JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

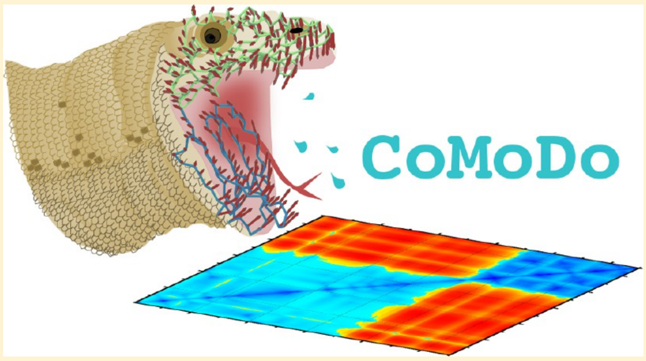
CoMoDo: Identifying Dynamic Protein Domains Based on Covariances of Motion

Silke A. Wieninger[‡] and G. Matthias Ullmann*

Structural Biology/Bioinformatics, University of Bayreuth, Universitätsstrasse 30, BGI, 95447 Bayreuth, Germany

S *Supporting Information*

ABSTRACT: Most large proteins are built of several domains, compact units which enable functional protein motions. Different domain assignment approaches exist, which mostly rely on concepts of stability, folding, and evolution. We describe the automatic assignment method CoMoDo, which identifies domains based on protein dynamics. Covariances of atomic fluctuations, here calculated by an Elastic Network Model, are used to group residues into domains of different hierarchical levels. The so-called dynamic domains facilitate the study of functional protein motions involved in biological processes like ligand binding and signal transduction. By applying CoMoDo to a large number of proteins, we demonstrate that dynamic domains exhibit features absent in the commonly assigned structural domains, which can deliver insight into the interactions between domains and between subunits of multimeric proteins. CoMoDo is distributed as free open source software at www.bisb.uni-bayreuth.de/CoMoDo.html.



INTRODUCTION

The term ’domain’ was originally introduced for the structurally separate regions of immunoglobulins.[1] With more structural data becoming available, the existence of such distinct structural regions and their impact on protein folding was hypothesized generally for globular proteins.[2] Although protein domains are intuitively thought of as compact structural units forming more connections within than across domains, standard definitions are hard to find. Most domain definitions rely on the three-dimensional protein structure, but domains can also be defined based on a functional description, like DNA binding domains or kinase domains in signaling proteins. Those functional domains may differ from the structural domains, because functional sites tend to lie at the interface between compact units.[3] Likewise, evolutionary defined domains, building blocks which can recombine on the genetic level to proteins with different functions,[4] do not necessarily coincide with structural domains, because discontinuous domains can only occur due to multiple insertion events.[5,6]

The assignment of structural domains is often performed manually, but there is also a wide range of automatic methods. Top-down approaches, the basis for most earlier methods, search for single cut points along the linear sequence, dividing the structure into continuous domains.[7,8] These methods need to make an additional effort to extend the domain identification to domains which consist of more than one segment.[9−13] In contrast, bottom-up methods, which cluster residues starting from a few residues,[14−18] secondary structure elements,[16,19] or hydrophobic cores,[20] may result in several segments per domain and often implement a postclustering merging step to reduce the number of small segments. More recent methods impose less restrictions on the tolerated number of segments per domain or on the splitting of secondary structure elements and even consider the possibility that domains may be composed of multiple chains.[16] An exhaustive comparison of structural domains, determined by various algorithmic methods, with manual domain assignments[21] showed that especially the number of assigned domains differs, while the domain boundaries usually agree well. Accordingly, hierarchical levels of domain partitioning, which should be interpreted in a context-dependent manner, were proposed by several authors.[13,15,16,21−23]

The obtained structural domains are frequently used to predict functional motions of proteins, because distinct domains can move in relation to each other at low energetic cost, undergoing hinge and shear motions;[24] but instead of inferring protein dynamics from structural domains, which are defined based on contact matrices,[9,10] distance matrices,[15] interaction energy matrices,[7,17] and graphs,[18,25] one can directly define dynamic domains based on concerted residue motions. Such dynamic domains can deviate from structural domains if the protein structure is not clearly divided into separate parts. Dynamic domains can be used to analyze potential large-scale protein motions or the effect of ligand binding and oligomerization on protein dynamics. In a previous study on

Received: February 16, 2015
Published: April 28, 2015

DOI: 10.1021/acs.jctc.5b00150

the enzyme aminoglycoside phosphotransferase 3′-IIIa,[26] we showed that binding of substrates between different dynamic domains leads to either more or less flexibility, depending on the architecture of the involved domains. Besides, dynamic domains can help to identify perturbation-sensitive sites of proteins, where addition or removal of a few interactions lead to large changes of protein dynamics. Several methods identify rigid domains based on coordinates given by two conformations, representing the open and the closed state of a protein[27,28] or based on snapshots from Molecular Dynamics simulations[22,29] and NMR ensembles.[30] Another common approach is to use principal components, calculated from a structural ensemble, or normal modes. These large-amplitude eigenvectors describe global protein movements and allow, based on the directions of motion, for identifying residues which belong to the same quasi-rigid domain.[31] Normal modes can be calculated by an elastic network model (ENM),[32,33] a coarse-grained method which uses purely topological constraints deduced from the protein structure to determine single-residue fluctuations and collective protein motions. In some approaches, only one[34,35] or several low-frequency normal modes[3,36] calculated by ENM are used. To consider the contribution of all normal modes, dynamic domains can be assigned based on covariances of motion.[33,37]

The here described program CoMoDo (**Co**variance of **Mo**tion **Do**mains) also groups covariances of residue motion using a clustering method called DomainClusterer to predict dynamic domains. In contrast to the work of Yesylevskyy et al.,[37] where the number of domains is determined based on the largest correlation difference between two clustering steps, CoMoDo implements a second method, called DomainTester, which checks whether a protein or protein part actually consists of several domains. As input data, CoMoDo only depends on the connectivity of the residues and on the covariance matrix, calculated by an optional simulation method. In this work, the covariance matrices are determined by a Gaussian network model (GNM),[32] one variant of the ENM. Other than most domain assignment methods, CoMoDo does not use any postprocessing steps to alter unexpected domain classifications after the actual assignment procedure. The dynamic domains are allowed to be discontinuous, and small fragments can belong to another dynamic domain than their sequential neighbors.

In the following, we describe the algorithms used by the programs DomainTester and DomainClusterer, as well as the overall workflow of CoMoDo and two alternative approaches, NormCoMoDo and FastCoMoDo. We compare our predictions to manual domain assignments for a data set of 135 proteins[21] to investigate the differing properties of dynamic and structural protein domains. On the examples of 4-hydroxyphenylacetate decarboxylase[38] and acetylene hydratase,[39] we demonstrate how the particular features of dynamic domains can help to understand mechanisms of protein action. Finally we study the influence of GNM parameters on the domain assignment and show the dynamic domains obtained with the alternative approaches FastCoMoDo and NormCoMoDo.

THEORY

CoMoDo identifies dynamic protein domains based on covariances of residue motion. It includes two programs written in C++ and shell scripts combining them into different CoMoDo approaches, which are described in the section CoMoDo, NormCoMoDo, and FastCoMoDo. CoMoDo is distributed as free open source software at www.bisb.uni-bayreuth.de/CoMoDo.html under the terms of the GNU Affero General Public License. The program DomainClusterer performs an agglomerative clustering of the residues into domains, while DomainTester checks if the protein or protein part consists of several domains. The dynamical input in the form of covariance matrices can be obtained from Normal Mode analysis or Principal Component analysis of structural ensembles. It is assumed that for the calculation of the covariances, each residue is represented by one pseudoatom, denoted as node in the following. Thus, for N residues, the covariance matrix is a symmetric $N \times N$ matrix. The matrix entries add up to zero, because translational and rotational motions are described by the eigenvectors with zero eigenvalues, which are excluded.[37] In contrast, the sum over all correlations, the normalized covariances, does in general not equal zero. Therefore, we use covariances instead of correlations as similarity measure in the agglomerative clustering procedure. Here, we calculate covariance matrices using a coarse-grained method, the Gaussian Network Model (GNM), which is described in the Methods section. In the following description of the algorithms, the term domain is only used for the final residue partition. The term cluster is used for preliminary groups of residues which have to be combined or split to become domains.

DomainTester: Differentiation between 1-Domain and Multidomain Proteins.

Distinguishing 1-domain proteins from multidomain proteins is a crucial part of the domain identification procedure. Figure 1 shows the obvious differences which exist between covariance matrices of 1-domain and multidomain proteins. We need to find rules that describe these differences and allow for a computational

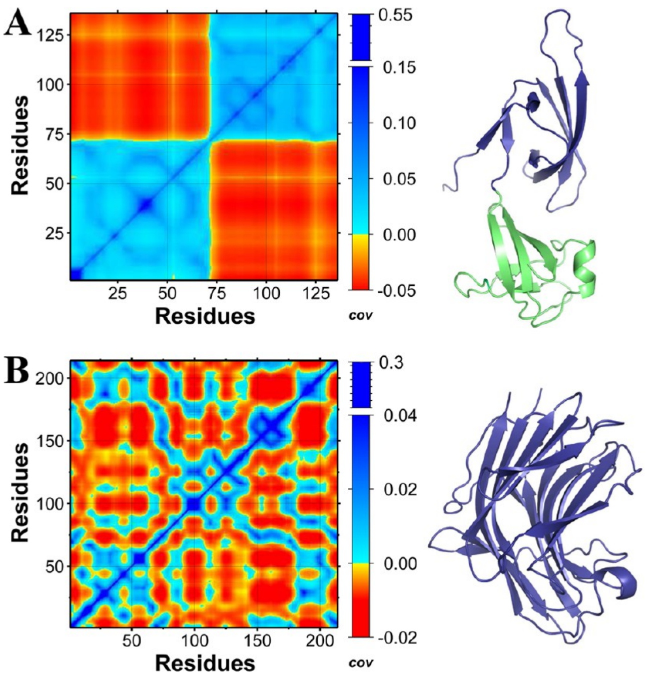


Figure 1. Covariance matrix and structure of a two-domain protein compared to those of a 1-domain protein. A) The covariance matrix of translation initiation factor 5A (PDB 1bkb[72]) has two separate positive-covariance segments. B) The covariance matrix of beta-glucanase (PDB 1byh[73]) shows no large area of only positive covariances.

