 Residue conservation

Residue conservation scores are assigned using the ConSurf server (Landau *et al.*, 2005) and used for examining the degree of sequence conservation at the PFSs. Conservation scores range from 1 (highly diverse) to 9 (highly conserved).

2.8 Residue correlations and dynamic domains

The residue–residue correlation map generated by PCA_NEST is a color-coded covariance matrix of $N \times N$ dimension, the *ij*-th element of which is defined as

$$\hat{C}_{Nij}^{(k)} = \frac{\mathbf{v}_i^{(k)} \bullet \mathbf{v}_j^{(k)}}{(\|\mathbf{v}_i^{(k)}\| \|\mathbf{v}_j^{(k)}\|)} \quad (9)$$

The covariance values vary in the range $[-1, 1]$. Positive values refer to pairs of residues moving in the same direction along mode

k , and negative values refer to those moving in opposite directions. Moreover, The plus and minus signs of the values in the most dominant eigenvector, obtained from SVD of $\hat{C}_{Nij}^{(k)}$, classifies the nodes into two motional subsets, which can be used to define the dynamic domains (see Supplementary Material).

3 RESULTS

The results are organized in three subsections. First, we show how the top-ranking PC modes satisfactorily describe the structural difference between X-ray- and NMR-characterized conformers. Second, we focus on particular enzymes to demonstrate how the PCA-predicted PFSs correlate with catalytic residues and/or evolutionarily conserved sites. Third, we present an overview of the newly designed resource, PCA_NEST, and its visualization and comparative analyses features.

3.1 Robust PC modes explain the conformational change between NMR and X-ray conformers

The top-ranking PC modes derived from a given NMR ensemble convey information on the most probable changes in conformations sampled by the examined molecule. We will explore the existence of a correlation between these conformational variations and a physical-model (ANM) derived dynamics. Also, due to the ease of moving along soft PC modes, we will examine to what extent the differences $\Delta\{\Phi\}$ in the NMR and X-ray conformers relate to reconfigurations along these PC modes.

3.1.1 Do structural differences $\Delta\{\Phi\}$ between NMR and X-ray conformers correlate with PC modes derived from NMR models?

The results for the examined 24 proteins are presented in Table 1, organized in three sets of columns: PCA results, ANM results and their correlations. We identified in each case the PC mode ($\mathbf{m}_{P_{\max}}$) that exhibits the highest correlation ($\alpha_{P_{\max}}$) with the experimentally observed conformational change $\Delta\{\Phi\}$ [see Equation (6)]. Notably, a correlation coefficient of 0.35 is achieved by this single mode, averaged over all examined cases (last row in Table 1), while in some cases the value can be above 0.6 (e.g. 1a67, 2cku and 1xqq). The single mode that exhibits the *highest correlation is the first mode* in 11/24 of the cases, the second mode in 5/24 and the third in 3/24 cases. Column 8 lists the cumulative correlation (δ_{PC}) resulting from all M modes. The corresponding average over all proteins is $\langle \delta_{PC} \rangle = 0.65$ between the PC modes and $\Delta\{\Phi\}$. Note that this correlation is achieved by a small subset ($\sim 6\%$) of the accessible $3N-6$ modes (see the average M and N in the last row of Table 1; $\langle M \rangle / (3 \langle N \rangle - 6) \approx 6\%$).

3.1.2 How does the physics-based model ANM describe $\Delta\{\Phi\}$?

CG normal mode analysis with ANM (Eyal et al., 2006; Yang and Chng 2008), using a 15Å cutoff, shows that it is possible to identify a single ANM mode for each protein, which also yields a comparable average correlation ($\langle \alpha_{A_{\max}} \rangle = 0.40$) with $\Delta\{\Phi\}$. That ‘single mode’ is again amongst the lowest frequency modes (first ANM mode in 9/24 cases, second in 4/24), although this tendency is not as pronounced as the PC modes. The cumulative effect of the slowest M modes (out of a total of $3N-6$ ANM modes) amounts to approximately the same average value $\langle \delta_{ANM} \rangle = 0.67$.

