Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Molecular Binding Sites Are Located Near the Interface of Intrinsic Dynamics Domains (IDDs)

Hongchun Li,[†,‡] Shun Sakuraba,[§] Aravind Chandrasekaran,[‡] and Lee-Wei Yang*[,‡]

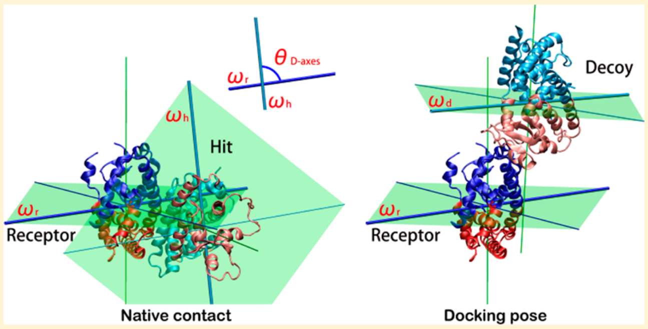
[†] Department of Chemistry, College of Chemistry and Chemical Engineering, and Key Laboratory for Chemical Biology of Fujian Province, Xiamen University, Xiamen, P. R. China

[‡] Institute of Bioinformatics and Structural Biology, National Tsing-Hua University, Hsinchu, Taiwan

[§] Quantum Beam Science Center, Japan Atomic Energy Agency, Kyoto, Japan

Supporting Information

ABSTRACT: We provide evidence supporting that protein–protein and protein–ligand docking poses are functions of protein shape and intrinsic dynamics. Over sets of 68 protein–protein complexes and 240 nonhomologous enzymes, we recognize common predispositions for binding sites to have minimal vibrations and angular momenta, while two interacting proteins orient so as to maximize the angle between their rotation/bending axes (>65°). The findings are then used to define quantitative criteria to filter out docking decoys less likely to be the near-native poses; hence, the chances to find near-native hits can be doubled. With the novel approach to partition a protein into "domains" of robust but disparate intrinsic dynamics, 90% of catalytic residues in enzymes can be found within the first 50% of the residues closest to the interface of these dynamics domains. The results suggest an anisotropic rather than isotropic distribution of catalytic residues near the mass centers of enzymes.



INTRODUCTION

Computational predictions of protein–protein docking (PPD) and protein–ligand docking sites are of great scientific and medicinal interest. Inspired by first-principles, protein-docking heuristics characterize protein interfaces taking two different approaches. The first one is to find the surfaces of two proteins that are complementary in terms of their geometry and electrostatics.[1−4] Efforts in parallel to this are to identify possible docking patches, the "hot spots", in a free protein (proteins in their unbound forms) regardless of its binding partners or whether it does form complexes with others. Predictors for potential PPD sites in isolated proteins, taking into account the linear combination of physiochemical properties including polarity and sequence conservation, can provide encouraging results for 256 homodimers.[5]

Our current study falls into the latter category except the focus is on the dynamics properties of the docking sites. Molecular binding inevitably involves entropic changes, usually a drop of molecular vibrational entropy (or configuration entropy of the side chains) that disfavors the binding free energy.[6−8] Currently, substantial efforts have been made to address the importance of conformational entropy in binding[9−11] including Jernigan group's recent use of the Gaussian network model[12] (GNM)-derived vibrational entropy[13] and Zou group's employment of orientational entropy to discern correct docking poses among the decoys.[14] These approaches require the calculations of both unbound and bound forms of proteins so that the difference of energies can be obtained. The entropy changes together with enthalpic changes, the difference in Gibbs free energy (ΔG), would determine the interaction strength between two proteins. Such calculations usually involve thousands of PPD "decoys" that are plausible poses in which two free proteins dock into each other, and the decoys with the largest negative ΔG or the highest scores based on shape/electrostatic complementarity could closely resemble experimentally determined orientations (native poses) of two proteins. Calculating entropic terms for all these decoys using molecular dynamics (MD) simulations[15,16] and other dynamics analyses[13,14] is not only computationally expensive but also frequently involves heuristic scaling when combining other enthalpic terms, which hopefully brings a qualitatively correct prediction. Here, instead of examining energetics of two proteins before and after binding, we study the "one-body" version of the docking problem, whereby entropically and topologically favored sites for molecular binding in *free proteins* (interchangeably termed as "unbound proteins") are identified.

In the current work, we use three physics models that describe respectively rigid-body (rb) rotation, vibrational, and mixed rb-rotation/vibrational motions of proteins and examine entropical preferences of the locations of PPD and enzyme active sites. Our data suggest that there are positional and

Received: May 1, 2014
Published: August 4, 2014

rotational predispositions of possible docking poses. With newly found criteria based on intrinsic dynamics domains (IDDs) of isolated proteins, we are able to better discern near-native PPD poses from decoys that assume merely good geometric complementarity. Applying the method to enzymes, we are able to locate 90% of the active site residues (out of 732 catalytic residues in 240 nonhomologous enzymes) within the residues that are the closest 50% to the interface of IDDs (domain (D)-plane; see Materials and Methods) or to the splitting-plane (S-plane; see Materials and Methods) that, roughly speaking, halves a protein crosswise. Random planes going through the mass centers of proteins do not achieve the same statistics as do the S-/D-planes. Further elaborated on in the Discussion, the current results suggest that the active sites and the PPD sites tend to stay with those residues having a minimal angular momentum.

■ MATERIALS AND METHODS

PPD Data Set. All the protein complexes are extracted from the popular protein−protein docking benchmark III (PPDB3).[17] In the benchmark, the proteins were divided into three categories (rigid-body, medium difficulty, and difficult), depending on how well the ZDOCK software[18] predicts correct docking poses for two proteins of interest. ZDOCK finds "difficulty" in proteins of the last category due to its algorithm to treat proteins as undeformable rigid-bodies and large conformational changes in the proteins upon forming complexes (see ΔR_J, Table S1, Supporting Information). "Difficult" proteins have a large interface root-mean-square deviation (I-RMSD) that is the RMSD between the residues at the docking interface in the bound forms and the same residues in the unbound forms after the free proteins have been superimposed onto their corresponding bound forms in the complex (see Figure S1, Supporting Information, for a schematic explanation).[17,19] We select the proteins with a resolution less than 3 Å and make sure protein conformational changes upon binding are larger than their diffraction resolution in crystals. The latter is to make sure that the observed conformational changes are discernible enough as compared to the resolution of positions of atoms. Hence, we select complexes that have a prominent distinguishable conformational change (DCC), defined as the size of observed conformational changes in proteins upon binding subtracted by the diffraction resolution of the concerned complex. A high DCC value would suggest "authentic" changes in protein conformation. From PPDB3, we select complexes having a DCC value larger than −0.25 Å, which results in 34 proteins from the "rigid-body" category and 26 proteins from the "medium difficulty" (14) and "difficult" (12) categories combined. All the "difficulty" (interchangeably termed as "Flexible") proteins have a positive DCC value, while 2 out of 14 "medium difficulty" proteins (interchangeably referred as "Medium") and 8 out of 34 "rigid-body" proteins (interchangeably termed as "Rigid") have negative DCC values (Table S2, Supporting Information).

To have an unbiased data set that contains an equal amount of "rigid-body" and "nonrigid-body (flexible)" proteins, we include all the rest of the "difficult" proteins in benchmark IV (PPDB4).[20] There are eight of them, solved at a resolution less than 3 Å and having positive DCC values. Our data set therefore comprises 68 complexes: 34 "Rigid", 14 "Medium", and 20 (12 + 8) "Flexible". The "Medium" and "Flexible" proteins can be divided by a cutoff I-RMSD of 2.26 Å, while "Rigid" and "Medium" can be divided by an I-RMSD of 1.48 Å with two exceptions in each of the categories (see Table S3, Supporting Information, for details).

In the docking study, 3600 decoys (possible PPD poses) have been generated for each complex in PPDB4[17,20] with a 15° sampling by ZDOCK 3.0.2. The decoys have been originally ranked by ZDOCK based on shape complementarity, desolvation energy, and electrostatics.[18] ZDOCK evaluates the docking quality of a given decoy pose by finding the RMSD between the interface residues identified in the native complex (the X-ray structure of the complex) and the same residues in the free proteins at a given decoy pose. In this study, we term such a RMSD as D-RMSD ("D" stands for "decoy"; Figure S1, Supporting Information), and therefore, the smaller the D-RMSD is the better the docking quality for a given decoy is. The clear definitions and distinctions of I-RMSD and D-RMSD can be found in Figure S1 of the Supporting Information, although both terms are referred as "I-RMSD" in ZDOCK. The decoys would be referred as "hits" if they have a D-RMSD less than 2.5 Å from the native pose. On average, ZDOCK finds 6.1 "hits" in the top-ranked 2000 decoys for each of the 36 complexes for which ZDOCK can find at least one hit.