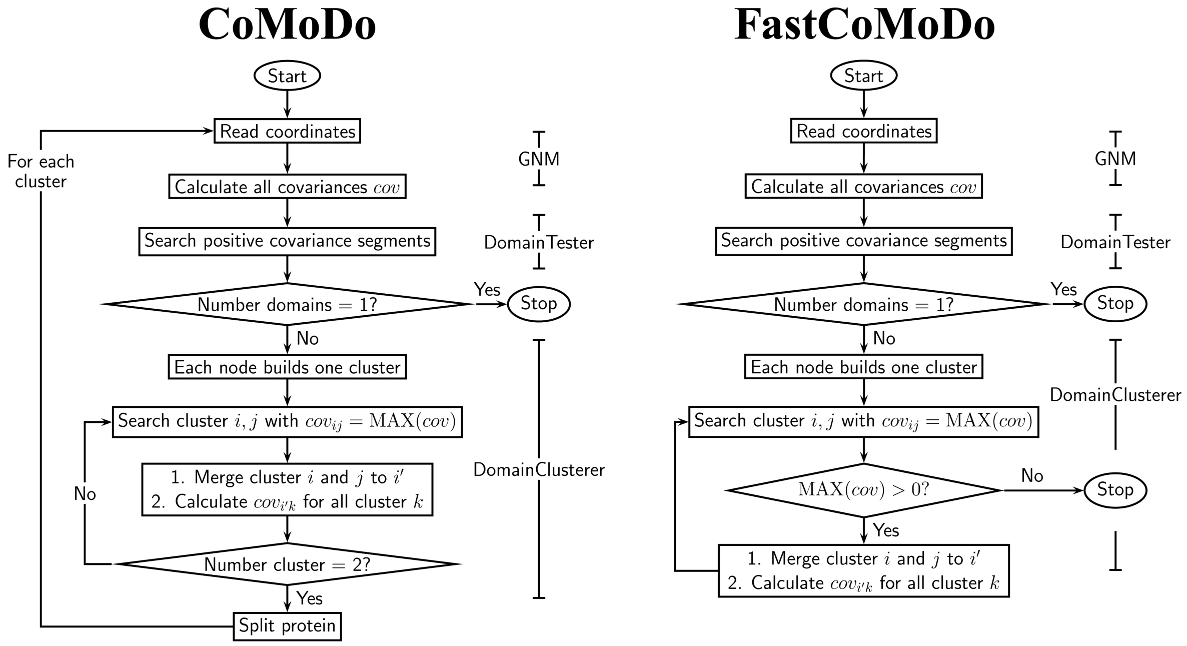

CoMoDo
Start
Read coordinates
For each cluster
Calculate all covariances cov
Search positive covariance segments
Number domains = 1?
Yes
Stop
No
Each node builds one cluster
Search cluster i, j with $cov_{ij} = \mathrm{MAX}(cov)$
1. Merge cluster i and j to i'
2. Calculate $cov_{i'k}$ for all cluster k
No
Number cluster = 2?
Yes
Split protein
GNM
DomainTester
DomainClusterer
FastCoMoDo
Start
Read coordinates
Calculate all covariances cov
Search positive covariance segments
Number domains = 1?
Yes
Stop
No
Each node builds one cluster
Search cluster i, j with $cov_{ij} = \mathrm{MAX}(cov)$
$\mathrm{MAX}(cov) > 0$?
No
Stop
Yes
1. Merge cluster i and j to i'
2. Calculate $cov_{i'k}$ for all cluster k
GNM
DomainTester
DomainClusterer

Figure 2. Overall workflow of CoMoDo and FastCoMoDo. FastCoMoDo uses a negative intercluster covariance as stopping criterion in the clustering procedure, while CoMoDo always clusters all nodes into two clusters and then employs DomainTester to test if the cluster can be further divided.

evaluation. Covariance matrices of multidomain proteins have large sequential areas of positive values, in contrast to covariance matrices of 1-domain proteins. Therefore, we search for regions of at least $s_{\min}$ nodes with positive covariance cov_{ij} between all pairs of nodes, that is

$$\mathrm{cov}_{ij} > 0 \text{ for all } i, j \text{ in } \{k, k+1, ..., l\} \text{ and } l - k + 1 \geq s_{\min} \tag{1}$$

We call these regions positive-covariance segments. The segments are usually overlapping, meaning that one node is part of several segments. α_{seg} is the fraction of nodes that belong to at least one positive-covariance segment. If at least $\alpha_{\mathrm{seg}}^{\min}$ of the nodes can be grouped into positive-covariance segments, the protein is considered as multidomain protein. Additionally, we request that the number of nonoverlapping positive-covariance segments, n_{seg}, is at least two. Thus, for being a multidomain protein, the following two criteria must be met:

$$n_{\mathrm{seg}} \geq 2 \text{ and } \alpha_{\mathrm{seg}} \geq \alpha_{\mathrm{seg}}^{\min} \tag{2}$$

DomainClusterer: Agglomerative Clustering of Covariances. In the beginning of the clustering process by DomainClusterer, every node builds one cluster. Then the two clusters with highest positive covariance to each other are merged into one cluster. We denote the two clusters with highest intercluster covariance as $c1$ and $c2$. They comprise N_{c1} and N_{c2} nodes, and their covariance is denoted as $\mathrm{cov}_{c1,c2}$. The covariance of cluster $c1$ to any other cluster d is $\mathrm{cov}_{c1,d}$. Let us denote the merged cluster consisting of nodes $c1$ and $c2$ as c. The covariance of the new cluster c to any other cluster d is calculated by averaging over the covariances of cluster $c1$ and $c2$ to cluster d:

$$\mathrm{cov}_{c,d} = \frac{\mathrm{cov}_{c1,d} \cdot N_{c1} + \mathrm{cov}_{c2,d} \cdot N_{c2}}{N_{c1} + N_{c2}} \tag{3}$$

The new cluster c consists of N_{c1}+N_{c2} nodes, and the total number of clusters is reduced by one. The intracluster covariance of the merged cluster is $\mathrm{cov}_{c,c}$. It is given by the average over all covariances within the cluster:

$$\mathrm{cov}_{c,c} = \frac{\mathrm{cov}_{c1,c1} \cdot N_{c1}^2 + \mathrm{cov}_{c2,c2} \cdot N_{c2}^2 + \mathrm{cov}_{c1,c2} \cdot 2N_{c1}N_{c2}}{N_{c1}^2 + N_{c2}^2 + 2N_{c1}N_{c2}} \tag{4}$$

The inter- and intracluster covariances equal the arithmetic mean over the covariances of all cluster nodes. An example of the clustering algorithm can be found in Figure S1A of the Supporting Information. With n giving the number of clusters, the relation

$$\sum_{a=1}^{n} \sum_{b=1}^{n} \mathrm{cov}_{a,b} \cdot N_a N_b = 0 \tag{5}$$

is true after each step, because the sum over all entries of the covariance matrix equals zero. The program stops either when a certain number of clusters is reached, or when the highest intercluster covariance is smaller than a given cutoff value, which is typically set to zero. DomainClusterer does not necessarily arrange all residues of positive-covariance segments determined by DomainTester into one cluster or domain. The clusters are allowed to be smaller than $s_{\min}$, the minimal size of positive-covariance segments, and can be discontinuous. In contrast to DomainTester, DomainClusterer neglects the sequential information.

CoMoDo, NormCoMoDo, and FastCoMoDo. CoMoDo and the alternative approaches NormCoMoDo and FastCoMoDo combine the programs DomainTester and DomainClusterer to predict dynamic domains (see Figure 2). All approaches start with calculating the covariance matrix of the whole protein and calling DomainTester to check if the protein consists of more than one domain. If so, DomainClusterer merges the residues until a stopping criterion is fulfilled, which depends on the applied approach. In CoMoDo, DomainClusterer merges residues until they are divided into two clusters. After each splitting, new covariance matrices of the clusters are calculated, and DomainTester checks if they are in turn composed of more than one cluster. This approach assumes that the motions of the dynamic domains are independent and can be calculated separately for each cluster. The final dynamic domains are arranged into different hierarchical levels, and one can easily see from the domain numbering which domains are split first and which domains are split in a later step, meaning that they are less anticorrelated to each other (see Figure S1B of the Supporting Information). Based on our calculations on a large protein set, we recommend the usage of CoMoDo whenever it is applicable; but the calculation of covariance matrices of splitted protein structures may be impossible, especially if full-atom methods are used for generating conformational ensembles. In such cases, the alternative approaches NormCoMoDo and FastCoMoDo can be used, which avoid recalculation of covariance matrices. In contrast to CoMoDo, they do not assume independency but consider the influence of residues of other clusters on the covariances. NormCoMoDo extracts the parts of the full covariance matrix which belong to residues of one cluster and renormalizes this smaller covariance matrix by subtracting the average covariance from all matrix elements. As a result, the sum over all cluster elements is zero again. The rest of the clustering procedure coincides with CoMoDo. FastCoMoDo differs from the other two approaches in the stopping criterion used for the clustering procedure. DomainClusterer stops merging residues when the largest intercluster covariance is smaller than a cutoff value, which is usually set to zero, and the final domain number corresponds to the number of remaining clusters. In FastCoMoDo, DomainTester is only used in the first step.

■ METHODS

Gaussian Network Model. The Gaussian Network Model (GNM) is a coarse-grained method which uses the atomic coordinates of the protein to build a network consisting of one or several nodes per residue.[26,32] The nodes are connected covalently if they represent sequential residues. Nodes representing nonsequential residues are only connected if their equilibrium distance d° is smaller than a given cutoff radius d_{cut}. The vector

$$\Delta\mathbf{R}_i = (x_i - x_i^\circ, y_i - y_i^\circ, z_i - z_i^\circ) = (\Delta x_i, \Delta y_i, \Delta z_i) \tag{6}$$

gives the difference between instantaneous and equilibrium positions of node i. The potential energy V of a network of N nodes is then given by

$$V = \sum_{i,j=1}^{N} \frac{k_{ij}}{2}\Gamma_{ij}((\Delta x_i - \Delta x_j)^2 + (\Delta y_i - \Delta y_j)^2 + (\Delta z_i - \Delta z_j)^2) \tag{7}$$

where k_{ij} equals the force constant for covalent or noncovalent interactions

$$k_{ij} = \begin{cases} k_{\text{cov}}, & \text{if linked covalently} \\ k_{\text{ncov}}, & \text{else} \end{cases} \tag{8}$$

and Γ_{ij} is the ijth element of the $N \times N$-dimensional Kirchhoff matrix $\mathbf{\Gamma}$, defined by

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } d_{ij}^\circ \leq d_{\text{cut}} \\ 0, & \text{if } i \neq j \text{ and } d_{ij}^\circ > d_{\text{cut}} \\ -\sum\limits_{k,k\neq i} \Gamma_{ik}, & \text{if } i = j \end{cases} \tag{9}$$

The energy function penalizes distortions from the equilibrium coordinates of the experimental structure by summation over pairwise energy terms. GNM allows calculation of variances $\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_i \rangle$ and covariances $\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle$ of residue fluctuations, which are evaluated from the diagonal and off-diagonal elements of the pseudoinverse Kirchhoff matrix $\tilde{\mathbf{\Gamma}}^{-1}$, respectively.

$$\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle = \frac{3k_{\text{B}}T}{k_{ij}} \tilde{\Gamma}_{ij}^{-1} \tag{10}$$

The absolute values of the force constants only change the absolute scale of fluctuations. Therefore, we define the force constant ratio K

$$K = \frac{k_{\text{ncov}}}{k_{\text{cov}}} \tag{11}$$

which influences the covariances of residue fluctuations. K, the cutoff radius d_{cut}, and the number of nodes per residue are parameters of the GNM which can be optimized by comparison of the theoretical atom flexibilities to experimental B-factors from X-ray crystallography.[40] Covariance matrices of residue motion indicate which residues tend to move simultaneously into the same direction (positive covariance), and which residues are anticorrelated to each other (negative covariance). As protein domains have only few connections to each other, they can move apart at low energetic cost, while residues of the same domain stick together. Therefore, covariance matrices can be used to identify the number and boundaries of domains present in a protein.