3.1.3 The modes that underlie the observed $\Delta\{\Phi\}$ are shared between PC and ANM modes A much higher correlation is observed between the PC modes and the ANM modes. Columns 8 and 9 list the mode numbers \mathbf{m}_{PA1} and \mathbf{m}_{PA2} corresponding to the most correlated PC and ANM mode-pairs, and their respective correlations, α_{PA1} and α_{PA2} . The averages over the 24 proteins yield $\langle \alpha_{PA1} \rangle = 0.61$ and $\langle \alpha_{PA2} \rangle = 0.51$. The lack of an even higher correlation is attributed to the anharmonic effects present in the slow PC modes (Kitao et al., 1998). Interestingly, the PC-ANM mode-pairs that exhibit the highest and second-highest correlations coincide with the particular modes $\mathbf{m}_{P_{\max}}$ and/or $\mathbf{m}_{A_{\max}}$ that best describe $\Delta\{\Phi\}$ in 20 out of 24 cases (PDB codes are written in boldface). In other words, *the modes that are closely reproduced by the two different methods correlate with $\Delta\{\Phi\}$* . For example, the seventh PC mode and the first ANM mode of bovine phospholipase (PDB code: 1bvm) are the most correlated pair while the first PC and the first ANM modes are the second most correlated. ANM mode 1 and PC modes 1 and 7 are therefore the most *robust* modes, confirmed by both PCA and ANM. Here, the first ANM mode explains $\Delta\{\Phi\}$ the most, among any other modes. The conformational changes that take place along these ‘soft’ modes incur, by definition, the lowest energy increase, and therefore lie along *the easiest paths of deformation*.

Growing evidence has shown that biologically relevant motions are accessed via a few structure-encoded low-frequency modes (Nicolay and Sanejouand, 2006). However, to determine exactly which mode predominantly accounts for such motions remains a difficult task. Also, the above analysis shows that the correlation between a single PC mode and $\Delta\{\Phi\}$ is usually low. Here, we propose to consolidate the NMR-derived PC modes by ANM analysis to identify the modes that significantly account for observed changes. The identified robust modes can be advantageously used in template-based structure refinement (Kidera and Gö 1992; Tama et al., 2004).

3.2 Potential Functional Sites (PFSs)

From the set of 64 NMR protein ensembles previously examined (Yang et al., 2007), we selected the subset of enzymes with known Enzyme Class (E.C.) numbers (from Enzyme Structures Database) and catalytic sites (from Catalytic Site Atlas (CSA; Porter et al., 2004) or literature). The $M_{12,i}$ profiles were examined for the resulting set of 12 enzymes. These profiles show that the catalytic sites exhibit a strong tendency to be *positioned at the minima* (Fig. 4), with an average mobility score of 2.9% over all enzymes (see Table 2; 100% for peak residues in the normalized $M_{12,i}$ profile; 0% for immobile residues), consistent with our earlier physics-based results (Yang and Bahar, 2005).

We also note that the profiles contain many (local) minima, and not all of them contain catalytic residues. Toward a further screening of the PFSs for selecting the most probable functional sites, we score-ranked them based on their relative solvent accessibility and spatial clustering properties (Bartlett et al., 2002; Gutteridge et al., 2003). The algorithm, referred to as *minima-screening algorithm*, depends on inter-residue contact topology exclusively, irrespective of residue type or evolutionary information such as conservation scores (see details in the Supplementary Material). We were able to locate the majority of catalytic residues within the top-ranking 6–8

