Hierarchical description and extensive classification of protein structural changes by Motion Tree

Ryotaro Koike (1), Motonori Ota (1), Akinori Kidera (2)

(1) Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

(2) Graduate School of Medical Life Science, Yokohama City University, Tsurumi, Yokohama 230-0045, Japan

Short title: Motion Tree for protein structural changes

Corresponding author: Ryotaro Koike

E-mail address: rkoike@is.nagoya-u.ac.jp

Tel: +81-52-789-4265

Fax: +81-52-789-4265

Postal address: Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Abstract

The crystal structures of the same protein, determined under different conditions, provide clues toward understanding the role of structural changes in the protein's function. Structural changes are usually identified as rigid body motions, which are defined using a particular threshold of rigidity, such as domain motions. However, each protein actually undergoes motions with various size and magnitude ranges. In this study, to describe protein structural changes more comprehensively, we propose a method based on hierarchical clustering. This method enables the illustration of a wide range of protein motions in a single tree diagram, named the "Motion Tree". We applied the method to 432 proteins exhibiting large structural changes, and classified their Motion Trees in terms of the characteristic indices of the trees. This classification of the Motion Trees revealed clear relationships to their protein functions. Especially, complex structural changes are significantly correlated with multi-step protein functions.

Keywords

Protein structural change; loop motion; domain motion; protein contact map; protein function.

Abbreviations

Protein Data Bank, PDB; adenylate kinase, ADK; Tree Depth, TD; Size Index, SI; catabolite activator protein, CAP; root mean square deviation, RMSD; Structural Classification of Proteins, SCOP.

Introduction

Proteins are flexible molecules. Their structures often change due to external stimuli, such as ligand binding and post-translational modification, to regulate their molecular functions ¹⁻⁵. Proteins do not necessarily function as a single structure, but utilize multiple structures before and after exposure to the external stimulus ⁶. Structural changes frequently play important roles in protein functions, such as in the regulation of protein expression ⁷, enzymatic activities ⁸ and metabolite transport ⁹. In these cases, the various functions involve different types of structural changes. In terms of the enzymatic activity, distinct structural changes were observed depending on the type of enzymatic reaction. Transferases tend to undergo large structural changes when binding ligands within their deep pockets, while hydrolases bind ligands mostly on their protein surfaces and scarcely change their structures ¹⁰. These observations highlighted the importance of a comprehensive description of protein structural changes for the purpose of understanding their functions.

In our analysis of protein structural changes, we compared two distinct structures of identical proteins determined under different conditions or functional states, deposited in the Protein Data Bank (PDB) ¹¹. The structural changes in the two structures are usually identified by searching for similar substructures behaving as a unit of rigid body motion ¹²⁻¹⁸, and their displacement represents the magnitude of the structural change. In the definition of rigid bodies, we encounter the problem of variations of size and magnitude in the rigid bodies, that is, the problem of the selection of the threshold to define the rigid body. The structural changes involve not only large rigid bodies exhibiting large displacements characterized as domain motions, but also small rigid bodies found in local motions, such as loop motions or shifts of secondary

structure elements. Many proteins undergo rigid body motions of various sizes and magnitudes within the molecule. As an example, in the next section we describe how adenylate kinase (ADK) exhibits large domain motions together with local motions of α -helices within a domain motion (see Fig. 2). Since each element of the structural changes has its own characteristic role in the function of ADK, all of the details of the structural changes, from small scale to large scale, must be described. These motions on different scales cannot be simply viewed as rigid body motions defined by a single threshold value, since a large threshold value may fail to detect local motion, while a small value will divide the motions into a number of small components. In other words, focusing on motions of a particular size and magnitude is not sufficient; instead, the various sizes and magnitudes of motions must be considered.

The comprehensive description of various motions occurring in a protein requires the systematic arrangement of every size of the rigid bodies, so that the sizes, magnitudes, and relationships of the motions may be fully illustrated in a single picture. One of the most straightforward solutions to this requirement is the hierarchical clustering approach resulting in a dendrogram, or a tree diagram, comprising various sizes of rigid bodies¹⁹. The width and the height of the tree correspond to the size and the magnitude of the motions, respectively. A larger/smaller rigid body is represented by a larger/smaller subtree, and a larger/smaller motion is depicted by a node placed near the root/leaves of the tree. The hierarchical description using the tree diagram also enables the description of the relationship between the motions allocated at each node. In this hierarchical clustering, a cluster at a node is placed as part of a larger cluster at its ancestral node, without any division or reorganization of the cluster. In other words, the motion at a node is interpreted as the internal motion occurring within a larger rigid

body at its ancestral node. In this study, we emphasize the importance of the “multi-scale” description of protein motions, rather than the strict definition of rigid bodies on a “single-scale”, which would be achieved by non-hierarchical clustering approaches using a single threshold value.