Computational Details. In the GNM, amino acids are represented by one node at the C_α position. If cofactors are present in the crystal, they influence the dynamics. Hence they are included in the elastic network by a number of nodes depending on the size of the molecule. In calculations analyzing the influence of GNM parameters on the domain assignment, the GNM cutoff radius d_{cut} is varied from 6 to 11 Å, while log K adopts values between −2 and 0 in steps of 0.25. As default parameters for DomainTester calculations, we use a minimal size for positive covariance segments of $s_{\text{min}} = 40$, and a minimal fraction of nodes which must belong to positive-covariance segments of $\alpha_{\text{seg}}^{\text{min}} = 0.5$. The covariance plots, protein images, and charts were produced using GMT,[41] PyMOL,[42] and gnuplot, respectively.

Protein Data Set and Evaluation. The domain identification algorithm is applied to proteins of the benchmark data set 3 from the pDomains Web site,[21] which offers domain information about 135 proteins, 55 of them being assigned as 1-domain protein. Table S3 of the Supporting Information gives the PDB codes and the number of assigned structural domains for all proteins of the data set. The data set is constructed based on domain assignments from methods which are not or at least not fully automatic. SCOP[43] relies on structural and evolutionary relationships between proteins, CATH[44] classifies proteins according to their structure by a combination of automatic and manual procedures, and AUTHORS collects

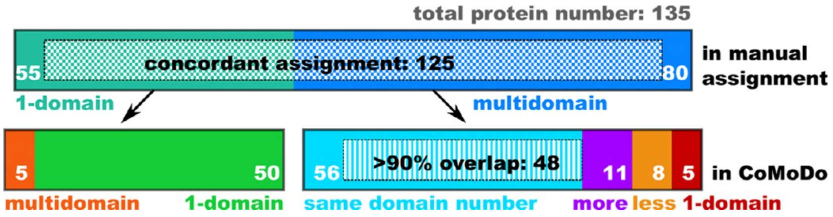


Figure 3. Comparison of domain assignments by CoMoDo with structural domains of 135 proteins of the pDomains data set.[21] 50 out of 55 proteins classified as 1-domain protein, and 75 out of 80 proteins classified as multidomain protein manually are assigned accordingly by CoMoDo. Of the 75 proteins classified as multidomain by both approaches, the same domain number is assigned to 56 proteins. More and less domains are assigned to 11 and 8 proteins, respectively. 48 of the 56 multidomain proteins with an equal number of structural and dynamic domains reach a domain overlap higher than 90%. In the calculations, a cutoff radius of 7 Å and force constant ratio of 0.5 were used.

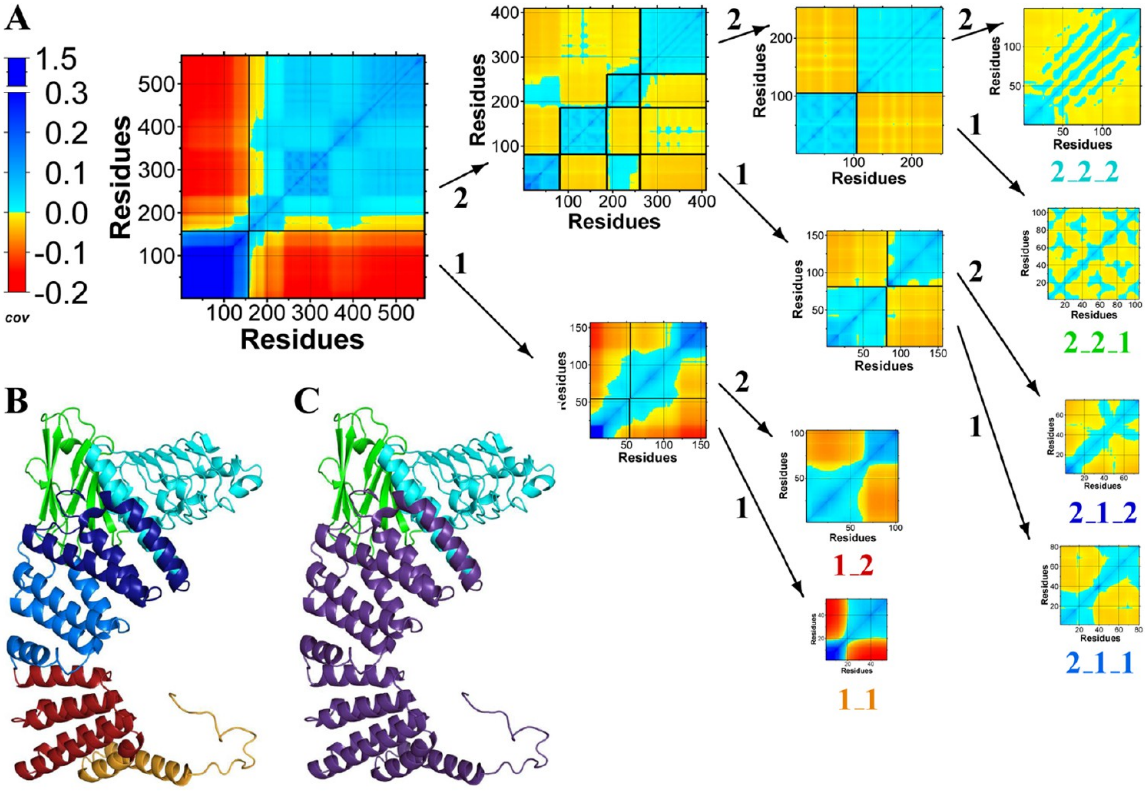


Figure 4. Clustering of Rab geranylgeranyltransferase (PDB 1dce[46]). A) Hierarchical clustering of the covariance matrix. First, the precursor cluster of the red and the orange domain is split from the rest. Then, the precursor cluster of the blue and the dark blue domain is split from the precursor of the green and cyan domain. B) Six dynamic domains assigned by CoMoDo. C) Three domains assigned manually.

assignments of the authors of protein structures. In the following, we refer to these methods as manual assignments and call the resulting domains structural domains. For proteins of the data set, the three manual methods agree about the number of domains and at least to 90% about the domain boundaries. For comparing dynamic domains to structural domains, we use the domain boundaries assigned by the authors of the structures. As measure for the similarity between domain boundaries assigned by different methods, the percentage of domain overlap[45] is used. For each dynamic domain, the number of common residues with all structural domains is determined. The best mapping of dynamic on structural domains corresponds to the combination with the highest sum of matching residues. If the domain numbers assigned by the two methods differ, the spare domains remain unpaired. The number of matching residues is divided by the total number of residues which are assigned both to a dynamic and to a structural domain. Residues may be unassigned by one of the approaches due to different reasons. More than one-quarter of the multidomain proteins of the data set has missing $C_\alpha$coordinates within the protein chain, which are not assigned to domains by CoMoDo, in contrast to some manual assignments. On the other hand, CoMoDo assigns each residue present in the ENM to a domain, while expert methods sometimes leave out residues. Also, cofactors are represented by several nodes in the ENM but not considered in manual assignments. In calculations comparing the GNM parameter choice based on B-factors to the usage of fixed parameters for all proteins, only proteins with available crystallographic B-factors and positive linear correlation coefficient between experimental and theoretical B-factors are used. In calculations studying the domain number of multidomain proteins in dependence on the GNM parameters, only proteins which are assigned as multidomain manually and by DomainTester for all parameter pairs are considered. The 60 multidomain proteins

DOI: 10.1021/acs.jctc.5b00150

with available crystallographic B-factors are highlighted in Table S3 of the Supporting Information.

■ PROPERTIES OF DYNAMIC PROTEIN DOMAINS

Dynamic versus Structural Protein Domains. We study the general properties of dynamic protein domains and their relationship to structural domains by applying CoMoDo to the 135 proteins of the pDomains data set.[21] The covariance matrices are calculated using the Gaussian network model (GNM) with a cutoff radius of 7 Å and force constants of 10 kcal(mol Å ^{2})$^{-1}$ for covalent interactions and 5 kcal(mol Å ^{2})$^{-1}$ for noncovalent interactions. As Figure 3 shows, 79% of the 135 proteins of the data set are split into the same number of dynamic domains as assigned manually, including the assignments as 1-domain protein. Considering only multidomain proteins, the same domain number is assigned to 70% of the 80 proteins. But 86% of those 56 multidomain proteins with according domain number reach an average percentage of domain overlap higher than 90%. These data show that for a known or preassigned domain number, the residue distribution is often quite obvious, while the domain number itself is more ambiguous.

To understand which structural properties of proteins can lead to large discrepancies between dynamic and structural domains, we study the protein Rab geranylgeranyltransferase,[46] shown in Figure 4, which is partitioned into three domains by the authors of the structure and into six dynamic domains by CoMoDo. For two domains, the CoMoDo assignment coincides with the manual assignment; but the third, largest structural domain consisting of more than 300 residues is divided into four dynamic domains. The six dynamic domains belong to different hierarchies and demonstrate how CoMoDo creates domains through iterative splitting of the structure and recalculation of covariance matrices. The first CoMoDo splitting step already cuts in the middle of the large structural domain. Although it can be classified as evolutionary domain, because its helical fold is also found in other proteins,[46] from a dynamical view its residues clearly belong to at least two different dynamic domains, as one can recognize in the covariance matrix of the whole protein. It seems that the bundling of α-helices in the structural domain makes it difficult to identify domain boundaries manually.