Enzyme Data Set. The Catalytic Site Atlas (CSA)[21] is a database containing positional and functional data of enzyme active sites for over 4000 enzymes. A total of 980 of these are manually curated from literature surveys, while information on catalytic residues for the rest of the enzymes are annotated by sequence alignment. Using the manually curated enzymes, we select the monomeric enzymes with resolutions higher than 3 Å, which results in 240 monomeric enzymes (listed in Table S4, Supporting Information). The monomeric state is assured if both the asymmetric unit and biological assembly files[22] have only one chain in the protein data bank (PDB).[23] Together, the total number of active sites from 240 proteins adds up to 732 with an average of 3.05 active residues per enzyme.

Introduction of IDDs. Despite the widely recognized importance of the entropic contribution to PPD problem, limited evidence (if any) has shown a clear dynamics pattern for PPD sites over a set of proteins with diverse biological functions. We initiated the study by searching such patterns using elastic network models,[24] mainly the Gaussian network model[25,26] and anisotropic network model[24,27] (ANM). Part of the results is shown in Figure S4 of the Supporting Information. Basically, the PPD sites, defined as the residues engaging a bimolecular contact (with a cutoff of 10 Å from the interface), are found to have an indiscriminate distribution at highly fluctuating, immobile, or transition regions in either low-frequency or high-frequency normal modes. No specific and conclusive pattern can be found in terms of dynamics. However, as we examined further, the regions of spatially close and dynamics-wise similar residues in two interacting proteins seem to orient with each other in specific ways; hence, the introduction of intrinsic dynamics domains (IDDs) characterizing such regions.

The IDDs, first coined in this article, are defined with a somewhat different physics nature from Hayward's dynamic domain (DynDom)[28,29] and Hinsen's rigid blocks.[30] Hayward et al. finds a screw axis for each residue involved in protein conformational changes and get the pairwise correlation, the inner product, of these screw axes in a rotation-orientation matrix. These correlations serve as the basis to cluster residues into "dynamics domains". Parameters such as minimal number of domains in a protein and the acceptable ratio of interdomain motions to intra-domain motions need to be defined. Dyndom is typically applied to two experimentally solved conformations of the same protein,[29] although it can be possibly applied to analyze dynamics derived from a single structure.[28] Hinsen's approach starts from characterizing the deformation energy for each residue and sets an energy threshold below which residues are considered "rigid". The translation and rotation motions for each "rigid" residue are determined, and the residues are clustered in a six-dimensional space according to their translation/rotational properties.

The main features of IDDs are summarized here. First, the "domains" in IDDs are clusters of residues in a monomeric or multimeric protein that move concomitantly or oppositely regardless of the magnitude of the motions. Such domains are characterized in *one-dimensional space*. That is to say each residue has only one value pertaining to its dynamics, instead of three values found for DynDom (to define the screw axis) or six values for Hinsen's rigid blocks (three translational and three rotational). We purposely reduce the dimension to make the dynamics further robust[31] and collective.[24] One dynamics value per residue allows residues to be clustered straightforwardly per their geometric proximity, which contrasts Hayward and Hinsen's approaches where the clustering is not carried out in the same three-dimentional space of atom positions, and therefore, residues of similar dynamics may not distribute with spatial adjacency. Hence, isolated parts in proteins could possibly be defined as one dynamics domain. Second, IDD has a uniform treatment to cluster residues into dynamics domains that are derived from either a single structure or multiple structures (equal or more than two), obtained either from experiments[24,32] or theoretical methods such as MD or Monte Carlo simulations. Third, the approach requires minimal number of parameters to be tuned. Once the covariance matrix $\mathbf{C}'_{N\times N}$ is formed (see below), the only parameter needed is a cutoff distance to define how spatially close two atoms that bear similar dynamics should be clustered. The cutoff distance of 4 Å, the upper bond of a possible hydrogen bond, is used in the study. However, we do not see any change in results for cutoff distances from 3.5 to 6 Å (data not shown).

In this study, the dynamic domains are defined to describe motions of proteins in their free forms. The detailed procedure to define dynamics domains is summarized in Figure 1 and eqs S1−S6 of the Supporting Information. We summarize the method's essence as follows. The positional covariance matrix $\mathbf{C}_{3N\times 3N}$ is first computed from either simulations or theoretical models. The covariance matrix tells the motions of residues are in concomitance or in opposition. The covariance matrix is then projected to a lower dimensional space to obtain the matrix $\mathbf{C}'_{N\times N}$ following the procedure described in eq S3 of the Supporting Information. The most dominant eigenvector (also referred as "the first mode") from the eigenvalue decomposition of $\mathbf{C}'_{N\times N}$ is used to classify residues into two dynamics categories based on the signs of the values in this eigenvector (one value for each residue), termed Res+ and Res− nodes (see eq S4 and Figure S2, Supporting Information, and Figure 1 for details). Residues in the same categories move in a concerted manner, and those in different categories move in opposite directions. The categorization leads to IDDs, and we retain only the two largest IDDs for each protein in the current study.

In the present research, the covariance matrix is computed from ENMs[24] for its computational efficiency. In ENMs, the protein dynamics are approximated as the motion of a network of spring-connected residues. Any two residues are connected if they are positioned spatially within a chosen cutoff distance. Such simple models are known to reveal functionally important motions in proteins.[24] We choose two computationally efficient ENMs,[24] GNM[12] and ANM,[27] to determine IDDs. Standard cutoff distances of 7.5 and 15 Å are used for GNM and ANM, respectively.

GNM-based IDDs are defined according to the first mode of the Γ^{-1} matrix in eq S6 of the Supporting Information. As for ANM-based IDDs, covariance $\mathbf{C}$ is first obtained according to eq S5 of the Supporting Information, and a dimension-reduced $\mathbf{C}'$ is subsequently derived based on eqs S3 and S4 of the

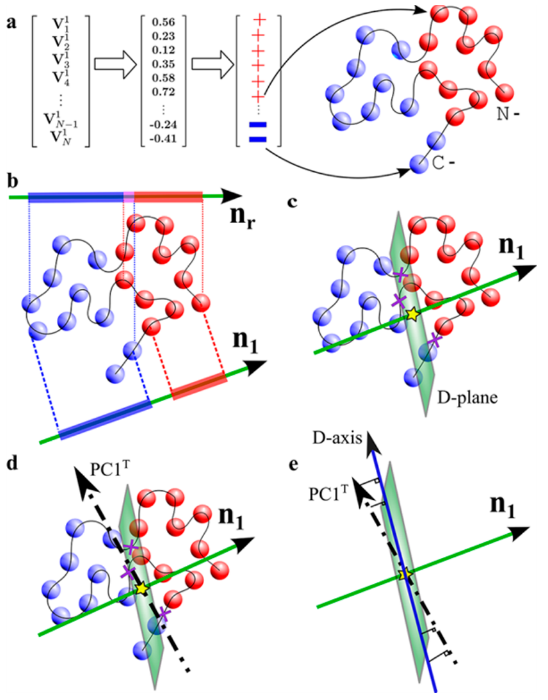


Figure 1. Procedures to determine IDDs, D-plane, and D-axis for a model protein. (a) The slowest GNM mode V_1 is derived from the matrix diagonalization of $\mathbf{C}'$ in eq S6, Supporting Information. Only the signs (±) of the numbers are kept and used to assign the dynamics properties of the protein residues. Residues belong to either Res+ (in red) or Res− (in blue). According to their spatial distributions, more than two clusters, defined as spatially close residues with the same signs (see Introduction of IDDs) can be formed (Figure S2, Supporting Information). Only two clusters with opposite signs will be kept for the following analyses. (b) To define a plane that best separates the two clusters, LDA (see Materials and Methods) is used. The method finds an axis (shown as n_1) on which the projections of Res+ (in red) and Res− (in blue) can be separated the widest as compared to any other possible axes (e.g., n_r on which the projections could overlap). (c) The axis n_1 found in (b) is called domain (D)-plane normal to define the D-plane (in light green). The normal as well as the plane is required to go through the geometric center of transition points, shown as a yellow star, which is usually close to the mass center (first panel, Figure S3, Supporting Information). The transition points (purple "x" sign) are the centers of any two consecutive residues that have different dynamics properties in the primary sequence. (d) To define the domain-axis (D-axis), we first find this vector $PC1^T$ (shown as the broken line),[32] onto which projections of transition points have the widest spread (see D-axis part in Materials and Methods). (e) The $PC1^T$ is then projected to the D-plane to obtain the D-axis (shown as the navy blue line). In our analyses, there is always one D-plane and one D-axis for every protein. See Figure S2, Supporting Information, for D-plane(s)/D-axis(es) defined for a real protein.