Hierarchical clustering is a method to analyze a matrix describing similarity/dissimilarity between target objects. Here, we use the distance difference matrix as a dissimilarity matrix; i.e., the difference between two distance matrices²⁰ describing two protein structures. The distance difference between two rigid bodies or between two clusters, defined in the hierarchical clustering, represents the magnitude of rigid body motion in the form of tree height. To describe protein structural changes optimally in a hierarchical manner based on the distance difference, we propose a method to generate such a tree diagram, called “Motion Tree”.

As the first application of the method, we attempted to obtain an overview of protein structural changes, based on a set of diverse proteins with multiple, distinct crystal structures in the PDB. After describing the structural changes in Motion Trees, we classified these trees using newly introduced indices, “Size Index” and “Tree Depth”, which capture the characteristics of Motion Trees. The results of the classification revealed the correlation between protein functions and structural changes.

Algorithm

Constructing Motion Tree

The Motion Tree is calculated by the hierarchical clustering method based on the distance difference matrix for Ca atoms, whose element is $e_{ij} = |d_{ij}^a - d_{ij}^b|$, where d_{ij}^a is the Euclidean distance between the Ca atoms of residues i and j in structure a ,

compared with d_{ij}^b in structure b of the identical protein. The clustering procedure is summarized in Figure 1. At the first step, each residue of the protein is assumed to constitute each cluster (A in Fig. 1). According to the dissimilarity measure calculated from e_{ij} 's (B, see the next paragraph), the most similar pair of clusters is selected (C). They are merged to generate a new cluster (F), if the spatial-proximity condition (see the paragraph after the next) is satisfied (D). Otherwise, the next most similar clusters are considered (E). For the new cluster, the dissimilarity measures against the other clusters are updated (B). This process (B to F) is repeated (G) to merge clusters progressively, and to finally attain a single cluster containing all residues (H). The result of the clustering is illustrated in a tree diagram.

To describe protein structural changes using the hierarchical clustering based on distance differences (e_{ij} 's), an optimal dissimilarity measure between clusters has to be defined. Since the dissimilarity measure implies the magnitude of motions between clusters, it may be most reasonable to use the maximum of the distance differences between the two clusters C_1 and C_2 , $\max\{e_{ij} : i \in C_1, j \in C_2\}$, as the dissimilarity measure in the complete-linkage clustering. However, we noticed that a single value of the maximum distance difference frequently overemphasized the motions of tiny portions, instead of characterizing the behaviors of the entire rigid body (for an example, see Supplementary Text). To alleviate this problem, we calculated the average of the 20 largest distance differences as the dissimilarity measure. Twenty was chosen according to a test calculation using a selected set of structural changes (see Supplementary Text). When the numbers of residues in the clusters ($|C_1|$ and $|C_2|$) are small, and the number of e_{ij} 's between clusters ($|C_1| \times |C_2|$) is less than 20, all e_{ij} 's are averaged.

In the procedure of merging two clusters, we introduce the spatial-proximity condition to ensure that a new cluster is a spatially connected unit in the protein structure: Grouping two clusters is accepted if the nearest Ca distances between the two clusters, measured respectively in structures *a* and *b* both, are less than 7 Å. The value, 7 Å, was chosen according to the definition of the inter-residue contact^{21,22}.

Interpretation of Motion Tree

A Motion Tree is constructed by clustering all residues (leaves of the tree) successively until they are joined into one cluster at the root. Conversely, the Motion Tree is interpreted by going down from the root to the leaves.