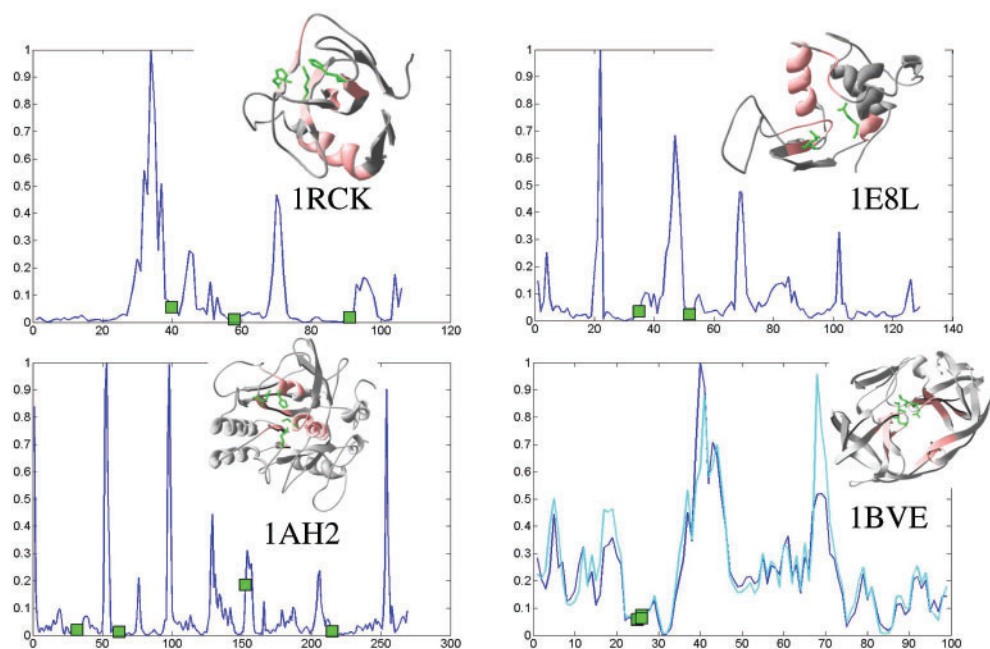


Fig. 4. Fluctuation profiles induced by dominant PC modes. Four examples are displayed, which illustrate how the enzyme active sites (green squares) lie at the minima of the normalized $M_{12,i}$ profiles (ordinate) based on PC modes 1 and 2, drawn a function of residue index (abscissa). The ribbon diagrams in the insets show the literature- and CSA-reported (Porter *et al.*, 2004) active sites for ribonuclease F1 (1rck), hen lysozyme (1e8l), serine protease (1ah2) and HIV-1 protease (1bve) in green ball-and-stick representation. The structural regions predicted to be most critically positioned are colored pink. The two curves in the last panel (1bve) refer to the profiles of the two monomers.

PFSs (see the ranks in parentheses in the column 8 in Table 2) by using this algorithm.

The aforementioned observations give insights into criteria underlying the design of active sites: (i) *enzymes tend to locate their key catalytic power in mechanically constrained (low-mobility), but solvent accessible, positions* and (ii) *the catalytic sites spatially cluster, irrespective of their separation along the sequence*.

As a further verification of the functional significance of PCA_NEST-predicted top-ranking PFSs (other than the catalytic residues), we examined their conservation properties. An average ConSurf conservation score (Con_Score; last column of Table 2) of 7.6 was obtained for top-ranking PFSs as opposed to an understandably higher score, 8.5, for catalytic residues and 5.4 for an average residue (A). If we separate the top-ranking PFSs into two sets, the ‘true positives’ that overlap with the catalytic sites (CAT) and the FPs that do not match any reported catalytic residues but are predicted by PCA_NEST to be potentially significant, and reexamine the Con_Score for FPs only, surprisingly, we found an average Con_Score of 7.5! This result lends strong support to the biological significance of PCA_NEST-predicted top-ranking PFSs.

3.3 PCA_NEST: a web tool for analyzing ensembles of structures

PCA_NEST uses as input structural ensembles (proteins/peptides, RNA/DNA or their complexes), either experimentally characterized or computationally generated (see for example Fig. S5 in the Supplementary Material). The examined ensemble is first subject to an iterative best-fitting (see Fig. 2 and Table S1 in Supplementary Material) and then to PCA that renders a series of output data.

The *raw output data* include two sets of results: CG and full atomic. CG nodes are centered at the C^α positions of the amino acids in proteins and at the phosphorus, sugar carbon C4* and base carbon C2 positions of the nucleotides in DNA/RNA structures (Yang *et al.*, 2006). Each set of results contains the following information (Fig. S4): (i) general specifications of the examined set (e.g. the number of nodes or atoms, number of frames/models, average RMSD between models), (ii) predicted directions and relative sizes of the principal changes in conformation, given by the PC eigenvectors $v^{(k)}$ for $1 \leq k \leq M$ (see Section 2), (iii) the PC eigenvalues ξ_k , in ascending order, starting from zero, which rescale the overall size of conformational change in each mode, (iv) newly reoriented models obtained after iterative best-fitting, (v) The average structural coordinates as well the particular model (among the members of the original ensemble) which is structurally closest to the computed mean structure, (vi) model coordinates in the space spanned by the PCs, $\sum^{1/2} U^T$ (Section 2) and (vii) effective frequencies of PC modes (Equation S1).