Supporting Information. The most dominant mode from the decomposition of **C**′ is derived to define IDDs. D-planes, the planes that best separate the IDDs, and Domain(D)-axes, the lines that thread through Res+/Res− transition points, are obtained according to the procedure described below and illustrated in Figure 1 and Figure S2 of the Supporting Information.

Find the D-plane Using Linear Discriminant Analysis (LDA). We define the D-plane as the plane that optimally separates Res+ and Res− nodes in the three-dimensional space (see a schematic presentation in Figure 1 and Figure S2, Supporting Information). To find such a plane, we employed the idea of LDA[33] where nodes are best separated when the variance between the sets, as projected to the normal vector **n** of a dividing plane, can be maximally separated as compared to the scattering of the nodes within each set. Specifically, it is the ratio of the between-set variance, defined as the variance of the averages of the sets, to the variance within the sets to be maximized.

$$F(\mathbf{n}) = \frac{\mathbf{n}^T\mathbf{C_B}\mathbf{n}}{\mathbf{n}^T\mathbf{C_W}\mathbf{n}} \tag{1}$$

In this study, **n** is also the normal vector of a domain (D)-plane (Figure 1b,c), which we are about to find. $\mathbf{C_B}$ is the between-set covariance matrix, and $\mathbf{C_w}$ is the within-set covariance matrix. The details of the methodology can be found in the Supporting Information. In this study, the D-plane normal is put at the mean position of transition points (Figure 1c), which is found close to the center of mass (COM) (Figure S3, Supporting Information).

Defining the Domain (D-)Axis between IDDs. Assuming there are two IDDs in the protein and the C_α atoms of residue *i* and residue *i*+1 on the primary sequence belong to two different IDDs, the midpoint between the C_α atoms of residues *i* and *i*+1 is defined as the transition point (Figure 1c). Given that there are *n* such transition points in the protein, the positional covariance matrix (assuming equal mass for every residue) of these transition points reads as

$$\mathrm{M} = \begin{bmatrix} \Delta r_{1x} & \Delta r_{2x} & & \Delta r_{nz} \\ \Delta r_{1y} & \Delta r_{2y} & \cdots & \Delta r_{nz} \\ \Delta r_{1z} & \Delta r_{2z} & & \Delta r_{nz} \end{bmatrix} \begin{bmatrix} \Delta r_{1x} & \Delta r_{2x} & & \Delta r_{nz} \\ \Delta r_{1y} & \Delta r_{2y} & \cdots & \Delta r_{nz} \\ \Delta r_{1z} & \Delta r_{2z} & & \Delta r_{nz} \end{bmatrix}^T \tag{2}$$

where $\Delta r_{i\alpha}$ is the deviation of the *i*-th residue from the mean position of the *n* transition points at α direction (α is *x*, *y*, or *z*). Diagonalization of **M** would result in three positive eigenvalues and corresponding eigenvectors. The eigenvector that associates the largest eigenvalue is the first principal component of the spatial distribution of the *n* points. Note that this eigenvector (referred as $\mathrm{PC1}^\mathrm{T}$ in Figure 1) as well as the normal vector of the D-plane are both pointing from the mean position of the transition points. The projection of the eigenvector onto the D-plane is defined as domain (D)-axis (Figure 1e).

■ S-PLANES AND AXES OF PROTEINS

The splitting (S)-plane is defined as the plane going through the protein mass center with the plane's normal being the largest principal component (PC1) of the spatial distribution of all the protein residues. The PC1 vector is derived similarly to D-axis. Instead of using transition points as inputs to eq 2, all the C_α atoms in a protein are used as $\Delta r_{i\alpha}$, and the resulting eigenvector $\mathrm{PC1}^\mathrm{T}$ forms the PC1 vector. Therefore, the S-plane divides a protein "crosswise" and separates the residues into two groups. If a line reposing on the S-plane is chosen to be the rb-rotation axis of the protein, the two groups would always move toward opposite directions (like the open arms of a spinning figure skater), which grants the two groups (one contains Res+ nodes and the other comprises Res− ones) similar physical nature as the IDDs. The splitting (S)-axis is then found following the same protocol to define D-axis.

Enrichment Factors for Hits (EFHs) and Enrichment Ratios (ERs) for Active Sites. Enrichment Factor for Hits is the value that describes the enrichment of chances to find hits among the set of decoys that fulfill the aforementioned criteria as compared to a random chance to find these hits among the 2000 top-scored decoys.

$$\mathrm{EFH} = \frac{N'_{\mathrm{hits}}/N'_{\mathrm{dec}}}{N_{\text{All-hits}}/N_{\text{All-dec}}} \tag{3}$$

where N'_{hits} and N'_{dec} are the number of hits and decoys, respectively, that fulfill our filtering criteria, such as "Cutting through and Angle >40°", based on IDD results. $N_{\text{All-hits}}$ and $N_{\text{All-dec}}$ are the total number of hits and decoys, respectively, for a given complex. $N_{\text{All-dec}}$ is 2000 in this study.

The enrichment ratio (ER) describes the enrichment of the chances to find active sites using IDD-based criteria as compared to a random guess. It is the ratio of the chance to find catalytic residues in the first *x* percentile of the residues that are the closest to the S-/D-planes to the probability to find them by chance.

$$\mathrm{ER} = \frac{N_{X\,\%\text{-act}}/N_{X\,\%\text{-res}}}{N_{\text{All-act}}/N_{\text{All-res}}} \tag{4}$$

where $N_{\text{All-res}}$ and $N_{\text{All-act}}$ are the total number of residues and catalytic residues in the 240 enzymes, respectively. $N_{X\%\text{-res}}$ and $N_{X\%\text{-act}}$ are the number of residues and active sites within the first *x%* of the residues that are the closest to the S-/D-planes, respectively (see Enzyme Active Sites of Proteins Stay Close to GNM-Defined D-Planes and PC1-Defined S-Planes section).

■ RESULTS

Position Restraints Imposed on PPD. IDDs of 68 protein complexes described in the PPD Data Set section are examined in three categories of distinctive flexibility (referred as "Flexible", "Medium", and "Rigid"). Exactly two proteins constitute each of the 68 complexes, and the protein (also its corresponding unbound form) that is bigger in size than its docking partner is termed "big" protein hereafter. The dynamics of the corresponding unbound forms of the proteins in the complexes are analyzed. Three physics models of different dynamics nature are used to examine the proteins' intrinsic dynamics: (1) The rb rotation is examined by PC1. (2) Vibrational motions are examined by ANM. (3) Mixed rb rotation and vibrational motions are examined by GNM. IDDs of the proteins are characterized for each of the underlying physics models. The IDDs and corresponding D-planes/D-axes (Figure 1) of the unbound forms are then superimposed onto the corresponding bound forms in the complexes. The positions of proteins relative to their docking partners' D-planes and the angle between two proteins' D-axes are examined.

Table 1. Angle and D-Plane "Cutting" Statistics for GNM, ANM, and PC1-IDDs[a]

IDDs defined by		GNM			ANM			PC1		
		θ	B	S	θ	B	S	θ	B	S
Rigid (34)	number	20	25	25	22	16	19	18	25	24
	ratio (%)	**59**	**74**	**74**	**65**	**47**	**56**	**53**	**74**	**71**
Medium (14)	number	9	11	12	5	8	10	7	11	10
	ratio (%)	**64**	**79**	**86**	**36**	**57**	**71**	**50**	**79**	**71**
Flexible (20)	number	17	18	16	12	11	12	11	17	17
	ratio (%)	**85**	**90**	**80**	**60**	**55**	**60**	**55**	**85**	**85**
All (68)	number	46	54	53	39	35	41	36	53	51
	ratio (%)	**68**	**79**	**78**	**57**	**51**	**60**	**53**	**78**	**75**

[a] θ, B, and S refer to $\theta_{\text{D-axes}} > 57.3°$, D-planes of big proteins cutting through paired small proteins, and D-planes of small proteins cutting through paired big proteins, respectively. The numbers in parentheses following each of the categories are the total number of proteins in that category. The detailed results for each protein can be found in Table S1 of the Supporting Information.

