JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

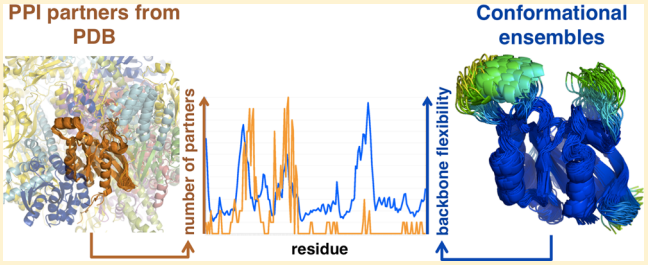
Specialized Dynamical Properties of Promiscuous Residues Revealed by Simulated Conformational Ensembles

Arianna Fornili,*[,†] Alessandro Pandini,[†,‡] Hui-Chun Lu,[†] and Franca Fraternali*[,†,§]

[†]Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, London SE1 1UL, United Kingdom
[‡]Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, London NW7 1AA, United Kingdom
[§]The Thomas Young Centre for Theory and Simulation of Materials, London, United Kingdom

Supporting Information

ABSTRACT: The ability to interact with different partners is one of the most important features in proteins. Proteins that bind a large number of partners (hubs) have been often associated with intrinsic disorder. However, many examples exist of hubs with an ordered structure, and evidence of a general mechanism promoting promiscuity in ordered proteins is still elusive. An intriguing hypothesis is that promiscuous binding sites have specific dynamical properties, distinct from the rest of the interface and pre-existing in the protein isolated state. Here, we present the first comprehensive study of the intrinsic dynamics of promiscuous residues in a large protein data set. Different computational methods, from coarse-grained elastic models to geometry-based sampling methods and to full-atom Molecular Dynamics simulations, were used to generate conformational ensembles for the isolated proteins. The flexibility and dynamic correlations of interface residues with a different degree of binding promiscuity were calculated and compared considering side chain and backbone motions, the latter both on a local and on a global scale. The study revealed that (a) promiscuous residues tend to be more flexible than nonpromiscuous ones, (b) this additional flexibility has a higher degree of organization, and (c) evolutionary conservation and binding promiscuity have opposite effects on intrinsic dynamics. Findings on simulated ensembles were also validated on ensembles of experimental structures extracted from the Protein Data Bank (PDB). Additionally, the low occurrence of single nucleotide polymorphisms observed for promiscuous residues indicated a tendency to preserve binding diversity at these positions. A case study on two ubiquitin-like proteins exemplifies how binding promiscuity in evolutionary related proteins can be modulated by the fine-tuning of the interface dynamics. The interplay between promiscuity and flexibility highlighted here can inspire new directions in protein−protein interaction prediction and design methods.



INTRODUCTION

The ability of proteins to interact with different partners is essential to the life of the cell. Indeed, the withdrawal or damage of proteins with a high number of connections in protein−protein interaction (PPI) networks (hubs) is often lethal[1] or associated with disease.[2]

Binding to multiple partners can be promoted by different factors,[3−9] including multiple domains, domain repeats, multiple interaction sites, intrinsically disordered regions, conformational changes, highly charged surfaces, post-translational modifications, and alternative splicing. Even if using different definitions,[4,10−12] hub proteins have been traditionally classified in two groups, according to their tendency to use the same ("date", "singlish-interface," or "sociable" hubs) or distinct ("party", "multiple interface," or "stable" hubs) interfaces to interact with different partners. Promiscuity in binding sites that are reused in different complexes has been often found to rely on the ability of the protein to adopt different conformations according to the partner,[12−16] with changes ranging from local side chain reorientations to global rearrangements and to disorder-to-order transitions.

The role of intrinsic disorder in hub proteins has been thoroughly investigated in the past.[3,4,9,15,17] However, there are many examples of multipartner proteins with an ordered structure, such as calmodulin, ubiquitin, or Ras proteins. In these cases, a possibility is that, in the absence of intrinsic disorder, promiscuity is promoted by specific dynamical properties pre-existing in the protein isolated state. Indeed, in the last decades an increasing number of experimental and theoretical studies[18−23] have shown that the conformational changes related with a change in the protein state (binding to ligands, post-translational modifications, interaction with light) can be sampled by the protein also in the initial unperturbed state. Some recent works highlighted that in specific cases the formation of multiple arrangements at a protein−protein interface can be assisted by conformational fluctuations in the unbound state in one of the partners.[16,24−27] However, only a few large-scale studies[28] of structure-based PPI networks have explicitly considered the inherent flexibility of proteins and

Received: June 9, 2013
Published: September 27, 2013

   dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127−5147

evidence of a general relationship between intrinsic dynamics and binding promiscuity in ordered proteins is still missing.

A first problem in the determination of a link between intrinsic dynamics and binding promiscuity is that, while classifications in PPI networks are usually made at the protein level (e.g., hub and nonhub), the regions in a protein that are involved in interactions with partners (semi-interfaces or simply interfaces in the following) can have a largely heterogeneous composition in terms of different properties, including mobility,[29−31] evolutionary conservation,[32] binding promiscuity,[13,33] and binding affinity.[34−36] Thus, a promiscuous protein can be composed of both nonpromiscuous and promiscuous residues, each with a different degree of intrinsic flexibility. Moreover, the flexibility of a protein can be measured in terms of different space and time scales, with previous works usually focusing on either backbone motions,[24,28,31,37] slow and highly collective, or side chain motions,[30] faster and more localized.

Here, we present the first large scale study of the intrinsic dynamical properties of promiscuous residues, where we mapped residue-based measures of conformational flexibility and binding promiscuity on interfaces of a large data set of proteins. Both backbone and side chain intrinsic motions were described with *in silico* generated conformational ensembles of the isolated proteins. The analysis revealed higher flexibility levels for promiscuous residues compared to nonpromiscuous ones. This additional flexibility was found to be highly organized in correlated motions both on a local and on a global scale, indicating that, when intrinsic disorder is not a major factor, an ordered form of flexibility can take over to promote binding promiscuity. Remarkably, the magnitude of intrinsic motions in promiscuous residues showed a reduced dependence from evolutionary conservation when compared with nonpromiscuous ones, providing an unprecedented indication that binding affinity and promiscuity can have opposite effects on residue dynamics. The functional importance of promiscuous residues was also confirmed with an investigation of the distribution of single nucleotide polymorphisms (SNPs) across interface regions, which suggests that residues in promiscuous positions have a reduced tolerance to genetic variations, related to the necessity to preserve their binding polyvalence.

METHODS

Data Set Preparation. A data set of proteins was generated using the PiSite database[38] (Figure 1). A nonredundant list of proteins from PiSite was used, composed of 7739 proteins with sequence identity <30% identified by clustering the sequences of 110325 proteins from 51482 PDB entries.[38] For a given protein, PiSite collects all the PDB entries containing the protein itself or very close homologues (>90% sequence identity). The original protein from the nonredundant PiSite list and its homologues define a sequence family. In the following, we will refer to sequence families as simply families and to the original proteins in the nonredundant PiSite list as Family Representatives (FRs). In PiSite, all the PDB complexes containing a member of the family are used to define the family binding properties (number of partners and binding states), which are then mapped onto the FR. The basic assumption is that, because of their high sequence similarity, members of the same family have a very similar behavior in terms of binding interactions, so that structural information on partners can be transferred from the family members to the FR. In this way, the partner annotation of FR is enriched compared to when only FR occurrences in the PDB are considered, and the bias due to the incompleteness of the PDB is reduced.[12]

For each family in nonredundant PiSite (Figure 1), we selected the members with known UniProtKB and SCOP IDs using the PDBSWS PDB/UniProt mapping[39] and the SCOP IDs using the Astral SCOP database[40] (v 1.75). Only the families with unambiguous SCOP domain classification across all the members and with at least one member with known UniProtKB ID were retained. We then selected 251 families that satisfied the following requirements:

(1) Their members have only one SCOP domain.
(2) The sequences of the resolved structures in the family cover at least 75% of the corresponding UniProtKB sequences.
(3) They have at least one partner with known structure.
(4) There is at least one structure in the family with no gaps in the resolved main chain. The ungapped X-ray structure with the best resolution in the family was selected as the structural representative (SR) to be used in the simulations and structural analyses. When no crystallographic structure with a complete main chain was found, an ungapped NMR structure was selected as SR if available.

The 251 families define our full data set S^{Full} (Supporting Information (SI) Table S1). Each family includes on average ∼20 members, for a total of 4917 PDB chains.

The PDB and chain IDs of partners from structural complexes (or 'structural partners') were extracted from PiSite for each family in S^{Full} and mapped to UniProtKB with the PDBSWS mapping. For consistency with the PiSite database, redundancies were removed by clustering the partner sequences with BLASTCLUST[41] (sequence identity ≤30%, sequence coverage ≥50%) to define the total number of structural partners (np_S) of the family. This was used to classify the families of S^{Full} as monopartner ($np_S = 1$, for a total of 151 families) and multipartner ($np_S \geq 2$, for a total of 100 families). Each of the main SCOP classes (α, β, α/β, $\alpha+\beta$) was represented in both groups with at least 14 entries. The proteins in the families have an average size of 217 (monopartner) and 167 (multipartner) residues, as measured from the SRs.

Family partners identified by nonstructural detection methods were extracted from the IntAct[42] database using the UniprotKB IDs of each family member. All the binary interactions were retained. The total number of unique partners (np) of each family was obtained using BLASTCLUST (see above) on the sequences of all the partners extracted from either IntAct or PiSite.

In the following, we will often refer to a family as simply a protein, with the FR as its sequence representative. Unless otherwise stated, structural properties and simulated conformational ensembles were calculated from the SR of the family, while binding multiplicity profiles and ensembles of experimental structures were built-up using all the family members (see the following and inset in Figure 1).

MD simulations (see below) were performed on a subset (S^{MD}) of S^{Full} composed of 6 monopartner and 6 multipartner proteins. The S^{MD} proteins were selected using stricter criteria to increase the confidence of their classification, so that multipartner proteins have at least 4 structural partners, the monopartner proteins have both np and $np_S = 1$, SCOP classes are equally sampled by the two groups and the average size of the proteins in the two groups is comparable (SI Tables S1 and S2).

A second data set (S^{Soc}) was considered to test and compare the findings observed on S^{Full}. S^{Soc} is composed of sociable proteins[12] (SI Table S3), defined as having at least 3 structural

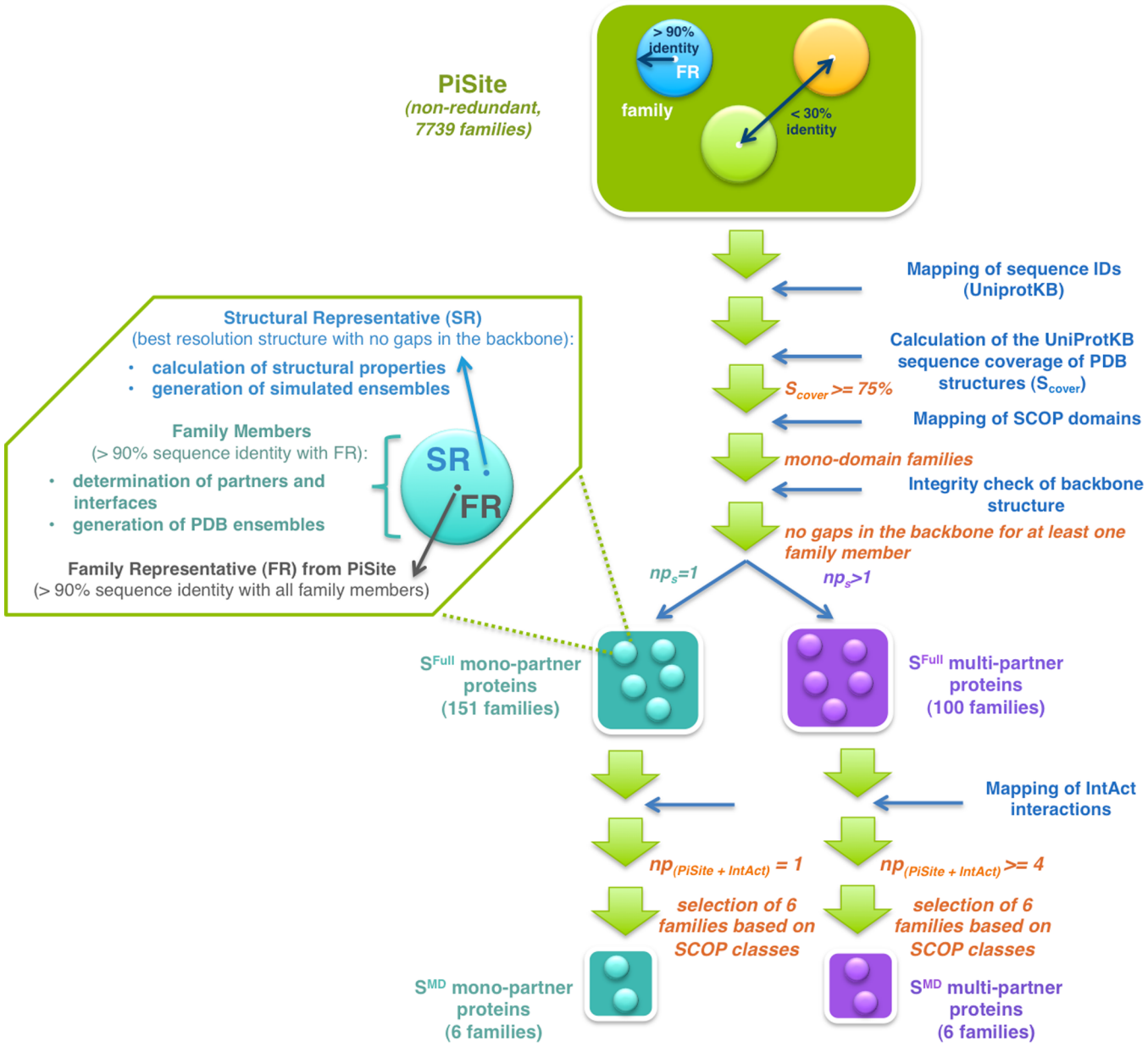


Figure 1. Generation of S^{Full} and S^{MD} data sets. The flowchart shows the steps followed in the generation of the S^{Full} and S^{MD} data sets from the nonredundant PiSite database. The operations indicated in blue on the right were performed on all the members of each family. The applied filters are indicated in orange. The inset shows how the components of each family are defined and used. See Methods in the main text for a full description of the procedures involved.

partners and 3 different binding states.[38] Starting from the original list of 102 sociable proteins,[12] 69 were selected so that the FR has a complete main chain structure. The FR was used as the structural representative SR for the simulations and structural analyses. Since the original data set of sociable proteins was not filtered for monodomain proteins, S^{Soc} includes entries with more than one SCOP domain (SI Table S3).

Only a limited number of proteins in the data sets (4 in S^{Full} and 5 in S^{Soc}) contained post-translational modifications (PTMs) in the PDB structure, amounting to less than 3% of the total. None of these was found in the S^{MD} subset. Thus, PTMs were not explicitly taken into account and residues with PTMs were modeled using the corresponding unmodified residue. A systematic study of the effects of PTMs on interface dynamics would require ad hoc data sets and a standardized framework including parameters for a large range of PTMs. A complete set of tools for the inclusion of PTMs in MD simulations has been recently proposed.[43,44]

Interface Definition. For each S^{Full} family *f*, interface residues were mapped on the SR by analyzing all the PDB binary complexes A^f:B (where A^f is a member of the family and B is one of its partners) listed in the PiSite database for the selected family members. For consistency with PiSite, the binary complexes were extracted from the PDB biological units. The POPSComp method[45] was used to determine the profiles of normalized solvent accessible surface area (SASA) that is buried upon complex formation for each chain A^f $\Delta SASA^n(i, A^f{:}B) = (SASA(i, A^f) - SASA(i, A^f{:}B))/ SASA(i, A^f)$, where $SASA(i, A^f)$ and $SASA(i, A^f{:}B)$ are the SASAs of residue *i* in the isolated and bound A^f molecule, respectively. All the $\Delta SASA^n$ profiles of the

dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127−5147

family were then mapped onto the SR sequence using A^f–SR sequence alignments given by the program ProFit.[46] The resulting $\Delta SASA^n(i, SR:B)$ values were used to define the SR interfaces. A SR residue i was considered to be part of a SR:B interface if upon complex formation it buries at least 10% of its SASA, that is, if $\Delta SASA^n(i, SR:B) > 0.1$. If i is found in interfaces involving n different partners, its binding multiplicity $b(i)$ is set to n. The binding multiplicity was used to partition the interface residues into three different binding classes. Residues with $b \geq 2$ were classified as multipartner (c_{multi}), while residues with $b = 1$ were grouped in the c_{mono} class if belonging to monopartner SRs and in the $c_{mono\_in\_multi}$ class if belonging to multipartner SRs (SI Table S4). Because of differences in the definition of the interface (distance-based in PiSite and SASA-based here) and of the selection process applied in the data set generation (see Data Set Preparation), the binding multiplicity used for the S^{Full} data set may differ in isolated positions from the one recorded in PiSite. To ensure full consistency in the analysis of the S^{Soc} data set, original b profiles from PiSite were used for sociable proteins.

