Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 18: PROTEIN DYNAMICS AND PCA

Bakan, A., & Bahar, I. (2009). *PNAS*, *106*(34), 14349–54.

THE INTRINSIC DYNAMICS OF ENZYMES PLAYS A DOMINANT ROLE IN DETERMINING THE STRUCTURAL CHANGES INDUCED UPON INHIBITOR BINDING.

ABSTRACT

Motivation: The conformational flexibility of target proteins continues to be a major challenge in accurate modeling of protein–inhibitor interactions.

Problem: A fundamental issue, yet to be clarified, is whether the observed conformational changes are controlled by the protein or induced by the inhibitor.

Solution Approach: The wealth of structural data for target proteins in the presence of different ligands now permits us to make a critical assessment of the balance between these two effects in selecting the bound forms. We focused on three widely studied drug targets, HIV-1 reverse transcriptase, p38 MAP kinase, and cyclin-dependent kinase 2. A total of 292 structures determined for these enzymes in the presence of different inhibitors and unbound form permitted us to perform an extensive comparative analysis of the conformational space accessed upon ligand binding, and its relation to the intrinsic dynamics before ligand binding as predicted by elastic network model analysis.

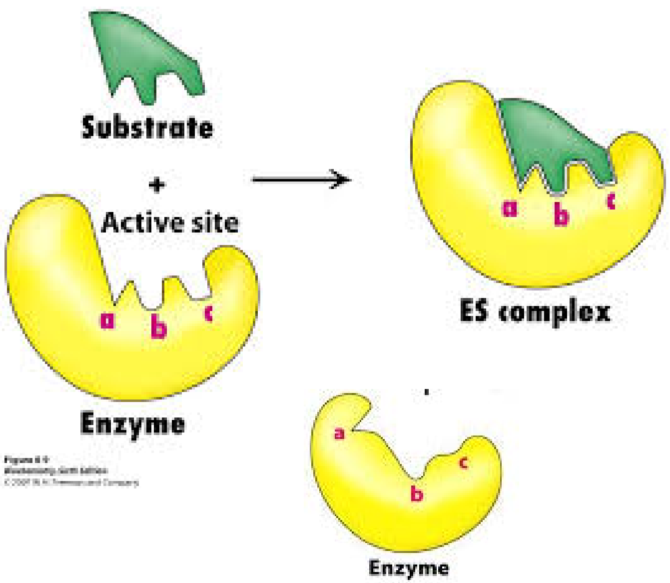
Results: Our results show that the ligand selects the conformer that best matches its structural and dynamic properties among the conformers intrinsically accessible to the protein in the unliganded form. The results suggest that simple but robust rules encoded in the protein structure play a dominant role in predefining the mechanisms of ligand binding, which may be advantageously exploited in designing inhibitors.

PROBLEM

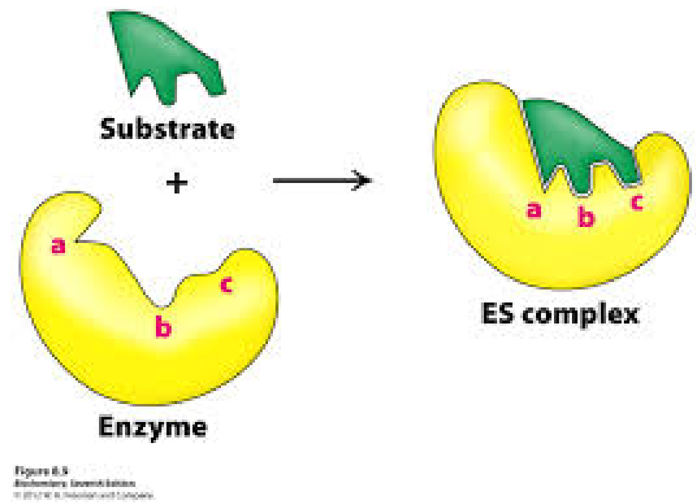
Are conformational changes controlled by

1. the protein native dynamics or
2. induced by the inhibitor

Protein native dynamics



Induced fit model



STRUCTURAL DYNAMICS OBSERVED VS THEORY

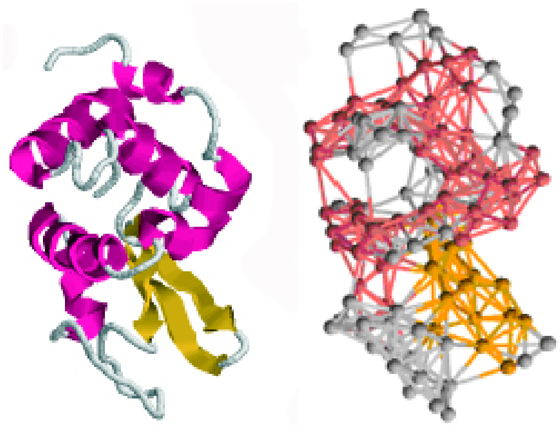
- Functional variations in structures observed experimentally
- Using NMR models

- Expected from a physical theory and method based on native contact topology.
- Using anisotropic network model (ANM)

Top-ranking PCA modes

In all three proteins, show how the ensembles of conformations observed in experiments (in the presence of different ligands) may be explained by the intrinsic dynamics of the protein (in the absence of ligands).

Top ranking ANM modes



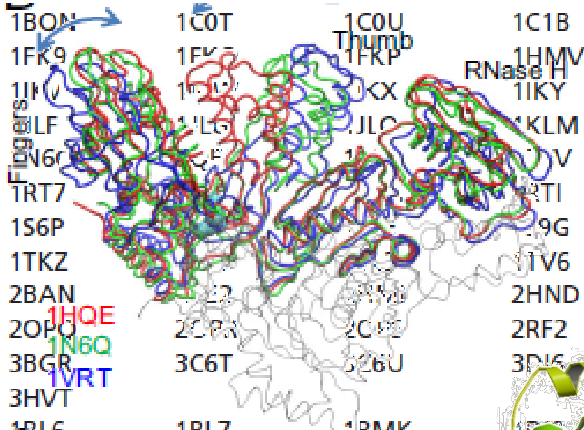
http://ignm.ccbb.pitt.edu/Dynomics.htm

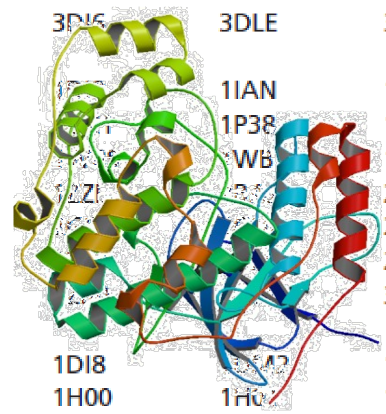
DATASET

HIV-1 reverse transcriptase (**HIV-1 RT**)

p38 MAP kinase

cyclin-dependent kinase 2.