The most striking results come from the relative position of GNM-defined IDDs. In 54 out of 68 cases, equivalent to a 79% probability, the D-planes from the "big" proteins (bigger in size than its docking partner) dissect through its smaller docking partners (Table 1 and examples in Figure S5, Supporting Information). This probability increases to 90% (18 out of 20) if only the Flexible category is considered or 85% (29 out of 34) if both Flexible and Medium categories are considered (Table 1). The significance of this result can be better understood as compared with the probability derived from random planes. We place a trial vector as the plane normal that points to a "random" direction (see Supporting Information for the details) and let the plane go through the center of mass (COM) of the protein. We first generate one random plane for each of the 68 big proteins, and then we examine how many "random planes" of these big proteins cut through their paired smaller partners and get the ratio. We repeat this procedure 1000 times to obtain 1000 of such ratios that have a mean of 58.5% and a standard error of 6.0%. The statistics reduce to 54.6 ± 6.0% if we put the random planes on randomly chosen C_α atoms. Contrasting the background chances, our data show that 51%, 78%, and 79% of the D-planes of the "big" proteins dissect through their docking partners when IDDs are defined by ANM, protein shape (or interchangeably referred as "PC1" hereafter), and GNM, respectively (Table 1). The PC1 and GNM methods demonstrate clear PPD preferences for sites on/near the S-/D-planes.

On the other hand, the dissecting probabilities for the "big" proteins being dissected by the D-planes of their smaller binding partners are found to be 60%, 75%, and 78% for ANM, PC1, and GNM, respectively (Table 1). The statistics are nonetheless not significant enough when comparing to the background probabilities that are found to be 78.2 ± 5.0% (per the random plane results).

When "Flexible" and "Rigid" categories are individually considered, higher percentage of D-planes of big proteins cutting through paired proteins is found for the "Flexible" category than that for the "Rigid" category. This is true for the IDDs defined by ANM (55% vs 47%), PC1 (85% vs 74%), and GNM (90% vs 74%) (Table 1). For the cases that D-planes of the small proteins cut through the paired bigger ones, the cutting percentage is also higher for the "Flexible" than for the "Rigid" in all the three methods. However, the differences are small and the values are not enough significant as compared to the background (random) chances.

Orientation Restraints Imposed on PPD. Taking only the acute angle (0−90°) between any randomly chosen two lines in space, one can analytically prove that the mean of the angle is 57.3° (or one radian) with a standard deviation of 21° and standard error of 2.6° for 68 trials (see Supporting Information for the proof). We are interested in knowing how the acute angle of D-axes from paired big and small proteins compares with the featureless angle, 57.3°. The most striking results again come from GNM-defined IDDs. In 32 out of the 68 complexes (14/34 in the Rigid category, 5/14 in the Medium category, and 13/20 in the Flexible category), D-axes angles ($\theta_{\text{D-axes}}$) are found to be larger than 70°. A total of 46 out of 68 have a $\theta_{\text{D-axes}}$ larger than 57.3° (Table 1). Only 2 out of 68 have a $\theta_{\text{D-axes}}$ less than 30° (Table S1, Supporting Information). GNM-derived $\theta_{\text{D-axes}}$ has a mean value ($<\theta_{\text{D-axes}}>$) of 64.9 ± 2.0° for the 68 complexes, and the value increases to 70.1 ± 2.8° if only the "Flexible" category is considered (Table S1, Supporting Information). On the other hand, $<\theta_{\text{D-axes}}>$ is found to be 60.2 ± 2.7° and 56.0 ± 2.6° for ANM and PC1, respectively, which deviate marginally from the featureless angle, 57.3° (Table S1, Supporting Information). The distributions of $\theta_{\text{S/D-axes}}$ for GNM, ANM, and PC1 can be found in Figure 2 and Figure S6 of the Supporting Information.

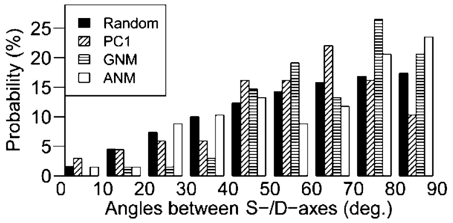
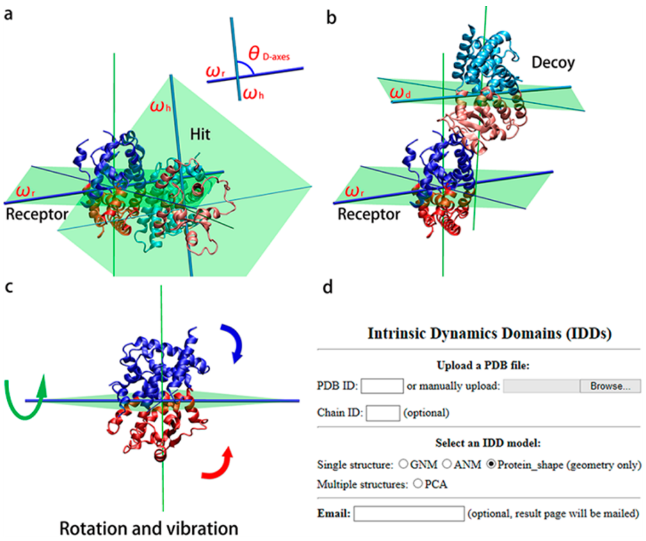


Figure 2. Distribution of $\theta_{\text{D-axes}}$ when the IDDs are derived by PC1, GNM, and ANM are compared with that of angles formed by two "random" vectors (filled solid bar; see Supporting Information for details to obtain random vectors).

The statistical differences of these distributions of the S-/D-axes angles and a random distribution are examined. Because none of the sampled angle distributions is Gaussian, we use a one-sample two-sided Kolmogorov−Smirnov test[34] to evaluate the differences of these distributions with a null hypothesis that the $\theta_{\text{S/D-axes}}$ distributions are sampled from a random distribution. The resulting *p*-values, the probabilities of the null hypothesis being correct, for ANM, PC1, and GNM are 0.43, 0.62, and 0.013, respectively. Thus, only the GNM-defined $\theta_{\text{D-axes}}$ has a significantly different distribution from the random one. In addition, we do not find any feature for angles between the normals of the D-planes of two interacting proteins. The D-plane normals meet at an angle of 54.6 ± 2.6° and 57.4 ± 2.5° for the cases of ANM and GNM, respectively (Table S3, Supporting Information). None of these statistically differ themselves from a random distribution.

Restraints Suggested by D-Planes/-Axes Help Discern Near-Native Docking Poses from ZDOCK Decoys. ZDOCK provides docking decoys for each complex, which are ranked based on the geometric and electrostatics complementarity between corresponding unbound proteins found in these poses. We use EFH defined in eq 3 to evaluate how well IDD helps enrich the chance to find near-native poses, "hits" (Figure 3), out of the top-ranked 2000 docking decoys for every complex of interest. The enriched chance to find "hits" due to the search is restricted to the decoys that fulfill one of the three IDD-based criteria: (1) The big protein's S-/D-plane cuts through its smaller docking partner (referred as "Cutting through" hereafter). (2) The angle between S-/D-axes of the two proteins is larger than 40° (referred as "Angle >40°" hereafter). (3) Both "Cutting through" and "Angle >40°" are satisfied (referred as "Cutting through and Angle >40°"). It is worth noting that no "hits" can be found in the decoys of the "Flexible" complexes, and therefore, no EFH values can be calculated for this category in Table 2 due to large conformational changes in proteins upon binding. Our results show that the highest EFH values for the "Rigid" and "Medium" proteins come from PC1-IDDs, followed by GNM-IDDs and then ANM-IDDs (Table 2). Among them, the criterion "Cutting through and Angle >40°" when IDDs are determined by PC1 brings the highest EFH value of 1.65. The enrichment further increases if only the proteins of Medium category (excluding rigid proteins) are concerned, where an EFH of 2.00 is obtained. The enrichment using the "Cutting through" criterion is better than "Angle >40°", while the combined criteria give the best results for GNM and PC1. The Fisher exact test[35] (Table2) indicates that PC1-/GNM-defined IDDs enrich the hits with statistical significances, especially for the combined criteria (p-values <0.05). They are statistically important in finding the near-native hits with improved rankings based on the criteria introduced herein (see Figure 3 and an example in Table S5, Supporting Information).