There was some redundancy in the PDB biological units of each family, so that very similar A^f:B complexes were observed. Interface redundancy was eliminated by clustering the SR-mapped $\Delta SASA$ profiles according to their overlap matrix $O_{ij} = (\Delta SASA_i(i \cap j) + \Delta SASA_j(i \cap j))/(\Delta SASA_i + \Delta SASA_j)$, where $\Delta SASA_i$ is the buried SASA of SR in interface i and $\Delta SASA_i(i \cap j)$ is the surface area of the SR residues involved both in interface i and j, calculated using the coordinates of interface i. The clustering was performed with the complete linkage method and the optimal number of clusters was chosen by maximizing the average silhouette[47] value s. An average s_{max} of 0.84 ($\pm$0.19) was obtained for the S^{Full} families, indicating a good partitioning of the interfaces into the clusters. The interfaces with the largest SASA within a given cluster were chosen as representatives to generate the final nonredundant set of 695 interfaces (SI Table S4).

The interfaces were partitioned into three classes according to the maximum binding multiplicity of their residues and the binding class of the corresponding proteins (SI Table S4). Consistently with the residue classification, interfaces containing at least one multipartner residue where defined as multipartner (c_{multi}), while the remaining interfaces were classified as c_{mono} and $c_{mono\_in\_multi}$ if belonging to mono- or multipartner proteins, respectively. The distributions of the total and relative hydrophobic interface sizes for c_{multi} (average $\Delta SASA$ = 1583 Å^2, average $\Delta SASA_{phob}{}^r$ = 62%) and c_{mono} (average $\Delta SASA$ = 1346 Å^2, average $\Delta SASA_{phob}{}^r$ = 61%) were not statistically different, while the $c_{mono\_in\_multi}$ interfaces were significantly smaller (average $\Delta SASA$ = 803 Å^2, Wilcoxon p-value $< 2 \times 10^{-6}$) and less hydrophobic (average $\Delta SASA_{phob}{}^r$ = 55%, Wilcoxon p-value $< 2 \times 10^{-4}$). The values of the relative hydrophobic interface size ($\Delta SASA_{phob}{}^r$) were calculated as the relative contribution to the total $\Delta SASA$ of the interface residues classified as hydrophobic in POPS.[48]

Interface Analysis. The physicochemical properties of interface residues in the different binding classes were analyzed in terms of propensities relative to all solvent exposed residues in the corresponding protein class.[33,49] The propensity $P_x(c_i)$ of property x in the binding class c_i was calculated as $P_x(c_i) = p_x(c_i)/p_x(\text{surf})$, where $p_x(c_i)$ and $p_x(\text{surf})$ are the fraction of residues with property x in c_i and at the surface, respectively. A $P_x(c_i)$ value >1 indicates a higher abundance of residues with a given property in c_i than in the rest of the surface. Solvent-exposed or surface residues were defined as having a relative accessibility $SASA^r \geq 15\%$[50] in the SR of the family. The $SASA^r$ values were calculated normalizing the SASA of each residue by that of the corresponding amino acid type X in the AXA tripeptide, where X is in an extended side chain conformation[51] selected from the Dunbrack and Cohen rotamer library[52] as implemented in PyMOL.[53] Propensities were calculated for the amino acid identity, the evolutionary conservation grade as determined by ConSurf,[54] the DSSP[55] secondary structure definition, the relative accessibility of the isolated protein $SASA^r$, and the extent of surface burial in the complex $\Delta SASA^n$. To provide a measure of the uncertainty associated with this calculation and to rule out the possibility that the observed propensities were biased by a few cases, confidence intervals (CI) at 95% were calculated by bootstrap resampling with 1000 replicates. The statistical significance of differences between the propensities was estimated with a Student's t test on the CI.[56,57] Resampling and statistical analyses were performed with R.[58]

Interaction hot spots were predicted using different methods, namely ANCHOR,[59] Robetta,[60] PISA,[61] HotPoint,[62] and KFC2a/b.[63] The prediction was performed on the nonredundant interfaces described above. The criteria used for the hot spot definition are summarized in SI Table S5. Robetta, HotPoint, and the KFC2a/b methods have been specifically parametrized for hot spot prediction against experimental alanine scanning data and the default criteria for residue classification defined in the original publications were used. The ANCHOR server[59] has instead been designed to scan for residues that, besides contributing significantly to the binding energy, are also 'anchors' (i.e., they have a large exposed surface in the monomeric state). In addition, we used PISA, an established method to determine the thermodynamic stability of protein assemblies and to distinguish biological interfaces from artifacts of crystal packing.[61] Here, the single contributions of the residues to the interface stability were used to estimate their importance in the complex formation. Thresholds on ANCHOR and PISA binding free energies were selected to obtain an overall hot spot fraction comparable to the other methods.

Hydration scores S_{hyd} were evaluated for each protein in the S^{MD} data set from the spatial distribution of water molecules observed in MD trajectories (see below). Water density maps $g(\mathbf{r})$[64−66] were calculated at discrete points $\mathbf{r}$ defined by a 0.5-Å spaced rectangular grid around the solute. Frames saved every 0.1 ps from the last 10 ns of the trajectory were superimposed to a reference using C^α atom positions to remove the overall roto-translational motion of the protein. The number density of the water oxygen atoms was then averaged at each grid point over the MD frames and normalized by the bulk density evaluated in the 6−8 Å shell around the solute. The hydration sites were then identified as local maxima of the density map with $g(\mathbf{r}) > 1$ and used to define the hydration score S_{hyd} as previously described.[66] Residues with a high S_{hyd} score were surrounded either by a large number of maxima or by maxima with a high density. The S_{hyd} value of each residue i of type aa was standardized by calculating the ratio $(S_{hyd} - \mu_{hyd}{}^{aa})/\sigma_{hyd}{}^{aa}$, where $\mu_{hyd}{}^{aa}$ and $\sigma_{hyd}{}^{aa}$ indicate the average and standard deviation of S_{hyd} calculated over residues of type aa.

Generation of Conformational Ensembles. Ensembles of conformations were generated for each SR in the S^{Full} and S^{Soc} data sets with the tCONCOORD[67] method. Given a starting structure, tCONCOORD samples alternative conformations by fulfilling a set of geometrical constraints as determined from the initial coordinates and interaction types (e.g., covalent bonds, hydrogen bonds, salt bridges, or hydrophobic interactions). Under-wrapped hydrogen bonds[68] are detected and modeled as

unstable. Each SR in the data sets was first energy-minimized using the OPLS-AA[69] force field with 250 steps of the steepest descent algorithm. Ensembles of 500 structures were then generated using standard tCONCOORD parameters.[67,70]

Molecular Dynamics (MD) simulations of the 12 SRs of the S^{MD} data set (SI Table S2) were performed with GROMACS 4.0.[71] The initial structures were solvated with a cubic box of SPC water molecules, setting the minimal distance between the protein and the box boundaries to 12 Å. Ionizable residues were modeled in their standard protonation state at pH 7. The systems were then neutralized adding the appropriate number of counterions. The final composition of the systems is detailed in SI Table S6. The GROMOS-96 force field was used with the 43a1 parameter set.[72] Periodic boundary conditions were imposed. All the bonds were frozen and a 2-fs time step was used. The Berendsen[73] algorithm was employed for temperature and pressure regulation, with coupling constants of 0.2 and 1 ps, respectively. The electrostatic interactions were calculated with the particle mesh Ewald method,[74] with a 14-Å cutoff for the direct space sums, a 1.2-Å FFT grid spacing, and a 4-order interpolation polynomial for the reciprocal space sums. For van der Waals interactions, a 14-Å cutoff was used. The neighbor list for noncovalent interactions was updated every 5 steps. The systems were first minimized with 1000 steps of steepest descent. Harmonic positional restraints with a force constant of 4.8 kcal·mol·Å^{-2} were imposed onto the protein heavy atoms and gradually reduced to 1.2 kcal·mol·Å^{-2} in 80 ps, while the temperature was increased from 200 to 300 K at constant volume. The system was then simulated at constant temperature (300 K) and pressure (1 bar) for 100 ps. After the removal of harmonic restraints, 2 ns of equilibration were run in NPT conditions. NPT production simulations were then run for 40 ns for each system. System coordinates were saved every 0.1 ps.

Gaussian Network Model (GNM) calculations were performed by representing each protein in S^{Full} as a network of C^{α} atoms linked by harmonic springs. Each node was assumed to fluctuate according to a Gaussian distribution around its equilibrium position, defined by the coordinates of the starting structure. Root mean square fluctuation (RMSF) values were derived for each C^{α} atom by inverting the Kirchhoff matrix[75] built-up using a unitary force constant for the springs and a 7 Å cutoff on the distance between C^{α} atoms.

The conformational variability within each family of S^{Full} was evaluated on the ensemble of PDB structures composed by all the family members as defined in Data Set Preparation. The ensembles were built-up by collecting either X-ray or NMR occurrences as listed in PiSite. Ensembles with more than one structure were found for 241 proteins in S^{Full} and 11 proteins in S^{MD}. In the generation of the ensembles, the PDB structures were first aligned with ProFit. Only the residues occurring in all the structures were retained, so that the resulting ensembles could be analyzed as pseudotrajectories with standard MD analysis tools. Each structure was also labeled as bound or unbound according to its binding state as reported in PiSite.

Analysis of Flexibility. C^{α} and side chain root-mean-square fluctuations (RMSF) of the residues from their average position were calculated for the generated tCONCOORD, MD, and PDB ensembles. C^{α} RMSF profiles were determined after removal of the overall roto-translational motion by best-fit superposition of the structures to all the C^{α} atoms of the starting experimental structure (SR) used as a reference. For side chains, the RMSF values were calculated after removal of the main chain roto-translation of the single residues,[30] using the experimental structure as a reference for the best-fit superposition.

The comparison of flexibilities between different proteins and types of ensemble was performed using standardized C^{α} and side chain RMSF profiles $Z\text{-score}(i) = (\mathrm{RMSF}_j(i) - \mu_j)/\sigma_j$, where $\mathrm{RMSF}_j(i)$ is the RMSF of residue i in protein j, while μ_j and σ_j are the average and standard deviation of the RMSF_j distribution.

Relevant rearrangements on protein interfaces could also arise from subtle conformational changes of local structures.[23,76] These changes are often difficult to detect by traditional flexibility analysis and require the isolation of local dynamics from the global motion.[23] To this end, the dynamics of local structures in the tCONCOORD and MD ensembles was analyzed with a fragment-based approach. It was previously shown that local conformational changes and their correlation are effectively described by means of a Structural Alphabet (SA) including prototypical geometries of backbone fragments.[23,77] In the present study, the M32K25 SA[77] was used. This SA comprises 25 representative fragments of 4 consecutive C^{α} atoms, and it was specifically designed to include the most typical local structures, as well as to correctly describe conformational transitions sampled by molecular simulations. Each fragment in the SA is labeled with a letter representing a prototypical conformational state. Therefore, any 4-residue-long segment in a protein structure can be annotated with a letter. The labeling is performed by identification of the most similar SA fragment in terms of root-mean-square deviation (RMSD)[23,77,78] between the protein segment and the letter. The conformation of a protein of n residues can be encoded into a structural string of length $n - 3$.[79] Following this procedure an ensemble of conformations is condensed to an alignment of structural strings. A column of this alignment summarizes the conformational states sampled by a protein segment within the simulation.

The correlation of conformational changes in a pair of protein segments (i,j) can be calculated as normalized Mutual Information (MI) between the associated columns in the alignment (eq 1):

$$I^n(C_i; C_j) = (I(C_i; C_j) - \varepsilon(C_i; C_j))/H(C_i, C_j) \quad (1)$$

where C_i and C_j are the relevant columns in the structural string alignment, $I(C_i;C_j)$ is the MI, $H(C_i,C_j)$ is the joint entropy,[80] and $\varepsilon(C_i;C_j)$ is the expected finite size error.[81] It was previously demonstrated that local correlated motions are instrumental for allosteric signal transmission and may be involved in conformational changes of interacting interfaces.[23] To this end, the distribution of statistically significant local correlations was compared for surface and interface fragments in the three binding classes. Statistically significant correlations were identified as previously reported.[23]

To analyze the flexibility of single 4-residue fragments, a fragment RMSF was calculated by defining $n-3$ sliding windows or fragments of 4 adjacent C^{α} atoms. For all the structures in the ensembles, each fragment was superimposed onto the reference experimental structure SR independently from the rest of the protein, to remove local roto-translational motions. The fragment RMSF was then calculated as the quadratic mean of the RMSF values of each C^{α} within the window.[77]

The correlation between single interfaces and global motions was investigated with the Functional Mode Analysis (FMA).[82] Given a functionally relevant property, FMA can be used to identify the protein collective motion that is maximally correlated with it (maximally correlated motion or MCM). This is usually expressed as a linear combination of principal components (PCs) derived from a principal component analysis (PCA)[83] of the system trajectory. We selected the FMA approach where the coefficients of the linear expansion β_i are determined by

maximizing the Pearson correlation coefficient between the time evolution of the functional property $F(t)$ and the projection of the trajectory onto the MCM. In this case, the variance of F during the dynamics var(F) can be approximated as[82] $\mathrm{var}(F) \approx \Sigma_{i=1,l} \beta_i^2\, \mathrm{var}(\mathrm{pc}_i)$, where pc_i is the projection of the trajectory onto the ith PC and l is the number of PCs considered in the MCM expansion. The relative contribution of PC_i to var(F) was thus evaluated as $\mathrm{pvar}_i = \beta_i^2\, \mathrm{var}(\mathrm{pc}_i)/\mathrm{var}(F)$.

The interface radius of gyration R_g^{IF} was considered in this study as a functional quantity related to the overall shape of the interface. R_g^{IF} was calculated over the C^α atoms of the interface residues for each nonredundant interface in S^{Full}. Using an alternative property to describe the interface geometry, namely the distance RMSD calculated over all the possible pairs of C^α atoms in the interface,[84] did not affect the conclusions described in Results. The MCM was expanded in the essential space of all C^α atoms, composed of the first l PCs accounting for the 90% of the total C^α fluctuation. An average essential space size of 15 (±7) PCs was observed in the whole data set. The first 450 structures of the tCONCOORD ensemble were used for the PCA analysis and as training set for the construction of the linear model. The remaining 50 structures were used for cross-validation.[82] On average, the optimized Pearson correlation coefficient between the MCM and the R_g^{IF} was 0.81 (±0.20) and 0.79 (±0.23) for the training and cross-correlation sets, respectively. This supports the validity of the linear model for the MCM and rules out the possibility of overfitting in the determination of the MCM. Since the motion along the MCM can be restricted in the underlying energy landscape,[82] while the PCs in the essential space represent the directions along which the protein motion has the largest amplitude, we also analyzed the single PC contributions pvar_i (see above) to the overall R_g^{IF} variance. The number $n\mathrm{PC}_{20}$ of distinct PCs with pvar ≥ 20% (deviating from the average value by ~1.6 σ) was determined. Different choices of the pvar threshold produced qualitatively similar results.

The comparison of the conformational spaces sampled by the MD, tCONCOORD and PDB ensembles was performed by calculating the normalized overlap[85] of the C^α covariance matrices as

$$\mathrm{CovMatOver(A, B)} = 1 - \mathrm{d(A, B)}/(\mathrm{trA} + \mathrm{trB})^{1/2} \qquad (2)$$

where A and B are the matrices to compare, tr is the trace operator and d(A,B) is the matrix difference: $\mathrm{d(A,B)} = (\mathrm{tr}[(\mathrm{A}^{1/2} - \mathrm{B}^{1/2})^2])^{1/2}$. The overlap ranges from 0 (no overlap) to 1 (identical matrices). The cumulative overlap between two sets of PCs was calculated as the average of the squared inner products between all the possible pairs of PCs from the two sets.[86] The cumulative overlap is 1 if the space spanned by the two sets is identical.