Similar observations can be made for nitrous oxide reductase[47] (Figure 5A1), split into three dynamic domains versus two structural domains. Again, the similar arrangement of secondary structure elements seems to lead to overestimation of the contacts between the residues. In contrast, iron−sulfur protein of carbon monoxide dehydrogenase (Figure 5A2) is assigned as 1-domain protein by CoMoDo but as 2-domain protein manually. The contacts between the two structural domains may be underestimated due to the sharp contrast between four-helix bundle architecture on one side and five-stranded β-sheet on the other side. Additionally, each structural domain binds one [2Fe-2S] cluster by a binding motif which also occurs in other proteins.[48] The examples above show that the assignment of structural domains is often influenced by evolutionary and functional aspects and can lead to domains of very different sizes. In contrast, the dynamic domains of one protein usually comprise a similar number of residues, an observation that was explained by Yesylevskyy et al. by increased domain stability.[37]

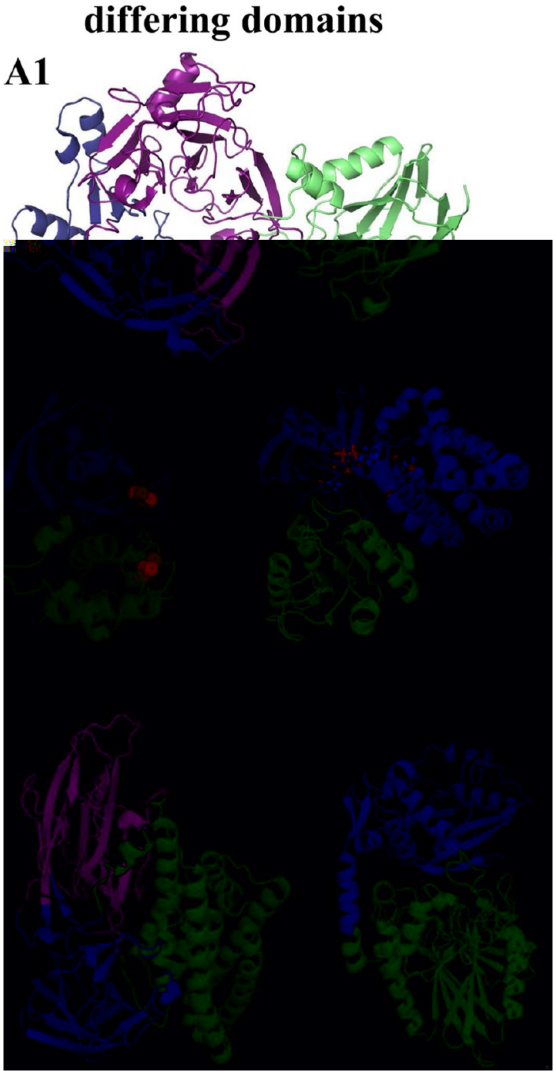


Figure 5. Comparison between dynamic and structural domains. A) Proteins with differing domain numbers. A1) Nitrous oxide reductase (PDB 1qni[47]) is divided into two domains manually but into three by CoMoDo. The blue and the magenta domain build one structural domain. A2) Iron−sulfur protein of carbon monoxide dehydrogenase (PDB 1ffu[48]) is assigned as 1-domain protein by CoMoDo but divided into two domains manually (colored in pale blue and pale green). The [2Fe-2S] clusters are shown as spheres. A3) Two domains are assigned to flavohemoglobin (PDB 1cqx[49]) by CoMoDo if FAD and heme are bound (shown as sticks). Manually, three domains are assigned. The blue and pale blue domain build one dynamic domain. Neglecting the ligands, CoMoDo also assigns three domains. B) Similar dynamic and structural domains. Arrows indicate residues with differing assignment. B1) In CryIA(a) toxin (PDB 1ciy[51]), residues at the domain interface are identified as intercalating segments coupled to the blue domain. The green domain is split off first. B2) In 5′-nucleotidase (PDB 1ush[52]), an α-helix is split between the dynamic domains.

But even if solely the protein structure is considered, the interrelation between the number of connections between two protein parts and the size and compactness of the protein parts themselves is often too complex to be predicted just by visual inspection. Protein parts which are anticorrelated to the rest of the protein are not always clearly visible as separate domains. At the same time, connections between compact regions can impede the independent movement with varying strength, depending on their position. At sensitive sites, small changes in the elastic network connections, like the binding of ligands, can have large effects on the covariances.[26] For example,

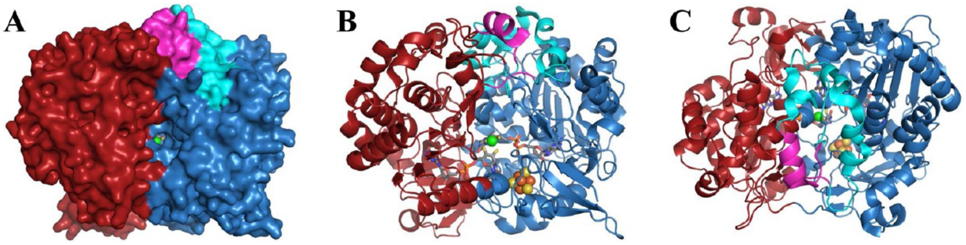


Figure 6. Dynamic domains 1 (blue, cyan) and 2 (red, magenta) of acetylene hydratase (PDB 2e7z[39]). The [4Fe-4S] cluster and the tungsten atom (green) are shown in VdW representation. The two molybdopterin guanine dinucleotide molecules are represented as sticks. A,B) View on the substrate channel which lies between the two dynamic domains. C) View on the alternative substrate channel found in other molybdenum and tungsten enzymes, which is sealed in acetylene hydratase by residues from both dynamic domains. Residues 329 to 370 and 384 to 393, shown in cyan, belong to domain 1. Residues 371 to 383, shown in magenta, belong to domain 2.

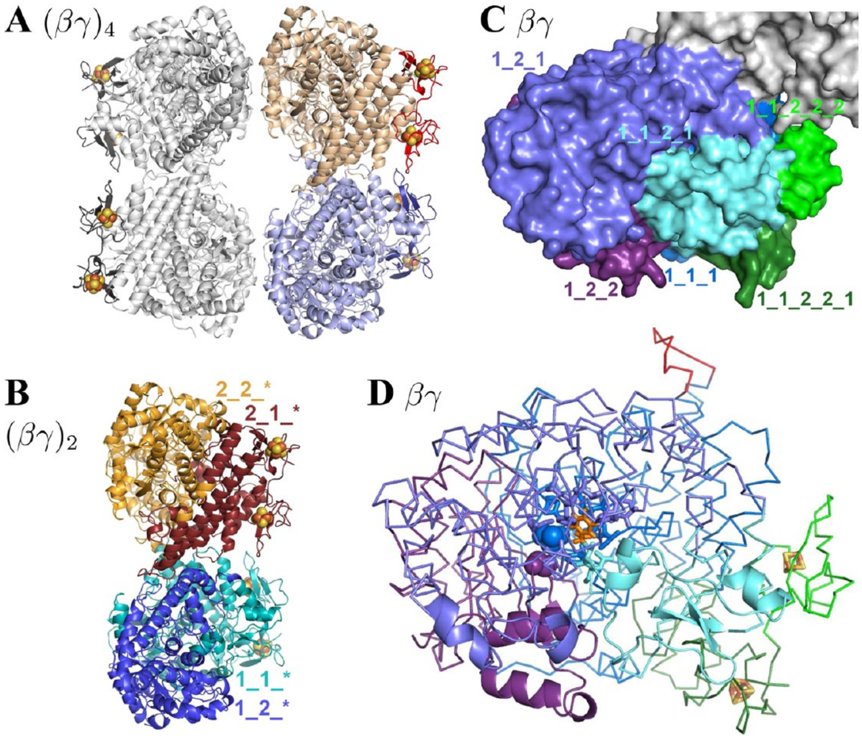


Figure 7. Dynamic domains of 4-hydroxyphenylacetate decarboxylase (PDB 2y8n[38]). A) Each of the four heterodimers consists of a β- (pale orange, pale blue, light gray) and a small γ- (red, blue, dark gray) subunit. The two identical homodimers (gray vs color) are separated by the first CoMoDo splitting step. The [4Fe-4S] clusters are shown in VdW representation. B) The first splitting of the heterodimers by CoMoDo occurs within the β-subunit and not at the subunit boundaries. C) Six dynamic domains of one heterodimer. The channel to the active site lies between the domains 1_2_1 and 1_1_2_1. The position of the neighboring heterodimer is indicated in gray. D) Radical segment (purple) and flexible segments (cyan, violet) of subunit β are shown in cartoon representation, colored according to their domain affiliation. Residues 252 to 263, colored in red, belong to the neighboring heterodimer instead of cluster 1_1_*. Residues of three dynamic domains, shown as sticks, are involved in ligand binding. The glycyl/thiyl radical dyad is shown in VdW representation, and the substrate 4-hydroxyphenylacetate (orange) and the [4Fe-4S] clusters are presented as sticks.

flavohemoglobin[49] (Figure 5A3) is assigned as a 2-domain protein by CoMoDo if FAD and heme are bound to it but as a 3-domain protein in the absence of cofactors. Manually, also three domains are assigned, but the cofactors should be included in the calculation, because they are part of the functional enzyme and influence the dynamic behavior. According to this, an Elastic Network Model study of high-resolution X-ray structures showed that adding ligands and cofactors to a GNM improves the correlation between theoretical and experimental B-factors.[50]

Another characteristic difference between dynamic and structural domains lies in the handling of small segments that protrude to the domain interface and are thus coupled dynamically to another domain than their sequential neighbors. These segments, called intercalating segments in the following, can also be seen in proteins with agreement in domain number and with high domain overlap. For example in CryIA(a) toxin from *Bacillus thuringiensis*[51] (Figure 5B1), a few loop residues but also small α-helices belong to another domain. Furthermore, α-helices and β-strands can be split between two dynamic domains, as in 5′-nucleotidase[52] (Figure 5B2). In manual domain assignments, such residues often remain unassigned, while other automatic domain assignment methods usually change the classification in a postprocessing step; but the location of intercalating segments can highlight interactions between domains and thus deliver information about protein functionality. To illustrate this statement, we analyze in the following the dynamic domains of two enzymes which are not part of the pDomains data set, acetylene hydratase and 4-hydroxyphenylacetate decarboxylase.