Intrinsic Dynamics Domains (IDDs)

Upload a PDB file:

PDB ID: ______ or manually upload: ______ Browse...

Chain ID: ______ (optional)

Select an IDD model:

Single structure: ○ GNM ○ ANM ● Protein_shape (geometry only)

Multiple structures: ○ PCA

Email: ______ (optional, result page will be mailed)

Figure 3. GNM-defined IDDs for the complex (1K74) RXR-α and PPAR-gamma can discern the near-native hits from the decoys. (a) The two IDDs of the receptor (1MZN) are colored in blue and red, while those of the ligand (1GZY) are in sky blue and salmon pink. The ligand protein in correctly docked pose, "Hit" (Materials and Methods), is cut through by the D-plane of the receptor, and its D-axis (ω_h) forms a $\theta_{\text{D-axes}}$ angle near 90° with that (ω_r) of the receptor. $\theta_{\text{D-axes}}$ is the acute angle between two D-axes of a docking pair (each protein has one D-axis). The two D-axes here do not intersect, but $\theta_{\text{D-axes}}$ is calculated by making the two D-axes vectors go through a common point. The green lines are the D-plane normals. (b) Incorrectly docked ligand protein, "Decoy", is not cut through by the D-plane of the receptor, and the angle of D-axes (ω_r and ω_d) is 13°. However, in the decoy sets, the hit nearest to panel (a) ranks 299th, while the decoy in panel (b) ranks 5th by the ZDOCK software.[18] (c) The IDDs moving toward opposite directions can result from a mix of rotation (e.g., rotating about the D-axis reposing on the D-plane) and a collective bending (vibration) over another axis that is also on the D-plane. The curved green arrow indicates the rb rotation, while the blue and red ones describe the bending motion. (d) The IDDs Web portal.

In the "Flexible" category, all the I-RMSDs between the superimposed unbound and bound structures are larger than 2.5 Å. As a result, the D-RMSDs of all the docking poses are also larger than 2.5 Å; hence, no "hits" can be found in this category. To assess whether or not our criteria help find the decoys that are closer to native poses than those that do not meet our criteria in the "Flexible" category, we define the metric $\Delta_{\text{D-RMSD}}$ such that $\Delta_{\text{D-RMSD}}$ = (D-RMSD between the native pose and the decoys that fulfill our criteria) − (D-RMSD between the native pose and the decoys that do not fulfill our criteria). As a result, large negative $\Delta_{\text{D-RMSD}}$ values would indicate the effectiveness of our criteria. We find that $\Delta_{\text{D-RMSD}}$ for the criterion 'Cutting through and Angle >40°' among the "Rigid", "Medium", and "Flexible" categories are −1.47, −2.33, and −1.40 Å for PC1-IDDs (Table 2). The corresponding values for GNM-IDDs are −0.83, −2.17, and −1.43 Å (Table 2). The decrease in $\Delta_{\text{D-RMSD}}$ values for the "Medium" and "Flexible" categories for both GNM-IDDs and PC1-IDDs are larger than that for the "rigid" category, indicating that IDD-derived criteria can arguably be employed to improve the decoy rankings for more flexible proteins.

Enzyme Active Sites of Proteins Stay Close to GNM-Defined D-Planes and PC1-Defined S-Planes. We examine whether there is a higher chance for active site residues (interchangeably termed as "catalytic residues") to be found near the S-/D-planes. Residues in 240 enzymes are ranked based on their distances from the corresponding S-/D-planes (0% is the closest and 100% is the farthest from the planes; see the x-axis of Figure 4) and the percentage of catalytic residues (y-axis of Figure 4) found in each x ranking percentile (see the caption of Figure 4 for an illustrative example). Our results reveal a large percentage of active sites located toward the S-plane and GNM-based D-planes but not the ANM-based D-planes.

The cumulative percentage, the percentage of the catalytic residues found in the first x percentile of the residues closest to the S-/D-planes, shows that 90%, 90%, 61%, and 53% of active sites can be found within top 50% of the protein residues having the shortest departure from the S-/D-planes defined by PC1, GNM, ANM, and random vectors, respectively (middle column of Figure 4 and Figure S7, Supporting Information). Knowing that half of the protein nearest the S-/D-planes can contain almost all the active sites, we further examine the ER in eq 4 for the first x% ranking percentile (to the right in Figure 4) that is defined as enrichment of the chance to find active sites among the residues within the first x% nearest these planes from the chance of a blind guess. We see the ER can go above

Table 2. Fisher Exact Test of the Top 2000 Decoys[a]

		Cutting through			Angle >40°			Cutting through and Angle >40°		
criteria		*p*-value	EFH	$\Delta_{\text{D-RMSD}}$	*p*-value	EFH	$\Delta_{\text{D-RMSD}}$	*p*-value	EFH	$\Delta_{\text{D-RMSD}}$
GNM	All	$\mathbf{5.6 \times 10^{-11}}$	1.34	**−1.96**	8.1×10^{-01}	1.06	0.01	$\mathbf{1.5 \times 10^{-06}}$	1.56	**−1.28**
	Rigid	$\mathbf{8.0 \times 10^{-14}}$	1.38	**−1.45**	4.5×10^{-01}	1.01	0.07	$\mathbf{1.2 \times 10^{-07}}$	1.56	−0.83
	Medium	2.5×10^{-01}	1.17	**−3.00**	6.3×10^{-02}	1.27	−0.10	$\mathbf{4.2 \times 10^{-02}}$	1.55	**−2.17**
	Flexible	N/A	N/A	**−2.12**	N/A	N/A	−0.02	N/A	N/A	**−1.43**
ANM	All	3.8×10^{-01}	1.16	−0.81	$\mathbf{1.0 \times 10^{-03}}$	1	−0.11	$\mathbf{8.0 \times 10^{-03}}$	1.08	−0.53
	Rigid	4.8×10^{-01}	0.83	−0.70	$\mathbf{9.8 \times 10^{-05}}$	1.05	−0.11	$\mathbf{2.0 \times 10^{-03}}$	0.98	−0.41
	Medium	5.0×10^{-01}	2.53	**−2.38**	4.3×10^{-01}	0.81	−0.04	6.5×10^{-01}	1.49	**−1.80**
	Flexible	N/A	N/A	0.14	N/A	N/A	−0.16	N/A	N/A	0.19
PC1	All	$\mathbf{1.0 \times 10^{-03}}$	1.41	**−2.33**	$\mathbf{1.3 \times 10^{-02}}$	1.08	−0.05	$\mathbf{6.8 \times 10^{-06}}$	1.65	**−1.63**
	Rigid	$\mathbf{3.4 \times 10^{-04}}$	1.35	**−2.09**	$\mathbf{1.9 \times 10^{-02}}$	1.07	−0.08	$\mathbf{3.9 \times 10^{-06}}$	1.56	**−1.47**
	Medium	1.6×10^{-01}	1.67	**−3.28**	2.8×10^{-01}	1.11	−0.18	7.3×10^{-02}	2.00	**−2.33**
	Flexible	N/A	N/A	**−2.04**	N/A	N/A	0.08	N/A	N/A	**−1.40**

[a] Criteria "Cutting through" and "Angle >40°" refer to "D-planes of big proteins cutting through paired small proteins" and "$\theta_{\text{S/D-axes}} > 40°$", respectively. Enrichment factor for hits (EFHs) value describes the enrichment of hits in the set of decoys that fulfill the certain criteria. The *p*-values smaller than 0.05 and the $\Delta_{\text{D-RMSD}}$ (in Å) values smaller than −1.00 Å are printed in boldface. $\Delta_{\text{D-RMSD}}$ is the difference between the mean D-RMSD of the decoys that fulfill these criteria and D-RMSD of those that do not.