Estimates of per−residue configurational entropy values in the MD ensembles were obtained using the heuristic Schlitter's formula[87,88]

$$S = \frac{1}{2} k_B \ln \det\left[1 + \frac{k_B T e^2}{\hbar^2} \mathrm{M}\sigma\right] \qquad (3)$$

where k_B is the Boltzmann's constant, T is the temperature, e is the Euler's number, $\hbar$ is the Plank's constant divided by 2π, M is a diagonal matrix with the atomic masses, and σ is the covariance matrix. Entropy contributions from the overall protein roto-translation were removed by superimposing each frame in the MD trajectories to the reference experimental structure using all the C^α atoms. For each residue, the covariance matrix σ was then calculated considering all the atoms in the residue or only the main chain atoms. In the all-atom case, to take into account differences in the side chain length, entropy values were normalized by the total number of atoms in the residue and multiplied by 10 (average number of atoms in a residue considering heavy atoms and polar hydrogen atoms). It is to be noted that the Schlitter's formula is an approximation to the upper bound of the entropy. Moreover, the decomposition into residue contributions is not exact. Indeed, when applying eq 3 to single residues the correlation between the residue and the rest of the protein is not explicitly taken into account.[88]

Flexibility analyses were performed on all frames of the tCONCOORD and PDB ensembles and on MD frames saved every 1 ps.

Comparison of Conformational Ensembles. Most of the findings presented in Results are based on the analysis of tCONCOORD conformational ensembles. The tCONCOORD method can be considered as a relatively fast method to explore the conformational landscape of a protein. As explained above, tCONCOORD generates alternative conformations of a protein by satisfying the constraints derived from a starting structure. An all atom representation of the protein is used, and even if energetic terms are not directly involved in the generation of structures, the parameters defining the upper and lower bounds of the constraints depend on the type of interaction represented by each constraint. These parameters were originally determined on the basis of MD simulations.[89] Subsequently, they were modified and combined with an estimate of hydrogen bond stability to increase the portion of conformational space explored by the ensemble.[67] Compared to MD simulations, tCONCOORD lacks an explicit representation of solvent and long-range interactions, and it does not provide direct time or energetic information on the simulated processes. However, it is not affected by convergence issues, and it can rapidly cover a large part of the accessible conformational space. Indeed, tCONCOORD ensembles have been shown to be in general representative of the structural variability of a protein on the basis both of MD simulations and experimental data.[67,89−91] In particular, they can be effectively used as predictors of the relative flexibility of protein residues, to distinguish rigid regions from flexible ones. Moreover, large conformational transitions between different states of a protein (e.g., from the open unbound to a close bound-like conformation of calmodulin[67]) can be successfully reproduced starting from a single state.

To test the dependence of the main results of the paper on the specific method used for the generation of the ensembles, we considered three more sources of information on the conformational variability of the proteins. First, we performed short (40 ns) equilibrium MD simulations in explicit water for 12 proteins (S^{MD} data set) selected from S^{Full} (see Data Set Preparation). These proteins were selected so that the main conclusions can be tested on them; thus, (1) they are representative of the variability of the original data set in terms of binding properties and fold, (2) their partner annotation is the most reliable among the S^{Full} proteins, and (3) they do not have a bias in terms of size or secondary structure composition. With respect to tCONCOORD, ensembles from short MD simulations can be expected to provide a more accurate representation of the thermal fluctuations around the starting state, but they are less likely to sample the conformational space farther from it, particularly when high-energy barriers are involved.

As a third method, we used the GNM method[75,92] to calculate equilibrium fluctuations around the SR structures from the complete S^{Full} data set. Elastic Network Models are often used when large data sets are involved because of their reduced computational cost.[28,93,94]

Differently from tCONCOORD and MD, the GNM adopts a reduced, coarse-grained representation of the protein, where usually only C^{α} atoms are considered. Their interactions are described by harmonic potentials, so that a Gaussian distribution of the fluctuations around the equilibrium positions is assumed. Despite its simplicity, the GNM has been found to accurately describe the protein global motions and in particular the equilibrium fluctuations in the neighborhood of the starting structure.[95−97]

At last, the conformational variability actually observed in experimental structures was analyzed by collecting all the PDB occurrences of a given protein or of close homologues (see Generation of Conformational Ensembles). These collections of PDB structures (PDB ensembles) can contain useful information on the protein conformational variability.[98] Indeed, flexibility indices from PDB ensembles have been compared in the past with RMSF values from simulations.[67,98] It is to be noted that PDB ensembles can contain structures solved with different techniques (X-ray and NMR), in different experimental conditions and in different binding states. Thus, in addition to conformational changes related to the protein intrinsic flexibility, PDB ensembles can include changes induced by external factors (e.g., binding to other proteins or ligands) or due to the different experimental conditions. Moreover, they can be biased by experimental errors and by the fact that the PDB covers only a fraction of the known states of a protein.

For each of the described ensembles, it is possible to calculate indices measuring the extent of the structural change sampled by each residue within the ensemble (Analysis of Flexibility). As explained above, flexibility indices from different ensembles include different types of contributions and derive from sampling of conformations on different space and time scales. In particular, the absolute magnitude of the structural changes observed in the selected ensembles can be expected to be different. For example, the comparison of the pairwise RMSD distributions from tCONCOORD and MD ensembles of S^{MD} proteins (SI Table S7) suggests that on average larger portions of the conformational space are sampled by tCONCOORD than by MD simulations.

The relationship between binding promiscuity and flexibility indices derived from the different ensembles will be discussed in detail in the Results. However, the findings presented in this paper depend mostly on the relative order of flexibility of the residues in a protein and on the shape of the collective modes of motion along which the protein is most free to move. Hence, it is useful to briefly report the direct comparison of the order of flexibility predicted by the different ensembles and the calculation of the overlap between the PCs extracted from them.

The calculation of the correlation coefficients between the RMSF profiles shows a generally good agreement among the tCONCOORD, MD, and GNM ensembles in the S^{MD} database (SI Table S8), with average correlations ranging from 0.69 (MD/GNM) to 0.75 (C^{α} tCON/MD). These values are in line with previous comparisons performed on different proteins.[89,97,99,100] A good agreement between tCONCOORD and MD simulations was also found for side chains fluctuations, with an average correlation coefficient of 0.71. Moreover, when the comparison between GNM and tCONCOORD C^{α} RMSF profiles is extended from the S^{MD} to the S^{Full} data set, similar values of average correlation coefficients are obtained (0.70 for S^{MD} and 0.75 for S^{Full}), indicating that S^{MD} is not biased toward cases with high correlation values. As expected considering the higher heterogeneity of PDB ensembles, smaller values were obtained for the average correlation values between simulated and PDB ensembles, ranging from 0.47 for GNM/PDB in S^{Full} to 0.63 for MD/PDB side chain profiles in S^{MD}.

The conformational spaces sampled by the tCONCOORD and MD ensembles of S^{MD} were also compared (SI Table S9) by calculating:

(a) The normalized overlap between the overall covariance matrices of C^{α} atoms (CovMatOver, see eq 2). This value depends on the similarity both of the shape of the collective modes of motion (eigenvectors of the covariance matrix) and of the amplitude of the fluctuations along these motions (eigenvalues).

(b) The cumulative overlap between the essential subspaces spanned by the first 15 PCs (CumOver15). A value of 15 was chosen for CumOver15 since it is the average size of the essential space used for the FMA analysis of tCONCOORD ensembles (Analysis of Flexibility section). This overlap index depends uniquely on the covariance matrix eigenvectors.

(c) The maximum inner product between selected pairs of PCs from the compared ensembles (MaxInpr). MaxInpr was calculated to measure how well a PC from the tCONCOORD essential space can reproduce any of the first 3 PCs from MD. Similarly to CumOver15, this index does not depend on the covariance matrix eigenvalues.

The comparison of tCONCOORD and MD covariance matrices (CovMatOver) indicates a partial overlap between the two types of ensembles, with an average value of 0.339 (SI Table S9). The observation of higher values for CumOver15 (average = 0.474) and MaxInpr (average = 0.566) indicates a better agreement in the shape of the collective motions (defined by the relative amplitude and direction of C^{α} motions) than in the absolute amplitude of the fluctuations along them. These results are in line with previous observations.[89]

The MaxInpr index was used also to compare the first 3 PCs from the PDB ensembles with the PCs from the essential space of tCONCOORD and MD ensembles. Smaller values were obtained for both the PDB/tCON (average = 0.421) and the PDB/MD (average = 0.373) pairs compared to the MD/PDB ones (average = 0.566), indicating that collective motions in the simulated ensembles are more similar to each other than to the principal modes of structural change observed in the PDB ensembles. The larger values found in general for PDB/tCON pairs of PCs than for PDB/MD ones indicate that PDB PCs are better reproduced by tCONCOORD ensembles for the set of proteins studied here. This might be related to a more complete coverage of large conformational changes by tCONCOORD than by MD simulations (SI Table S7).

Analysis of SNPs. Human homologues of proteins in the S^{Full} and S^{Soc} data set were identified running NCBI-BLAST v.2.2.26+[101] (cutoff E-value = 1×10^{-2}) against a self-compiled library composed of 2583 human proteins with solved or homologous structures, annotated partner interactions and SNP mapping. Information on nonsynonymous SNPs was retrieved from the dbSNP database (build 315, ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/database/b135_archive/organism_data/b135_SNPContigLocusId_37_3.bcp),[102] while disease-related SNPs (DisSNPs) were extracted from the Online Mendelian Inheritance in Man (OMIM) database (www.omim.org/downloads);[103] therefore, only genetically inherited diseases variations are considered here. SNPs and DisSNPs were mapped back on the original sequences of the S^{Full} and S^{Soc} proteins using the BLAST sequence alignment. This procedure produced SNP annotation for 38 proteins of S^{Full} and 25 proteins of S^{Soc}.

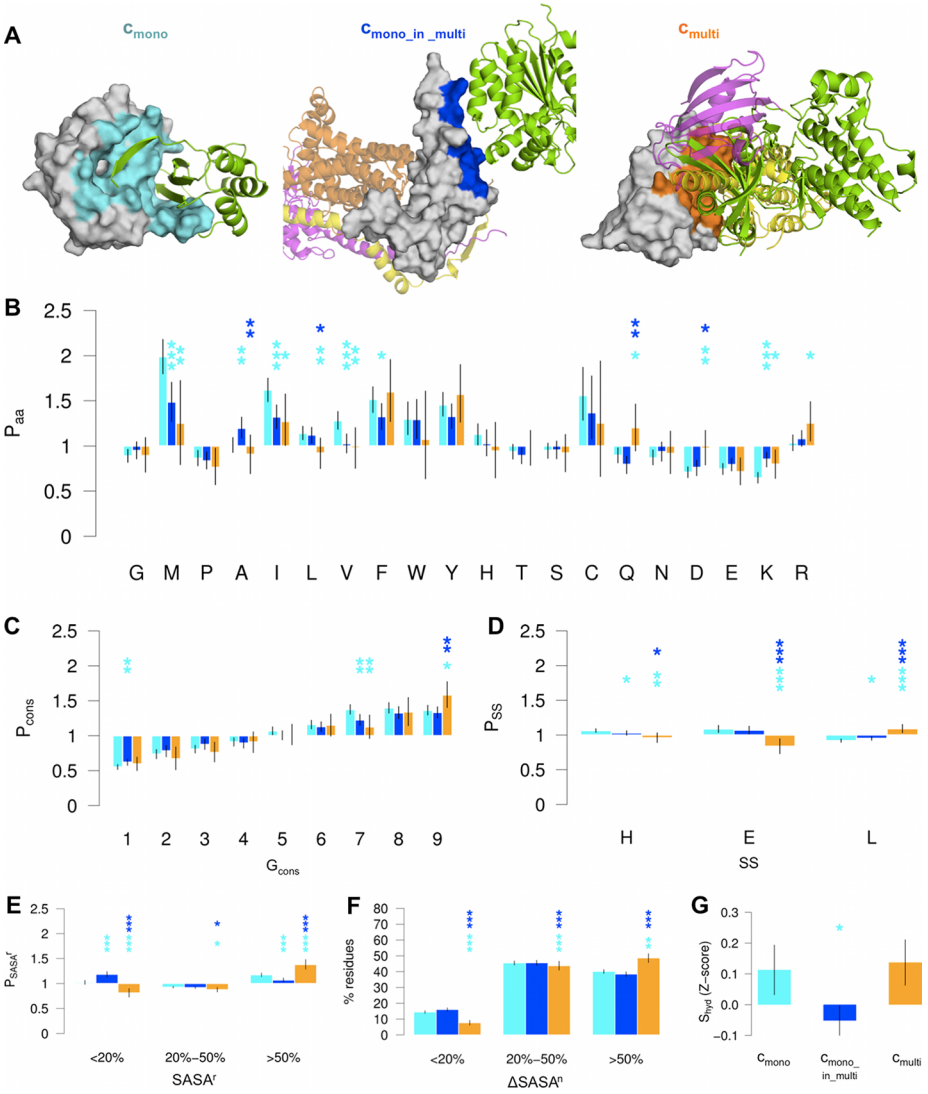


Figure 2. Physicochemical properties of interface residues in the c_{mono}, $c_{mono\_in\_multi}$, and c_{multi} binding classes. (A) Examples of c_{mono} (left), $c_{mono\_in\_multi}$ (middle), and c_{multi} (right) interfaces. The residues involved in each interface are color-mapped onto the surface in cyan (c_{mono}), blue ($c_{mono\_in\_multi}$), and orange (c_{multi}). Selected partners are represented as cartoon. Left: PDB ID 1x8d, chains D/C. Middle: PDB ID 1pp9, chains S/B. Right: PDB ID 1b6c, chains E/F. (B–E) Different types of calculated propensities relative to the surface of c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), and c_{multi} (orange) residues of S^{Full}. Error bars and significance levels were estimated with bootstrap resampling. Stars are drawn above $c_{mono\_in\_multi}$ and c_{multi} distributions, indicating the significance levels (*** p-value <0.001, ** 0.001 ≤ p-value <0.01, * 0.01 ≤ p-value <0.05) of the comparison with c_{mono} (cyan) and $c_{mono\_in\_multi}$ (blue). (B) Amino acid propensity P_{aa}. (C) Conservation propensity P_{cons}. The conservation is expressed as ConSurf conservation grade (G_{cons}, ranging from 1, less conserved, to 9, most conserved). (D) DSSP secondary structure (SS) propensity P_{SS}. The three SS groups collect positions annotated as helix (H = "H" + "G" + "I" in the DSSP dictionary), strand (E = "E" + "B") and loop (L = blank + "S" + "T"). (E) Relative solvent accessibility (SASA^r) propensity P_{SASA^r}. Three levels of SASA^r are considered, with SASA^r < 20%, 20% ≤ SASA^r < 50%, and SASA^r ≥ 50%. (F) Fraction of S^{Full} interface residues with a small (<20%), medium (20–50%) or large (≥50%) normalized buried SASA (ΔSASA^n). The maximum ΔSASA^n value is used for residues involved in more than one complex. (G) Standardized hydration score S_{hyd} in the S^{MD} binding classes.

Per-protein propensity values were calculated for each binding class and for each protein with the formulas described in Interface Analysis, using as reference the SNPs and DisSNPs abundance found at the surface of the single proteins.

RESULTS

A data set of 251 monodomain proteins (S^{Full}) was extracted from the PDB and partitioned into 151 monopartner and 100 multipartner proteins using the PiSite database[38] (Methods). The composition of the data set in terms of 7 general protein function categories (SI Figure S1) was obtained using a functional annotation of SCOP superfamilies.[104,105] As expected,[2,12,33] monopartner proteins (cyan) showed an enrichment in the metabolism and general categories, while multipartner proteins (magenta) were particularly rich in the categories related to extra- and intracellular processes, information, and regulation.

Interface residues for each protein in S^{Full} were extracted from its PDB complexes and partitioned into binding classes according to their binding multiplicity b (Methods). In particular, two classes c_{mono} and $c_{mono\_in\_multi}$ were defined for monopartner residues (b = 1) belonging to monopartner and multipartner proteins, respectively. Examples of both types of residues are given in Figure 2A, left (c_{mono}) and middle ($c_{mono\_in\_multi}$) panels. Residues with $b \geq 2$ were assigned to the c_{multi} class (Figure 2A, right panel). A total of 12 622 interfaces residues were found,

dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127–5147

with c_{mono}, $c_{mono\_in\_multi}$, and c_{multi} accounting for 54, 36, and 10%, respectively, of the overall population (SI Table S4). The results obtained on the S^{Full} data set were compared with those from a second smaller data set S^{Soc}, a collection of highly promiscuous proteins (Methods). As for S^{Full}, the 4690 interface residues of S^{Soc} were partitioned into $c_{mono\_in\_multi}$ (59%) and c_{multi} (41%) residues (SI Table S4).