The Motion Tree of adenylate kinase (ADK) is shown in Fig. 2, as an example. ADK is an enzyme that transfers a phosphoryl group from ATP to AMP, and has two lids to bind ATP and AMP, respectively²³. The remainder of the protein is called the core domain. In the substrate-free state, both lids are in the open conformation²³. On the other hand in the substrate-bound state, ADP and AMP are at the binding sites for ATP and AMP, respectively, and the lids are in the closed conformation²⁴. The Motion Tree of the structural change was calculated for these two states of ADK^{23,24}. The root denotes the cluster composed of all residues in the protein. The cluster is then divided into two sub-clusters; i.e., two rigid bodies corresponding to the ATP lid and the rest of the protein, respectively, at the first node (node A in Fig. 2). The height from the bottom is the largest, indicating that the relative motion of the ATP lid to the rest of the protein is the largest in the protein. The largest magnitude also means the largest threshold for the definition of a rigid body; that is, the rigid bodies contain internal motions occurring among smaller rigid bodies at the descendant nodes. At the second node (node B), the

rest of the protein was further divided into the AMP lid and the core domain, to represent the relative motion of the AMP lid to the core domain. The height of node B, which is lower than node A, implies that the magnitude of motion of the AMP lid is smaller than that of the ATP lid. The third node (node C) represents a much smaller internal motion within the AMP lid, where a helix moves against two other helices. The fourth node is also in the AMP lid, the fifth and sixth nodes are in the core domain, and the seventh node is in the ATP lid. These observations in the Motion Tree suggest that the AMP lid is more flexible than the ATP lid or the core domain. In fact, the Ca RMSD (root mean square deviation) values within the AMP lid, the ATP lid, and the core domain between the two structures are 2.50, 0.98 and 1.17Å, respectively.

The process of the division of rigid bodies is repeated $N-1$ times (N is the number of residues), and the entire protein is finally divided into N residues that are the smallest rigid body with a magnitude of motion defined as zero.

As shown in the above example, Motion Tree successfully illustrated the complicated protein structural changes in detail, in a single picture. In the following, as the first application of Motion Tree, we will provide an overview of the protein structural changes based on the classification of Motion Trees for a set of diverse proteins in the PDB. For this purpose, we introduced two indices, “Tree Depth” and “Size Index”, together with the definition of an “effective node” for these indices.

Effective node. For the analysis of diverse proteins, we focus only on large rigid body motions, in terms of both size and magnitude, by defining effective nodes. At an effective node, a cluster consisting of more than 30 residues is divided into two sub-clusters, each containing more than five residues. The magnitude of the effective node is more than 5.0 Å. In the Motion Tree of ADK, the effective nodes are nodes A, B

and C.

Tree Depth. We occasionally observe complex structural changes, where a number of distinct portions of a protein move individually in different directions. To measure such complexity, we introduce Tree Depth (TD), to extend the conventional notion of the depth of a tree²⁵. The value of TD is the maximum number of ancestral nodes counted at every effective node in the Motion Tree. The TD value of ADK is two, defined at node C.

Size Index. The distribution of the sizes of rigid bodies is another characteristic of Motion Tree. We label a rigid body as “L” when it contains more than 30 residues, otherwise it is “S”. Accordingly, the relative motion between the two rigid bodies defined at an effective node is categorized into “L-L”, “L-S” or “S-S”. A motion labeled as “L-L” may represent a domain motion, and “L-S” infers a local motion. The latter means that the motion of a small, localized portion is recognized as part of a larger rigid body. The other type of motion, “S-S”, is rarely observed, but often indicates motion between secondary structure elements (see details in Supplementary Text). The Size Index (SI) is defined by a three-dimensional vector, representing the presence (1) or absence (0) of the three types of motions, “L-L”, “L-S” or “S-S”. In the case of ADK, SI is (1, 1, 0); two L-L motions at nodes A and B, respectively, one L-S motion at node C, and no S-S motions.

Results

Classification of protein structural changes

Motion Trees were calculated for a non-redundant dataset composed of 432 proteins, exhibiting large structural changes of more than 2 Å Ca RMSD in the PDB (see Dataset

section for details). All Motion Trees thus calculated are summarized in a Web site (<http://idp1.force.cs.is.nagoya-u.ac.jp/rk1/mtntntr/classification/>). Except for 8 proteins only containing the motions of small portions of five residues or less, and lacking an effective node, 424 proteins contained at least one effective node, and their Tree Depth and Size Index values were calculated.