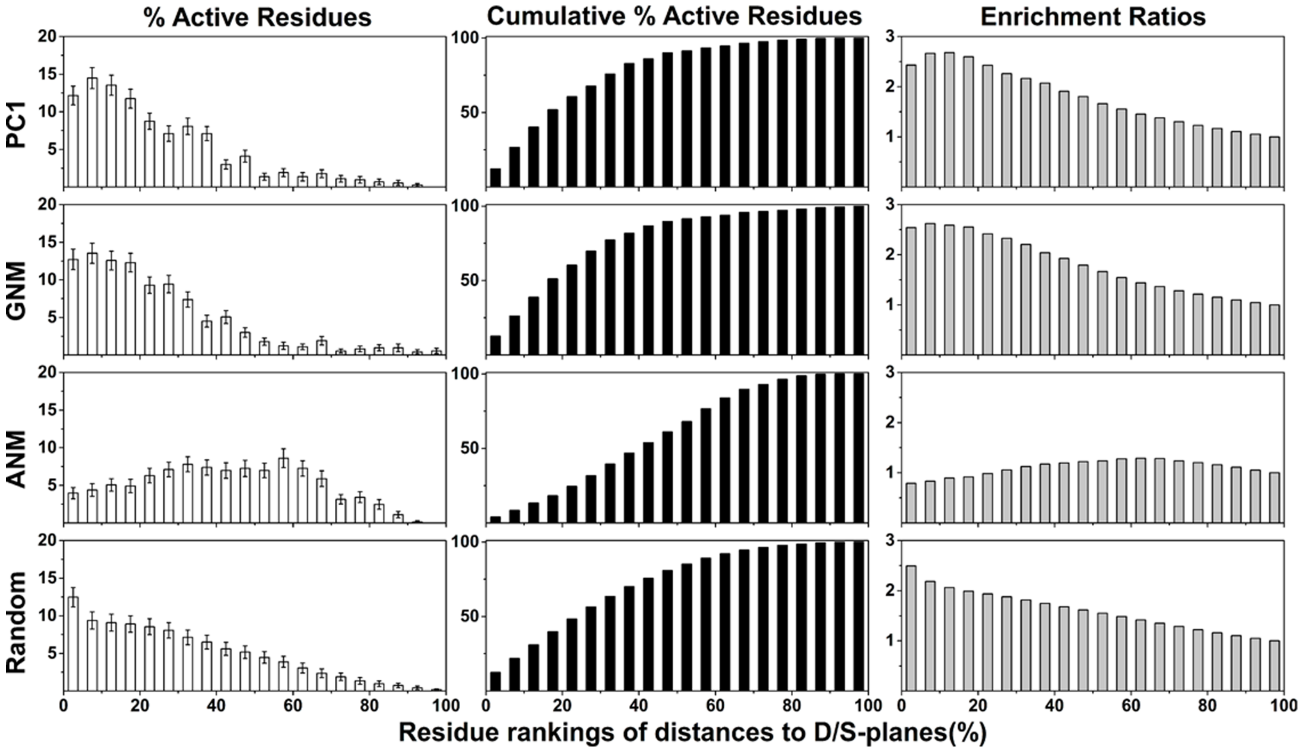


Figure 4. Percentage of active sites (white), cumulative percentage active sites (black), and enrichment ratios (gray) found in the *x* percentile of the residues ranked based on how far they are away from the S-/D-planes defined by PC1, GNM, ANM, and random vectors, respectively. The calculations are based on 240 enzymes containing 732 catalytic residues, and the definition of "Random" can be found in the Results. For example, enzyme beta mannase (PDB: 1bqc) has 302 residues and five catalytic residues. The top-ranked 15 residues that are closest to the S-plane (defined by the PC1) belong to the first 5% percentile in the *x*-axis of the topmost panel in the left column (302 × 5% ≈ 15). Within these 15 residues, presented by the bin of 0−5% in *x*, two catalytic residues can be found, which constitutes 40% (2/5) of the active site, shown in the *y*-axis. Such data from 732 catalytic residues of 240 enzymes are shown in the left most column. Panels in the middle column are simply the accumulative counts (in percentile) of the corresponding panels on the left. In the above example, the enrichment ratio (ER; Materials and Methods) is 8.1, calculated from the chances of finding a catalytic residue within the first 5% in the *x*-axis, 2/15, over the odds to find a catalytic residue by a random guess, 5/302 (upper right panel). ERs from different physics models are shown in the panels to the right, except that the data are for the entire enzyme set and are accumulative (that is to say 10% in *x*-axis meaning 0−10% and 15% in *x*-axis meaning 0−15%, so on and so forth). See Figure S7 of the Supporting Information for more details.

2.5 for PC1 and GNM within the first 20%. Although random planes (that penetrate COM) can have an ER of ~2.5 for the first 5% in *x*, their enrichment ratios start to fall behind the GNM's and PC1's from the 10% point on, indicating that *GNM/PC1-based active site search* (hence in specific directions) *is superior to a spherical search near the COM, except for catalytic residues immediately adjacent to the COM* (<5% from the planes). The gain of the enrichment ratios over a random spherical search for GNM and PC1, within different cumulative rank percentiles, is plotted in Figure S8 of the Supporting Information. The maximal gain over random for PC1 is 1.30 and for GNM is 1.28 at the top 20% rank. A two-sample one-sided Kolmogorov−Smirnov test is carried out on the cumulative distribution functions, and p-values are tabulated in Table S6 of the Supporting Information. The results show that the cumulative distribution functions derived from PC1- and GNM-based S-/D-planes are larger than those from ANM-based planes and random planes, with a significance p-value less than 0.001, while the results from PC1 and GNM are statistically indistinguishable.

Online Service. An online service has also been implemented for users to obtain the proteins' IDDs, D-plane normal, D-axes, and top 50% of the residues closest to the D-planes for the structures of their interest (http://dyn.life.nthu.edu.tw/IDD/IDD.php) (Figure 3d). IDDs of a single structure can be calculated using GNM, ANM, and protein shape. Multiple structures obtained from NMR spectra or MD simulations are accepted for the IDDs determination using one of the first six principle components of $\mathbf{C}'$.[32]

■ DISCUSSION

Entropic and Topological Restraints on Docking Position/Orientation. Taking the protein as a rigid-body and assuming the protein rotates about an axis (say, S-axis) that lies on the splitting-plane where PC1 is the normal, the protein rotates with the largest possible moment of inertia (I). For all the 240 enzymes we studied, the moment of inertia derived from proteins rotating about axes lying on S-planes (and also going through the COM of the proteins) is statistically larger ($p < 0.05$) than the moment of inertia derived from any other random axes passing through the COM not reposing on the S-plane. The average difference ($I_{\text{axes\_on\_S-planes}} - I_{\text{random\_axes}}$) is a positive value of $(2.6 \pm 0.3) \times 10^5$ a.m.u. Å^2. According to equipartition of energy, the average kinetic energy for the aforementioned rotation is $(1/2)k_{\text{B}}T$ that is also equal to $1/2 \times I \times \omega^2$ (I is the moment of inertia; ω is the angular velocity). Hence, the largest I, obtained from the protein rotating about axes lying on a S-/D-plane, results in the *slowest possible* ω. The residues far away from the rotating axis on the S-/D-plane would have the largest motion due to coupled rotation and the slowest molecular vibrational motions, where the ones close to the D-plane have the least motions. For a residue that has a small distance r from a S-/D-plane, its angular momentum is $\sim mr^2\omega$. We argue that *this is the smallest possible angular momentum a residue can have* (for the small r and the slowest ω) as compared to other possible angular momentum values. Our data suggest that the active sites as well as the PPD sites tend to stay close to *where a minimal angular momentum is ensured.*

As for why active sites and PPD sites tend to stay close to where has small angular momentum, our working hypotheses on its kinetic and thermodynamic causes are summarized as follows. From the kinetics point of view, it has been reported that molecular recognition involves a transient encounter through nonspecific interactions before a more specific search via surface/charge complementarity.[36] We speculate that binding near the D-planes has at least two kinetic advantages as compared to binding at the "tips" that are far away from the D-planes. First, the high kinetic energy at the tips of the molecule due to molecular rotation and vibration provide very stringent orientation window for molecules to have an effective collision. Once a contact attempt fails, the high kinetic energy of the tips would drive the engaging partner away instead of keeping the partner close. On the contrary, regions near D-planes would not have such a stringent requirement allowing collisions only in an absolutely correct contact orientation, and the region would not necessarily drive the binding partners away once the failure of contact attempts. Second, D-planes are where molecules have opposite motions either in rotation or low-frequency vibrations. Therefore, if binding partners stay near the D-planes, the molecular "clamping" and rotation motions driven by thermal energy would allow an exploratory binding surface search while binding taking place at the tips would not.