Visualization modules that use animations, 2D-plots or 3D diagrams to comprehensively present these data are available for assisting viewers as shown in Figure 3. *Analyzers* that utilize the aforementioned raw data for advanced analyses include four modules: (i) *PFSs finder* that lists the PFSs based on the two top-ranking PC modes and sorted after the minima-screening algorithms described in the Supplementary Material, (ii) *cross-correlation maps* (see Fig. 3b and Supplementary Material), (iii) *decoy Generator and DynDom Analyzer* (Fig. 3e) and (iv) *computed order parameters* that permit a fast evaluation of amide bond order parameters (S^2) based on the best-fitted structural ensemble or selected conformers

Table 2. Identification of PFSs based on the PCA of NMR ensembles

NMR	Name	M	N	E.C.No.	CATH ^a	Top-ranking PFSs ^b	Catalytic residues ^c	$\langle M_{12,cat} \rangle$ $\times 100^d$	$\langle \text{Con_Score} \rangle^e$ CAT/PFS/FP/All
1AO8	Dihydrofolate reductase	21	163	1.5.1.3	3.40.430.10	<i>A97,V40,F3,T34,A6,D26,Y46,H28</i>	<i>L19,D26(6)</i>	4.4	7.5/8.0/7.8/5.1
1Y8B	Malate synthase G	10	723	2.3.3.9	1.10.1860.10	<i>L117,V535,S537,D270,F336,L298,E108,3.20.20.360; G296</i>	<i>D270(4),E272(4),R338(5),D631</i>	3.7	9.0/8.5/8.4/5.5
1AX3 ^f	Glucose permease 2a domain	16	162	2.7.1.69	2.70.70.10	<i>P130,F84,I19,A44,S54,V150,F34</i>	<i>T86,H68,H83(2),G85(2)</i>	1.1	9.0/6.7/6.5/5.2
1BVM	Phospholipase A2	20	123	3.1.1.4	1.20.90.10	<i>R100,H48,F5,C45,M8,C29,I104,I95</i>	<i>G30(6),H48(2),D99(1)</i>	1.8	9.0/8.7/8.3/5.4
1T3K	Cdc25-like tyrosine phosphatase	20	132	3.1.3.48	N/A	<i>V69,I26,D28,C82,G108,R78,L58,M3</i>	<i>H71(1),C72,R78(6)</i>	1.7	6.7/7.0/7.3/5.0
1GIO ^f	Angiogenin	10	125	3.1.27.-	3.10.130.10	<i>F116,F46,T12,V105,P113,I81,F120,I57</i>	<i>H14(3),K41,H115(1),F116(1)</i>	0.4	8.0/7.7/8.0/5.4
1RCK	Ribonuclease F1	42	106	3.1.27.3	3.10.450.30	<i>V78,E58,I89,A20,Y42,V16,F48</i>	<i>H40(5),E58(2),H91(3)</i>	2.7	9.0/6.2/3.7/5.5
1E8L	Hen lysozyme	50	129	3.2.1.17	1.10.530.10	<i>N57,V92,K96,V99,N106,Y53,K33</i>	<i>E35(7),D52(6)</i>	2.9	9.0/6.1/6.6/5.7
1AH2	Serine protease	18	269	3.4.21.62	3.40.50.200	<i>P219,G223,L122,A149,V30,S215,A226,G63</i>	<i>D32(5),H62(8),N153,S215(6)</i>	5.8	9.0/8.1/7.8/5.2
1BVE	HIV-1 protease	28	198	3.4.23.16;	2.40.70.10	<i>B:184,A:L24,A:185,B:L24,A:G49,B:G49,3.1.26.4</i>	<i>D25(2),T26(2)</i>	6.3	9.0/8.7/8.5/5.8
						<i>A:T31,B:T31</i>			
1AYK ^f	Collagenase	30	169	3.4.24.7	3.40.390.10	<i>E119,G121,F38,I34,W41,L126,A84,V30</i>	<i>E119(1)</i>	0.7	9.0/7.3/8.4/5.4
1NMV	Mitotic regulator hPin1	10	111	5.2.1.8	2.20.70.10;	<i>I159,L61,G120,S58,A107,,I156,Y92,F134</i>	<i>H59(2),C113,H157(1)</i>	3.0	7.3/7.7/9.0/5.0
Average					3.10.50.40			2.9	8.5/7.6/7.5/5.4

^aThe ensembles listed here are of distinct structure folds and topologies, based on their CATH number.