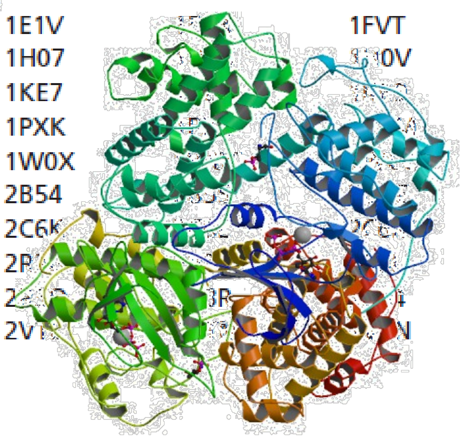


Table S1. Datasets: HIV-1 RT*, p38 MAP kinase†, and Cdk2‡ structures

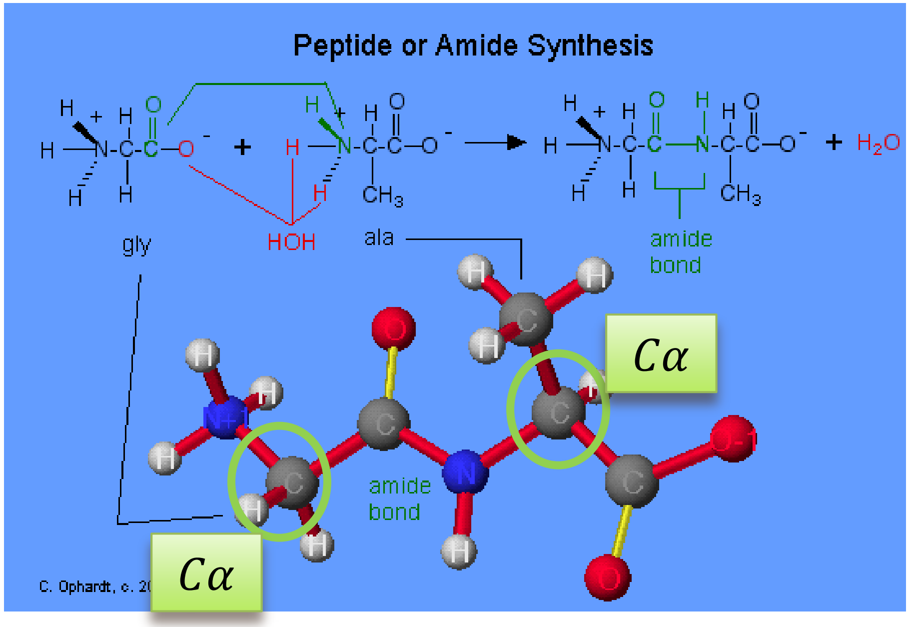
RT	1BQM	1BQN	1C0T	1C0U	1C1B	1C1C	1DLO	1DTQ	1DTT	1EET
	1EP4	1FK9	[illegible]	1FKP	1HMV	1HNI	1HNV	1HPZ	1HQE	1HQU
	1HYS	[illegible]	[illegible]	[illegible]	1IKY	1J5O	1JKH	1JLA	1JLB	1JLC
	1JLE	[illegible]	[illegible]	[illegible]	1KLM	1LW0	1LW2	1LWC	1LWE	1LWF
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1RT1	1RT2	1RT3	1RT4	1RT5
	1RT6	1RT7	[illegible]	[illegible]	[illegible]	1RTJ	1S1T	1S1U	1S1V	1S1W
	1S1X	1S6P	[illegible]	[illegible]	[illegible]	1SUQ	1SV5	1T03	1T05	1TKT
	1TKX	1TKZ	[illegible]	[illegible]	[illegible]	1TVR	1UWB	1VRT	1VRU	2B5J
	2B6A	2BAN	[illegible]	[illegible]	2HND	2HNY	2HNZ	2I5J	2IAJ	2IC3
	2OPP	2OPQ	[illegible]	[illegible]	2RF2	2RKI	2VG5	2VG6	2VG7	2ZD1
	2ZE2	3BGR	3C6T	[illegible]	[illegible]	3DLE	3DLG	3DM2	3DMJ	3DOK
	3DOL	3HVT								
p38	1A9U	1BL6	1BL7	1BMK	[illegible]	1IAN	1KV1	1KV2	1LEW	1LEZ
	1M7Q	1OUK	1OUY	1OVE	[illegible]	1P38	1R39	1R3C	1W7H	1W82
	1W83	1W84	1WBN	1WBO	[illegible]	[illegible]	1WBV	1WBW	1WFC	1YQJ
	1YW2	1YWR	1ZYJ	1ZZ2	[illegible]	[illegible]	2BAK	2BAL	2BAQ	2EWA
	2FSL	2FSM	2FSO	2FST	[illegible]	[illegible]	2GHM	2GTM	2GTN	2I0H
	2NPQ	2OKR	2OZA	2P5A	[illegible]	[illegible]	2PTO	2PUU	2PV5	2PV8
	2QD9	2RG5	2RG6	2ZAZ	[illegible]	[illegible]	3BV2	3BV3	3BX5	3C5U
	3CG2	3CTQ	3D7Z	3D83						
Cdk2	1AQ1	1B38	1B39	1CKP	1DI8	[illegible]	1E1V	[illegible]	1FVT	1G5S
	1GIH	1GII	1GIJ	1GZ8	1H00	[illegible]	1H07	[illegible]	[illegible]	1H0W
	1HCK	1HCL	1JSV	1JVP	1KE5	1KE6	1KE7	[illegible]	[illegible]	1OIQ
	1OIR	1OIT	1P2A	1PW2	1PXI	1PXJ	1PXK	[illegible]	[illegible]	1PXN
	1PXP	1PYE	1R78	1URW	1V1K	1VYZ	1W0X	[illegible]	[illegible]	1Y8Y
	1Y91	1YKR	2A0C	2A4L	2B52	2B53	2B54	[illegible]	[illegible]	2BHH
	2BTR	2BTS	2C5Y	2C68	2C69	2C6I	[illegible]	[illegible]	[illegible]	2C6O
	2CLX	2DS1	2DUV	2EXM	2J9M	2R3F	[illegible]	[illegible]	[illegible]	2R3J
	2R3K	2R3L	2R3M	2R3N	2R3O	2R3P	[illegible]	[illegible]	[illegible]	2UZN
	2UZO	2V0D	2VTA	2VTH	2VTI	2VTJ	[illegible]	[illegible]	[illegible]	2VTO
	2VTP	2VTQ	2VTR	2VTS	2VV9	2W06				

STRUCTURAL DATA ANALYSIS PROCEDURE : STEP 1

The experimental structural data are analyzed as follows:

1. The ensemble of structures are superimposed using the Kabsch algorithm in an iterative procedure (see *SI Text*),

- mean positions $\langle \boldsymbol{R}_i \rangle \ [\langle x_i \rangle \langle y_i \rangle \langle z_i \rangle]^T$ are determined for α-carbons $1 \leq i \leq \ N$ (or those with known coordinates),



ITERATIVE SUPERIMPOSITION METHOD

Iterative Procedure for Optimal Superimposition of Ensembles of Structures.

(i) Each structure in the ensemble is first pairwise superimposed onto a randomly selected reference structure

(ii) An average set of coordinates is calculated for the superimposed set obtained in *i*, referred to as the ‘‘average model,’’

(iii) all structures are pairwise superimposed on the newly generated ‘average model’

(iv) steps *ii-iii* are repeated until the average model generated in two successive iterations changes by less than the threshold RMSD of 0.001 Å.