Acetylene Hydratase: Substrate Channel between Dynamic Domains. Acetylene hydratase catalyzes the hydration of acetylene to acetaldehyde.[39,53] Catalysis occurs by a water molecule bound to a bis-molybdopterin guanine dinucleotide-ligated tungsten atom. The water molecule is activated by an aspartate residue, Asp13, whose deprotonation is shifted to unusually high pH values by interaction with a nearby [4Fe-4S] cluster. Figure 6 shows the two dynamic domains assigned to acetylene hydratase by CoMoDo. For the calculation of the covariance matrix by GNM, the tungsten atom and the [4Fe-4S] cluster were each represented by one node in the elastic network, while each molybdopterin guanine dinucleotide molecule was represented by five nodes. No substrate nodes were included. Choosing the best GNM parameters based on comparison of B-factors out of a reduced parameter set (see section Choosing GNM and CoMoDo Parameters and Table S1 of the Supporting Information), we used a force constant of 0.56 kcal(mol Å ^{2})$^{-1}$ and a cutoff radius of 10 Å. Asp13, the [4Fe-4S] cluster, the tungsten atom, and one molybdopterin molecule belong to domain 1, while the second molybdopterin molecule belongs to domain 2. The substrate channel lies between the two dynamic domains, as is seen frequently in proteins and may lead to an easier entry of the substrate acetylene. Interestingly, in all other known enzymes of the DMSO reductase family of molybdenum and tungsten enzymes,[54,55] a different position of the channel to the active site is found. In acetylene hydratase, this alternative channel is sealed by a lid consisting of the residues 328 to 393 (see Figure 6C). While most of the lid residues (329 to 370 and 384 to 393) belong to domain 1, a few interjacent residues (371 to 383) are allocated to domain 2, which shows their strong interaction with residues of domain 2. Thus, the lid over the original substrate channel is connected to different dynamic domains, instead of being a flexible structure which could easily move away.

4-Hydroxyphenylacetate Decarboxylase: Dynamic Domains of a Multimer. All proteins of the pDomains data set are single subunits, although several of them are part of a larger protein complex in their active form. The splitting reflects the assumption that domains do not spread over several subunits; but from a dynamical point of view, residues of different subunits can belong to the same dynamic domain, which may be of functional importance in protein−protein interactions. In the following, the domain assignment method is demonstrated on the multimer 4-hydroxyphenylacetate decarboxylase (HPD). HPD is a glycyl radical enzyme which catalyzes the chemically difficult decarboxylation of 4-hydroxyphenylacetate to *p*-cresol.[38,56,57] The $(\beta\gamma)_4$ tetramer consists of heterodimers built of a catalytic β-subunit harboring a glycyl/thiyl dyad (Gly873, Cys503) and a small γ-subunit with two [4Fe-4S] clusters. The γ-subunit is not present in all glycyl radical enzymes and is proposed to be involved in regulation of the oligomeric state and catalytic activity of HPD.[58] In the GNM calculation, the [4Fe-4S] clusters are represented by one node each, lying in the center of the cluster. The substrate 4-hydroxyphenylacetate is not included. We use the parameter pair d_{cut} = 11 Å and log $((k_{\mathrm{ncov}})/(k_{\mathrm{cov}})) = -1.75$, which leads to the highest correlation between crystallographic and theoretical B-factors of the full parameter set and is as well part of the reduced parameter set (see Table S1 of the Supporting Information). The tetramer is split into 26 dynamic domains (see Figure 7). The first CoMoDo splitting step results in two identical $(\beta\gamma)_2$ clusters. The next splitting step separates the two heterodimers of one $(\beta\gamma)_2$ cluster from each other, but the separation is not symmetric and not complete, because the residues 252 to 263 of one heterodimer build an intercalating segment which is grouped into cluster 2_* instead of cluster 1_*, whereas all residues of the other heterodimer belong to cluster 2_*. Also the final domain number of the two heterodimers is not identical, one consisting of six, the other of seven dynamic domains. Thus, small differences in the coordinates and accordingly the covariances can lead to different assignments, but the main conclusions about the functional implications of the dynamic domain architecture are equivalently valid for both heterodimers.

First, the channel to the active site is flexible due to its location between two dynamic domains (1_2_1 and 1_1_2_1), as in acetylene hydratase. Second, residues interacting with the substrate 4-hydroxyphenylacetate belong to three different domains, domain 1_1_1 (R223, S344, G345, F405, E505), 1_2_1 (F214, I219, H536, F537 and E637, I750, V752), and 1_1_2_1 (V399 and L400). Their distribution onto several domains with uncorrelated motion could ensure the flexibility required to arrange the active site residues after substrate binding and prebuild the transition state of the reaction. Third, the interaction between the β- and the γ-subunits is quite strong, which can be deduced from the fact that the first splitting event occurs in the middle of the β-subunit. Only in a later step, the cluster 1_1_2_2_*, which consists mostly of residues of the γ-subunit, is split from domain 1_1_2_1, containing residues of the β-subunit. Furthermore, parts of the β-subunit and the γ-subunit belong to the same dynamic domain, showing the strong interaction between them. In the heterodimer shown in Figure 7C, domain 1_1_2_2_1 consists of residues 1−41 of the γ-subunit and residues 88−97 and 285−312 of the β-subunit. In the other heterodimer, a single dynamic domain includes all residues of the γ-subunit and 59 residues of the β-subunit. Finally, HPD interacts with an activating enzyme (AE) to generate the radical on Gly873. The radical segment containing Gly873 is flanked by two peptide sequences that are weakly structured in X-ray crystallography and are postulated to open upon complex formation with the AE.[38] Reductive cleavage of *S*-adenosylmethionine in the AE generates a transient 5′-deoxyadenosyl radical which then generates the Gly873 radical. The two interacting residues of the radical dyad belong to dynamic domains 1_1_1 and 1_2_2, whereas the flexible stretch including residues N672 to E700 belongs to domain 1_2_1 and the flexible stretch including residues Q121 to K167 belongs to domain 1_1_2_1. Thus, the flexible sequences are dynamically decoupled from the radical domain and can move away, which could lead to conformational changes in the radical domain.

■ CHOOSING GNM AND COMODO PARAMETERS

The number of dynamic domains and their boundaries determined by CoMoDo depend on the covariance matrices used and on the parameters of CoMoDo itself. First we investigate the influence of the GNM parameters by varying the cutoff radius from 6 to 11 Å and the logarithm of the ratio of noncovalent to covalent force constant, log *K*, from −2 to 0 in steps of 0.25, which results in 54 parameter pairs. Of the 55

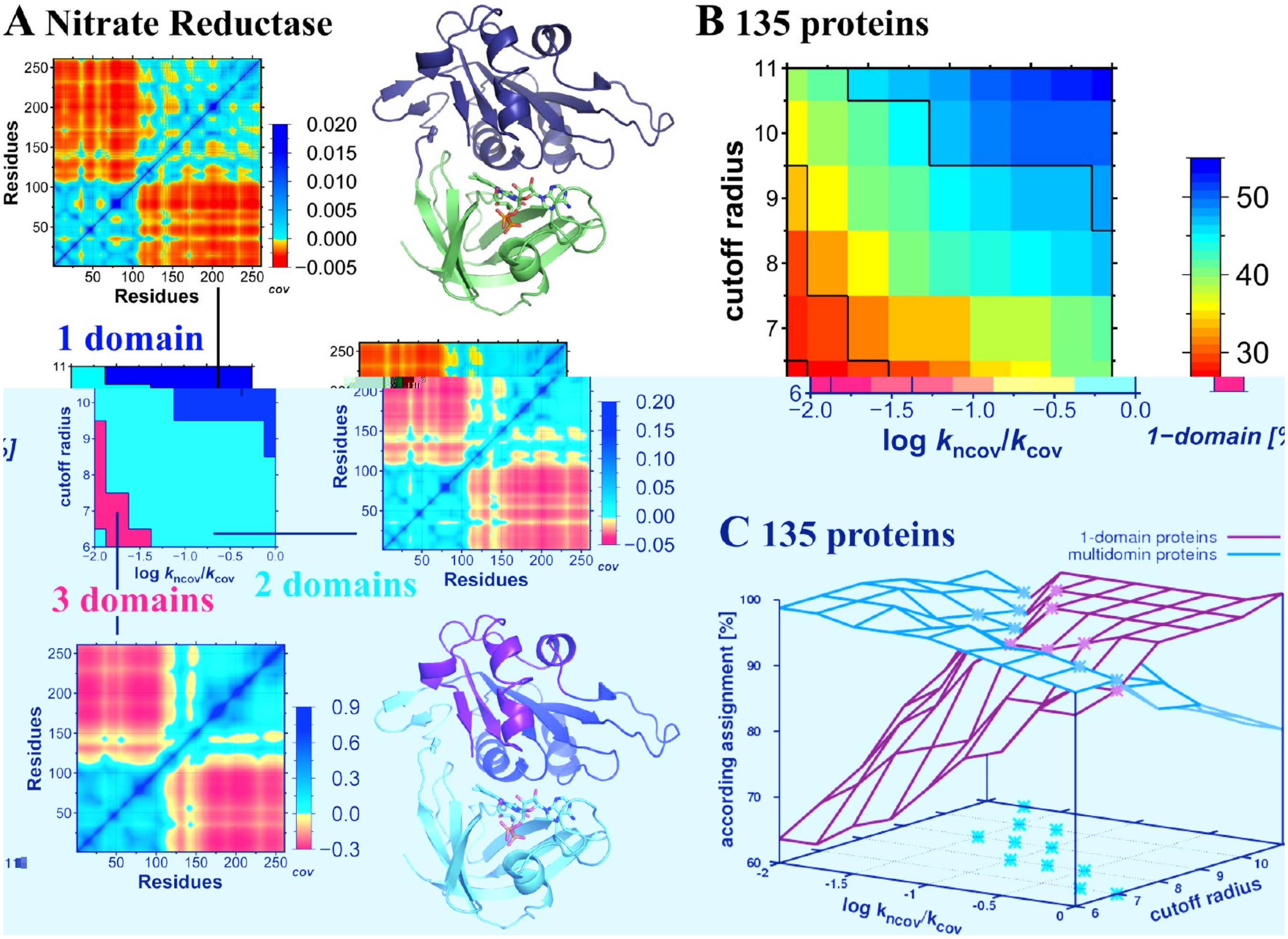


Figure 8. Influence of GNM parameters on the number of dynamic domains. A) One to three domains are assigned to nitrate reductase (PDB 1cne[59]). Manually, two domains are assigned. The covariance matrices, shown for exemplary parameter pairs, illustrate that small cutoff radii in combination with a small force-constant ratio underestimate the nonbonded interactions and lead to a broad zone of positive covariance along the diagonal, which results in the assignment of three domains. In contrast, large cutoff radii and large noncovalent force constants result in the assignment as 1-domain protein. B) The percentage of proteins of the pDomains data set assigned as 1-domain protein by DomainTester. A percentage of 40.7 corresponds to the percentage of 1-domain proteins assigned manually. The black lines indicate the domain borders of nitrate reductase from A). C) Percentage of the 80 multidomain proteins and the 55 1-domain proteins, assigned manually, which are classified accordingly by DomainTester. Eleven GNM parameter pairs, called reduced set, which lead to an accordance of at least 90% between DomainTester and manual predictions for both 1-domain and multidomain proteins are highlighted by stars.

proteins assigned as 1-domain protein in the pDomains data set, 20 are assigned as 1-domain protein by DomainTester for all 54 parameter pairs. Of the 80 proteins assigned as multidomain in the pDomains data set, the domain number is always the same for 14 proteins and agrees with the manual assignment except for two of them. For the rest of the proteins, the number of dynamic domains depends on the GNM parameters used. Although for some proteins we expect the number of dynamic domains to deviate from the number of structural domains, as discussed in the first section, the manual assignments can serve as a guideline for the choice of GNM parameters, because common conceptions of size and compactness of protein domains are adopted.