^bPredicted by PCA_NEST (top scoring eight minima centers identified in the $M_{12,i}$ profile of the enzyme).

^cExperimentally determined, accessible from the literature or from manually compiled databases (e.g., CSA; Porter *et al.*, 2004). Residues in boldface are in top-ranking minima predicted by the minima-screening algorithm (rank is indicated in parentheses); those written in italic lie at the minima of the $M_{12,i}$ curve.

^dPercent mobility of catalytic residues $\langle M_{12,CAT} \rangle$ averaged over all catalytic residues for each protein.

^eAverage sequence conservation score evaluated using ConSurf ('9' for most conserved residues; '1' for the least; Landau *et al.*, 2005) for catalytic residues (CAT), for PCA-predicted potential functional sites (PFS), for the false positives predictions (FP) and for all residues (All). The score varies from 1 (most diverse) to 9 (most conserved).

^fTruncated proteins. The respective first 13, 5 and 6 N-terminal residues of 1AX3, 1GIO and 1AYK were excluded.

that suitably represent the conformational space sampled near native state (see Fig. S6 in the Supplementary Material).

4 DISCUSSION AND CONCLUSIONS

4.1 Native ensembles provide information on intrinsic dynamics

Overall our results suggest that the experimentally determined ensembles of conformers do exhibit features characteristic of the particular proteins' dynamics under native state conditions. Notably, the RMSD between NMR models and the difference $\Delta\{\Phi\}$ between NMR and X-ray structures yield an average correlation of 0.70 (Table S1 in Supplementary Material). This observation is consistent with the recent finding that a handful of experimentally known conformers, or simply collections of structural homologs, satisfactorily reproduce NMR order parameters (Best *et al.*, 2006). We further argue that the inferred motions are structure-induced properties 'intrinsic' to the examined proteins (Bahar *et al.*, 2007; Gutteridge and Thornton, 2005; Lange *et al.*, 2008), i.e. a dynamic signature of a given molecular topology, which can be described by physics-based analytical models (e.g. ANM). The dynamic significance of the native ensembles of structures was also suggested by our earlier study where the deviations from mean positions of residues within NMR ensembles were shown to be consistent with the equilibrium fluctuations predicted by the ANM (Yang *et al.*, 2007). It should be noted that the NMR ensembles in

the PDB are modeled by combining experimentally determined restraints (typically, inter-proton distances extracted from NOESY data and peptide torsional angles from measurements of three-bond J-couplings) with an empirical force field containing contributions from local and non-local interactions, including non-linear effects; and the deviations among the NMR models reflect the uncertainties in the atomic coordinates optimally assessed by this elaborate procedure. ANM modes, on the other hand, represent unique solutions, analytically derived based on a single representative structure; the predicted equilibrium dynamics is fully controlled by the $3N \times 3N$ Hessian matrix where *uniform* force constants are adopted for elastic springs that are assumed to connect spatially close residues. In view of these methods/approximations, the observed correlation of 0.61 between ANM and NMR-PCA modes (Table 1) essentially point to the dominant effect of the common component, inter-residue (ANM) or inter-atomic (NMR) contact topology inherently defining the robust modes of motions, in both cases.

4.2 X-ray structure is a conformational state along the dominant PCs derived from NMR models

Despite the structural (Andrec *et al.*, 2007) and dynamic (Yang *et al.*, 2007) differences in biomolecules independently determined by X-ray and NMR, these differences are shown here to be nonetheless explained by the M PC modes, with a correlation of 0.66 averaged over all examined protein pairs. A significant fraction of this

correlation is usually contributed by a *single* mode (see Table 1, $\alpha_{P_{\max}}$ and $\alpha_{A_{\max}}$ values). Despite of the difficulty in predicting the conformational changes between NMR and X-ray structures, which originates from the nature of the structures as being characterized in distinct environments (temperature, buffer/ligand used and crystal forms, etc.), we have shown that the observed structural differences essentially reflect the changes in conformation along the most *robust modes* reproduced independently and consistently by PCA and ANM. Also, the positional covariance derived by PCA_NEST can serve as a guidance in iterative structure refinement (Tama *et al.*, 2004) or ligand-induced conformation changes based on linear response theory (Ikeguchi *et al.*, 2005).