STEP 2

2. Departures from their mean positions,

$$\Delta \boldsymbol{R}_i^s = [\Delta x_i^s \; \Delta y_i^s \; \Delta z_i^s]^T \qquad \text{where } \Delta x_i^s = x_i^s - \langle x_i \rangle$$

are organized in a ***3N*-dimensional deformation vector**

$$\Delta \boldsymbol{R}^s \text{ where } (\Delta \boldsymbol{R}^s)^{\boldsymbol{T}} = [(\Delta \boldsymbol{R}_1^s)^{\boldsymbol{T}} (\Delta \boldsymbol{R}_2^s)^{\boldsymbol{T}} \ldots (\Delta \boldsymbol{R}_N^s)^{\boldsymbol{T}}],$$

for all structures, *S*, in the dataset;

and their cross-correlations, averaged over the entire set are combined in a *3N* x *3N* covariance matrix **C**

STEP 3

3. C is diagonalized to determine the principal modes of structural variations, ***p****(i)*, observed in experiments.

The principal modes (*m* of them, for an ensemble of $m < 3N - 6$ structures) are **rank-ordered**:

PCA mode 1 (PC1), $\boldsymbol{p}^{(1)}$, refers to the direction of maximal variance, succeeded by PC2, etc.

Of interest is to view the distribution of dataset structures in the subspace spanned by PC1 and PC2, which permit us to discriminate, or cluster, the conformations based on their most distinctive structural similarities and/or dissimilarities.

CALCULATION OF THE COVARIANCE MATRIX

The covariance matrix **C** is a *3N* X *3N* matrix for a protein of *N* residues (with known coordinates), which may be written in terms of a set of *N* x *N* submatrices $\boldsymbol{C}^{ij}$ $(1 \leq i, j \leq N)$, each of size 3x3, given by

$$\mathbf{C}^{(ij)} = \begin{bmatrix} \langle \Delta x_i \Delta x_j \rangle & \langle \Delta x_i \Delta y_j \rangle & \langle \Delta x_i \Delta z_j \rangle \\ \langle \Delta y_i \Delta x_j \rangle & \langle \Delta y_i \Delta y_j \rangle & \langle \Delta y_i \Delta z_j \rangle \\ \langle \Delta z_i \Delta x_j \rangle & \langle \Delta z_i \Delta y_j \rangle & \langle \Delta z_i \Delta z_j \rangle \end{bmatrix}$$

$\langle \Delta x_i \Delta y_j \rangle$ represents the cross correlation between (*i*) the X-component of the fluctuation vector $\boldsymbol{\Delta R}_i^S$ representing the departure of the *i*th residue from its mean position, and (*ii*) the Y-component of $\boldsymbol{\Delta R}_j^S$ representing the departure of the *j*th residue from its mean position, averaged over all structures $(1 \leq s \leq m)$ in the examined dataset

OBTAINING PRINCIPAL MODES

- Decomposing the covariance matrix **C** for each dataset as

$$\boldsymbol{C}\boldsymbol{p}^{(i)} = \sigma_i \boldsymbol{p}^{(i)}$$

where $\boldsymbol{p}^{(i)}$and σ_i, are the respective *i*th eigenvalue and eigenvector of **C**, σ_1 corresponding to the largest variance component.

- The **fractional contribution** of $\boldsymbol{p}^{(i)}$ to structural variance in the dataset is given by

$$f_i = \sigma_i / \sum_j \sigma_j$$

where the summation is performed over all *m* components.

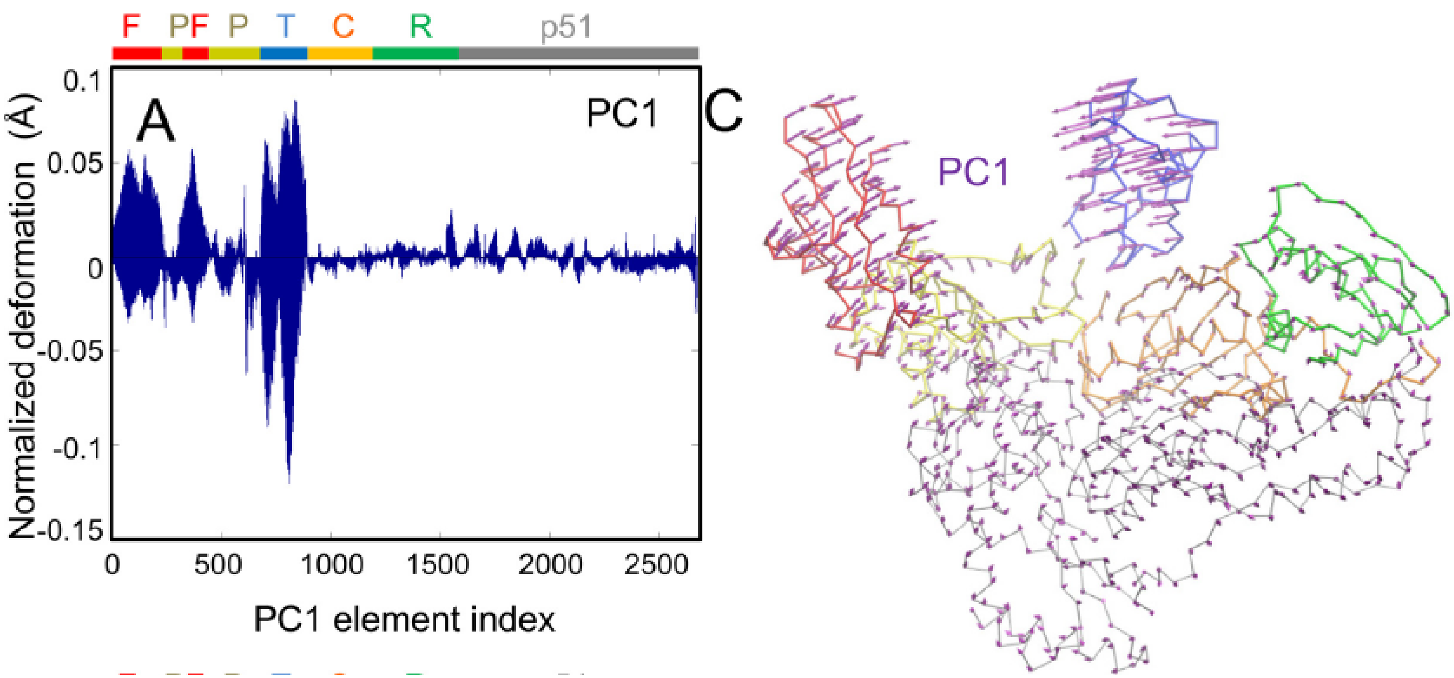
- The **square displacement** of the *k*th residue along ***p****(1)* and ***p****(2)* (or PC1 and PC2) is

$$(\Delta \boldsymbol{R}_k)^2_{1\le i\le 2} = tr\left\{\left[\sum_{i=1}^{2} \sigma_i \boldsymbol{p}^{(i)} \boldsymbol{p}^{(i)T}\right]_{\boldsymbol{kk}}\right\}$$

where the subscript *kk* denotes the *kth* diagonal element (a 3X3 matrix) of the 3*N*x3*N* matrix enclosed in square brackets.