Based on the Size Index values, most of the structural changes, observed in 401 of the 424 proteins, were categorized into three major groups (inner circle in Fig. 3), (0,1,0) containing only L-S or local motions, (1,0,0) only L-L or domain motions, and (1,1,0) both L-S and L-L (referred to as mixed motions). Among the Tree Depth (TD) values, we found that about half of the protein structural changes (197/424) were characterized by simple motions with TD values of zero, containing only one effective node (tinted areas of the outer circle in Fig. 3). The domain motions are especially simple: about 90 % of them (104/117) had TD values of zero. Motion Trees with TD values of at least two (deeply colored areas of outer circle) accounted for 17 % of the total (71/424), and more than half of them (40/71) exhibited mixed motions.

Protein function and structural change

For 410 of the 424 proteins in Fig. 3, the protein functions were successfully assigned according to the UniProt “keyword” annotations²⁶. The correlation between the protein function and the type of structural change was evaluated by the keywords that were specific to a type of structural change based on the p -value (see Materials and Methods for details). Table 1 lists the three keywords with the lowest p -values (largest specificity) for each type of structural change (all p -values are listed in Supplementary Table 1). The specific keywords ($p < 0.05$) are “hydrolase” and “protease” in the local motion, “DNA-binding” and “repeat” in the domain motion, and “transferase”, “kinase”

and “ATP-binding” in the mixed motion. The correlations between these keywords and the types of structural changes can be explained as follows.

Previously, we reported that hydrolases undergo small structural changes upon ligand binding¹⁰. This fact indicates that the motions of hydrolases are subtle and localized, if any. This is also true for proteases, since they are members of the hydrolases. Therefore, it is reasonable to find the preferences of local motion in hydrolases and proteases.

Most of the “DNA-binding” proteins undergoing domain motions are multi-domain proteins frequently consisting of the “winged helix DNA-binding domain” (SCOP superfamily²⁷). The domain motions annotated with the keyword “repeat” are explained by the observation that the two domains causing the motions evolved by gene duplication.

Transferases bind at least two ligand molecules, to transfer a functional group from one to the other. We reported that transferases undergo large structural changes to insulate the ligand molecules from water¹⁰. To satisfy these functional requirements, forming the atomic configuration for the transfer reaction and shielding the ligand from the water environment, the transferase may undergo complicated structural changes and exhibit both local and domain motions. About half of the proteins exhibiting mixed motion with the annotation “ATP-binding” are kinases, which are members of the transferases, and about half of the rest are proteins involved in biomolecular synthesis. The ATP-binding proteins associated with biomolecular synthesis may also undergo complicated motions to catalyze complex, multi-step chemical reactions²⁸. The *p*-values of these specific keywords in the mixed motion are remarkably lower than those of the keywords in the local and domain motions (Table 1). This indicates a strong

correlation between the mixed motion and the protein functions described by these keywords.

In the following, we will show several representative examples of Motion Trees for local motion, domain motion and mixed motion, to relate the structural changes of the proteins to their functions.

Local motion (L-S)

In Fig. 4a, the Motion Tree of human cytosolic sialidase Neu2, a glycoside hydrolase with the six-bladed b-propeller fold, is shown as an example of a local motion. The tree was calculated for the inhibitor-free (chain A in PDB entry 1so7²⁹) and inhibitor-bound (chain A in 1vcu²⁹) forms. The tree has only one effective node (node A in Fig. 4a), representing the motion of the long loop (residues 109-119) containing Glu111 responsible for the substrate binding; actually, Glu111 binds the inhibitor, DANA, in the bound form²⁹. The second and third nodes (nodes B and C) produce rigid bodies consisting of two and four residues, respectively, and thus are not classified as effective nodes. The second node indicates a disorder (the free form) to order (the bound form) transition of the loop containing the catalytic residue, Asp46²⁹. The greater part of this loop (residues 42-48 in the loop 42-50) is missing in the free form, and thus the magnitude of the motion in the second node is underestimated in the tree.

All of these motions were found on the catalytic side of the b-propeller (Fig. 4b). This means that the two sides of the b-propeller have completely different dynamic properties. As reported previously²⁹, the two sides have quite different amino acid compositions: the catalytic side is negatively charged, while the other side is hydrophobic.