Simulated ensembles of accessible conformations were generated for each isolated protein in S^{Full} and S^{Soc} using the tCONCOORD method (Methods). To test the dependence of results from the method used for the generation of the ensembles, we performed also GNM calculations on all S^{Full} proteins and Molecular Dynamics (MD) simulations on selected cases (S^{MD} data set, Methods). The intrinsic flexibility of the residues was described measuring the RMSF from average positions either of the C^{α} or of the side chain atoms. Correlations in residue motions were calculated using either C^{α} Cartesian coordinates, providing collective motions,[83] or the fragment encoding from a Structural Alphabet (SA), providing local correlated motions[23] (Methods). The same analyses were performed on the ensemble of experimental structures generated by collecting all the occurrences of a given protein in the PDB (Methods).

The results section is organized as follows. A first characterization of the physicochemical properties of interface residues will be followed by the presentation of the central results on intrinsic flexibility. In particular, the distribution of flexibility among the different binding classes, its dependence from evolutionary conservation, and the relationship between correlated motions and interface shape modulation will be discussed. The results obtained from the simulated ensembles will then be compared with the conformational variability found in the experimental structures. The findings from the investigation of the relationship between binding promiscuity and SNPs occurrence will be then introduced. The section will be closed by a case study exemplifying the general conclusions.

Multipartner Residues Have Distinctive Physicochemical Properties. In this section, we will characterize the residues in the three binding classes according to different physicochemical properties. In particular, amino acid identity, evolutionary conservation, solvent accessibility, and secondary structure were determined and compared with results from different data sets.[2,6,8,12,13,33]

The overall interface composition showed a previously documented[33,51,106] enrichment with respect to the surface in large aromatic (F, W, Y), hydrophobic (M, A, I, L, V) and specific polar/charged (C, Q, R) amino acids (Figure 2B). When compared to the c_{mono} and $c_{mono\_in\_multi}$ classes (cyan and blue), c_{multi} residues (orange) were found to be richer in the polar/charged Q, D, and R and, less significantly, in the aromatic F and Y amino acids. On the other side, c_{multi} turned out to be particularly poor in hydrophobic amino acids (M, A, I, L, V). These findings are consistent with the more pronounced polar character generally found in hub interfaces.[2,6,12] Moreover, the polyvalence of the Q amino acid (which can be either an hydrogen bond donor or acceptor) and the long or bulky side chains of R, Y, and F are particularly suitable to adapt to different interfaces with different local arrangements.[4,6,8] The analysis on S^{Soc} (SI Figure S2A) confirmed the c_{multi} enrichment in Q, R, F, and Y, together with a preference, specific to the sociable c_{multi} residues, for the I and M amino acids.

As expected,[7] the conservation propensity P_{cons} indicated that interface regions have a higher abundance of conserved residues than the overall surface (Figure 2C). Indeed, P_{cons} was >0 for ConSurf[54] conservation grades (G_{cons}) > 5 in all the binding classes. Interestingly, the c_{multi} group, while poorer in residues with intermediate conservation (G_{cons} = 7, 8) was found to be richer in highly conserved residues (G_{cons} = 9) than the c_{mono} and $c_{mono\_in\_multi}$ groups, both for the S^{Full} (Figure 2C) and the S^{Soc} (SI Figure S2B) data set. No large deviations from the average surface values were observed for the abundance of secondary structure elements in the three interface residue classes of S^{Full} ($P_{ss} \sim 1$ in Figure 2D). However, when compared with the two c_{mono} (cyan) and $c_{mono\_in\_multi}$ (blue) classes, c_{multi} residues showed a tendency to be richer in loops and poorer in strands. A similar preference for loops has been found for 'overlapping regions' in date hub proteins.[13] These regions are conceptually close, even if differently defined, to our multipartner residues. No significant differences in the secondary structure composition were found between $c_{mono\_in\_multi}$ (green) and c_{multi} (red) residues in the S^{Soc} data set (SI Figure S2C), which both showed a general enrichment in α-helices with respect to the surface.

P_{SASA} (Figure 2E) gives the propensity of an interface residue to expose a small (<20%), medium (20−50%), or large (>50%) fraction of its solvent accessible surface area (SASA, see Methods), as measured on the isolated protein. Interestingly, very exposed residues (>50%) were found to be particularly frequent in the c_{multi} group, as compared with both the c_{mono} and $c_{mono\_in\_multi}$ groups, either for the S^{Full} (Figure 2E) and the S^{Soc} (SI Figure S2D) data set. Similarly, if the fraction of SASA buried upon complexation is considered (Figure 2F), a higher proportion of residues with a large relative ΔSASA (>50%) was found for c_{multi} residues than c_{mono} and $c_{mono\_in\_multi}$ ones. An analogous propensity for burying large portions of SASA has been found in overlapping regions of date hub proteins.[13]

For the proteins in the S^{MD} subset of S^{Full}, it was possible to evaluate the hydrophilicity (S_{hyd}, Methods) of the residues from the distribution of water molecules around the solute in the MD simulations. To eliminate the dependence from the amino acid identity, each S_{hyd} value was standardized against the S_{hyd} distributions of the corresponding amino acid type. Even if a reduced statistical significance was observed, probably due to the smaller size of the S^{MD} data set, the comparison of the S_{hyd} Z-scores showed that monopartner residues in multipartner proteins tend to be less hydrated (Figure 2G) than the other two classes. Indeed, the presence of water molecules has been observed in PDB complexes involving multipartner interfaces.[107] Our findings would indicate that the tendency to strongly coordinate water molecules is mainly due to the promiscuous part of the interface.

To summarize these results, we found that multipartner residues have characteristic physicochemical features that distinguish them from monopartner residues belonging either to mono- or multipartner proteins. In particular, they tend to be richer in specific charged/polar (R, Q, D) and aromatic (Y, F) amino acids and in loops, and to be more conserved, more solvent exposed in the isolated protein, and more buried in the complex. Moreover, within multipartner proteins, multipartner residues seem to be preferentially hydrated compared to monopartner ones.

Multipartner Residues Have a Higher Intrinsic Conformational Flexibility. In the following sections, we will analyze the correlation between intrinsic conformational flexibility and binding multiplicity. We measured the intrinsic flexibility of S^{Full} residues in terms of RMSF Z-score values of either C^{α} or side chain atoms. The distribution of flexibility values observed for c_{multi} residues in tCONCOORD ensembles was compared with that of c_{mono} and $c_{mono\_in\_multi}$ residues (averages

are reported in Figure 3A and B, light colors). All the pairwise comparisons indicated a significant difference between the distributions (p-values <0.01), with an average flexibility order $c_{multi} > c_{mono} > c_{mono\_in\_multi}$ for both C^{α} (Figure 3A) and side chain (Figure 3B) RMSF values. A similar analysis on the S^{Soc} data set (SI Figure S3A and B) confirmed that multipartner residues (red) in sociable proteins are on average more flexible than the monopartner ones (green).

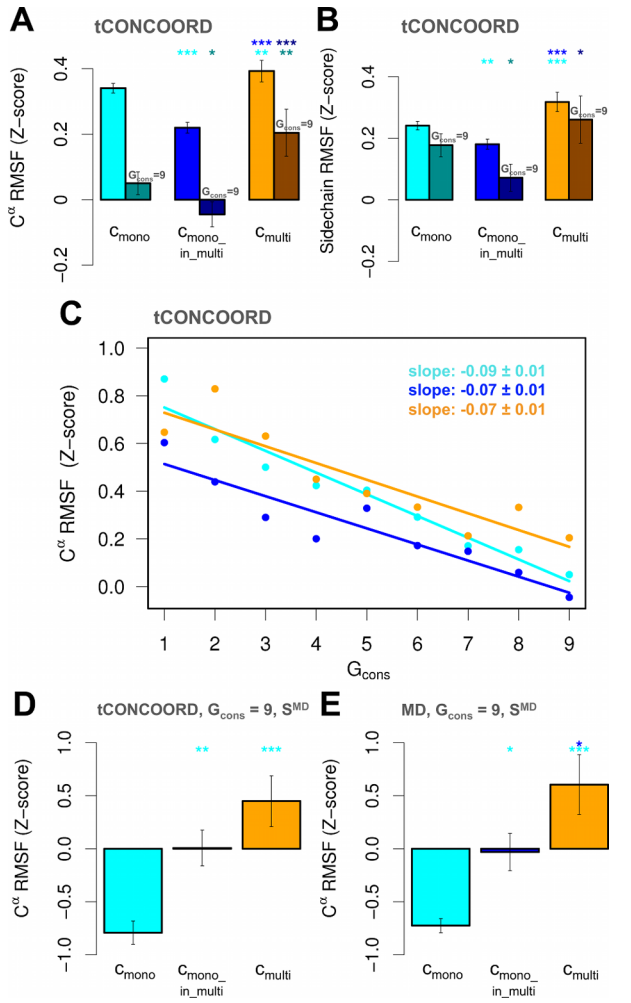


Figure 3. Conformational flexibility of interface residues in S^{Full} and S^{MD}. (A/B) Average C^{α} (A) and side chain (B) RMSF Z-scores (tCONCOORD) calculated over c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), and c_{multi} (orange) residues in the S^{Full} data set. The standard error of the mean is represented with an error bar. The significance levels from pairwise Wilcoxon comparison tests are reported with a star code (see Figure 2 legend). Stars are drawn above $c_{mono\_in\_multi}$ and c_{multi} bars indicating the significance levels of the comparison with c_{mono} (cyan) and $c_{mono\_in\_multi}$ (blue). Averages and significance levels calculated considering only the residues with the highest ConSurf conservation grade (G_{cons} = 9) are also reported in dark colors. (C) Dependence of tCONCOORD average C^{α} RMSF Z-scores (dots) from evolutionary conservation for c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), c_{multi} (orange) residues. Residues are partitioned into 9 groups according to their ConSurf conservation grade (G_{cons}). A best-fit linear regression is also reported for each binding class. (D/E) Average C^{α} RMSF Z-scores calculated from tCONCOORD (D) and MD (E) ensembles over highly conserved (G_{cons} = 9) c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), and c_{multi} (orange) residues in the S^{MD} data set. The standard error of the mean is represented with an error bar. The significance levels from pairwise Wilcoxon comparison tests are reported with a star code (see Figure 2 legend).

The intrinsic flexibility of interface residues in the different classes was compared also using alternative conformational ensembles (Methods). In particular, equilibrium fluctuations around the native structure were evaluated using full-atom MD simulations and coarse-grained GNM calculations. Remarkably, multipartner residues showed the largest average mobility among the binding classes for all the considered ensembles and flexibility indices (C^{α}, side chain and fragment RMSF), in both the complete S^{Full} data set and the S^{MD} subset (SI Figure S4). Also, the relative flexibility order of the c_{mono} and $c_{mono\_in\_multi}$ residue groups was generally preserved, except for C^{α} flexibility in S^{MD}. In many cases, similar significance levels in the distribution comparison were obtained from the different ensembles. This indicates that, despite the differences in the methods used for the generation of the ensembles, in all of them, multipartner residues have flexibility properties distinct from the monopartner ones.

To assess the possible impact of these flexibility differences on binding energetics, the configurational entropy of single residues was estimated with the Schlitter's formula (Methods) applied to the MD ensembles of multipartner proteins in S^{MD} (SI Table S10). The per-residue entropy term TS (where T = 300 K and S is calculated according to eq 3) of multipartner residues was on average higher than monopartner ones by 1.04 kcal/mol for the whole residue (considering an average residue size of 10 atoms) and ~0.44 kcal/mol for the residue main chain, with maximum differences of 1.52 and 0.65 kcal/mol, respectively (SI Table S10). It has to be noted that the Schlitter's formula gives only an approximate estimate of the entropic term (Methods). Moreover, accurate measures of the impact of flexibility on the binding free energy would require that protein partners are explicitly taken into account in the calculation.

A further measure of the relationship between flexibility and binding was obtained by calculating the correlation coefficient between profiles of tCONCOORD RMSF and binding multiplicity b for each protein in S^{Full} (SI Figure S5). The observation of large correlation coefficients in this analysis would require not only a difference in flexibility between mono- and multipartner residues but also that, within multipartner residues, higher b values correspond to higher RMSF values. This makes this test more stringent than the previous one. Indeed, small average correlations were found for both mono- and multipartner proteins, using either C^{α} (SI Figure S5A) or side chain (SI Figure S5B) RMSF values. However, the distribution of multipartner proteins (magenta) was shifted toward significantly higher values than monopartner ones (cyan) in both cases, with some multipartner proteins showing correlation coefficients as high as 0.6.

The presented results show that, compared to monopartner residues, multipartner residues have an average 'excess' of intrinsic flexibility, which could be used by the residues to adapt to the different environments provided by the different partners when binding occurs. In the subsequent section, we will further refine this analysis by focusing on the interface residues that are most important for the interaction with the partner.

Opposite Effects of Evolutionary Conservation and Binding Promiscuity on Intrinsic Flexibility. Previous studies highlighted an increased rigidity for hot spots[30,31,108] and in general for evolutionary conserved interface residues, which have been found to prefer preorganized bound-like conformation even in the unbound state.[29] Thus, we analyzed the relationship between evolutionary conservation and conformational flexibility in our data set, to highlight possible differences in the behavior of the three residue binding classes.

A small but significant anticorrelation between flexibility from tCONCOORD ensembles and conservation was found by plotting the average C^{α} RMSF Z-scores of interface residues in S^{Full} against the ConSurf conservation grade value G_{cons} (Figure 3C). A best-fit linear regression showed that this dependence is slightly more pronounced for the c_{mono} class (cyan), producing an increased difference between c_{mono} and c_{multi} C^{α} RMSF distributions when only the most conserved residues are compared (dark cyan and dark orange bars in Figure 3A). The same slope was found for the $c_{mono\_in\_multi}$ (blue) and c_{multi} (orange) linear models (Figure 3C), so that $c_{mono\_in\_multi}$ average C^{α} flexibility was consistently smaller than c_{multi} for all the G_{cons} values. Similar findings, but with reduced differences between the residue binding classes, were observed when considering side chain RMSF values (Figure 3B and SI S3F).

To check if these results are affected by the method used to generate the conformational ensembles, we compared tCONCOORD and MD RMSF distributions for the most conserved residues in the three binding classes of the S^{MD} subset (Figure 3D and E). A fully consistent picture was obtained from the two methods, confirming the higher average flexibility of the most conserved multipartner residues (orange) with respect to the monopartner ones (cyan and blue). Moreover, the generality of the results was checked on the S^{Soc} data set (SI Figure S3C/E), where c_{multi} residues (red) showed an even smaller dependence of C^{α} RMSF from the conservation grade than all the other classes.

We further restricted our analysis to hot spot interface residues, that is, the residues that contribute most to the binding energy upon complex formation.[7,34] Owing to the limited availability of experimental information on hot spots, especially for multipartner proteins,[109] we identified candidate hot spots using different prediction methods on the nonredundant set of interfaces of S^{Full}. For each protein and each method, a given residue was classified as hot spot if it satisfies the method criteria (Methods and SI Table S5) in at least one of the interfaces in which it is involved. In agreement with previous findings,[35,109] the majority of multipartner hot spots (from 79 to 88% depending on the specific method) was predicted as such only in one interface (SI Table S11), indicating that hot spots are partner-specific.[35] Hot spot populations generated by different predictors had only a partial overlap (SI Figure S6), as expected from the differences in the strategies adopted by the different methods (Methods). In spite of this, the comparison of the conformational flexibility of mono- and multipartner hotspots produced surprisingly consistent results (Figure 4). Indeed, the highest average tCONCOORD flexibility was observed for c_{multi} hot spots (orange) in almost all the cases for both C^{α} atoms (Figure 4A) and side chains (Figure 4B).

The results presented in this section show that binding promiscuity has a counteracting effect on the loss of flexibility expected for more conserved positions. As a consequence, the preference of promiscuous residues for a higher mobility is even more pronounced when considering only the residues that are more likely to be determinant for the binding to the partner.