PROJECTION OF CONFORMATIONS ONTO THE SUBSPACE SPANNED BY THE PCS

The projection of a given conformational change *R*s onto *p(i)*.
The points in the Figs represent the projection of each structure s onto PC1 and PC2. In the extreme case of (*R*s)T perfectly aligned along *p(i)*,

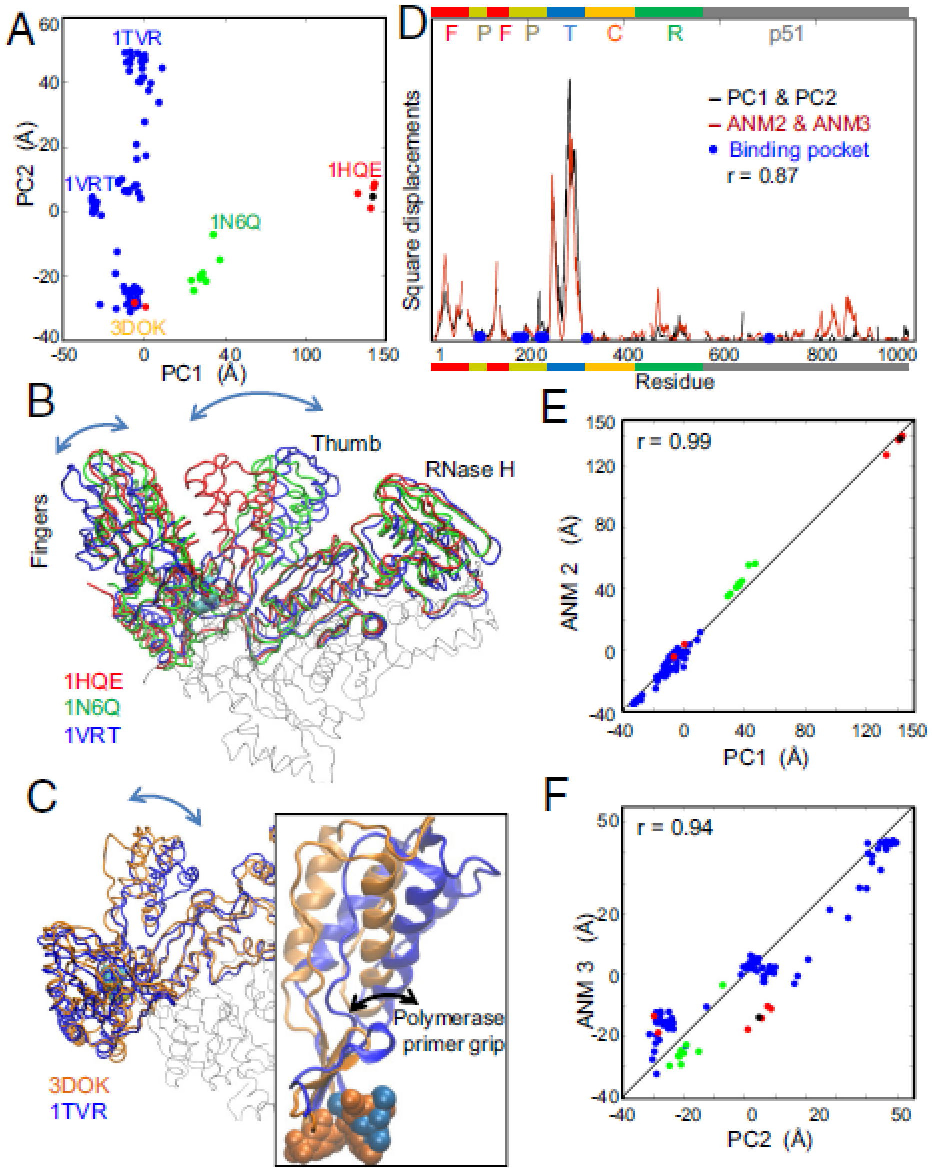


RESULTS FOR HIV-1 RT

Projection of 6 unliganded (red), 97 NNRTI bound (blue), 8 dsDNA/RNA-bound (green), and 1 ATP-bound (black) RT structures onto PC1 and PC2

PC1: The most distinctive feature is the large movement of the thumb and anti-correlated displacements of the fingers and thumb

PC2 describes the out-of-plane fluctuations of the thumb

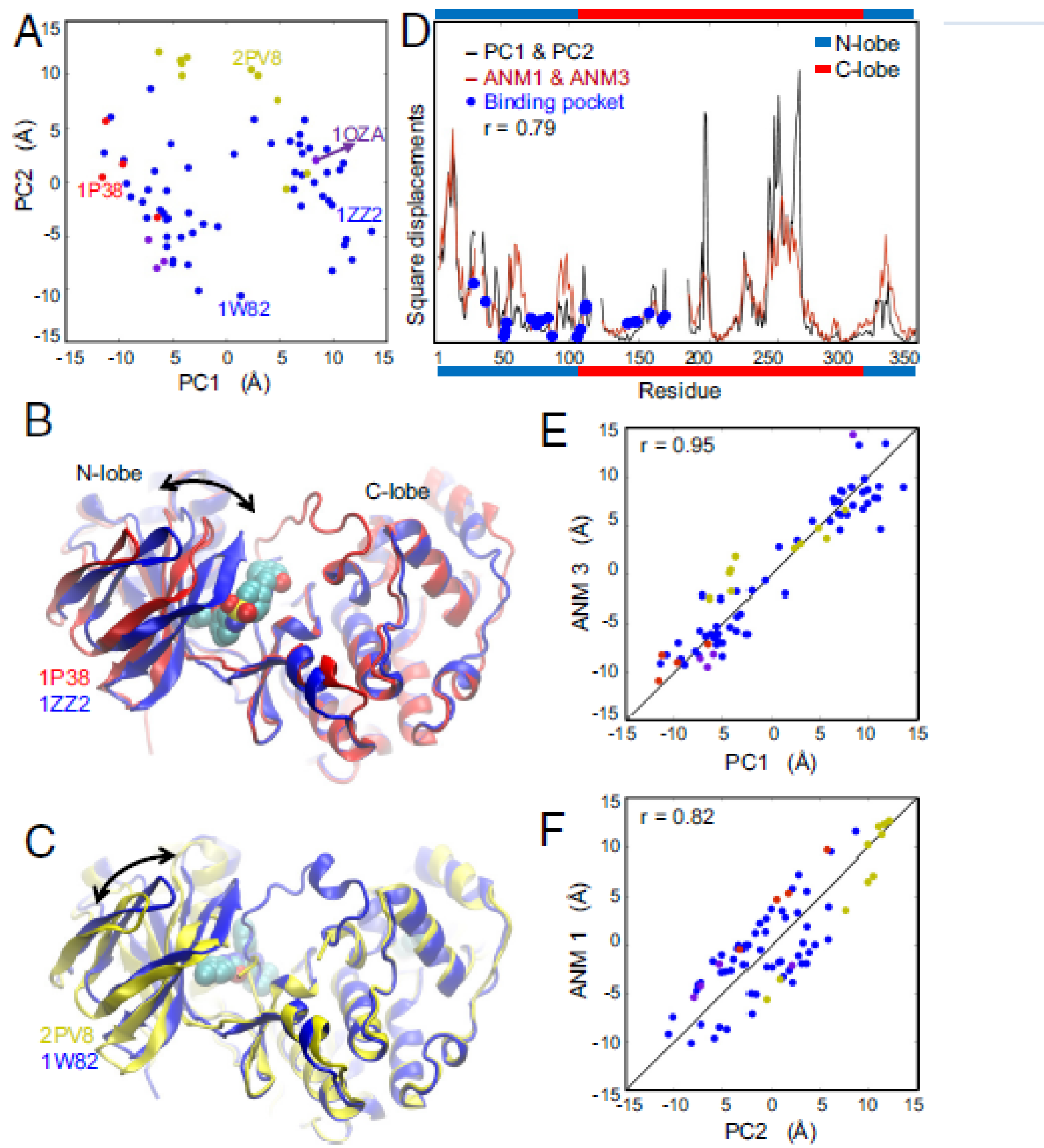


RESULTS FOR P38 MAP KINASE

Projection of 4 unliganded (red dots), 56 inhibitor-bound (blue), 10 glucoside-bound (yellow), and 4 peptide-bound (violet) p38 structures onto PC1 and PC2.