The thermodynamics end of the binding is less straightforward than the kinetics end. We first discuss the ligand−protein binding before extending the arguments to protein−protein cases. In following discussions, the comparison is made between binding near the D-planes and binding at "tips". It is well acknowledged that the binding entropy mainly consists of changes in molecular rigid-body component[37] (translation/rotation) and molecular vibrational entropy[11,37,38] including configurational entropy of residue side-chains.[39] It has been shown by Murray and Verdonk that the change in rigid-body entropy of binding for the protein is <0.1 kJ/mol for a protein of ∼300 residues and a ligand with molecular weight of 300.[40] Also, a study from Gilson's group has shown that the entropy changes in HIV1 protease and the AIDS drug Amprenavir upon binding are mainly from configurational entropy (24.6 kcal/mol), while the translation and rotation entropy alone contribute to just 1.8 kcal/mol. On the other hand, we speculate that tips have higher side-chain configurational entropy than that of same side-chains near the D-planes due to a higher kinetic energy fluctuation resulting from large molecular vibrational motions coupled with rotation. Hence, the side-chains "frozen"[39] at one configuration upon ligand binding would result in more entropic loss at the tips than that near the D-plane belt. Big entropy loss at tips goes against favorable binding Gibbs free energies and therefore leads to observed bias of enzyme active sites/ligand binding sites distributed near the D-planes. In the above reasoning, the solvent entropy gain upon binding is assumed the same for both at tips and near D-planes.

The aforementioned rationales on the thermodynamic aspect apply to binding partners having a big difference in sizes, e.g., protein and ligand, or two proteins where one is much bigger than another. When binding partners are proteins with similar sizes, the effects of entropy changes upon binding for situations at tips or at D-planes are less straightforward. On one hand, binding at the tips of proteins would result in a moment of inertia of the complex larger than that when binding near the D-planes, which in turn causes a larger rotational entropy, $(3/2)R \ln(\varphi I_a I_b I_c)^{1/2}$ (where φ is a constant and I_a, I_b, and I_c are the moments of inertia for the molecule rotating about one of the three principal components[37]) and favors binding at tips. On the other hand, binding at the tips of proteins would suffer from entropy loss more so than binding near D-planes simply because molecular tips tend to have a higher molecular vibration and side-chain configurational entropy than D-planes' proximity before the binding, which disfavors the binding at the tips. MD simulations showed that molecular association involves rotational/translational entropy loss of ∼9 kcal/mol at room temperature,[41] while NMR experiments revealed that conformational entropy loss in calmodulin association with a few calmodulin-binding domains ranges from 5 to 17 kcal/

mol.[42] This suggests similar but opposite entropy contributions to a preferential binding near D-planes. However, Frederick et al. also found that changes in the corresponding conformational entropy are linearly proportional to the changes in overall binding entropy (including the solvent effect).[42] As a result, conformational entropy may play a more vital role than external entropy in binding, which supports observed dominance of protein−protein association near the D-planes.

In our study, D-axes of docking partners are found to meet at a large angle (toward 90°) and almost never stay in parallel (toward 0°). That they disfavor a small crossing angle (66/68 have $\theta_{\text{D-axes}} > 30°$) implies that the docking partners have minimal inclination to "bend" or "rotate" in the same direction (see illustrations in Figure 5). That $\theta_{\text{D-axes}}$ tends to hold a large angle may not be attributed to rigid-boy rotation alone because $\langle\theta_{\text{S-axes}}\rangle$, the angle between favorite rotation directions of two paired proteins, for all three categories are found featureless (close to 57.3°). Rather, this phenomenon could attribute to a delicate interplay between rigid-body rotation and vibration, as demonstrated by the first GNM mode.[26] ANM gives a $\langle\theta_{\text{D-axes}}\rangle$ value of 60.2 ± 2.7° (62.9 ± 4.8° for the Flexible category), statistically larger than the featureless angle, indicating vibrational motion can be essential for causing a large $\theta_{\text{D-axes}}$. Hence, favorable docking orientations are those that allow IDDs to exhibit their lowest frequency movements even in the bound state. In other words, proteins seek possible movement throughout the docking process so as to alleviate inevitable entropy loss (Figure 5). This model is consistent with an early study suggesting that complex constitution does not fully restrain the conformational freedom of the binding partners as a whole, instead it leads to a redistribution of dynamics.[43] Our results suggest that two proteins do not bind in an orientation via which they interfere each other's favorite motions.

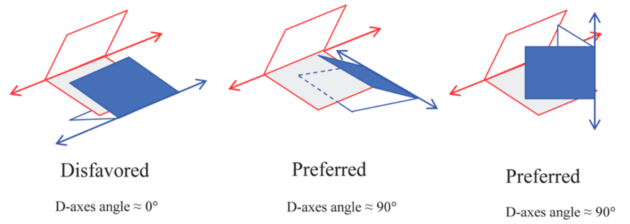


Figure 5. Favored and disfavored docking orientations illustrated in three models with different spatial arrangements of IDDs. In the figure, the blue proteins dock into their red counterparts. Each protein comprises two faces, inferring its two IDDs, which are connected along the D-axes about which the two IDDs bend and/or rotate. "Disfavored" encounter adopts an orientation via which two molecules interfere each other's favorite motions.

Better Ranking/Enrichment of Hits in PPD Decoys. There have been reports that PPD interfaces are located near the protein mass center[44] or could be penetrated by one of the principal components spanning the protein shape.[45] Here, we not only demonstrate that the PPD interface and ligand binding sites distribute along the molecule with a clear directionality (e.g., more catalytic residues can be found along the direction of PC1 or D-plane normal than a "spherical" search by random planes) but also take a step further to verify that the chances of finding near-native hits can be enriched from sets of decoys using our criteria based on S-/D-planes of receptors and $\theta_{\text{D-axes}}$. These findings suggest a pervasive principle for molecular interactions including but not limited to PPD and enzyme−substrate recognitions.[7,8]

Applicability of Our Methods and Observed Rules in Molecular Interactions. Table S7 of the Supporting Information shows that the rules we found do not apply to a particular shape or size group of proteins, although minor differences can be found in "averages" of concerned quantities. For instance, only looking at the "averages", we found that less extended ($\lambda1/\lambda2\_b = 2.3$) proteins with smaller sizes (34.6 kD) tend to follow our rules better than extended ($\lambda1/\lambda2\_b = 3.0$) and larger (38.2 kD) proteins (Table S7, Supporting Information). However, these differences are not contrasting with statistical significance ($p = 0.05$). On the other hand, our data show that "flexible" proteins comply to herein found rules more so than the other two protein categories for its high ratio of "D-planes cutting through docking partners" and the larger $\langle\theta_{\text{D-axes}}\rangle$ (for GNM only)(Table 1). However, as for the rules applied to find sets of decoys having smaller D-RMSD from the native hits, the rules herein found do not necessarily perform better in flexible category than the other two categories. For instance, with the filters "D-plane of big protein cutting through the small protein" and "$\theta_{\text{D-axes}} > 40$" GNM defined IDDs filtered decoys have $\Delta_{\text{D-RMSD}}$ values of −2.17 and −1.43 Å for Medium and Flexible categories, respectively (Table 2). In addition, the rules can be applied to proteins of various functions with minor differences between functional categories. For instance, in our data set comprising 40 enzymes and 28 nonenzymes, there are 67.5%, 75.0%, and 72.5% of the enzymes fulfilling the conditions of θ, B, and S, respectively (according to the definitions in Table 1), while the numbers are 67.9%, 85.7%, and 85.7% for the 28 nonenzymes. For condition "B", nonenzymes are 12.7% higher than enzymes even though the nonenzymes are found larger than enzymes (39.0 vs 32.9 kD for big proteins).

GNM Results Are Closer to PC1's than to ANM's. The slowest GNM mode is a strongly coupled vibrational and rotational motion of the protein, where the ANM modes describe purely vibrational motions.[25,26] We show in this study that this rotation motion has its rotation axis on the S-/D-planes. In this study, both PC1 (rotation only) and GNM (vibration coupled with rotation) provide distinctive dynamics features for binding sites and PPD poses more so than ANM. It can be found that the center of transition points defined by GNM is closer to the mass center (where S-plane is placed) than that defined by ANM (Figure S3, Supporting Information). Also, the GNM-defined D-plane normal is close to PC1 with a crossing angle of ∼10°, whereas such an angle between the ANM-defined D-plane normal and PC1 is >26° (Figure S3, Supporting Information). That PPD orientation conforms to GNM features more than ANM ones would mean that the immobile residues defined by the GNM (immobile in both rotation and vibration) can be more functionally distinctive than those found by the ANM (immobile in vibration only).