Multipartner Proteins Use Different Global Motions to Modulate Different Interfaces. In the previous sections, we analyzed the flexibility of residues in terms of their equilibrium fluctuations from the average position in the simulated ensembles. Here, we extend this investigation to the correlation of these motions, to highlight possible differences between monopartner and multipartner interfaces. In particular, we aimed at detecting to which extent global motions modulate the interface dynamics in the different classes.

For each protein in S^{Full}, a set of nonredundant interfaces was extracted from all its PDB complexes. Each interface was then assigned to a binding class c_{mono}, $c_{mono\_in\_multi}$, or c_{multi} according

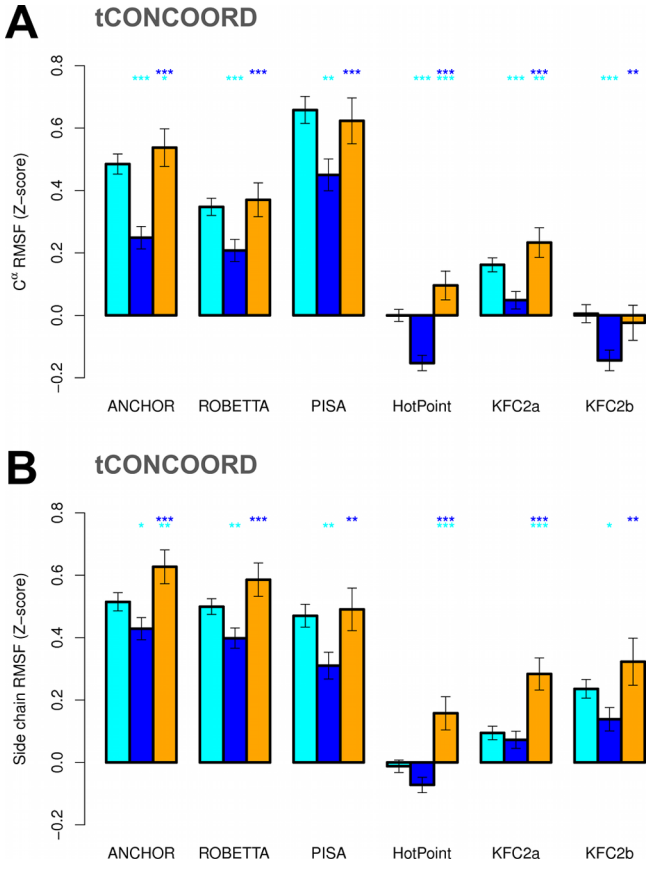


Figure 4. Conformational flexibility of hot spots. Average C^{α} (A) and side chain (B) RMSF Z-scores (tCONCOORD) of predicted hotspots in the three binding classes c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), and c_{multi} (orange) of the S^{Full} data set. The standard error of the mean is represented with an error bar. The significance levels from pairwise Wilcoxon comparison tests between the RMSF distributions are reported with a star code (see Figure 2 legend). For a given hotspot prediction method, stars are drawn above $c_{mono\_in\_multi}$ and c_{multi} bars indicating the significance levels of the comparison with c_{mono} (cyan) and $c_{mono\_in\_multi}$ (blue).

to the maximum binding multiplicity of its residues (Methods and SI Table S4). The collective motions of each S^{Full} protein were extracted by a principal component analysis (PCA) of its tCONCOORD conformational ensemble. The first l principal components (PCs) accounting for the 90% of the overall fluctuation of C^{α} atoms were selected. While other choices would be possible to represent collective motions, using PCs ensured that the motions considered in the analysis were allowed by the underlying energy landscape. The identification of the PCs mostly correlated with each interface was performed through a functional mode analysis (FMA),[82] using as functional property the radius of gyration of the interface C^{α} atoms (R_g^{IF}). The linear combination of PCs that best correlates with the R_g^{IF} of each interface (maximally correlated motion or MCM) was then calculated. The contribution of single PCs to the MCM, together with the value of the MCM-R_g^{IF} correlation, determines the fraction of the R_g^{IF} variance that is due to the motion along a given PC (pvar). A PC was considered to be correlated with an interface if pvar $\geq$ 20%.

The number of R_g^{IF}-correlated PCs ($n\mathrm{PC}_{20}{}^{IF}$) was calculated for each interface (Figure 5A). All the interfaces turned out to be correlated on average with no more than 1 PC. The behavior of c_{mono} interfaces (cyan) was similar to that of the multipartner ones (orange), while the $c_{mono\_in\_multi}$ interfaces (blue) seemed to

dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127−5147

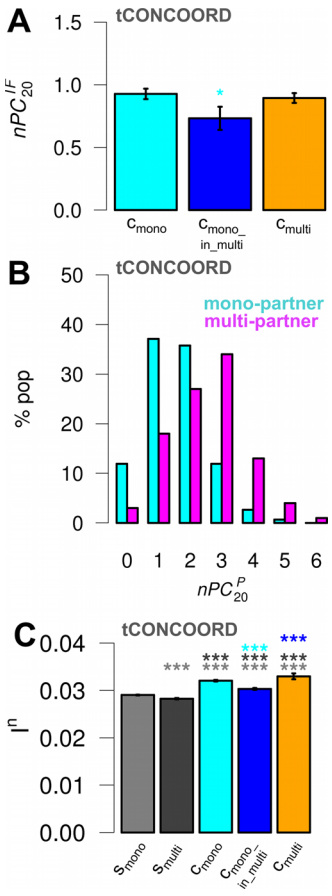


Figure 5. Analysis of correlated motions in S^{Full}. (A/B) Correlation between global and interface motions by functional mode analysis (FMA) on tCONCOORD ensembles from the S^{Full} data set. (A) Number of PCs accounting for at least 20% of the R_g^{IF} variance of a given interface (nPC_{20}^{IF}). Average values calculated over interfaces of the three binding classes c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), and c_{multi} (orange) are reported. (B) Distributions of nPC_{20} per protein (nPC_{20}^{P}), calculated as the number of unique PCs accounting for at least 20% of the R_g^{IF} variance of any of the protein interfaces. Values for monopartner (cyan) and multipartner (magenta) proteins are reported. (C) Distributions of normalized MI (I^n) calculated over pairs of surface residues in monopartner (light gray, s_{mono}) and multipartner (dark gray, s_{multi}) proteins, and over pairs of interface c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), and c_{multi} (orange) residues. In A and C, the standard error of the mean is represented with an error bar. The significance levels from pairwise Wilcoxon comparison tests are reported with a star code (see Figure 2 legend).

be slightly less affected by the protein global motions. If the number of unique R_g^{IF}-correlated PCs is summed over all the interfaces of a given protein (nPC_{20}^{P} in Figure 5B), a significant difference is found between monopartner (cyan) and multipartner (magenta) proteins, indicating that the latter use a larger number of independent global motions to modulate the shape of their interfaces. Since for multipartner proteins each interface is on average correlated with ~0.9 PC and each protein has, on average, four nonredundant interfaces, the peak at 3 for multipartner nPC_{20}^{P} implies that different interfaces tend to be correlated with different PCs. Qualitatively similar results were obtained when performing the FMA analysis on MD ensembles of the S^{MD} data set (SI Figure S7).

The PCs used in the FMA represent the main global or collective motions of a protein. Correlations between residues can also be analyzed by considering their local dynamics, that is, by removing the overall roto-translation of the protein from their motion. Indeed, it has been shown that analyzing local dynamics can highlight communication pathways within the protein that are difficult to identify solely from the collective motions involving the entire structure.[23] To this end, local motions were analyzed with a fragment-based approach[23,110] and the extent of correlation was estimated by normalized mutual information I^n (eq 1 in Methods). Correlations within interface residues of each binding class (cyan, blue and orange in Figure 5C) calculated from the tCONCOORD ensembles were found to be significantly higher than those within mono- and multipartner surface residues (light and dark gray), suggesting a higher level of communication between interface residues. Interestingly, the c_{multi} distribution (orange) was found to be significantly higher than the $c_{mono\_in\_multi}$ one (blue), indicating that, within multipartner proteins, multipartner residues are on average more correlated than monopartner ones. No significant differences were instead found between c_{multi} (orange) and c_{mono} (cyan) distributions.

Multipartner Residues Have a Higher Conformational Variability within Experimental Structures. In this section, we will investigate if the higher intrinsic flexibility found for multipartner residues correlates with the variability observed in the experimental structures. Indeed, many proteins in the S^{Full} data set have a relatively large number of occurrences in the PDB (>10 for ~50% of the proteins). These PDB ensembles contain information on the protein conformational variability[98] that can be extracted and compared with that derived from the simulated ensembles.

The overall variability within the PDB ensemble of each protein was decomposed into three different contributions according to the binding state of the structures that are compared: unbound−unbound (U−U), bound−bound (B−B), and unbound−bound (U−B). The highest structural variability (as measured by the maximum C^α RMSD calculated over all the structure pairs of a given protein) was found on average for the U−B pairs (2.66 ± 0.36 Å), followed by U−U (1.81 ± 0.28 Å) and B−B (1.34 ± 0.15 Å). The U−U RMSD values can be considered as related to the intrinsic plasticity of the isolated protein, which can have different accessible states ('pre-existing equilibrium'), while the U−B values include also the structural changes caused by the binding of the partners ('induced fit'). The smaller variability observed within bound structures (B−B) reflects the higher number of constraints in the complexes.

For each of these different contributions, multipartner proteins showed on average a higher conformational variability (U−U_{multi}, B−B_{multi}, and U−B_{multi}, dark colors in Figure 6A) than monopartner ones (U−U_{mono}, B−B_{mono}, and U−B_{mono}, light colors). This suggests that the higher intrinsic flexibility observed for multipartner residues in isolated proteins could be used to enhance pre-existing equilibrium or to assist induced-fit changes. Indeed, the single contributions of the residues to the overall C^α fluctuation within the PDB ensemble (Figure 6B−C) showed a similar picture as that observed for the simulated RMSF (see also SI Figure S4 for a direct comparison). A higher conformational variability (Figure 6B, light colors) and a weaker dependence from conservation (Figure 6C) were found for multipartner residues, resulting in an increased RMSF difference between highly conserved mono and multipartner residues (Figure 6B, dark colors). The c_{mono} and $c_{mono\_in\_multi}$ groups were instead more similar to each other than in the simulated case, indicating a comparable degree of variation within the PDB structures. Consistently with the simulated ensembles, reduced differences

dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127−5147

were found between the different binding classes when the side chain variability was considered (SI Figure S8 and S4).

Multipartner Residues Have a Lower Propensity for SNPs. In this section, we provide further insight on the functional relevance of promiscuous residues by analyzing the distribution of Single Nucleotide Polymorphisms (SNPs) across the different binding classes considered in the previous sections.

Recent large-scale studies of the human genome such as the International HapMap Project[111] and the 1000 Genomes Project[112] have produced a large number of data that can be used to accurately assess the relationship between genotype and phenotype. In particular, SNPs are single nucleotide variations observed at a specific location of the genome in at least 1% of the population.[113] The mapping of SNPs, and in particular nonsynonymous SNPs, to protein regions is currently exploited in a wide range of applications, from disease association studies to pharmacogenomics.[114] In this study, only missense non-synonymous SNPs are considered.

In order to study the relationship between binding promiscuity and human SNPs, human homologues of S^{Full} and S^{Soc} proteins were identified. We recorded the occurrence of nonsynonymous SNPs from the dbSNP database[102] and of SNPs with a known association with disease (DisSNPs) from the OMIM database[103] (Methods). The SNP and DisSNP positions were then mapped back to the original S^{Full} and S^{Soc} proteins.

The comparison of the SNP propensities of the different binding classes with respect to the surface (Figure 7A, light colors) showed that promiscuous positions in S^{Full} (orange) are less rich in SNPs than both classes of monopartner residues (cyan and blue). Even if with reduced statistical significance, this trend was confirmed by S^{Soc} proteins (green and red). As an example, the survivin protein, already in the human form in the S^{Full} data set, is shown in Figure 7B. While promiscuous (red shades surface) and nonpromiscuous (light blue surface) residues are present in survivin in almost equal proportion, 5 out of 6 interface SNPs (spheres) were found in nonpromiscuous regions, mainly involved in the interaction with the survivin partner borealin (green cartoon).

A possible explanation of these findings is that the human equivalent of the promiscuous positions considered here tend to be less tolerant to variation, probably due to the higher number of constraints that they are experiencing to preserve effective binding. Mutations at these sites might be more prone to yield a lethal phenotype and are thus less likely to be viable.[115]

The analysis of DisSNPs (Figure 7A, dark colors) was strongly affected by the small number of observations (SI Table S12). The large uncertainty associated with the average propensities, especially for the S^{Soc} proteins, prevented the observation of statistically significant differences between the binding classes. It is thus evident that these results, while indicating an interesting and unexplored relationship between binding diversity and variability in the human genome, will need to be confirmed on a larger data set of human proteins.

Case Study: Two Ubiquitin-like Proteins. In this section, we will exemplify the relationship observed between binding and flexibility analyzing two related multipartner proteins: Neddylin and the small ubiquitin-related modifier 2 (SUMO-2).

Neddylin is a ubiquitin homologue (56% sequence identity, Figure 8A) with the characteristic ubiquitin-like fold (Figure 8B/D). The analysis of its complexes highlighted that, as for ubiquitin,[116] the Neddylin main interface is centered at an hydrophobic patch at the C-terminus of the β5 strand (Figure 8B, green circle), involving different hydrophobic residues from the β1-β2 loop (L8) and the

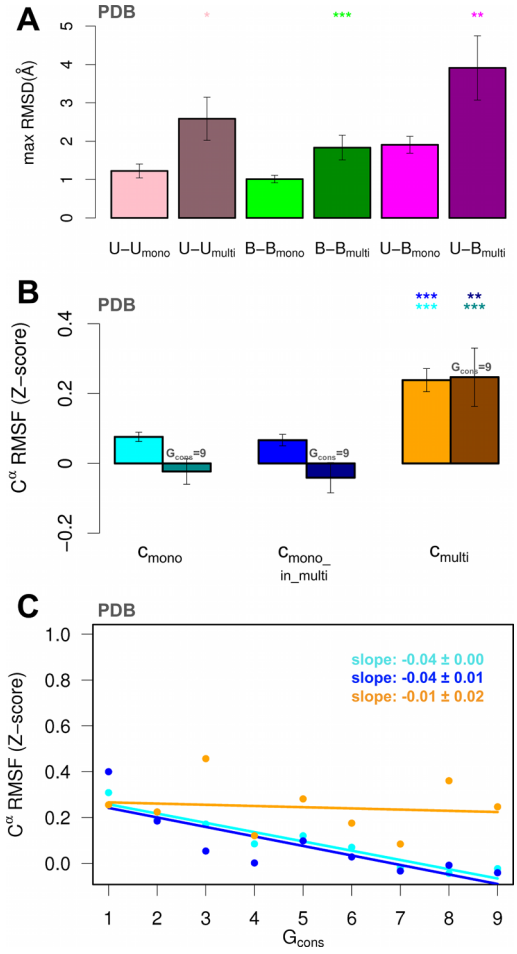


Figure 6. Analysis of the conformational variability within PDB ensembles in S^{Full}. (A) Averages of maximum C^{α} RMSD values calculated over all the possible pairs of PDB structures within the unbound (pink) and bound (green) ensembles, and between bound and unbound structures (magenta) of each mono- (light color) and multi- (dark color) partner protein. The standard error of the mean is represented with an error bar. The significance levels from pairwise Wilcoxon comparison tests between (mono, multi) pairs of distributions are reported with a star code (see Figure 2 legend). (B) Average C^{α} RMSF Z-scores (PDB ensembles) calculated over c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), and c_{multi} (orange) residues in the S^{Full} data set. The standard error of the mean is represented with an error bar. The significance levels from pairwise Wilcoxon comparison tests are reported with a star code (see Figure 2 legend). Averages and significance levels calculated considering only the residues with the highest ConSurf conservation grade (G_{cons} = 9) are also reported in dark colors. (C) Dependence of PDB average C^{α} RMSF Z-scores (dots) from evolutionary conservation for c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), c_{multi} (orange) residues. Residues are partitioned into 9 groups according to their ConSurf conservation grade (G_{cons}). A best-fit linear regression is also reported for each binding class.

β3–5 strands (I44, V70, L71). This region is highly promiscuous and rich of hot spots (sticks in Figure 8B). A secondary interaction site, previously observed also in ubiquitin,[116] was found at the α1 C-terminus (Figure 8B, yellow circle). This does not show any hot spot and is mainly composed of monopartner residues.