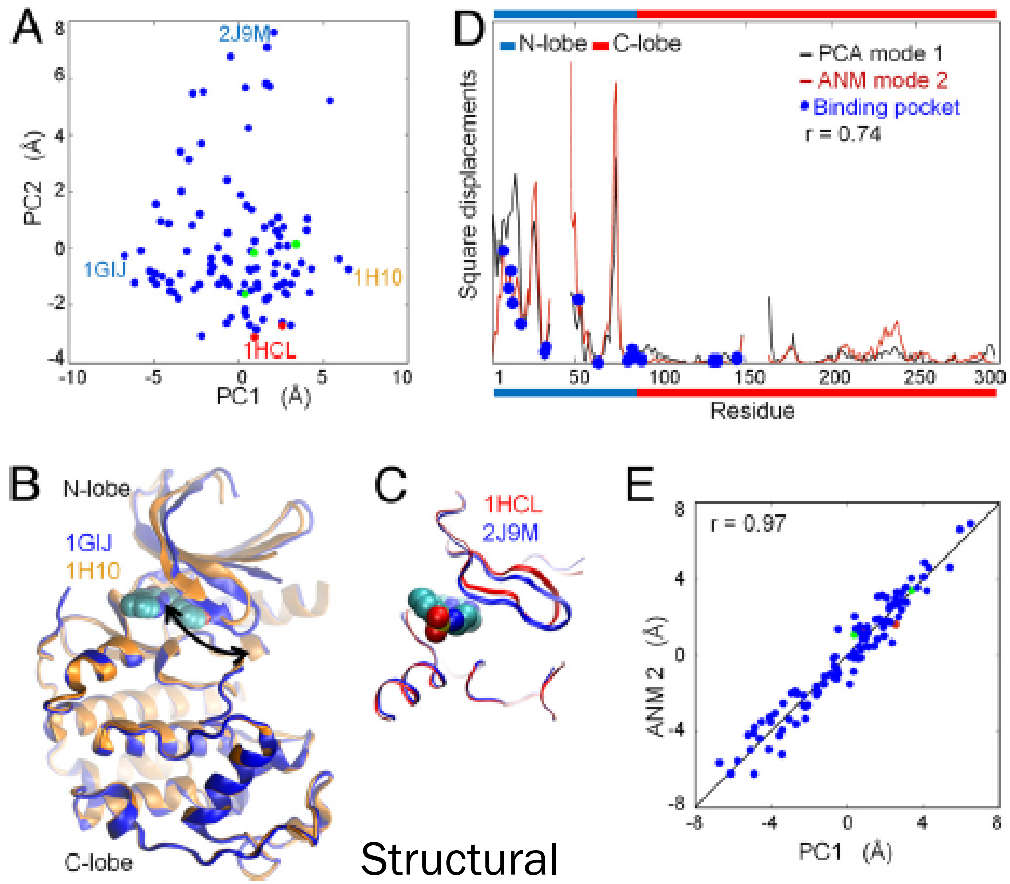
Structural variation along PC1

Structural variation along PC2



RESULTS FOR CDK2

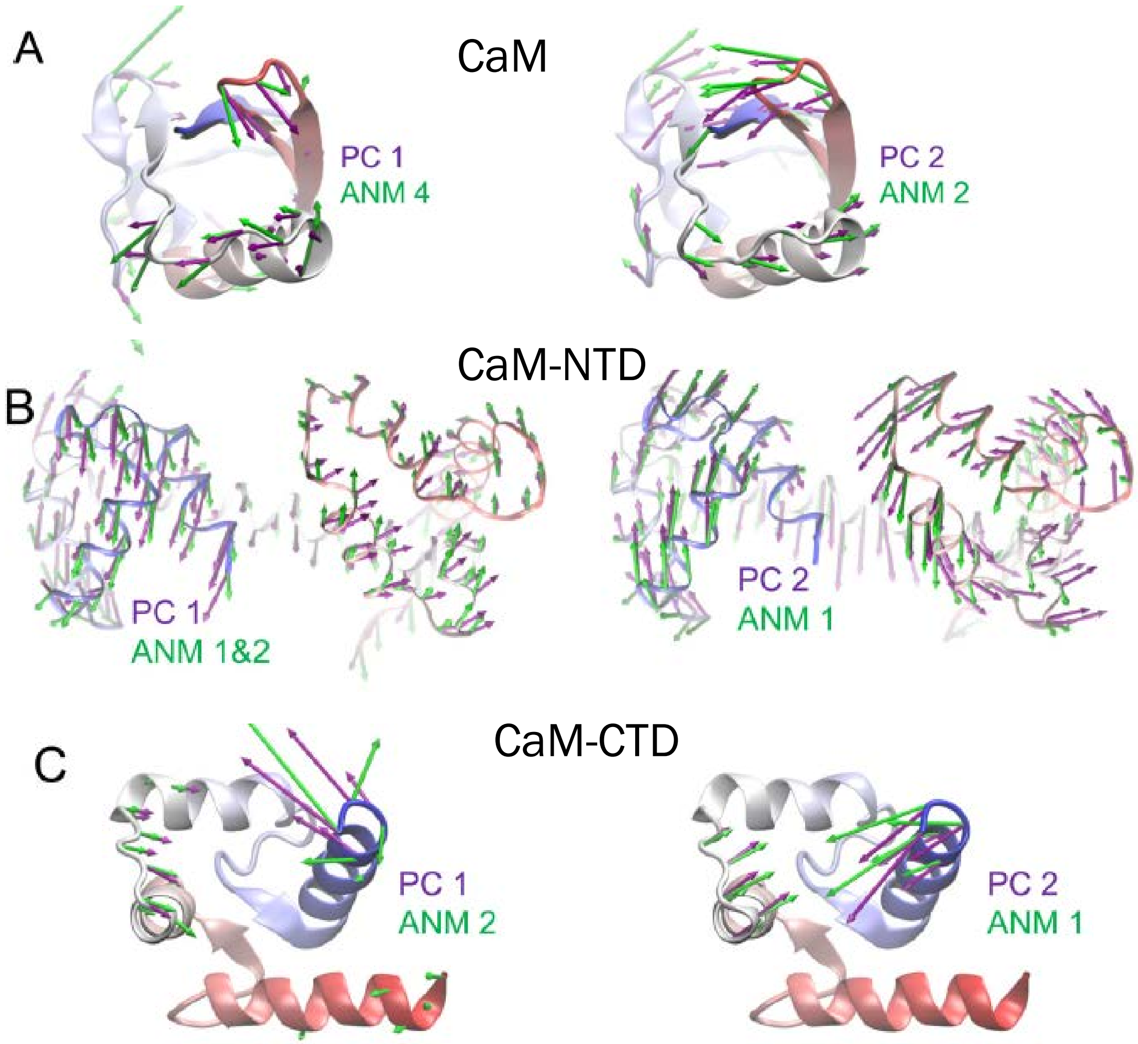
Projection of 2 unliganded (red), 3 ATP-bound (green), and 101 inhibitor-bound (blue) Cdk2 structures onto PC1 and PC2.