We analyze the assignment of one to three dynamic domains to nitrate reductase[59] (see Figure 8A), which is assigned as two-domain protein by manual methods. For combinations of small cutoff radii with small K, the covariance matrices are less scattered and there is a broad zone of positive covariance along the diagonal, which allows for the detection of many positive-covariance segments. On the contrary, for high cutoff radii and large K, very small absolute values of covariance lead to strong fragmentation by alternation between negative and positive values and subsequently to classification as 1-domain protein. As Figure 8B shows, larger cutoff radii and high K generally lead to a higher ratio of proteins assigned as 1-domain. The GNM parameter pairs giving the expected domain number of two for nitrate reductase coincide with parameters leading to a reasonable ratio of 1-domain proteins when applying DomainTester to the full pDomains data set. To check if the two methods classify the same set of proteins as 1-domain protein, Figure 8C shows the percentage of 1-domain or multidomain proteins, according to manual predictions, which are assigned correspondingly by DomainTester. Obviously, at a higher ratio of 1-domain proteins, more proteins which are assigned as 1-domain manually are also assigned as 1-domain by CoMoDo. The opposite is true for multidomain proteins. For good agreement between manual methods and CoMoDo, a compromise between the contrary trends must be found.

Eleven GNM parameter pairs which lead to an accordance of at least 90% for both 1-domain and multidomain proteins lie at the intersection between the two curves in Figure 8C. Those parameters, called reduced set, are given in Table S1 of the Supporting Information. As Figure 9B shows, the parameter pairs of the reduced set are also well-suited for the determination of the number of dynamic domains of multidomain proteins, with nine of them leading to an agreement with the number of structural domains of at least 65%.

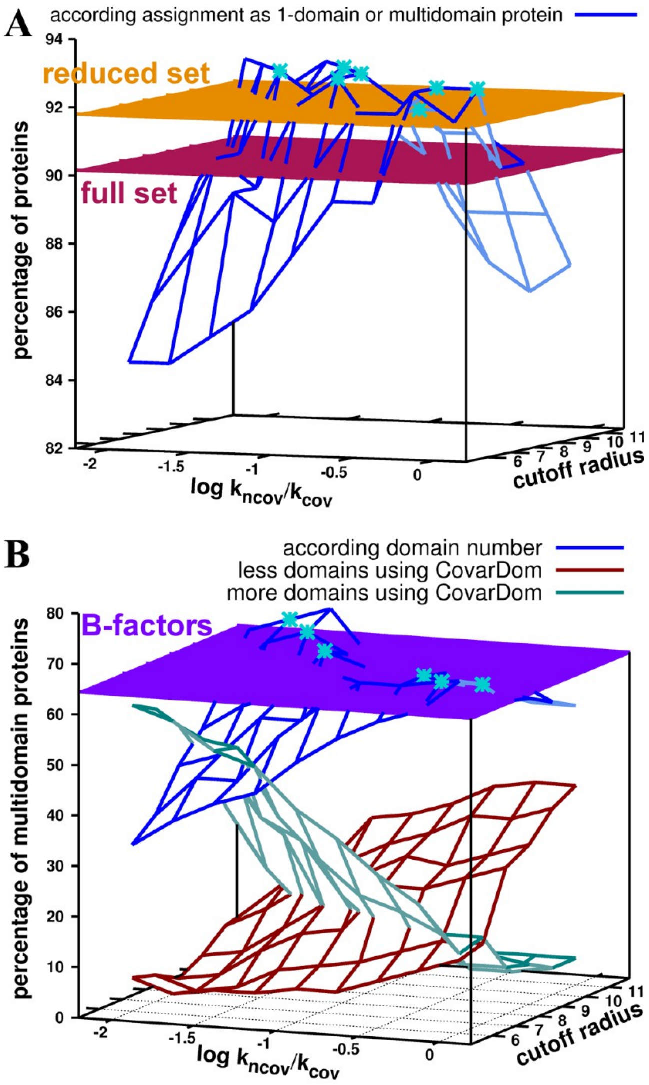


Figure 9. Accordance between the number of dynamic and structural domains. The parameter pairs of the reduced set are highlighted by cyan stars. Calculations are performed on all proteins of the pDomains data set with available experimental B-factors. A) Percentage of proteins which are assigned accordingly by DomainTester and by manual methods as 1-domain or multidomain protein. The planes situated at 90.2% and 91.8% give the accordance if the GNM parameter pairs are chosen separately for each protein from the full or the reduced parameter set by comparison to experimental B-factors. B) Percentage of the 76 proteins assigned as multidomain in the pDomains data set for which CoMoDo assigns the same, a smaller or a larger number of domains than given by manual predictions. The B-factor approach leads to agreement of 64.5% for both the full and the reduced set.

Instead of using one fixed GNM parameter pair, one can determine the parameters separately for each protein by comparing theoretical to crystallographic B-factors.[60−62] Table S3 of the Supporting Information gives the GNM parameter pairs determined by the so-called B-factor approach out of the reduced and the full set for each protein and the corresponding number of dynamic domains assigned by CoMoDo; but the B-factor approach turns out to perform less well than the best parameter pairs, in the distinction between 1-domain and multidomain proteins just as in the determination of the domain number of multidomain proteins (see Figure 9A). Applying the B-factor approach to the full set bears a risk of selecting parameter pairs which generally lead to the assignment of too many 1-domain proteins or too many domains in multidomain proteins. The parameter pair d_{cut} = 11 Å and log $(k_{ncov})/(k_{cov}) = -2$ is selected most often, for 11 of 122 proteins, followed by the parameter pair d_{cut} = 11 Å and log $(k_{ncov})/(k_{cov}) = 0$, selected ten times. Both parameter pairs are not part of the reduced set. Besides, the linear correlation coefficient between crystallographic and theoretical B-factors is low for many proteins of the pDomains data set. One possible reason is the high fraction of proteins which were crystallized as larger complexes. While 63% of the proteins crystallized as monomers have a linear correlation coefficient of at least 0.6, this number decreases to 51% for multimeric proteins. Besides, the correlation between B-factors is usually higher if the theoretical B-factors are calculated considering the crystal environment of the protein.[63,64] We recommend to apply the B-factor approach to the reduced set only or to simply use one of the nine parameter pairs leading to high accordance for both 1-domain and multidomain proteins (see Table S1 of the Supporting Information). A cutoff radius of 7 Å, as employed in the first results section, corresponds to the typical value chosen in GNM to include the interactions in the first shell of neighbors.[65,66] Several studies proposed the usage of stronger force constants for covalent than for noncovalent interactions[50,62] or distance-dependent force constants.[67] As our analysis shows, the nonbonded interactions should however not be underestimated, because noncovalent force constants which are a hundred times weaker than the covalent force constant lead to the assignment of too many dynamic domains. Overall it may be advisable to run repeated calculations to analyze which results are stable over a large set of GNM parameters.

Besides the influence of cutoff radius and force constant on the calculation of the covariance matrix, the identification of dynamic protein domains also depends on the parameters used by DomainTester. Although default values work well for most proteins, it may be necessary to adapt them for certain protein architectures. For example, cytochrome f is assigned as 1-domain protein by CoMoDo, but consists of two structural domains according to the authors.[68] By visual inspection of the structure and the covariance matrix (Figure 10), one would agree that cytochrome f consists of two dynamic domains, because two separated, clearly anticorrelated protein parts exist. The residues of the smaller structural domain are highly positively correlated and show a strong anticorrelation to most of the residues of the larger structural domain. Also the residues of the larger structural domain are dynamically coupled, but the sequence of residues building the central β-sheet is disrupted by long loops, short α-helices, and the residues of the small domain. Therefore, the corresponding positive-covariance segment includes only 43 residues. In total, the positive-covariance segments comprise 43% of the residues, which leads

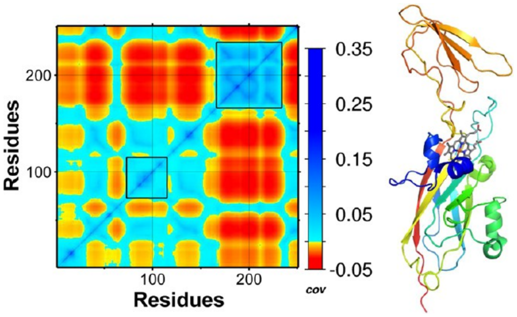


Figure 10. Structure and covariance matrix of cytochrome f (PDB 1e2v[68]). DomainTester detects two positive-covariance segments, which comprise 43% of the residues. These residues are encircled in the covariance plot. If the required fraction of nodes in positive-covariance segments is lowered to 0.4, two dynamic domains are assigned. The structure of cytochrome f in cartoon representation is colored by sequence to show that the smaller domain is inserted between two β-strands belonging to the larger domain.

to the classification as 1-domain protein if default values are used in CoMoDo. By lowering the required total fraction of nodes in positive-covariance segments from 0.5 to 0.4, also CoMoDo assigns two domains to cytochrome f.