A positive correlation was observed between the profiles of binding multiplicity (orange) and C^{α} RMSF from the tCONCOORD ensemble (blue) of Neddylin (Figure 8F), with a correlation coefficient calculated over the exposed residues of 0.36. Correspondingly, a higher average C^{α} RMSF Z-score

dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127–5147

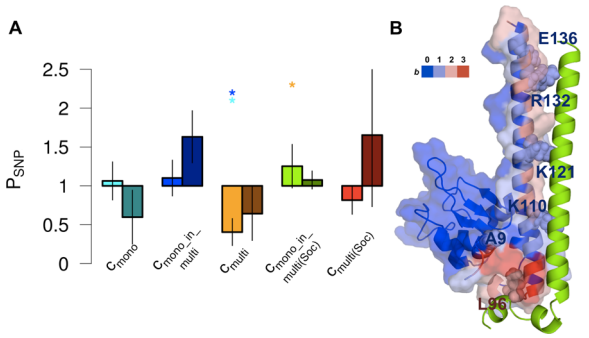


Figure 7. SNP and DisSNP propensity in the S^{Full} and S^{Soc} data sets. (A) Propensities of SNPs (light colors) and DisSNPs (dark colors) relative to the surface for c_{mono} (cyan), $c_{mono\_in\_multi}$ (blue), c_{multi} (orange), $c_{mono\_in\_multi(Soc)}$ (green) and $c_{multi(Soc)}$ (red) interface residues. The propensity is calculated per protein. The reported values are averages over S^{Full} monopartner proteins for c_{mono}, S^{Full} multipartner proteins for $c_{mono\_in\_multi}$ and c_{multi}, and S^{Soc} proteins for $c_{mono\_in\_multi(Soc)}$ and $c_{multi(Soc)}$. The error bars represent the standard error of the mean. The significance levels from pairwise Student's t tests are reported with a star code (see Figure 2 legend). (B) SNPs in the human survivin protein. SNPs found in the interface region of survivin are labeled and represented as van der Waals spheres. The protein surface (transparent) is colored according to the binding multiplicity b from blue ($b = 0$, non-interface) to red ($b = 3$). A survivin binding partner (borealin) is also represented as green cartoon (PDB ID: 2RAW).

(SI Table S13) was found for multipartner residues (0.79 ± 1.48) than monopartner ones (0.03 ± 0.72). Indeed, many multipartner residues (Figure 8B) are located at or in close proximity of high-flexibility regions (Figure 8D), namely the $\beta1$-$\beta2$, $\alpha1$-$\beta3$, and $\beta3$-$\beta4$ loops, and the C-terminus. The intrinsic backbone mobility of promiscuous locations in the simulated ensemble is paralleled by a high conformational variability at these same positions in the different PDB complexes where Neddylin interacts with different partners (SI Table S13). Additionally, the FMA analysis (Figure 9A) showed that the backbone flexibility of each of the two representative interfaces mainly correlates with one specific collective motion. In particular, PC4 (Figure 9B) accounts for 54% of the R_g^{IF} variance of the main interface (green bars in Figure 9A). A similar collective motion, involving a 'pincer-like' movement of the $\beta1$-$\beta2$ and $\alpha1$-$\beta3$ loops, has been observed in NMR ensembles of ubiquitin representing solution dynamics up to the μs time scale.[24]

Multipartner residues of Neddylin seemed to rely less on side chain flexibility than backbone mobility to adapt to different environments, since the average side chain RMSF was comparable to that of monopartner residues for both the tCONCOORD and PDB ensembles (SI Table S13). Indeed, while a few multipartner residues adopted different rotamers in different complexes (e.g., R42 in SI Figure S9B), others relied either on the backbone flexibility (L8 in SI Figure S9A) or on their capacity to support different interaction geometries without changing their conformation (e.g., H68 in SI Figure S9C, where it interacts with an aromatic ring in either a T-shaped or parallel stacking interaction).

The SUMO-2 protein, which is a homologue of ubiquitin in spite of its low sequence identity (14%, Figure 8A), showed important differences with respect to Neddylin in binding modes and dynamics. This is consistent with this protein belonging to a separate sequence subgroup[117] of ubiquitin homologues, characterized by the replacement of the key ubiquitin residues Q41 and Y59 (yellow arrows and sticks in Figure 8A and D, respectively) with two hydrophobic residues that are no longer able to form hydrogen-bonds with the nearby loops (sticks in Figure 8E). This is reflected in both the protein structure and dynamics. In particular, compared to Neddylin the tCONCOORD SUMO-2 ensemble presented a higher C^{α} flexibility at the $\alpha1$ C-terminus and the $\alpha1/\beta3$ loop (Figure 8E).

The increment in flexibility was paralleled by an increment in binding promiscuity in this region. A multipartner interaction site, typical of SUMO proteins, was found at the 'groove' defined by the $\alpha1$ C-terminus and the $\beta2$ strand, which is correspondingly enriched in multipartner hot spots with respect to Neddylin (Figure 8C and G). The mapping of tCONCOORD C^{α} and side chain RMSF onto the SUMO-2 surface (Figure 10C/D), shows that the α/β groove is flanked by residues with high flexibility either at C^{α} ($\beta1$-$\beta2$ loop) or side chain atoms ($\beta2$) or both ($\alpha1$ C-terminus). Moreover, the water density map from the MD simulation showed that the region surrounding the groove is richer in high-density hydration sites than the nonpromiscuous regions (SI Figure S10).

The analysis of the SUMO-2 PDB structures (SI Table S14) highlighted that, differently from Neddylin, side chain conformational changes are more important for multipartner residues (average side chain RMSF Z-score = 0.56 ± 1.13) than for monopartner ones (-0.06 ± 0.98). Particularly relevant is the contribution of the R34 and Q35 hot spots at the $\alpha1$ C-terminus, with a side chain RMSF Z-score of 2.63 and 1.32, respectively. Correspondingly, these residues adopt significantly different rotamers in the different PDB complexes (Figure 10A/B).

This example suggests that a fine-tuning of the intrinsic dynamical properties of the interface can be central in modulating the binding specificity in evolutionary related proteins.

DISCUSSION

In this work, we studied the intrinsic conformational flexibility of multipartner proteins to assess its role in promoting diversity of binding. Through the generation of simulated conformational ensembles from a starting experimental structure, we measured the tendency of the isolated proteins to sample different conformations independently from the interactions with their partners. The conformational flexibility that we considered here is a different concept from the intrinsic disorder analyzed in other works on hub proteins,[9,12,15,17,118] characterized by the absence of a definite structure in all or part of the protein. It is also different from the conformational plasticity[12−14] as measured by the conformational changes observed when multiple experimental structures are available. Indeed, if these structures correspond to different states, such as bound and unbound conformations, the conformational variability is likely to include also induced fit and allosteric effects in addition to intrinsic flexibility.

The choice to consider the isolated protein is justified by many studies indicating that the intrinsic dynamics of a protein is correlated with its function.[19,20,26,119−122] Indeed, even when isolated, proteins have been shown to sample functionally relevant states, which can then be stabilized or selected by interactions with the environment such as post-translational modifications or interactions with ligands.[18] Many of these works used elastic network models (ENM) to characterize the equilibrium fluctuations within the native structure basin.[28,95,120,123] In spite of their simplicity, these methods have been shown to provide results in good agreement with both experiments and more sophisticated approaches such as MD. For this study, we chose a method of intermediate complexity and computational cost, tCONCOORD. While allowing for anharmonicities and providing a full atom

dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127−5147

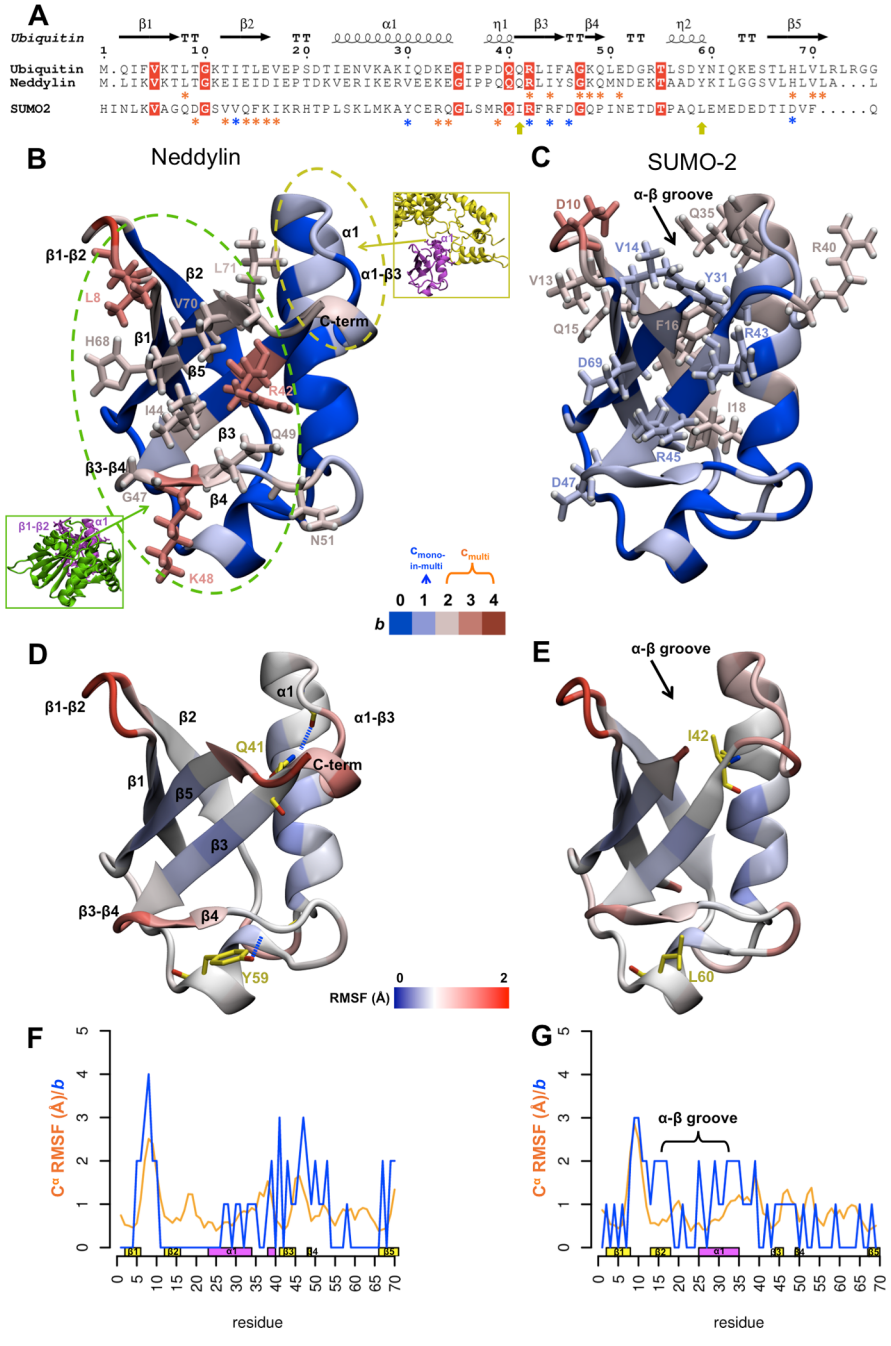


Figure 8. Comparison of two ubiquitin-like proteins. (A) Sequence alignment of Ubiquitin (UniProt/KB ID: P0CG48), Neddylin (Q15843), and SUMO-2 (P61956). The sequences were aligned using T-COFFEE v 9.02[130] with default parameters. The DSSP secondary structure of Ubiquitin (PDB ID: 1ubi) is shown (top). Conserved residues are highlighted in red. Stars below Neddylin and SUMO-2 sequences indicate $c_{mono\_in\_multi}$ (blue) and c_{multi} (orange) residues identified as hotspots by at least two prediction methods. (B/C) Cartoon representation of Neddylin (B, PDB ID: 1ndd) and SUMO-2 (C, PDB ID: 1wm3) structures. Residues are colored according to the binding multiplicity b from blue (b = 0, non-interface) to red (b = 4). Residues identified as hotspots by at least two prediction methods are represented as sticks. The approximate boundaries of two representative interfaces of Neddylin are indicated in B with a green (main interface) and yellow (secondary interaction site) dashed line. The insets show the structures of the representative complexes (PDB ID 1xt9, chains B/A, and PDB ID 1r4n, chains J/C in the green and yellow insets, respectively). (D/E) Mapping of tCONCOORD C^{α} RMSF values onto Neddylin (D) and SUMO-2 (E) structures. The structures are colored according to RMSF values from blue (0 Å) to red (2 Å). Selected residues (highlighted with yellow arrows in panel A) are represented as yellow sticks, while hydrogen bonds in D are indicated as dotted blue lines. (F/G) Profiles of binding multiplicity (b, blue) and C^{α} RMSF from tCONCOORD (orange) for Neddylin (F) and SUMO-2 (G). DSSP annotation of the secondary structure is reported as magenta (helices) and yellow (strands) blocks.

description of the protein where both backbone and side chains are included, it is still faster than MD simulations.[67]

The interaction data on each protein of our S^{Full} data set were derived from a structural PPI database, PiSite,[38] where partners from all the PDB complexes involving homologues within a high (>90%) sequence identity family are mapped onto a family representative. Structural PPI databases generally contain higher-confidence interaction data than interactome networks derived from a range of different experimental methods, resulting in degree distributions with shorter tails.[11] In our S^{Full} data set, this

dx.doi.org/10.1021/ct400486p | J. Chem. Theory Comput. 2013, 9, 5127−5147

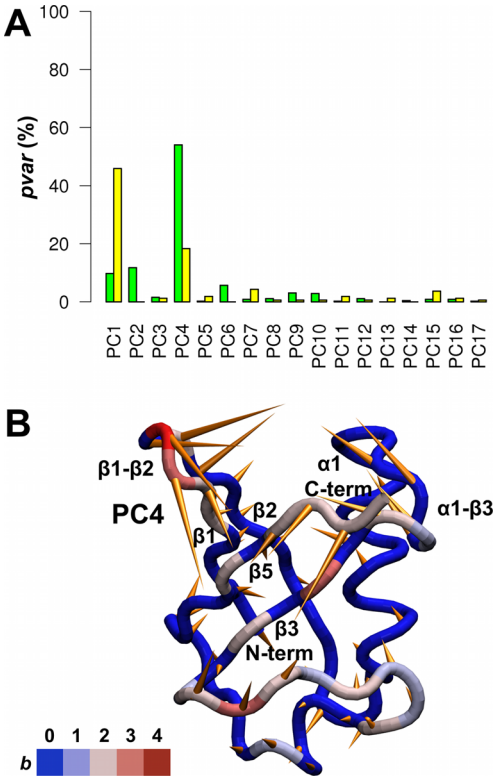


Figure 9. FMA analysis of Neddylin. (A) Decomposition of the R_g^{IF} variance of the two representative interfaces of Neddylin into percentage contributions (pvar) from the first 17 PCs of the tCONCOORD ensemble. Green and yellow bars represent the pvar values for the interfaces shown in the green and yellow insets in Figure 8B. (B) Porcupine representation of the 4th PC of the Neddylin tCONCOORD ensemble. Direction and relative amplitude of the motion of each C^{α} atom along the PC is represented by orange spikes. The residues involved in the main Neddylin interface (Figure 8B, green inset) are color-mapped according to their binding multiplicity b onto the tube representation of the average tCONCOORD structure.

is reduced also by the fact that only monodomain proteins were considered, so that a maximum number of 12 nonredundant partners per protein was found. The investigation was limited to monodomain proteins to exclude cases where multiple-partner binding is simply achieved by using different domains or a different relative arrangement of the domains.[6]

The relationship between conformational flexibility and ability to bind multiple partners was analyzed primarily at the residue level. Indeed, the composition of either proteins or interfaces in terms of binding multiplicity turned out to be highly heterogeneous, with 40% of multipartner proteins in S^{Full} presenting both mono- and multipartner interfaces and 91% of multipartner interfaces including both mono- and multipartner residues. Thus, we aimed mostly at identifying the dynamical features specific of multipartner residues and not at characterizing multipartner proteins as a whole.

Even if we did not attempt a classification of the S^{Full} multipartner proteins, more than 60% of them have at least 10% of their interface residues involved in interactions with different partners, while only 8% multipartner proteins have no overlapping interfaces. On the other hand, as found in different data sets,[124] very large overlaps between interfaces seem to be avoided (only 5% of the multipartner proteins have more than 60% of their interface interacting with multiple partners). The alternate data set S^{Soc} used to validate our results, is composed of proteins previously classified as 'sociable' or 'transient' hubs,[12] that can be related, even if with a different definition, to date hubs.