4.3 The role of conformational dynamics in residue conservation

It has been known for more than half a century that amino acid side chains are critical to the functions of enzymes. Hence, it comes as a surprise that the PCA results based on CG-nodes (C^α atoms only) succeed in closely predicting the position of the enzyme active site. The high stability of the active site in the *collective* motions (low PC modes) suggests that enzymes are evolutionarily inclined to locate their functional residues at sites that enable a precise positioning necessary for the chemical reaction. These same sites coincide with, or closely neighbor, hinge sites corresponding to collective motions, which further suggest a possible coupling between local catalytic events to global (e.g. domain) motions.

Another interesting result was to observe the high level of conservation of the top-ranking PFSs (other than the known catalytic sites; see columns 7 and 10 (last) of Table 2) inferred from the PCA of native ensembles of structures. Overall, the PCA of native ensembles of proteins provides hypotheses about important sites to be further tested and established by experiments and/or more detailed computations.

4.4 When PCA_NEST does not work?

In general, if the ensemble of models does not provide an adequate sampling of the conformational space accessible to the examined protein, the results from PCA_NEST may be equally biased. Improper sampling is usually, but not necessarily, due to a small number of models. The Supplementary Material Section V presents and discusses the results from PCA using different ensembles of NMR models, applied to the prediction of order parameters from NMR relaxation experiments performed for ubiquitin. The ensembles containing smaller numbers of models lead to impaired predictions. However, a small number (as few as 12) of models may be sufficient to restore accuracy, provided that these models adequately sample the subspace spanned the first few PC modes.

4.5 Closing remarks

PCA_NEST provides a gateway for the PCA of collections of structures, determined experimentally or theoretically, for a given biomolecule, with supporting intuitive interfaces to best present various types of results. Our analyses herein exemplify how the web-based tool can be utilized to gain biologically and physically meaningful results. Notably, PCA_NEST offers standardized procedures to process ensembles such as iterative best-fitting, coarse-graining, PFS identification and other calculations

that allow for comparing results on a common ground and drawing unbiased conclusions based on statistically meaningful data.

The robust modes shared by PCA and ANM indicate the directions of conformational changes which incur a minimal energy cost as the molecule travels in the neighborhood of its global energy minimum. The fact that such energetically favored reconfigurations correlate with the conformational mechanisms involved in biological function suggests that structures are designed to move so due to *functional reasons*. Particular structures survive in evolution by performing biologically correct movements, i.e. they set in motion certain domains/loops and arrest the movements in other parts (e.g. global hinges). Evidence is presented here that catalytic residues usually occupy positions that are highly constrained, almost 'static', in these robust modes that are conserved/selected. This is in support of the concept of *structural predisposition* of enzymes to enable the functional coupling between conformational mechanics and catalytic activity (Yang and Bahar, 2005). Finally, the residues identified to be immobile in these essential PC modes (both the global hinges and other sites) are noted to be evolutionarily conserved, closing the loop {structure \rightarrow dynamics \rightarrow function \rightarrow structure}, i.e. *structure* induces *dynamics*, which enable *function*, which, in turn, selects/conserves the *structure*.

ACKNOWLEDGEMENTS

We thank Drs Guoying Qi and Steven Hayward for giving comments and providing crucial technical support on the PCA_NEST interlink with DynDom. We also thank Mr Christopher Jon Jursa for the help with the 2D plotter applets, Mr Shun Sakuraba for useful discussion on fundamentals on Kabsch's algorithm and Dr Choon-Peng Chng for providing MD trajectory files of flagellin. We thank Drs Scott Showalter and Robert Best for providing experimentally obtained (model-free based) order parameters of ubiquitin and HIV-1 protease, respectively. L.W.Y. also gives heartfelt thanks to Dr Kazuhiro Takemura for insightful discussions on order parameter calculations.

Funding: Japan Society for the Promotion of Science (fellowship and grant-in-aid to L.W.Y.); National Institutes of Health (grants 5R01GM086238-02 and 5R01LM007994-06 to I.B.).

Conflict of Interest: none declared.