■ ALTERNATIVE COMODO APPROACHES

After each splitting step, CoMoDo recalculates covariance matrices of isolated clusters and uses DomainTester to decide whether they can be further divided. When covariance matrices of the splitted protein cannot be determined, for example because molecular dynamics is used instead of an ENM, or if the influence of the whole protein on cluster dynamics should be considered, other approaches can be used. Instead of combining nodes until only two clusters remain, FastCoMoDo uses a cutoff value of the intercluster covariance as stopping criterion of DomainClusterer, which is set to zero in the following. As the intercluster covariance is usually positive until a small number of clusters is reached, the resulting domain numbers are similar to those calculated by CoMoDo. Still, FastCoMoDo assigns less often two and more often three dynamic domains to proteins than CoMoDo and agrees less often with the domain numbers assigned manually (see Figure 11A). The additional domains assigned by FastCoMoDo are often small fragments lying between larger domains, as for example in TAFII250[69] (Figure 11B) and endonuclease I-DmoI[70] (Figure 11C). The small domains have a negative intercluster covariance to all other domains and are therefore not added to them by FastCoMoDo. In CoMoDo, the precursor cluster containing the small domain and a larger domain is classified as one domain by DomainTester, because no positive-covariance segment is found in the corresponding covariance matrix. Thus, the small fragments are not separated from the large dynamic domain using CoMoDo. For some proteins which have two domains that are strongly anticorrelated, but consist of further dynamic domains themselves, FastCoMoDo also assigns less dynamic domains than CoMoDo. Figure 12 shows the covariance matrices of elongation factor Tu,[71] calculated using a noncovalent force constant of 5 kcal(mol Å ^{2})$^{-1}$ and a cutoff radius of 7 Å. In the last step of FastCoMoDo, DomainClusterer merges domains 2_1 and 2_2 with a positive intercluster covariance of 0.0028. Thus, if an intercluster covariance of zero is chosen as stopping criterion, the final domain number of elongation factor Tu is two, although from the covariance matrix it is obvious that three dynamic domains are present. CoMoDo allows for the assignment of three domains, because the removal of the highly anticorrelated domain 1 from the elastic network shifts the intercluster covariance between domains 2_1 and 2_2 to negative values, as the sum over all covariances is always zero.[37]

Another possibility is to renormalize the parts of the original matrix which belong to one cluster using NormCoMoDo. In contrast to CoMoDo, where the dynamic domains are considered as being independent from each other, the influence of the residues of the other dynamic domains are still present in the covariances. Figure 12 shows the difference between the renormalized part of the covariance matrix calculated for the whole protein and the covariance matrix newly calculated only for cluster 2 on the example of elongation factor Tu. The differences, which give the deviation from independent behavior of domain 1 and cluster 2, are rather small with

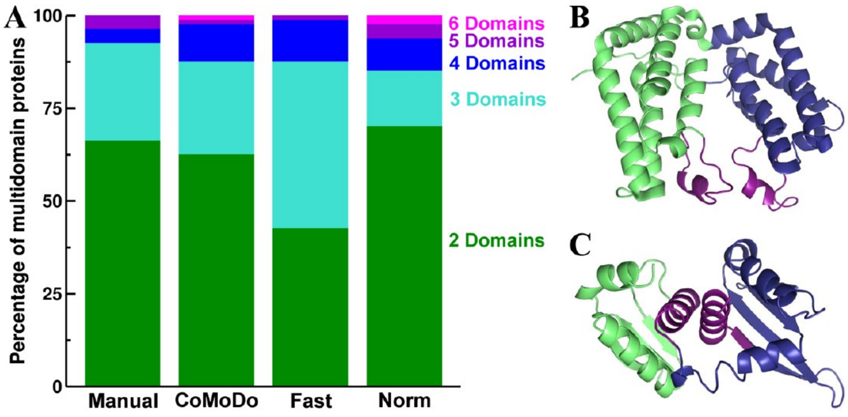


Figure 11. Dynamic domains assigned by alternative CoMoDo approaches. A) Comparison of the domain numbers assigned to multidomain proteins manually, by CoMoDo, FastCoMoDo, and NormCoMoDo. For all approaches, the total number of proteins assigned as multidomain is 80, but the protein sets are not the same, because DomainTester assigns five proteins assigned as multidomain manually as 1-domain proteins and five proteins assigned as 1-domain manually as multidomain proteins. The bars indicate the percentage of the multidomain proteins with the specified domain number. The proteins B) TAFII250 (PDB 1eqf[69]) and C) endonuclease I-DmoI (PDB 1b24[70]) are assigned as 2-domain proteins by CoMoDo and as 3-domain proteins by FastCoMoDo. The colors indicate the dynamic domains assigned by FastCoMoDo. Blue and magenta domain together build one CoMoDo domain.

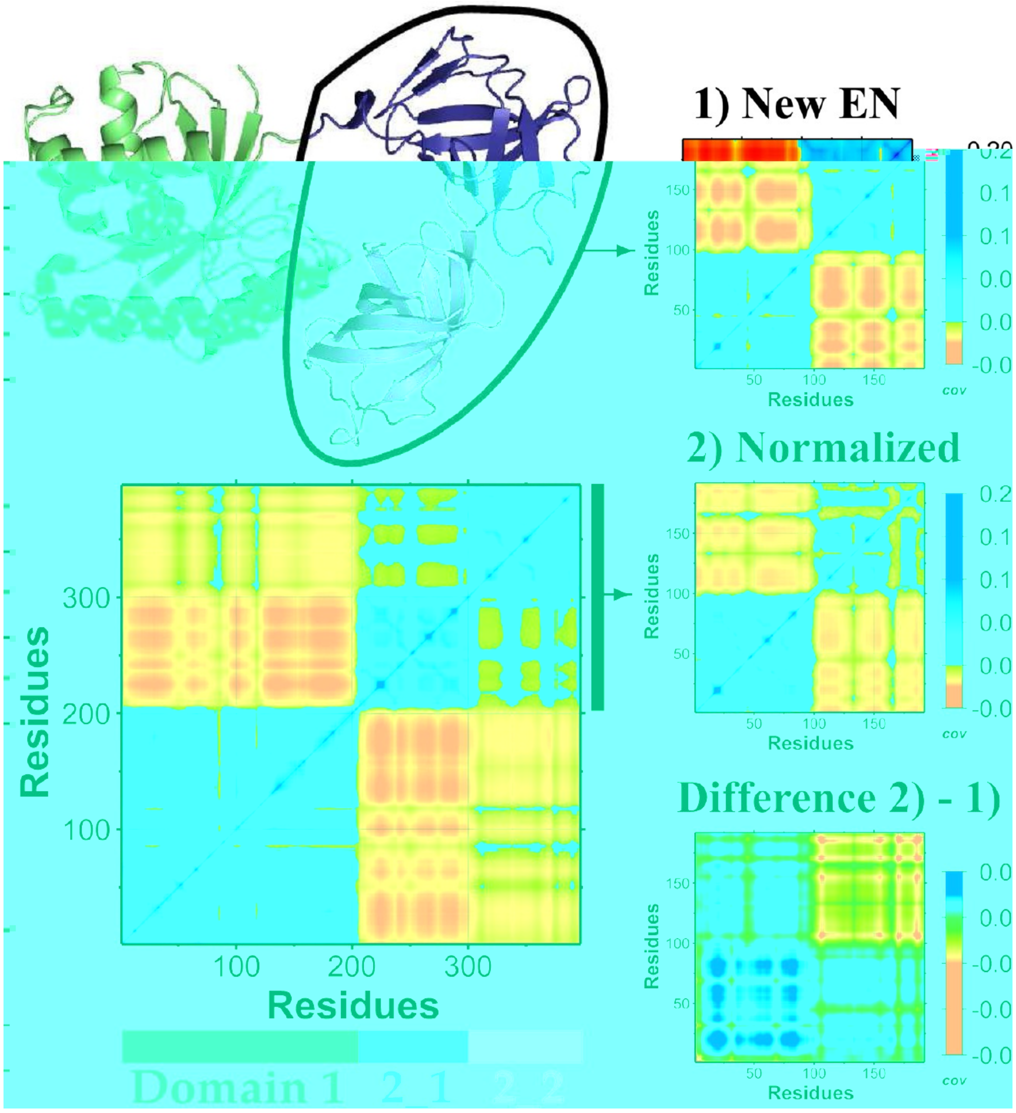


Figure 12. Dynamic domains of elongation factor Tu (PDB 1tui[71]) assigned by alternative CoMoDo approaches. Elongation factor Tu is assigned as 3-domain protein by CoMoDo but as 2-domain protein by FastCoMoDo or NormCoMoDo. Using FastCoMoDo, the domains 2_1 and 2_2 are merged into one dynamic domain, because their intercluster covariance is positive. Using NormCoMoDo, only one positive-covariance segment is found in cluster 2_*. Covariance matrix 1) of cluster 2 is calculated only for the residues of cluster 2, whereas covariance matrix 2) is calculated for the whole protein and then renormalized. The difference between the two covariance matrices indicates the deviation from independent behavior of domain 1 and cluster 2, that is it shows the influence of the residues of domain 1 on the residues of cluster 2.

values between −0.06 to +0.02. Nevertheless, using renormalized covariances results in the assignment of two instead of three dynamic domains to elongation factor Tu, because DomainTester detects only one positive-covariance segment in cluster 2 comprising the residues 7−97 of domain 2_1. Domain 2_2 has more interactions with Domain 1, which leads to less independent movement and thus to partly negative covariances of its residues. Employing the renormalization strategy to all multidomain proteins of the pDomains benchmark shows that NormCoMoDo, just as FastCoMoDo, leads to less overall agreement with manual domain assignments than CoMoDo.

■ CONCLUSIONS

Dynamic domains have direct functional relevance, because the functionality of a protein is tightly connected to its dynamics. For example, ligand binding sites often lie between dynamic domains. The uncorrelated motion of different domains can allow for an easier entry of the substrates and a perfect arrangement of the active site residues. Besides, sites lying between anticorrelated domains are often perturbation-sensitive, such that ligand binding has a large effect on the dynamics of the protein, potentially leading to allosteric behavior. In contrast to structural domains, dynamic domains are often sequentially discontinuous, and the location of intercalating segments highlights the residues which mediate interdomain relations. Additional information about the strength of interactions between dynamic domains is given by their hierarchical organization created by CoMoDo. For multimeric proteins, all subunits can be included in the calculation. Just as for different domains, the interactions between different subunits can be deduced from the existence of intersubunit domains and the order of the splitting events. The assignment of dynamic protein domains by CoMoDo is not influenced by human conception but purely based on previously calculated dynamical data. Still, the automatic domain assignment is influenced by the choice of both GNM and CoMoDo parameters and should only be used as a guideline which is followed by manual inspection. Repeated calculations can be performed to analyze which results are stable over a wide range of GNM parameters. Using a cutoff radius of 7 Å, a standard value for GNM calculations, and a noncovalent force constant which is smaller than the covalent force constant, but not by several orders of magnitude, works well for most proteins. Instead, the elastic network parameters can also be chosen by comparison of theoretical to crystallographic B-factors from a reasonable set of cutoff radii and ratios between covalent and noncovalent force constants.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

GNM parameter pairs of the reduced set, accordance values between dynamic and structural domains for the assignment as 1-domain or multidomain protein, an exemplary calculation of inter- and intracluster covariances and the number of dynamic domains assigned to all proteins of the pDomains data set, determined for different GNM parameters. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00150.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Matthias.Ullmann@uni-bayreuth.de.