CONCLUSIONS

We have demonstrated in this study that the PPD follows a positional/orientational predisposition that is regulated by protein shape and intrinsic dynamics. With the developed method, intrinsic dynamics domains that characterize the most robust dynamics features of a protein, we are able to examine the PPD phenomenon in a consistent framework. In this framework, we are able to reveal that protein−protein and protein−small ligand docking has a preference to occur at places that have already inert dynamics, both in their rotational and vibrational degrees of freedom.

In this study, we also give explicit criteria to consider PPD patches only in the "docking belts" of the big proteins while maintaining an angle of crossing D-axes larger than 40° and to consider enzymes' active sites be within 50% of the residues closest to the S-/D-plane. By considering only favored docking orientations/positions suggested in this study, the conformational search space for the docking interface can be much reduced, both translationally (by D-plane results) and rotationally (by D-axes results), thus making possible a more efficient and hopefully more accurate PPD algorithm.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Seven tables, nine figures, supplemental results, and discussion and methods for both PPD and enzyme catalytic residues studies. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +886-3-574-2467. Fax: +886-3-571-5934. E-mail: lwyang@life.nthu.edu.tw.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. H. Li and S. Sakuraba have contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to the National Center for High-performance Computing for computer time and facilities and to the National Science Council for Grant NSC 102B0251 V9. We also thank Dr. Akio Kitao for discussions pertaining to this topic.

■ ABBREVIATIONS

IDD, intrinsic dynamics domains; PPD, protein−protein docking; GNM, Gaussian network model; ANM, anisotropic network model; PC1, first principal component; MD, molecular dynamics; rb, rigid-body; D-plane, domain-plane; S-plane, splitting-plane; D-axis, domain-axis; S-axis, splitting-axis; COM, center of mass; EFH, enrichment factor for hits; ER, enrichment ratio; D-RMSD, decoy root-mean-square deviation; I-RMSD, interface root-mean-square deviation; DCC, distinguishable conformational change; CSA, Catalytic Site Atlas

■ REFERENCES

(1) Katchalskikatzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular-surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2195−2199.

(2) Ten Eyck, L. F.; Mandell, J.; Roberts, V. A.; Pique, M. E. Surveying Molecular Interactions with DOT. In *Proceedings of the 1995 ACM/IEEE Supercomputing Conference*, San Diego, CA, 1995, pp 506−517.

(3) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **1997**, *272*, 106−120.

(4) Chen, R.; Weng, Z. P. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* **2002**, *47*, 281−294.

(5) Murakami, Y.; Jones, S. SHARP(2): Protein−protein interaction predictions using patch analysis. *Bioinformatics* **2006**, *22*, 1794−1795.

(6) Yang, L. W.; Bahar, I. Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. *Structure* **2005**, *13*, 893−904.

(7) Shrivastava, I. H.; Bahar, I. Common mechanism of pore opening shared by five different potassium channels. *Biophys. J.* **2006**, *90*, 3929−3940.

(8) Dutta, A.; Bahar, I. Metal-binding sites are designed to achieve optimal mechanical and signaling properties. *Structure* **2010**, *18*, 1140−1148.

(9) Meirovitch, H. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181−186.

(10) Andricioaei, I.; Karplus, M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289−6292.

(11) Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **1993**, *215*, 617−621.

(12) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **1997**, *79*, 3090−3093.

(13) Zimmermann, M. T.; Leelananda, S. P.; Kloczkowski, A.; Jernigan, R. L. Combining statistical potentials with dynamics-based entropies improves selection from protein decoys and docking poses. *J. Phys. Chem. B* **2012**, *116*, 6725−6731.

(14) Huang, S.-Y.; Yan, C.; Grinter, S. Z.; Chang, S.; Jiang, L.; Zou, X. Inclusion of the orientational entropic effect and low-resolution experimental information for protein−protein docking in Critical Assessment of PRedicted Interactions (CAPRI). *Proteins* **2013**, *81*, 2183−2191.

(15) Takemura, K.; Burri, R. R.; Ishikawa, T.; Ishikura, T.; Sakuraba, S.; Matubayasi, N.; Kuwata, K.; Kitao, A. Free-energy analysis of lysozyme−triNAG binding modes with all-atom molecular dynamics simulation combined with the solution theory in the energy representation. *Chem. Phys. Lett.* **2013**, *559*, 94−98.

(16) Takemura, K.; Guo, H.; Sakuraba, S.; Matubayasi, N.; Kitao, A. Evaluation of protein−protein docking model structures using all-atom molecular dynamics simulations combined with the solution theory in the energy representation. *J. Chem. Phys.* **2012**, *137*, 215105.

(17) Hwang, H.; Pierce, B.; Mintseris, J.; Janin, J.; Weng, Z. Protein−protein docking benchmark version 3.0. *Proteins* **2008**, *73*, 705−709.

(18) Chen, R.; Li, L.; Weng, Z. ZDOCK: An initial-stage protein-docking algorithm. *Proteins* **2003**, *52*, 80−87.

(19) Mintseris, J.; Wiehe, K.; Pierce, B.; Anderson, R.; Chen, R.; Janin, J.; Weng, Z. Protein−protein docking benchmark 2.0: An update. *Proteins* **2005**, *60*, 214−216.

(20) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein−protein docking benchmark version 4.0. *Proteins* **2010**, *78*, 3111−4.

(21) Porter, C. T.; Bartlett, G. J.; Thornton, J. M. The Catalytic Site Atlas: A resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **2004**, *32*, D129−33.

(22) Krissinel, E.; Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **2007**, *372*, 774−97.

(23) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−42.

(24) Bahar, I.; Lezon, T. R.; Yang, L. W.; Eyal, E. Global dynamics of proteins: Bridging between structure and function. *Annu. Rev. Biophys.* **2010**, *39*, 23−42.

(25) Thorpe, M. F. Comment on elastic network models and proteins. *Phys. Biol.* **2007**, *4*, 60−63.

(26) Yang, L. W. Models with energy penalty on interresidue rotation address insufficiencies of conventional elastic network models. *Biophys. J.* **2011**, *100*, 1784−1793.

(27) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **2001**, *80*, 505−515.

(28) Hayward, S.; Kitao, A.; Berendsen, H. J. C. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins* **1997**, *27*, 425−437.

(29) Hayward, S.; Lee, R. A. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J. Mol. Graph. Model.* **2002**, *21*, 181−183.

(30) Hinsen, K.; Thomas, A.; Field, M. J. Analysis of domain motions in large proteins. *Proteins* **1999**, *34*, 369−382.

(31) Nicolay, S.; Sanejouand, Y. H. Functional modes of proteins are among the most robust. *Phys. Rev. Lett.* **2006**, 96.

(32) Yang, L. W.; Eyal, E.; Bahar, I.; Kitao, A. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics. *Bioinformatics* **2009**, *25*, 606−14.

(33) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugenic.* **1936**, *7*, 179−188.

(34) Massey, F. J. The Kolmogorov−Smirnov Test for goodness of fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68−78.

(35) Fisher, R. A. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **1922**, *85*, 87−94.

(36) Jen-jacobson, L.; Engler, L. E.; Ames, J. T.; Kurpiewski, M. R.; Grigorescu, A. Thermodynamic parameters of specific and nonspecific protein−DNA binding. *Supramol. Chem.* **2000**, *12*, 143−160.

(37) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Herndon, VA, 2000; Chapter 8, pp 134−136.

(38) Bahar, I.; Wallqvist, A.; Covell, D. G.; Jernigan, R. L. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry* **1998**, *37*, 1067−75.

(39) Bromberg, S.; Dill, K. A. Side-chain entropy and packing in proteins. *Protein Sci.* **1994**, *3*, 997−1009.

(40) Murray, C.; Verdonk, M. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 741−753.

(41) Minh, D. D. L.; Bui, J. M.; Chang, C.-e.; Jain, T.; Swanson, J. M. J.; McCammon, J. A. The entropic cost of protein−protein association: A case study on acetylcholinesterase binding to fasciculin-2. *Biophys. J.* **2005**, *89*, L25−L27.

(42) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **2007**, *448*, 325−329.

(43) Grünberg, R.; Nilges, M.; Leckner, J. Flexibility and conformational entropy in protein−protein binding. *Structure* **2006**, *14*, 1205.

(44) Nicola, G.; Vakser, I. A. A simple shape characteristic of protein−protein recognition. *Bioinformatics* **2007**, *23*, 789−92.

(45) Foote, J.; Raman, A. A relation between the principal axes of inertia and ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 978−83.