We found a significant difference in the intrinsic conformational flexibility of mono- (c_{mono} and $c_{mono\_in\_multi}$) and multi- (c_{multi}) partner residues. In particular, the comparison between $c_{mono\_in\_multi}$ and c_{multi} residues suggests a nonuniform distribution of the flexibility across the interface of multipartner proteins, with multipartner residues being on average more flexible than monopartner ones. This holds when considering either backbone or side chain motions, suggesting that an enhanced ability to sample different conformations, either globally or locally, can indeed be exploited to support diversity of interactions. Remarkably, a more pronounced tendency of promiscuous residues to sample different conformations is observed in all the different types of conformational ensembles analyzed in this study and in two different protein data sets. This confirms the generality and robustness of the present findings. Correlated motions were also considered and their functional importance assessed. The flexibility of multipartner proteins was found to modulate the shape of single interfaces in a highly specific way, by using different collective motions for different interfaces within the same protein. Moreover, within multipartner proteins a stronger correlation was found between the local motions of multipartner residues, suggesting that promiscuous regions are connected by preferential communication pathways.

Similarly to disorder,[125] a higher flexibility at the interface of isolated molecules is associated with a higher entropic penalty upon binding. This entropy increment has been suggested to be exploited by disordered multipartner proteins to decouple binding affinity from binding specificity.[126] On the other side, residues considered to be important in the interaction with the partner, either because of their high evolutionary conservation or for being predicted as hot spots, have been shown to be more rigid than the average.[24,29−31] Here, we found that (a) there is a clear anticorrelation between evolutionary conservation and intrinsic flexibility and (b) this trend is reduced in the case of multipartner residues, suggesting that the higher flexibility of conserved multipartner residues is the result of a balance between the counteracting effects of preserving binding strength and allowing for binding diversity. In this context, the higher propensity for conservation found for multipartner residues in ordered proteins could reflect the higher number of evolutionary constraints deriving from the necessity to optimize this balance.

The findings on intrinsic flexibility are confirmed by the analysis on the conformational variability observed in the experimental structures. Previous works using PDB ensembles of different data sets have shown that overlapping regions in date hub interfaces tend to visit more different side chain rotamers than nonoverlapping ones[13] and that sociable/hub proteins sample more different overall backbone conformations than nonsociable/nonhub ones.[12] Here, we unified these results by performing a systematic analysis on both side chain and backbone changes at the residue level. Moreover, we decomposed the overall variability observed in the PDB into contributions from pairs of structures with the same and with different binding states. Interestingly, we found a higher plasticity for multipartner proteins not only when considering changes between unbound and bound structures but also when analyzing ensembles of unbound and bound structures separately. This is highly consistent with the larger intrinsic flexibility found for the multipartner residues in isolated proteins, since it implies a more

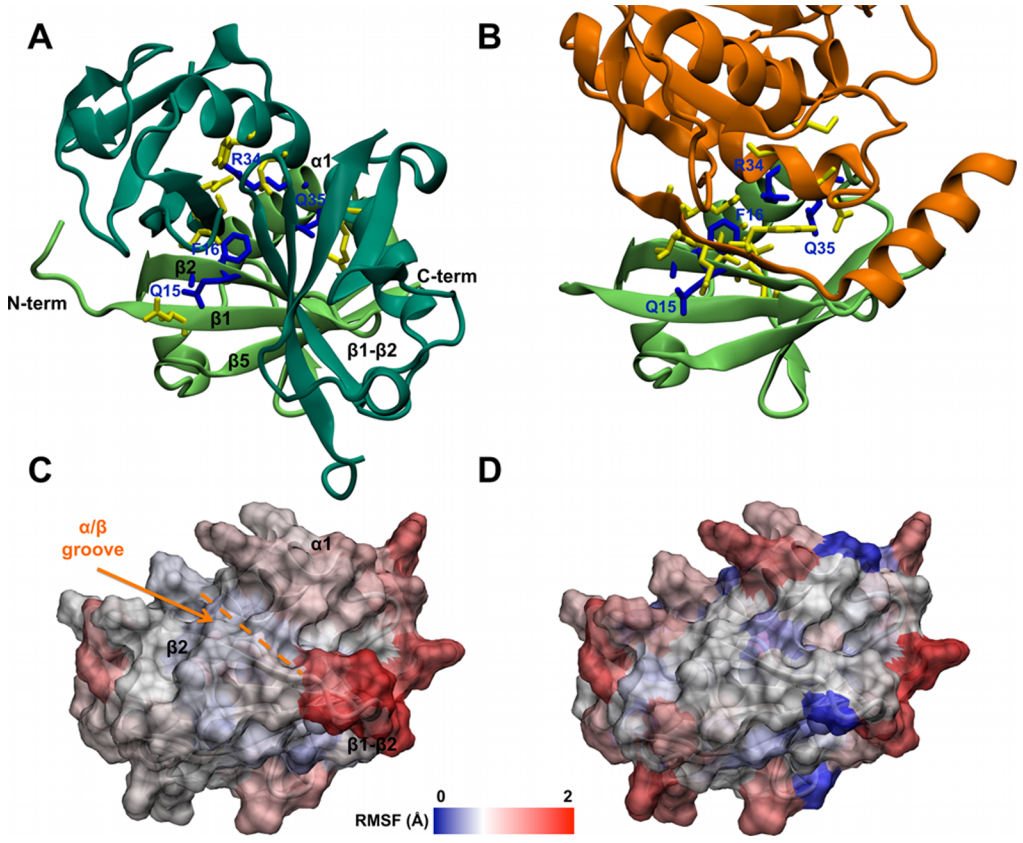


Figure 10. Binding properties and conformational flexibility of the α/β groove in SUMO-2. (A/B) Cartoon representation of two different complexes of SUMO-2 (light green). Selected multipartner hot spots of SUMO-2 (blue) and interacting residues from partners (yellow) are shown as sticks. (A) SUMO-2 homotrimer (PDB ID: 1wm2). The other two SUMO-2 proteins are shown in dark green. (B) Complex with Thymine-DNA glycosylase (orange, PDB ID: 2d07). (C/D) C^{α} (C) and side chain (D) RMSF values mapped onto SUMO-2 surface (PDB ID: 1wm3). Residues are colored according to their RMSF values from blue (0 Å) to red (2 Å).

pronounced tendency to explore different conformations independently from their binding state.

The higher propensity to conservation and surface burial upon complexation of multipartner residues suggests that they have an important role in defining the binding energetics. The comparison of the intrinsic flexibility of predicted hot spot residues from different binding classes confirmed the results obtained on the whole set of interface residues. Indeed, multipartner hot spots turned out to be on average more flexible than monopartner ones independently from the specific method used for hot spot prediction. This rules out the possibility that the observed higher flexibility of multipartner residues in the whole interface originates from residues that are only marginally involved in the interaction with partners. Moreover, the observation that promiscuous regions are depleted in SNPs, even if to be confirmed on larger data sets of human proteins, provides further evidence to the essentiality of these regions.

The findings presented in this work have potential applications to methods for the prediction of PPIs,[127] whose accuracy has been recently brought to levels comparable to high-throughput experiments.[128] In particular, the introduction of per-residue flexibility indices could be used for the identification of promiscuous regions in protein interfaces. This could also be relevant for the detection of druggable regions, which could be targeted by small molecules or other proteins to inhibit interactions with specific partners.[129] A further possible application is in protein design, where the specificity or promiscuity of proteins could be enhanced by modifying the distribution of flexibility across the interfaces.[6,8]

■ ASSOCIATED CONTENT

Supporting Information

Figure S1: functional annotation of S^{Full}. Figure S2: physicochemical properties of interface residues in the $c_{mono\_in\_multi}$ and c_{multi} binding classes of S^{Soc}. Figure S3: conformational flexibility of interface residues in S^{Full} and S^{Soc} (C^{α} and side chains). Figure S4: comparison of conformational flexibility of S^{Full} and S^{MD} interface residues from simulated ensembles (tCONCOORD, MD, GNM) and experimental structures (PDB). Figure S5: correlation between RMSF and binding multiplicity profiles in S^{Full}. Figure S6: pairwise comparison of hotspot prediction methods. Figure S7: analysis of correlated motions in S^{MD}. Figure S8: analysis of the conformational variability within PDB ensembles in S^{Full} (side chains). Figure S9: examples of multipartner residues in Neddylin. Figure S10: water distribution around the SUMO-2 protein. Table S1: list of the 251 proteins in the S^{Full} data set. Table S2: list of the 12 proteins in the S^{MD} data set. Table S3: list of the 69 sociable proteins in the S^{Soc} data set. Table S4: distribution of residues and interfaces over the three binding classes c_{mono}, $c_{mono\_in\_multi}$, and c_{multi} for the S^{Full}, S^{MD} and S^{Soc} data sets. Table S5: criteria used for the different hot spot prediction methods. Table S6: Composition of the 12 systems used for MD simulations of the S^{MD} data set. Table S7: Analysis of pairwise RMSD distributions of tCONCOORD and MD ensembles. Table S8: comparison

between pairs of RMSF profiles from the tCONCOORD, MD, GNM and PDB ensembles (C^{α} and side chains). Table S9. Overlap between the tCONCOORD, MD and PDB essential spaces. Table S10. Average per-residue entropy for $c_{mono\_in\_multi}$ and c_{multi} residues in S^{MD}. Table S11: percentage of c_{multi} residues classified as hot spots in 1 to 4 different interfaces. Table S12: SNP and DisSNP propensity in the S^{Full} and S^{Soc} data sets (values per protein). Table S13: C^{α} and side chain RMSF Z-scores of $c_{mono\_in\_multi}$ and c_{multi} interface residues of Neddylin. Table S14: C^{α} and side chain RMSF Z-scores of $c_{mono\_in\_multi}$ and c_{multi} interface residues of SUMO-2. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*Phone: +4402078486329. Fax: +4402078486435. E-mail: arianna.fornili@kcl.ac.uk.

*Phone: +4402078486843, fax: +4402078486435. E-mail: franca.fraternali@kcl.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Jens Kleinjung and Prof. Julián Echave for insightful discussions. We are also very grateful to Prof. Türkan Haliloğlu for kindly providing software and guidance for GNM calculations. This research was supported by the Leverhulme Trust (F/07 040/AL to A.F. and F.F.) and the British Heart Foundation (FS/12/41/29724 to A.F. and F.F.).

REFERENCES

(1) Jeong, H.; Mason, S. P.; Barabasi, A. L.; Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **2001**, *411*, 41−42.

(2) Kar, G.; Gursoy, A.; Keskin, O. Human cancer protein−protein interaction network: A structural perspective. *PLoS Comput. Biol.* **2009**, *5*, e1000601.

(3) Ekman, D.; Light, S.; Björklund, Å. K.; Elofsson, A. What properties characterize the hub proteins of the protein−protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol.* **2006**, *7*, R45.

(4) Patil, A.; Kinoshita, K.; Nakamura, H. Hub promiscuity in protein−protein interaction networks. *Int. J. Mol. Sci.* **2010**, *11*, 1930−1943.

(5) Tsai, C.-J.; Ma, B.; Nussinov, R. Protein−protein interaction networks: How can a hub protein bind so many different partners? *Trends Biochem. Sci.* **2009**, *34*, 594−600.

(6) Erijman, A.; Aizner, Y.; Shifman, J. M. Multispecific recognition: Mechanism, evolution, and design. *Biochemistry* **2011**, *50*, 602−611.

(7) Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R. Principles of protein−protein interactions: What are the preferred ways for proteins to interact? *Chem. Rev.* **2008**, *108*, 1225−1244.

(8) Humphris, E. L.; Kortemme, T. Design of multi-specificity in protein interfaces. *PLoS Comput. Biol.* **2007**, *3*, e164.

(9) Dosztányi, Z.; Chen, J.; Dunker, A. K.; Simon, I.; Tompa, P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* **2006**, *5*, 2985−2995.

(10) Han, J.-D. J.; Bertin, N.; Hao, T.; Goldberg, D. S.; Berriz, G. F.; Zhang, L. V.; Dupuy, D.; Walhout, A. J. M.; Cusick, M. E.; Roth, F. P.; Vidal, M. Evidence for dynamically organized modularity in the yeast protein−protein interaction network. *Nature* **2004**, *430*, 88−93.

(11) Kim, P. M.; Lu, L. J.; Xia, Y.; Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **2006**, *314*, 1938−1941.

(12) Higurashi, M.; Ishida, T.; Kinoshita, K. Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci.* **2008**, *17*, 72−78.

(13) Dasgupta, B.; Nakamura, H.; Kinjo, A. R. Distinct roles of overlapping and non-overlapping regions of hub protein interfaces in recognition of multiple partners. *J. Mol. Biol.* **2011**, *411*, 713−727.

(14) Bhardwaj, N.; Abyzov, A.; Clarke, D.; Shou, C.; Gerstein, M. B. Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein Sci.* **2011**, *20*, 1745−1754.

(15) Kim, P. M.; Sboner, A.; Xia, Y.; Gerstein, M. The role of disorder in interaction networks: A structural analysis. *Mol. Syst. Biol.* **2008**, *4*, 179.

(16) Nussinov, R.; Ma, B. Protein dynamics and conformational selection in bidirectional signal transduction. *BMC Biology* **2012**, *10*, 1−5.

(17) Fong, J. H.; Panchenko, A. R. Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Mol. Biosyst.* **2010**, *6*, 1821.

(18) Boehr, D. D.; Nussinov, R.; Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5*, 789−796.

(19) del Sol, A.; Tsai, C.-J.; Ma, B.; Nussinov, R. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* **2009**, *17*, 1042−1050.

(20) Dobbins, S. E.; Lesk, V. I.; Sternberg, M. J. E. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein−protein docking. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10390−10395.

(21) Volkman, B. F.; Lipson, D.; Wemmer, D. E.; Kern, D. Two-state allosteric behavior in a single-domain signaling protein. *Science* **2001**, *291*, 2429−2433.

(22) Haliloglu, T.; Erman, B. Analysis of correlations between energy and residue fluctuations in native proteins and determination of specific sites for binding. *Phys. Rev. Lett.* **2009**, *102*, 088103.

(23) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. Detection of allosteric signal transmission by information—theoretic analysis of protein dynamics. *FASEB J.* **2012**, *26*, 868.

(24) Lange, O. F.; Lakomek, N. A.; Fares, C.; Schroder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B. L. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **2008**, *320*, 1471−1475.

(25) Peters, J. H.; de Groot, B. L. Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. *PLoS Comput. Biol.* **2012**, *8*, e1002704.

(26) Keskin, O. Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: A case study of antibodies. *BMC Struct. Biol.* **2007**, *7*, 31.

(27) Münz, M.; Hein, J.; Biggin, P. C. The role of flexibility and conformational selection in the binding promiscuity of PDZ domains. *PLoS Comput. Biol.* **2012**, *8*, e1002749.

(28) Stein, A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P. A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. *Structure* **2011**, *19*, 881−889.

(29) Yogurtcu, O. N.; Erdemli, S. B.; Nussinov, R.; Turkay, M.; Keskin, O. Restricted mobility of conserved residues in protein−protein interfaces in molecular simulations. *Biophys. J.* **2008**, *94*, 3475−3485.

(30) Rajamani, D.; Thiel, S.; Vajda, S.; Camacho, C. J. Anchor residues in protein−protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 11287−11292.

(31) Kuttner, Y. Y.; Engel, S. Protein hot spots: The islands of stability. *J. Mol. Biol.* **2012**, *415*, 419−428.

(32) Bouvier, B.; Grünberg, R.; Nilges, M.; Cazals, F. Shelling the Voronoi interface of protein−protein complexes reveals patterns of residue conservation, dynamics, and composition. *Proteins* **2009**, *76*, 677−692.

(33) Davis, F. P.; Sali, A. The overlap of small molecule and protein binding sites within families of protein structures. *PLoS Comput. Biol.* **2010**, *6*, e1000668.

(34) Bogan, A. A.; Thorn, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **1998**, *280*, 1−9.