Structural variation along PC1

Structural variation along PC2

RESULTS CONT



CONCLUSION

- presented a detailed analysis of conformational changes experimentally observed for three enzymes upon binding a broad range of ligands, and those predicted by simple physics-based models based on their native fold contact topology
- First principal mode of structural change, PC1, observed in experiments exhibits a correlation of 0.78 0.1 with a top ranking mode (ANM1-ANM3) intrinsically preferred by the unliganded protein.
- The three PCs describe between 50% (Cdk2) and 80% (RT) of the structural variance observed in the datasets of enzymes.

Maisuradze, G. G., Liwo, A., & Scheraga, H. a. (2009).
Journal of molecular biology, *385*(1), 312–29

PRINCIPAL COMPONENT ANALYSIS FOR PROTEIN FOLDING DYNAMICS.

ABSTRACT

Protein folding is considered here by studying the dynamics of the folding of the triple β-strand WW domain from the Formin-binding protein 28. Starting from the unfolded state and ending either in the native or nonnative conformational states, trajectories are generated with the coarse-grained united residue (UNRES) force field. The effectiveness of principal components analysis (PCA), an already established mathematical technique for finding global, correlated motions in atomic simulations of proteins, is evaluated here for coarse-grained trajectories. The problems related to PCA and their solutions are discussed. The folding and nonfolding of proteins are examined with free-energy landscapes. Detailed analyses of many folding and nonfolding trajectories at different temperatures show that PCA is very efficient for characterizing the general folding and nonfolding features of proteins. It is shown that the first principal component captures and describes in detail the dynamics of a system. Anomalous diffusion in the folding/nonfolding dynamics is examined by the mean-square displacement (MSD) and the fractional diffusion and fractional kinetic equations. The collisionless (or ballistic) behavior of a polypeptide undergoing Brownian motion along the first few principal components is accounted for.

DATA

Data set: various fold/unfold states of small 37-residue protein, triple β-strand WW domain from the Formin-binding protein 28 (FBP28) (1E0L in Protein Data Bank notation1).

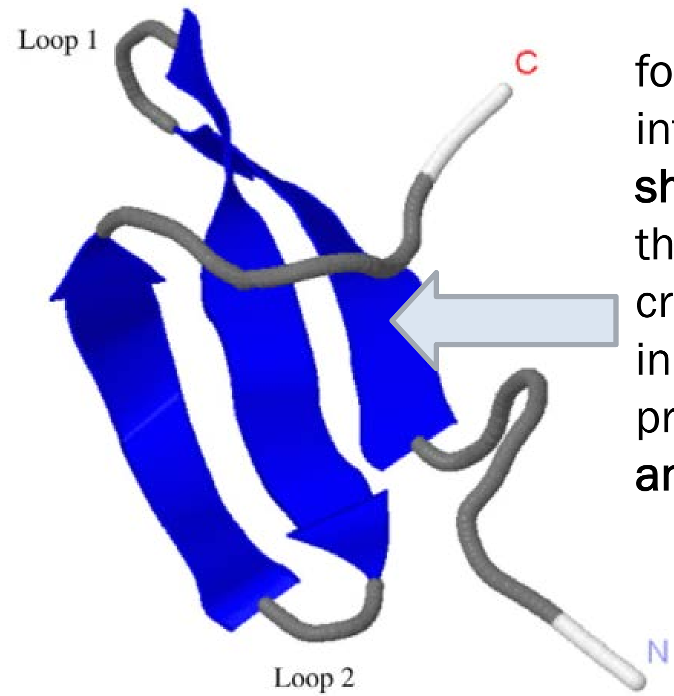


Fig. 1. Experimental NMR structure[1] of the triple β-strand WW domain from FBP28 (1E0L).

formation of intermolecular **β-sheets** is thought to be a crucial event in the initiation and propagation of **amyloid diseases**

PRINCIPLE COMPONENTS ANALYSIS INPUT

Model: using **coarse-grained models** to carry out **molecular dynamics** simulations employing physics-based **united-residue (UNRES)** force field generating trajectories starting from the unfolded state to native state at different **temperatures**

Principal component analysis

The PCA method is based on the covariance matrix with elements C_{ij} for coordinates i and j

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \quad (3)$$

where $x_1, \ldots, x_{3N}$ are the mass-weighted Cartesian coordinates of an N-particle system and $\langle \rangle$ is the average over all instantaneous structures sampled during the simulations.

The symmetric $3N \times 3N$ matrix **C** can be diagonalized with an orthonormal transformation matrix **R**:

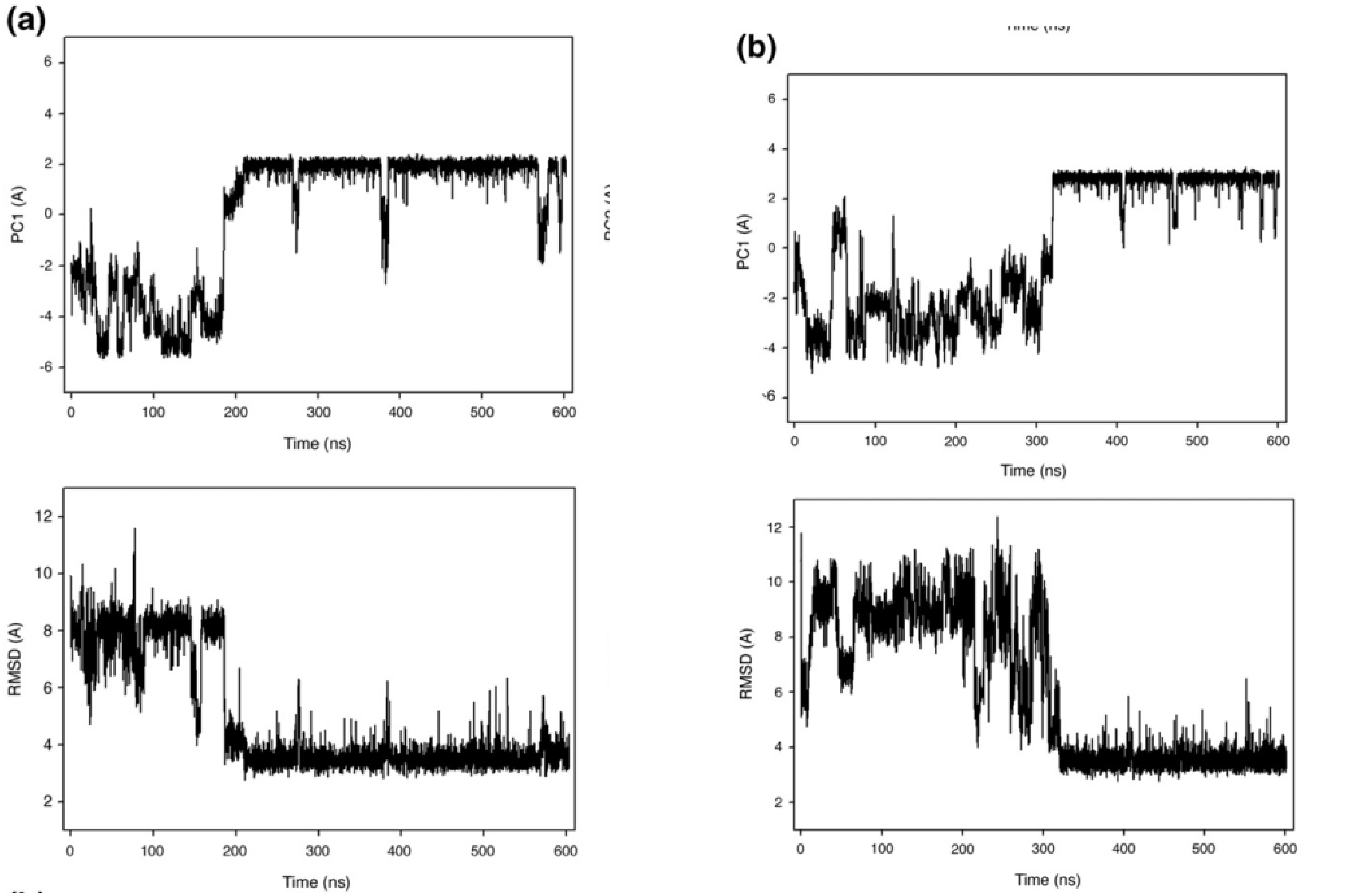
$$\mathbf{R}^{\mathbf{T}}\mathbf{C}\mathbf{R} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \lambda_{3N}), \tag{4}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{3N}$ are the eigenvalues, and $\mathbf{R}^{\mathbf{T}}$ is the transpose of **R**. The columns of **R** are the eigenvectors, or the principal modes; the trajectory can be projected onto the eigenvectors to give the principal components $q_i(t)$, $i = 1, \ldots, 3N$:

$$\mathbf{q} = \mathbf{R}^{\mathbf{T}}(\mathbf{x}(t) - \langle\mathbf{x}\rangle) \tag{5}$$

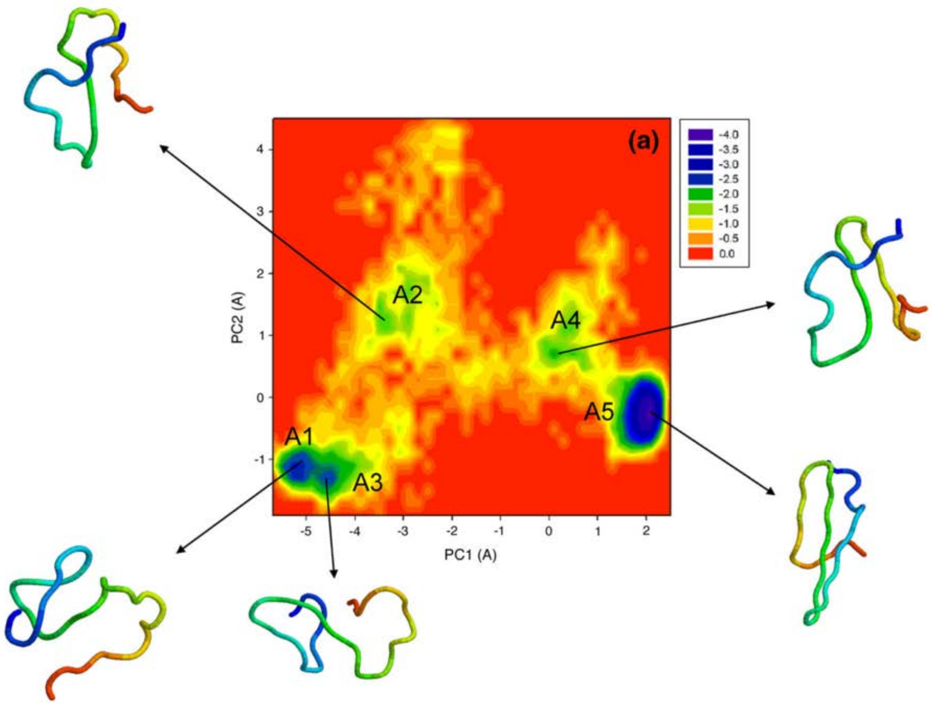
The eigenvalue λ_i is the mean-square fluctuation in the direction of the principal mode. The first few PCs typically describe collective, global motions of the system, with the first PC containing the largest mean-square fluctuation.

RESULTS

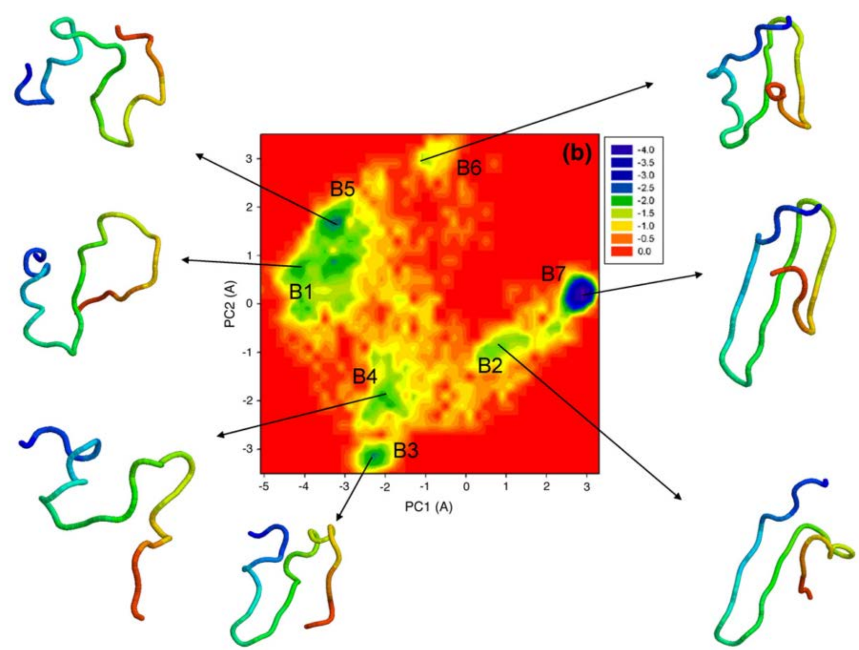


The first principal component and rmsd from the native structure of fast- (a) and slow- (b) MD trajectories at 330 K for 1E0L.

RESULTS



Free-energy landscapes (in kilocalories per mole) for 1E0L with representative structures at the minima of fast-(a) and slow-(b) MD trajectories at 330K. A1–A5, and B1–B7 are the minima on the free-energy landscapes.



Yang, L.-W., Eyal, E., Bahar, I., & Kitao, A. (2009).
Bioinformatics, 25(5), 606–14

PRINCIPAL COMPONENT ANALYSIS OF NATIVE ENSEMBLES OF BIOMOLECULAR STRUCTURES (PCA_NEST): INSIGHTS INTO FUNCTIONAL DYNAMICS.