Domain motion (L-L)

Figure 5a shows the Motion Trees of catabolite activator protein (CAP), a transcriptional regulator for catabolite genes, as an example of a domain motion. CAP forms a homo-dimer, and binds to specific DNA sites in the presence of cAMP³⁰. To investigate the structural changes in CAP upon DNA binding, we examined the CAP dimers, one binding only cAMP molecules and the other binding both cAMP and DNA. In the DNA-free state, a protomer binds anti-cAMP at the N-terminal domain and syn-cAMP at the interface between the N-terminal and C-terminal domains. The other protomer binds only anti-cAMP at the N-terminal domain (PDB entry 1i6x, see the upper panel of Figure 5b). A significant structural change occurs upon syn-cAMP binding (see the discussion in Section 3 of Supplementary Text). In the DNA-bound state, each protomer binds anti-cAMP at the N-terminal domain (1zrd³¹, the lower panel of Figure 5b), and their structures are almost identical (Ca RMSD = 0.8 Å). In Figure 5a, the upper panel shows the Motion Tree for the DNA-bound and DNA-free protomers with only anti-cAMP (B chains colored red and blue in Figure 5b), and the lower panel is also the tree for the DNA-bound and DNA-free protomers, but with the DNA-free protomer binding two cAMPs (A chains colored gray in Figure 5b). The upper tree has one effective node (node A), representing a domain motion between two rigid bodies. The smaller rigid body (residues 136-207, colored red in Figure 5b) corresponds to the C-terminal domain classified as the “winged helix DNA-binding domain”. DNA binding by proteins with the winged helix DNA-binding domain is frequently accompanied by domain motions (Table 1). The second node (B) represents the motion of a small loop (residues 54-57) located at the interface of the two domains. On the other hand, the lower tree does not show the domain motion, but the motion of

the small loop identified by node B in the upper tree is also found at the first node (A'). These trees, together with the tree in Supplementary Text, indicate that DNA binding and syn-cAMP binding cause almost the same domain motion. However, there is a significant difference in the motion of the small loop, indicating that the small loop has three distinctive structures, the cAMP-free form, the cAMP-bound form, and the DNA-bound form. This loop is referred to as “activation region 3”, related to transcription activation, and binds domain 4 of the RNA polymerase σ^{70} subunit³².

Mixed motion (L-S and L-L)

In Fig. 6a, the Motion Tree of Src kinase (Csk), a signaling protein related to cell growth, is shown as an example of mixed motion. Csk has three domains, the SH3, SH2 and kinase domains, and the kinase domain is further divided into two sub-domains, N-lobe and C-lobe (see Fig. 6b). The residues crucial for substrate binding and catalytic activity are located at the boundary between these two lobes³³. The rotation of the SH2 domain is the most characteristic motion observed between the catalytically active and inactive states³³. The Motion Tree was calculated according to the structural change from the inactive form (chain C in PDB entry 1k9a³³) to the active form (chain A in 1k9a) (Fig. 6a), which were obtained as multiple configurations in an asymmetric unit³³. The tree contains four effective nodes, representing three domain motions and a local motion. The first node (node A in Fig. 6a) indicates the largest motion of the SH2 domain, which is related to the transition to the active state. At the second node (node B), the N-lobe of the kinase domain and the SH3 domain comprise a rigid body, whose motion occurs against the C-lobe to form the catalytic site. The third node (node C) exhibits motion of the loop (residues 201-207) containing the ATP binding site. At the fourth node (node D), the SH3-SH2 and SH2-kinase linkers form a rigid body together

with the SH3 domain, and move relatively to the b-sheet in the N-lobe. This motion induces the formation of the salt-bridge between Glu236 and Lys 222, which connects the SH2-kinase linker with the N-lobe b-sheet, and contributes to the recognition of the ATP molecule ³³.

Discussion

The hierarchical clustering approach was previously utilized in other studies for the description of protein structural changes. Daily and Gray used the clustering method to define rigid bodies and their relationships ¹³. Boutonnet *et al.* presented hierarchical pictures of protein structural changes, based on their multiple linkage clustering method ³⁴. However, their clustering methods are fundamentally different from the one in the present study, specifically in terms of the use of the RMSD of superimposed atoms for the dissimilarity measure in clustering, instead of the distance differences. When the RMSD is employed in the hierarchical clustering, the RMSD value is calculated at each node by each different superimposition, depending on the number and the configuration of the atoms selected at the node. In consequence, the displacement between a pair of corresponding atoms, a component in the RMSD value, differs from one node to another and is not constant throughout the tree. Hence, clustering based on such variable dissimilarity measures may complicate the interpretation of the tree structure, thus precluding an understanding of the protein structural changes. On the other hand, the distance difference is a standard definition of the dissimilarity measure in hierarchical clustering. It is constant throughout all nodes and allows us to interpret the tree structure on the basis of the constant dissimilarity measure.