Present Address

‡Unité de Bioinformatique Structurale, Institut Pasteur, 25-28 rue du docteur Roux, 75015 Paris, France.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Lars Ackermann for implementing the program TEN and Anncharlott Berglar for the artistic design of the graphical abstract. This work was supported by the BioMedTec International Graduate School "Lead Structures of Cell Function" of the Elitenetwork Bavaria, the DFG grant UL 174/8, and the Bavarian program for the promotion of equal opportunities for women in research and teaching.

■ REFERENCES

(1) Cunningham, B. A.; Gottlieb, P. D.; Pflumm, M. N.; Edelman, G. M. In *Progress in immunology*; Amos, B., Ed.; Academic Press: New York, 1971; pp 3–24.

(2) Wetlaufer, D. B. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70*, 697–701.

(3) Potestio, R.; Pontiggia, F.; Micheletti, C. *Biophys. J.* **2009**, *96*, 4993–5002.

(4) Doolittle, R. F. *Annu. Rev. Biochem.* **1995**, *64*, 287–314.

(5) Russell, R. B. *Protein Eng.* **1994**, *7*, 1407–1410.

(6) Majumdar, I.; Kinch, L. N.; Grishin, N. V. *PloS One* **2009**, *4*, e5084.

(7) Zhou, H.; Xue, B.; Zhou, Y. *Protein Sci.* **2007**, *16*, 947–955.

(8) Emmert-Streib, F.; Mushegian, A. *BMC Bioinf.* **2007**, *8*, 237.

(9) Holm, L.; Sander, C. *Proteins* **1994**, *19*, 256–268.

(10) Islam, S. A.; Luo, J.; Sternberg, M. J. *Protein Eng.* **1995**, *8*, 513–525.

(11) Siddiqui, A. S.; Barton, G. J. *Protein Sci.* **1995**, *4*, 872–884.

(12) Wernisch, L.; Hunting, M.; Wodak, S. J. *Proteins* **1999**, *35*, 338–352.

(13) Berezovsky, I. N. *Protein Eng.* **2003**, *16*, 161–167.

(14) Xuan, Z. Y.; Ling, L. J.; Chen, R. S. *Eur. Biophys. J.* **2000**, *29*, 7–16.

(15) Carugo, O. *J. Appl. Crystallogr.* **2007**, *40*, 778–781.

(16) Feldman, H. J. *BMC Bioinf.* **2012**, *13*, 286.

(17) Genoni, A.; Morra, G.; Colombo, G. *J. Phys. Chem. B* **2012**, *116*, 3331–3343.

(18) Ansari, E. S.; Eslahchi, C.; Pezeshk, H.; Sadeghi, M. *Proteins* **2014**, *82*, 1937–1946.

(19) Sowdhamini, R.; Blundell, T. L. *Protein Sci.* **1995**, *4*, 506–520.

(20) Swindells, M. B. *Protein Sci.* **1995**, *4*, 103–112.

(21) Holland, T. A.; Veretnik, S.; Shindyalov, I. N.; Bourne, P. E. *J. Mol. Biol.* **2006**, *361*, 562–590.

(22) Yesylevskyy, S. O. *Biopolym. Cell* **2010**, *26*, 146–151.

(23) Koike, R.; Ota, M.; Kidera, A. *J. Mol. Biol.* **2014**, *426*, 752–762.

(24) Gerstein, M.; Lesk, A. M.; Chothia, C. *Biochemistry* **1994**, *33*, 6739–6749.

(25) Guo, J. T.; Xu, D.; Kim, D.; Xu, Y. *Nucleic Acids Res.* **2003**, *31*, 944–952.

(26) Wieninger, S. A.; Serpersu, E. H.; Ullmann, G. M. *J. Mol. Biol.* **2011**, *409*, 450–465.

(27) Wriggers, W.; Schulten, K. *Proteins* **1997**, *29*, 1–14.

(28) Hayward, S.; Berendsen, H. J. *Proteins* **1998**, *30*, 144–154.

(29) Bernhard, S.; Noé, F. *PLoS One* **2010**, *5*, e10491.

(30) Romanowska, J.; Nowinski, K. S.; Trylska, J. *J. Chem. Theory Comput.* **2012**, *8*, 2588–2599.

(31) Hayward, S.; Kitao, A.; Berendsen, H. J. *Proteins* **1997**, *27*, 425–437.

(32) Bahar, I.; Atilgan, A. R.; Erman, B. *Folding Des.* **1997**, *2*, 173–181.

(33) Keskin, O.; Durell, S. R.; Bahar, I.; Jernigan, R. L.; Covell, D. G. *Biophys. J.* **2002**, *83*, 663–680.

(34) Kundu, S.; Sorensen, D. C.; Phillips, G. N. *Proteins* **2004**, *57*, 725–733.

(35) Sistla, R. K.; Brinda, K. V.; Vishveshwara, S. *Proteins* **2005**, *59*, 616–626.

(36) Hinsen, K.; Thomas, A.; Field, M. J. *Proteins* **1999**, *34*, 369–382.

(37) Yesylevskyy, S. O.; Kharkyanen, V. N.; Demchenko, A. P. *Biophys. J.* **2006**, *91*, 670–685.

(38) Martins, B. M.; Blaser, M.; Feliks, M.; Ullmann, G. M.; Buckel, W.; Selmer, T. *J. Am. Chem. Soc.* **2011**, *133*, 14666–14674.

(39) Seiffert, G. B.; Ullmann, G. M.; Messerschmidt, A.; Schink, B.; Kroneck, P. M.; Einsle, O. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 3073–3077.

(40) Kundu, S.; Melton, J. S.; Sorensen, D. C.; Phillips, G. N. *Biophys. J.* **2002**, *83*, 723–732.

(41) Wessel, P.; Smith, W. H. F. *EOS, Trans. Am. Geophys. Union* **1998**, *79*, 579.

(42) DeLano, W. L. DeLano Scientific LLC Palo Alto: CA, USA, 2008. http://www.pymol.org.

(43) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536–540.

(44) Sillitoe, I.; Cuff, A. L.; Dessailly, B. H.; Dawson, N. L.; Furnham, N.; Lee, D.; Lees, J. G.; Lewis, T. E.; Studer, R. A.; Rentzsch, R.; Yeats, C.; Thornton, J. M.; Orengo, C. A. *Nucleic Acids Res.* **2013**, *41*, D490–D498.

(45) Veretnik, S.; Bourne, P. E.; Alexandrov, N. N.; Shindyalov, I. N. *J. Mol. Biol.* **2004**, *339*, 647–678.

(46) Zhang, H.; Seabra, M. C.; Deisenhofer, J. *Structure* **2000**, *8*, 241–251.

(47) Brown, K.; Tegoni, M.; Prudêncio, M.; Pereira, A. S.; Besson, S.; Moura, J. J.; Moura, I.; Cambillau, C. *Nat. Struct. Biol.* **2000**, *7*, 191–195.

(48) Hänzelmann, P.; Dobbek, H.; Gremer, L.; Huber, R.; Meyer, O. *J. Mol. Biol.* **2000**, *301*, 1221–1235.

(49) Ermler, U.; Siddiqui, R. A.; Cramm, R.; Friedrich, B. *EMBO J.* **1995**, *14*, 6067–6077.

(50) Kondrashov, D. A.; Cui, Q.; Phillips, G. N. *Biophys. J.* **2006**, *91*, 2760–2767.

(51) Grochulski, P.; Masson, L.; Borisova, S.; Pusztai-Carey, M.; Schwartz, J. L.; Brousseau, R.; Cygler, M. *J. Mol. Biol.* **1995**, *254*, 447–464.

(52) Knöfel, T.; Sträter, N. *Nat. Struct. Biol.* **1999**, *6*, 448–453.

(53) Schink, B. *Arch. Microbiol.* **1985**, *142*, 295−301.

(54) Kisker, C.; Schindelin, H.; Rees, D. C. *Annu. Rev. Biochem.* **1997**, *66*, 233−267.

(55) Dobbek, H.; Huber, R. *Met. Ions Biol. Syst.* **2002**, *39*, 227−263.

(56) D'Ari, L.; Barker, H. A. *Arch. Microbiol.* **1985**, *143*, 311−312.

(57) Andrei, P. I.; Pierik, A. J.; Zauner, S.; Andrei-Selmer, L. C.; Selmer, T. *Eur. J. Biochem.* **2004**, *271*, 2225−2230.

(58) Yu, L.; Blaser, M.; Andrei, P. I.; Pierik, A. J.; Selmer, T. *Biochemistry* **2006**, *45*, 9584−9592.

(59) Lu, G.; Lindqvist, Y.; Schneider, G.; Dwivedi, U.; Campbell, W. *J. Mol. Biol.* **1995**, *248*, 931−948.

(60) Eyal, E.; Yang, L. W.; Bahar, I. *Bioinformatics* **2006**, *22*, 2619−2627.

(61) Zheng, W.; Brooks, B. R.; Thirumalai, D. *Biophys. J.* **2007**, *93*, 2289−2299.

(62) Yang, Q.; Sharp, K. A. *Proteins* **2009**, *74*, 682−700.

(63) Song, G.; Jernigan, R. L. *J. Mol. Biol.* **2007**, *369*, 880−893.

(64) Hinsen, K. *Bioinformatics* **2008**, *24*, 521−528.

(65) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534−552.

(66) Halle, B. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 1274−1279.

(67) Hinsen, K. *Proteins* **1998**, *33*, 417−429.

(68) Sainz, G.; Carrell, C. J.; Ponamarev, M. V.; Soriano, G. M.; Cramer, W. A.; Smith, J. L. *Biochemistry* **2000**, *39*, 9164−9173.

(69) Jacobson, R. H.; Ladurner, A. G.; King, D. S.; Tjian, R. *Science* **2000**, *288*, 1422−1425.

(70) Silva, G. H.; Dalgaard, J. Z.; Belfort, M.; Van Roey, P. *J. Mol. Biol.* **1999**, *286*, 1123−1136.

(71) Polekhina, G.; Thirup, S.; Kjeldgaard, M.; Nissen, P.; Lippmann, C.; Nyborg, J. *Structure* **1996**, *4*, 1141−1151.

(72) Peat, T. S.; Newman, J.; Waldo, G. S.; Berendzen, J.; Terwilliger, T. C. *Structure* **1998**, *6*, 1207−1214.

(73) Keitel, T.; Simon, O.; Borriss, R.; Heinemann, U. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 5287−5291.