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.ccbb.pitt.edu/oPCA_Online.htm

PRINCIPAL COMPONENT ANALYSIS

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

$$\mathbf{C} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} \tag{4}$$

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

$$\Delta\mathbf{q}_i^f = \frac{\mathbf{q}_i^f - \overline{\mathbf{q}_i}}{\sqrt{M-1}} = \frac{1}{\sqrt{M-1}}\left(\Delta x_i^f, \Delta y_i^f, \Delta z_i^f\right)^{\mathrm{T}} \tag{5}$$

describes the deviation of atom i from its mean position $\overline{\mathbf{q}_i}$.

PRINCIPAL COMPONENT ANALYSIS CONT

$$\mathbf{C} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{V}\,\Sigma\,\mathbf{V}^{\mathrm{T}} = \left(\mathbf{U}\,\Sigma^{1/2}\mathbf{V}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{U}\,\Sigma^{1/2}\mathbf{V}^{\mathrm{T}}\right) \qquad (6)$$

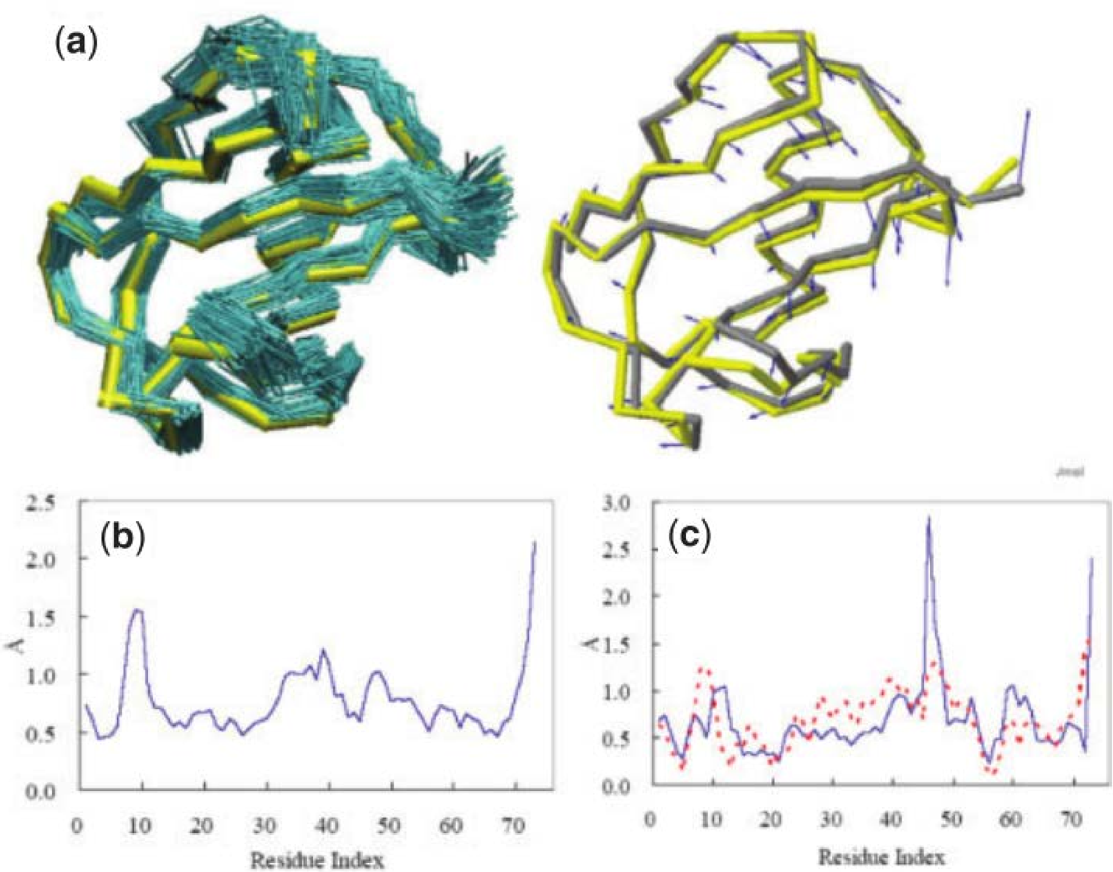
where $\mathbf{V}$ is the matrix of the $3N$-dimensional eigenvectors $\mathbf{v}^{(k)}$ $(1 \leq k \leq M)$ associated with the M non-zero PC modes, and $\mathbf{\Sigma}^{1/2}$ is the diagonal matrix of the square root $\xi_k^{1/2}$ of the corresponding eigenvalues, obtained from the singular value decomposition (SVD)

INTERPRETING PRINCIPAL COMPONENTS

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 $\mathbf{v}_i^{(k)} (1 \leq i \leq N)$

and the $\xi_k^{1/2}$ represents the weight of the mode k, the modes being rank-ordered as $\xi_1 \geq \xi_2 \geq \ldots \geq \xi_M$. 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

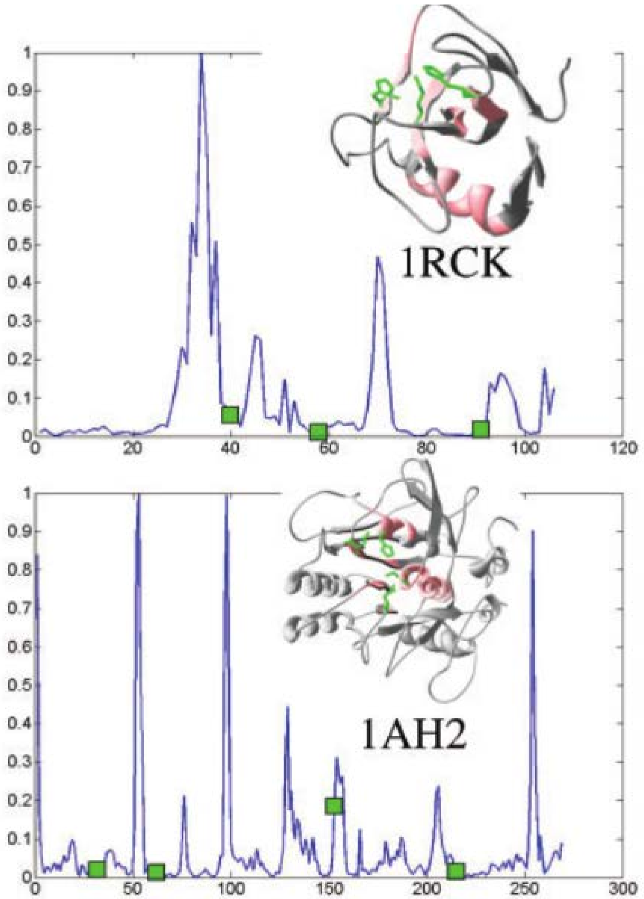
RESULTS



(a)Anensemble of NMR models (teal) for ubiquitin (1xqq) and corresponding X-ray structure (1ubq; yellow).

The mean structure of the NMR ensemble (gray) moves towards its X-ray counterpart (yellow) along the first PC mode.

RESULTS



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

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

reflecting the weighted sum of the top-ranking two PC modes