In this study, we have presented a method to obtain a hierarchical description of

protein structural changes in the form of Motion Trees. As shown in the above examples, a single picture of a Motion Tree successfully exhibits all of the details of the functional motions. Extensions of the method to the problems of quaternary motions in protein complexes, as well as comparisons and classifications of numerous structures generated by molecular dynamics simulations, are in progress.

Materials and Methods

Dataset

We selected proteins with crystal structures determined at resolutions better than 3.0 Å (102,101 chains) from the PDB¹¹. Small (< 50 amino acids) and large (> 500) proteins were excluded from the analysis (93,729 chains remained). Proteins were classified as identical when the sequence identity was more than 95%. The proteins were grouped according to the SCOP concise classification strings at the family level (sccs;²⁷). The pair of proteins with the largest Ca RMSD value was chosen as the representative pair from each of the groups (2,259 protein pairs). From this non-redundant set, 432 proteins with Ca RMSD values larger than 2.0 Å were subjected to the Motion Tree analysis.

Estimation of p -value based on the binomial distribution

Proteins annotated by the keyword, “3D structure” in UniProt²⁶ were prepared as the population set (total number: M). The frequency of proteins, m_i , annotated by the keyword i was counted to estimate the probability p_i as m_i/M . When we focused on a sampling set of N proteins that exhibit structural changes (local, domain or mixed motion), the p -value to find n_i proteins annotated by keyword i is calculated by $P(x \geq n_i)$

with

$$P(x = n_i) = \frac{N!}{n_i! (N - n_i)!} p_i^{n_i} (1 - p_i)^{(N - n_i)} \quad (1).$$

Among the 424 proteins used in the analysis, 410 were annotated by certain keywords (local (171), domain (112) and mixed motions (104)). The 20 keywords that were most frequently found were examined. Their p -values are shown in Supplementary Table 1.

Acknowledgements

This work was funded by the Japan Science and Technology Agency, Institute for Bioinformatics Research and Development (JST-BIRD), and MEXT SPIRE Supercomputational Life Science.

References

1. Amemiya, T., Koike, R., Fuchigami, S., Ikeguchi, M. & Kidera, A. (2011). Classification and annotation of the relationship between protein structural change and ligand binding. *J Mol Biol* **408**, 568-84.
2. Amemiya, T., Koike, R., Kidera, A. & Ota, M. (2012). PSCDB: a database for protein structural change upon ligand binding. *Nucleic Acids Res* **40**, D554-8.
3. Hayward, S. (2004). Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. *J Mol Biol* **339**, 1001-21.
4. Qi, G. & Hayward, S. (2009). Database of ligand-induced domain movements in enzymes. *BMC Struct Biol* **9**, 13.
5. Xin, F. & Radivojac, P. (2012). Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics* **28**, 2905-13.
6. Fuchigami, S., Fujisaki, H., Matsunaga, Y. & Kidera, A. (2011). Protein Functional Motions: Basic Concepts and Computational Methodologies. *Advancing Theory for Kinetics and Dynamics of Complex, Many-Dimensional Systems: Clusters and Proteins: Advances in Chemical Physics, Vol 145* **145**, 35-82.
7. Kercher, M. A., Lu, P. & Lewis, M. (1997). Lac repressor-operator complex. *Curr Opin Struct Biol* **7**, 76-85.
8. Hammes, G. G. (2002). Multiple conformational changes in enzyme catalysis. *Biochemistry* **41**, 8221-8.
9. Anderson, B. F., Baker, H. M., Norris, G. E., Rumball, S. V. & Baker, E. N. (1990). Apolactoferrin structure demonstrates ligand-induced conformational change in transferrins. *Nature* **344**, 784-7.
10. Koike, R., Amemiya, T., Ota, M. & Kidera, A. (2008). Protein structural change upon ligand binding correlates with enzymatic reaction mechanism. *J Mol Biol* **379**, 397-401.
11. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
12. Abyzov, A., Bjornson, R., Felipe, M. & Gerstein, M. (2010). RigidFinder: a fast and sensitive method to detect rigid blocks in large macromolecular complexes. *Proteins* **78**, 309-24.
13. Daily, M. D. & Gray, J. J. (2009). Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput Biol* **5**, e1000293.