(35) Carbonell, P.; Nussinov, R.; del Sol, A. Energetic determinants of protein binding specificity: Insights into protein interaction networks. *Proteomics* **2009**, *9*, 1744−1753.
(36) Keskin, O.; Ma, B.; Nussinov, R. Hot regions in protein−protein interactions: The organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* **2005**, *345*, 1281−1294.
(37) Qin, H.; Lim, L.; Song, J. Protein dynamics at Eph receptor−ligand interfaces as revealed by crystallography, NMR, and MD simulations. *BMC Biophysics* **2012**, *5*, 2.
(38) Higurashi, M.; Ishida, T.; Kinoshita, K. PiSite: A database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.* **2009**, *37*, D360−D364.
(39) Martin, A. C. R. Mapping PDB chains to UniProtKB entries. *Bioinformatics* **2005**, *21*, 4297−4301.
(40) Chandonia, J.-M.; Hon, G.; Walker, N. S.; Lo Conte, L.; Koehl, P.; Levitt, M.; Brenner, S. E. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* **2004**, *32*, D189−92.
(41) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(42) Aranda, B.; Achuthan, P.; Alam-Faruque, Y.; Armean, I.; Bridge, A.; Derow, C.; Feuermann, M.; Ghanbarian, A. T.; Kerrien, S.; Khadake, J.; Kerssemakers, J.; Leroy, C.; Menden, M.; Michaut, M.; Montecchi-Palazzi, L.; Neuhauser, S. N.; Orchard, S.; Perreau, V.; Roechert, B.; van Eijk, K.; Hermjakob, H. The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* **2010**, *38*, D525−531.
(43) Margreitter, C.; Petrov, D.; Zagrovic, B. Vienna-PTM web server: A toolkit for MD simulations of protein post-translational modifications. *Nucleic Acids Res.* **2013**, *41*, W422−W426.
(44) Petrov, D.; Margreitter, C.; Grandits, M.; Oostenbrink, C.; Zagrovic, B. A systematic framework for molecular dynamics simulations. *PLoS Comput. Biol.* **2013**, *9*, e1003154.
(45) Kleinjung, J.; Fraternali, F. POPSCOMP: An automated interaction analysis of biomolecular complexes. *Nucleic Acids Res.* **2005**, *33*, W342−W346.
(46) Martin, A. C. R.; Porter, C. T. *ProFit*, 2.6; 2008. Available online: http://www.bioinf.org.uk/software/profit/.
(47) Rousseeuw, P. Silhouettes—A graphical aid to the interpretation and validation of cluster-analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.
(48) Fraternali, F.; Cavallo, L. Parameter optimized surfaces (POPS): Analysis of key interactions and conformational changes in the ribosome. *Nucleic Acids Res.* **2002**, *30*, 2950−2960.
(49) Dong, Q.; Wang, X.; Lin, L.; Guan, Y. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC Bioinformatics* **2007**, *8*, 147.
(50) Tuncbag, N.; Gursoy, A.; Keskin, O. Prediction of protein−protein interactions: Unifying evolution and structure at protein interfaces. *Phys. Biol.* **2011**, *8*, 035006.
(51) Jones, S.; Thornton, J. M. Protein−protein interactions: A review of protein dimer structures. *Prog. Biophys. Mol. Bio.* **1995**, *63*, 31−65.
(52) Dunbrack, R. L.; Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **1997**, *6*, 1661−1681.
(53) DeLano, W. *The PyMOL Molecular Graphics System*, DeLano Scientific: San Carlos, CA, 2002.
(54) Goldenberg, O.; Erez, E.; Nimrod, G.; Ben-Tal, N. The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.* **2009**, *37*, D323−D327.
(55) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(56) Wolfe, R.; Hanley, J. If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *Can. Med. Assoc. J.* **2002**, *166*, 65−66.
(57) Cumming, G.; Fidler, F.; Vaux, D. L. Error bars in experimental biology. *J. Cell Biol.* **2007**, *177*, 7−11.
(58) R-Development-Core-Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2010.
(59) Meireles, L. M. C.; Dömling, A. S.; Camacho, C. J. ANCHOR: A web server and database for analysis of protein−protein interaction binding pockets for drug discovery. *Nucleic Acids Res.* **2010**, *38*, W407−11.
(60) Kortemme, T.; Baker, D. A simple physical model for binding energy hot spots in protein−protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 14116−14121.
(61) Krissinel, E.; Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **2007**, *372*, 774−797.
(62) Tuncbag, N.; Keskin, O.; Gursoy, A. HotPoint: Hot spot prediction server for protein interfaces. *Nucleic Acids Res.* **2010**, *38*, W402−406.
(63) Zhu, X.; Mitchell, J. C. KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* **2011**, *79*, 2671−2683.
(64) Lounnas, V.; Pettitt, B. M.; Phillips, G. N. A global model of the protein−solvent interface. *Biophys. J.* **1994**, *66*, 601−614.
(65) De Simone, A.; Dodson, G. G.; Verma, C. S.; Zagari, A.; Fraternali, F. Prion and water: Tight and dynamical hydration sites have a key role in structural stability. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7535−7540.
(66) Fornili, A.; Autore, F.; Chakroun, N.; Martinez, P.; Fraternali, F. Protein−water interactions in MD simulations: POPS/POPSCOMP solvent accessibility analysis, solvation forces, and hydration sites. *Methods Mol. Biol.* **2012**, *819*, 375−392.
(67) Seeliger, D.; Haas, J.; de Groot, B. L. Geometry-based sampling of conformational transitions in proteins. *Structure* **2007**, *15*, 1482−1492.
(68) Fernández, A.; Berry, R. S. Extent of hydrogen-bond protection in folded proteins: A constraint on packing architectures. *Biophys. J.* **2002**, *83*, 2475−2481.
(69) Jorgensen, W.; Tirado-Rives, J. The OPLS potential functions for proteins—Energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(70) Bobay, B. G.; Thompson, R. J.; Hoch, J. A.; Cavanagh, J. Long range dynamic effects of point-mutations trap a response regulator in an active conformation. *FEBS Lett.* **2010**, *584*, 4203−4207.
(71) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(72) Van Gunsteren, W.; Billeter, S.; Eising, A.; Hunenberger, P.; Kruger, P.; Mark, A.; Scott, W.; Tironi, I. *Biomolecular Simulation: The GROMOS96 Manual and User's Guide*; Hochschuleverlag AG an der ETH Zurich: Zurich, Switzerland, 1996.
(73) Berendsen, H.; Postma, J.; Van Gunsteren, W.; Di Nola, A.; Haak, J. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(74) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(75) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **1997**, *2*, 173−181.
(76) Daily, M. D.; Gray, J. J. Local motions in a benchmark of allosteric proteins. *Proteins* **2007**, *67*, 385−399.
(77) Pandini, A.; Fornili, A.; Kleinjung, J. Structural alphabets derived from attractors in conformational space. *BMC Bioinformatics* **2010**, *11*, 97.
(78) Park, B. H.; Levitt, M. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **1995**, *249*, 493−507.
(79) Pandini, A.; Bonati, L.; Fraternali, F.; Kleinjung, J. MinSet: A general approach to derive maximally representative database subsets by using fragment dictionaries and its application to the SCOP database. *Bioinformatics* **2007**, *23*, 515−516.
(80) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley-Interscience: New York, 1991.
(81) Roulston, M. Estimating the errors on measured entropy and mutual information. *Phys. D* **1999**, *125*, 285−294.
(82) Hub, J. S.; de Groot, B. L. Detection of functional modes in protein dynamics. *PLoS Comput. Biol.* **2009**, *5*, e1000480.

(83) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **1993**, *17*, 412−425.
(84) Bordogna, A.; Pandini, A.; Bonati, L. Predicting the accuracy of protein−ligand docking on homology models. *J. Comput. Chem.* **2011**, *32*, 81−98.
(85) Hess, B. Convergence of sampling in protein simulations. *Phys. Rev. E* **2002**, *65*.
(86) Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **2000**, *62*, 8438−8448.
(87) Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **1993**, *215*, 617−621.
(88) Schäfer, H.; Smith, L. J.; Mark, A. E.; van Gunsteren, W. F. Entropy calculations on the molten globule state of a protein: Side-chain entropies of α-lactalbumin. *Proteins* **2002**, *46*, 215−224.
(89) de Groot, B. L.; van Aalten, D. M.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. Prediction of protein conformational freedom from distance constraints. *Proteins* **1997**, *29*, 240−251.
(90) Eyrisch, S.; Helms, V. What induces pocket openings on protein surface patches involved in protein−protein interactions? *J. Comput.-Aided Mol. Des.* **2008**, *23*, 73−86.
(91) de Groot, B. L.; Vriend, G.; Berendsen, H. J. Conformational changes in the chaperonin GroEL: New insights into the allosteric mechanism. *J. Mol. Biol.* **1999**, *286*, 1241−1249.
(92) Haliloglu, T.; Gul, A.; Erman, B. Predicting important residues and interaction pathways in proteins using Gaussian network model: Binding and stability of HLA proteins. *PLoS Comput. Biol.* **2010**, *6*, e1000845.
(93) Zen, A.; Micheletti, C.; Keskin, O.; Nussinov, R. Comparing interfacial dynamics in protein−protein complexes: An elastic network approach. *BMC Struct. Biol.* **2010**, *10*, 26.
(94) Zen, A.; Carnevale, V.; Lesk, A. M.; Micheletti, C. Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families. *Protein Sci.* **2008**, *17*, 918−929.
(95) Bahar, I.; Rader, A. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struc. Biol.* **2005**, *15*, 586−592.
(96) Meireles, L.; Gur, M.; Bakan, A.; Bahar, I. Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins. *Protein Sci.* **2011**, *20*, 1645−1658.
(97) Micheletti, C.; Carloni, P.; Maritan, A. Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and Gaussian models. *Proteins* **2004**, *55*, 635−645.
(98) Kohn, J. E.; Afonine, P. V.; Ruscio, J. Z.; Adams, P. D.; Head-Gordon, T. Evidence of functional protein dynamics from X-ray crystallographic ensembles. *PLoS Comput. Biol.* **2010**, *6*, e1000911.
(99) Doruker, P.; Atilgan, A. R.; Bahar, I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor. *Proteins* **2000**, *40*, 512−524.
(100) Liu, L.; Koharudin, L. M. I.; Gronenborn, A. M.; Bahar, I. A comparative analysis of the equilibrium dynamics of a designed protein inferred from NMR, X-ray, and computations. *Proteins* **2009**, *77*, 927−939.
(101) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and applications. *BMC Bioinformatics* **2009**, *10*, 421.
(102) Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308−311.
(103) Hamosh, A.; Scott, A. F.; Amberger, J. S.; Bocchini, C. A.; McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **2005**, *33*, D514−7.
(104) Vogel, C.; Chothia, C. Protein family expansions and biological complexity. *PLoS Comput. Biol.* **2006**, *2*, e48.
(105) Vogel, C.; Berzuini, C.; Bashton, M.; Gough, J.; Teichmann, S. A. Supra-domains: Evolutionary units larger than single protein domains. *J. Mol. Biol.* **2004**, *336*, 809−823.
(106) Keskin, O.; Tsai, C.-J.; Wolfson, H.; Nussinov, R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.* **2004**, *13*, 1043−1055.
(107) Keskin, O.; Nussinov, R. Similar binding sites and different partners: Implications to shared proteins in cellular pathways. *Structure* **2007**, *15*, 341−354.
(108) Li, X.; Keskin, O.; Ma, B.; Nussinov, R.; Liang, J. Protein−protein interactions: Hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: Implications for docking. *J. Mol. Biol.* **2004**, *344*, 781−795.
(109) Cukuroglu, E.; Gursoy, A.; Keskin, O. Analysis of hot region organization in hub proteins. *Ann. Biomed. Eng.* **2010**, *38*, 2068−2078.
(110) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. GSATools: Analysis of allosteric communication and functional local motions using a Structural Alphabet. *Bioinformatics* **2013**, *29*, 2053−2055.
(111) Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Dermitzakis, E.; Schaffner, S. F.; Yu, F.; Peltonen, L.; Dermitzakis, E.; Bonnen, P. E.; Altshuler, D. M.; Gibbs, R. A.; de Bakker, P. I. W.; Deloukas, P.; Gabriel, S. B.; Gwilliam, R.; Hunt, S.; Inouye, M.; Jia, X.; Palotie, A.; Parkin, M.; Whittaker, P.; Yu, F.; Chang, K.; Hawes, A.; Lewis, L. R.; Ren, Y.; Wheeler, D.; Gibbs, R. A.; Muzny, D. M.; Barnes, C.; Darvishi, K.; Hurles, M.; Korn, J. M.; Kristiansson, K.; Lee, C.; McCarrol, S. A.; Nemesh, J.; Dermitzakis, E.; Keinan, A.; Montgomery, S. B.; Pollack, S.; Price, A. L.; Soranzo, N.; Bonnen, P. E.; Gibbs, R. A.; Gonzaga-Jauregui, C.; Keinan, A.; Price, A. L.; Yu, F.; Anttila, V.; Brodeur, W.; Daly, M. J.; Leslie, S.; McVean, G.; Moutsianas, L.; Nguyen, H.; Schaffner, S. F.; Zhang, Q.; Ghori, M. J. R.; McGinnis, R.; McLaren, W.; Pollack, S.; Price, A. L.; Schaffner, S. F.; Takeuchi, F.; Grossman, S. R.; Shlyakhter, I.; Hostetter, E. B.; Sabeti, P. C.; Adebamowo, C. A.; Foster, M. W.; Gordon, D. R.; Licinio, J.; Manca, M. C.; Marshall, P. A.; Matsuda, I.; Ngare, D.; Wang, V. O.; Reddy, D.; Rotimi, C. N.; Royal, C. D.; Sharp, R. R.; Zeng, C.; Brooks, L. D.; McEwen, J. E. Integrating common and rare genetic variation in diverse human populations. *Nature* **2010**, *467*, 52−58.
(112) Abecasis, G. R.; Auton, A.; Brooks, L. D.; DePristo, M. A.; Durbin, R. M.; Handsaker, R. E.; Kang, H. M.; Marth, G. T.; McVean, G. A. An integrated map of genetic variation from 1,092 human genomes. *Nature* **2012**, *491*, 56−65.
(113) Brookes, A. J. The essence of SNPs. *Gene* **1999**, *234*, 177−186.
(114) Kim, S.; Misra, A. SNP Genotyping: Technologies and biomedical applications. *Annu. Rev. Biomed. Eng.* **2007**, *9*, 289−320.
(115) Ferrer-Costa, C.; Orozco, M.; de la Cruz, X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* **2002**, *315*, 771−786.
(116) Winget, J. M.; Mayor, T. The diversity of ubiquitin recognition: Hot spots and varied specificity. *Mol. Cell* **2010**, *38*, 627−635.
(117) Kiel, C.; Serrano, L. The ubiquitin domain superfold: Structure-based sequence alignments and characterization of binding epitopes. *J. Mol. Biol.* **2006**, *355*, 821−844.
(118) Patil, A.; Nakamura, H. Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.* **2006**, *580*, 2041−2045.
(119) Bahar, I.; Chennubhotla, C.; Tobi, D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr. Opin. Struc. Biol.* **2007**, *17*, 633−640.
(120) Zheng, W.; Brooks, B. R.; Thirumalai, D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7664−7669.
(121) Maguid, S.; Fernández-Alberti, S.; Echave, J. Evolutionary conservation of protein vibrational dynamics. *Gene* **2008**, *422*, 7−13.
(122) Pandini, A.; Mauri, G.; Bordogna, A.; Bonati, L. Detecting similarities among distant homologous proteins by comparison of domain flexibilities. *Protein Eng. Des. Sel.* **2007**, *20*, 285−299.
(123) Echave, J.; Fernández, F. M. A perturbative view of protein structural variation. *Proteins* **2010**, *78*, 173−180.
(124) Tyagi, M.; Shoemaker, B. A.; Bryant, S. H.; Panchenko, A. R. Exploring functional roles of multibinding protein interfaces. *Protein Sci.* **2009**, *18*, 1674−1683.

(125) Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197−208.
(126) Mittag, T.; Kay, L. E.; Forman-Kay, J. D. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* **2010**, *23*, 105−116.
(127) Kuzu, G.; Gursoy, A.; Nussinov, R.; Keskin, O. Exploiting conformational ensembles in modeling protein−protein interactions on the proteome scale. *J. Proteome Res.* **2013**, *12*, 2641−2653.
(128) Zhang, Q. C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C. A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; Maniatis, T.; Califano, A.; Honig, B. Structure-based prediction of protein−protein interactions on a genome-wide scale. *Nature* **2012**, *490*, 556−560.
(129) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein−protein interfaces. *Nature* **2007**, *450*, 1001−1009.
(130) Poirot, O.; O'Toole, E.; Notredame, C. Tcoffee@igs: A web server for computing, evaluating, and combining multiple sequence alignments. *Nucleic Acids Res.* **2003**, *31*, 3503−3506.