REFERENCES

- Alexandrov, V. *et al.* (2005) Normal modes for predicting protein motions: a comprehensive database assessment and associated web tool. *Protein Sci.*, **14**, 633–643.
- Amadei, A. *et al.* (1993) Essential dynamics of proteins. *Proteins*, **17**, 412–425.
- Andrec, M. *et al.* (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins*, **69**, 449–465.
- Atilgan, A.R. *et al.* (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**, 505–515.
- Bahar, I. and Rader, A.J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 1–7
- Bahar, I. *et al.* (2007) Intrinsic dynamics of enzymes in the unbound state, and relation to allosteric regulation. *Curr. Opin. Struct. Biol.*, **17**, 633–640.
- Barrett, C.P. *et al.* (2004) Dynamite: a simple way to gain insight into protein motions. *Acta Crystallogr.*, **D60**, 2280–2287.
- Bartlett, G. *et al.* (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

- Best,R.B. et al. (2006) Relation between native ensembles and experimental structures of proteins. *PNAS*, **103**, 10901–10906.
- Doytchinova,I.A. et al. (2004) Identifying human MHC supertypes using bioinformatic methods. *J. Immunol.*, **172**, 4314–4323.
- Eyal,E. et al. (2006) Anisotropic network model: systematic evaluation and a new interface. *Bioinformatics*, **22**, 2619–2627.
- Garcia,A.E. (1992) Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, **68**, 2696–2699.
- Gutteridge,A. and Thornton,J. (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. *J. Mol. Biol.*, **346**, 21–28.
- Gutteridge,A. et al. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330**, 719–734.
- Hayward,S. et al. (1993) Effect of solvent on collective motions in globular proteins. *J. Mol. Biol.*, **234**, 1207–1217.
- Hayward,S. and Berendsen,H.J.C. (1998) Systematic Analysis of Domain Motions in Proteins from Conformational Change: New Results on Citrate Synthase and T4 Lysozyme. *Proteins*, **30**, 144–154
- Howe,P.W.A. (2001) Principal components analysis of protein structure ensembles calculated using NMR data. *J. Biomol. NMR*, **20**, 61–70.
- Ikeguchi,M. et al. (2005) Protein structural change upon ligand binding: linear response theory. *Phys. Rev. Lett.*, **94**, 078102-1–078102-4.
- Jolliffe,I.T. (2002) *Principal Component Analysis*. 2nd edn. Springer, New York, XXIX, p. 487.
- Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, **A32**, 922–923.
- Kenzaki,H. and Kikuchi,M. (2008) Free-energy landscape of kinesin by a realistic lattice model. *Proteins*, **71**, 389–395.
- Kidera,A. and Gö,N. (1992) Normal mode refinement: crystallographic refinement of protein dynamic structure. I. Theory and test by simulated diffraction data. *J. Mol. Biol.*, **225**, 457–475.
- Kitao,A. et al. (1991) The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulation of melittin in water and in vacuum. *Chem. Phys.*, **158**, 447–472.
- Kitao,A. et al. (1998) Energy landscape of a native protein: jumping-among-minima model. *Proteins*, **33**, 496–517.
- Landau,M. et al. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.
- Lange,O.F. et al. (2008) Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science*, **320**, 1471–1475.
- Lindorff-Larsen,K. et al. (2005) Simultaneous determination of protein structure and dynamics. *Nature*, **433**, 128–132.
- Ma,J. (2005) Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, **13**, 373–380.
- Nicolay,S. and Sanejouand,Y.H. (2006) Functional modes of proteins are among the most robust. *Phys. Rev. Lett.*, **96**, 078104-1–078104-4.
- Porter,C.T. et al. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
- Qi,G. et al. (2005) A comprehensive and non-redundant database of protein domain movements. *Bioinformatics*, **21**, 2832–2838.
- 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. 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.
- Yang,L.-W. and Bahar,I. (2005) Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure*, **13**, 893–904.
- Yang,L.-W. and Chng,C.P. (2008) Coarse-grained models reveal functional dynamics - I. elastic network models – theories, comparisons and perspectives. *Bioinform. Biol. Insights*, **2**, 25–45.
- Yang,L.-W. et al. (2006) oGNM: a protein dynamics online calculation engine using the Gaussian Network Model. *Nucleic Acids Res.*, **34**, W24–W31.
- Yang,L.-W. et al. (2007) Insights into equilibrium dynamics of proteins from comparison of NMR and X-Ray data with computational predictions. *Structure*, **21**, 741–749.