14. Hayward, S. & Berendsen, H. J. (1998). Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins* **30**, 144-54.
15. Hayward, S., Kitao, A. & Berendsen, H. J. (1997). Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins* **27**, 425-37.
16. Hinsen, K., Thomas, A. & Field, M. J. (1999). Analysis of domain motions in large proteins. *Proteins* **34**, 369-82.
17. Nichols, W. L., Rose, G. D., Ten Eyck, L. F. & Zimm, B. H. (1995). Rigid domains in proteins: an algorithmic approach to their identification. *Proteins* **23**, 38-48.
18. Wriggers, W. & Schulten, K. (1997). Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins* **29**, 1-14.
19. Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster Analysis, 5th Ed.* Wiley Series in Probability and Statistics, Wiley.
20. Nishikawa, K., Ooi, T., Isogai, Y. & Saito, N. (1972). Tertiary Structure of Proteins. I. Representation and Computation of the Conformations. *Journal of the Physical Society of Japan* **32**, 1331-1337.
21. Tudos, E., Fiser, A. & Simon, I. (1994). Different sequence environments of amino acid residues involved and not involved in long-range interactions in proteins. *Int J Pept Protein Res* **43**, 205-8.
22. Yuan, C., Chen, H. & Kihara, D. (2012). Effective inter-residue contact definitions for accurate protein fold recognition. *BMC Bioinformatics* **13**, 292.
23. Muller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4**, 147-56.
24. Berry, M. B., Bae, E., Bilderback, T. R., Glaser, M. & Phillips, G. N., Jr. (2006). Crystal structure of ADP/AMP complex of Escherichia coli adenylate kinase. *Proteins* **62**, 555-6.
25. Aho, A. V., Ullman, J. D. & Hopcroft, J. E. (1983). *Data Structures and Algorithms*, Addison-Wesley.
26. UniProt, C. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* **41**, D43-7.
27. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40.

28. Delarue, M. (1995). Aminoacyl-tRNA synthetases. *Curr Opin Struct Biol* **5**, 48-55.
29. Chavas, L. M., Tringali, C., Fusi, P., Venerando, B., Tettamanti, G., Kato, R., Monti, E. & Wakatsuki, S. (2005). Crystal structure of the human cytosolic sialidase Neu2. Evidence for the dynamic nature of substrate recognition. *J Biol Chem* **280**, 469-75.
30. Busby, S. & Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). *J Mol Biol* **293**, 199-213.
31. Napoli, A. A., Lawson, C. L., Ebright, R. H. & Berman, H. M. (2006). Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: Recognition of pyrimidine-purine and purine-purine steps. *Journal of Molecular Biology* **357**, 173-183.
32. Rhodius, V. A. & Busby, S. J. (2000). Interactions between activating region 3 of the Escherichia coli cyclic AMP receptor protein and region 4 of the RNA polymerase sigma(70) subunit: application of suppression genetics. *J Mol Biol* **299**, 311-24.
33. Ogawa, A., Takayama, Y., Sakai, H., Chong, K. T., Takeuchi, S., Nakagawa, A., Nada, S., Okada, M. & Tsukihara, T. (2002). Structure of the carboxyl-terminal Src kinase, Csk. *J Biol Chem* **277**, 14351-4.
34. Boutonnet, N. S., Rooman, M. J. & Wodak, S. J. (1995). Automatic analysis of protein conformational changes by multiple linkage clustering. *J Mol Biol* **253**, 633-47.

Figure legends

Figure 1

Flowchart to construct Motion Tree. The resultant tree is formed from leaves (A, residues of the protein) to the root (H, the protein). The dissimilarity measure is calculated as the average of the 20 largest distance differences (B). The most similar pair of clusters is selected (C) to make a new cluster (F), but the pair has to satisfy the spatial-proximity condition (D). If not, the next most similar clusters are chosen (E), and evaluated whether they meet the condition, until the new cluster is obtained (F). This process (B-F) is repeated (G). For details, see the text.

Figure 2

The Motion Tree of adenylate kinase (ADK). The structures in the ligand-free form (chain A in PDB entry 4ake) and the ligand-bound form (the ligands are AMP and ADP; chain A in 2eck) were compared. The three effective nodes, A-C, are denoted by dots with labels on the tree. The A node divides the rigid body defined at the ancestral node further into two rigid bodies, and the smaller rigid body is shown in the upper part of the diagram (this depiction mode was applied to all Motion Trees). The smaller rigid bodies are residues 117-164 at the first node (A), 30-80 at the second node (B), and 60-80 at the third node (C). Note that rigid bodies are not necessarily composed of consecutive residues; instead, they generally consist of several fragments. The motion at each effective node is illustrated by the two structures superimposed on the Ca atoms of the larger rigid bodies (cyan: free-form; blue: bound-form), to highlight the shifts of the smaller rigid bodies (orange: free-form; red: bound-form). The magnitude axis indicates the dissimilarity measure of clusters. The bars on the left side of the tree indicate the sizes of the rigid bodies at nodes A-C, which are blue for the larger rigid body and red

for the smaller rigid body. All protein structures in this paper were rendered with PyMol.

Figure 3

Classification of the 424 Motion Trees based on Tree Depth and Size Index. The inner circle shows the distribution of the three major groups, local (cyan), domain (green), and mixed motions (orange), and others (gray). The outer circle shows the categories of Motion Trees in each group, according to the Tree Depth. A deeper color represents a larger Tree Depth. The number of Motion Trees classified in the category is provided in parentheses.

Figure 4

Local motions in human cytosolic sialidase Neu2. (a) Motion Tree for the free (chain A in PDB entry 1so7) and the inhibitor-bound forms (The inhibitor is DANA; chain A in 1vcu). The effective node A is denoted by a black dot. The illustrations of the motions are given not only for the effective node but also for the second and third nodes, B-C, which are not effective. All of the motions are localized on one side of the b-propeller structure; the smaller rigid bodies colored red or orange are residues 108-116 for the first node (A), 49-50 for the second node (B), and 15-18 for the third node (C), respectively; the larger rigid bodies are colored blue or cyan. The disordered region (residues 42-48) in the free form is colored green on the inhibitor-bound structure. (b) The structure of cytosolic sialidase Neu2 in the inhibitor-bound form (the inhibitor is not depicted). The smaller portions undergoing local motions at the first, second and third nodes (A-C) are colored red, violet, and orange, respectively. As stated above, the green loop is disordered in the inhibitor-free form.

Figure 5

Motions in the dimer of catabolite activator protein (CAP) upon DNA binding. (a) Upper: Motion Tree of the protomers between the DNA-free form (binding anti-cAMP; chain B in PDB entry 1i6x) and the DNA-bound form (binding anti-cAMP and DNA; chain B in 1zrd). The motions are illustrated for the effective node A and for the second node B. The effective node (A) represents the domain motion between the smaller rigid body (residues 136-207, colored red or orange) and the larger rigid body (the remaining residues, colored blue or cyan). The second node (B) represents the motion of the small portion (residues 54-57, colored red or orange). Lower: Motion Tree of the protomers between the DNA-free form (binding syn- and anti-cAMPs; chain A in 1i6x) and the DNA-bound form (binding anti-cAMP and DNA; chain A in 1zrd). This tree does not have any effective node. The motion of the small portion (residues 54-56) at the first node A', colored red or orange, is illustrated. (b) The structures of the DNA-free and DNA-bound forms of the CAP dimer. The CAP dimer, cAMPs and DNA duplex are depicted by cartoon, CPK and molecular surface representations, respectively. Protomers for the upper tree are colored blue (the N-terminal domain) and red (the C-terminal domain).

Figure 6

Mixed motions in Src kinase (Csk). (a) Motion Tree for the active form (chain A in PDB entry 1k9a) and the inactive form (chain C in 1k9a). This tree contains four effective nodes, A-D, and their motions are illustrated in the same manner as above. The first, second and fourth nodes (A, B and D) represent the domain motions, where the smaller rigid bodies are residues 80-174 at the first node, residues 6-79/175-241/251-271/323-324 at the second node and residues

187-200/208-228/251-271/323-324 at the fourth node. On the other hand, the motion at the third node (C) is localized, and its smaller rigid body includes residues 201-207. (b) The structure of Csk in the active form. There are four domains: the SH3 domain (blue), the SH2 domain (green), and the N-lobe (yellow) and the C-lobe (orange) of the kinase domain. The linkers between SH3 and SH2, and between SH2 and the kinase domain, are colored violet. The region responsible for ATP binding is colored red.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

(a)

(b)

Graphical Abstract