[Supplementary Data]

MDContactCom: A tool to identify differences of protein molecular dynamics from two MD simulation trajectories in terms of interresidue contacts

Chie Motono[1,2,*], Shunsuke Yanagida[3], Miwa Sato[3], Takatsugu Hirokawa[1,4,5,*]

[1] Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

[2] Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), AIST, Waseda University, Tokyo 169-0072, Japan

[3] Mitsui Knowledge Industry Co., Ltd, 2-5-1 Atago, Minato-ku, Tokyo, 105-6215, Japan

[4] Division of Biomedical Science, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan.

[5] Transborder Medical Research Center, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan.

*To whom correspondence should be addressed.

Appendix 1: Input and output of MDContactCom

To run the MDContactCom, users provide two MD trajectories in one of the following formats: multi-frame PDB format, Amber (Case *et al.*, 2021), CHARMM (Brooks *et al.*, 2009), Desmond (Schrödinger, 2021), GROMACS (Abraham *et al.*, 2015) or NAMD (Phillips *et al.*, 2020).

The default outputs consist of (i) a text file providing similarity indices (Tanimoto coefficient and Euclidean distance), (ii) a plot of data in (i), (iii) PDB files for displaying residues with large differences in similarity indices and their interresidue contacts on the protein structure.

Appendix 2: Depiction of algorithms of MDContactCom

When two MD trajectories A and B are inputted to MDContactCom, they are processed as follows:

(i) Contact frequency calculation. Interresidue contacts are detected for each structure frame in a MD trajectory (Fig. 1a). Two residues are in contact if the distance between any two heavy atoms of these residues is less than threshold (default distance 5.0 Å).

Contact frequency f_{ij} between residue i and residue j is calculated as

$$f_{ij} = n_{ij} / N,$$

where n_{ij} is the number of frames in which residues i and j were in contact and N is the total number of frames in a trajectory. Thus, $f_{ij} = 1$ for a contact present in all frames.

MDContactCom supports four definitions of residue-residue contact. Two residues are in contact if the distance between:

1 any two heavy atoms of these residues is less than threshold (default distance 5.0 Å),
2 any two atoms of these residues is less than threshold (default distance 4.5 Å),
3 any two Cα atoms of these residues is less than threshold (default distance 10.0 Å),
4 any two Cβ atoms of these residues is less than threshold (default distance 8.0 Å).

Users can specify freely the threshold of the atomic distance.

(ii) Comparison of contact frequencies between two trajectories. To compare two trajectories A and B, similarity coefficients S_{iAB} of residue i are calculated using the following equations:
for Tanimoto coefficient

$$S_{iAB} = \frac{\sum_{j=1}^{N} f_{ijA} f_{ijB}}{\sum_{j=1}^{N} (f_{ijA})^2 + \sum_{j=1}^{N} (f_{ijB})^2 - \sum_{j=1}^{N} f_{ijA} f_{ijB}}$$

for Euclidean distance,

$$S_{iAB} = \left[\sum_{j=1}^{N} (f_{ijA} - f_{ijB})^2 \right]^{1/2}$$

(iii) Visualization of residues with large differences in similarity indices. All residues are sorted by Euclidean distance S_{iAB} and selected (top 5, 10, and 20% of residues in default mode). PDB files are created to show the selected residues and their contacts on the protein 3D structure. In the PDB files, Ca atoms of the selected residues are added with a new residue name "PSD" in the ATOM records. Interresidue contacts among the selected residues are also added in the CONNECT records. Selected residues and contacts are highlighted by rendering them in a molecular visualization system. The selected residues and contacts can be displayed in PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC), Molecular Operating Environment (MOE) (Chemical Computing Group ULC), UCSF Chimera (Pettersen *et al.*, 2004), and Maestro (Schrödinger, LLC, New York, NY, 2020). The information is useful for identifying the first region of the protein to focus on after MD simulations are performed.

Appendix 3: An example of the application of MDContactCom to analyze MD trajectories of Cyclophilin A and its variant V29L

Allosteric regulation is a crucial feature of biochemical pathways. Ligand binding to an allosteric site or mutation of the allosteric site modulates the enzymatic activity at a distant functional site. Examples of allostery without conformational change have recently been reported, such as PDZ domain (Kumawat and Chakrabarty, 2017), CAP dimer (Tzeng and Kalodimos, 2009), and met repressor (Stacklies *et al.*, 2009). These examples have triggered the concept of dynamic allostery, where side-chain dynamics are modulated upon ligand binding (Tsai *et al.*, 2008; Liu and Nussinov, 2017).

Another example of dynamic allostery are mutations in Cyclophilin A (CypA, Figure

S1) (Holliday *et al.*, 2017). CypA is a ubiquitous protein belonging to the immunophilin family that has peptidyl prolyl cis-trans isomerase activity and regulates protein folding and trafficking.

To validate our comparison method, we performed 400 ns-long MD simulations of both wild-type and V29L variant of CypA and applied MDContactCom to compare the last 200 ns trajectories. The analysis revealed that residues with the highest similarity scores were mostly located in the ligand binding sites, which is over 15 Å from the mutated residue L29. We tracked the interresidue contacts of these residues and elucidated the path that propagates the mutation effect.

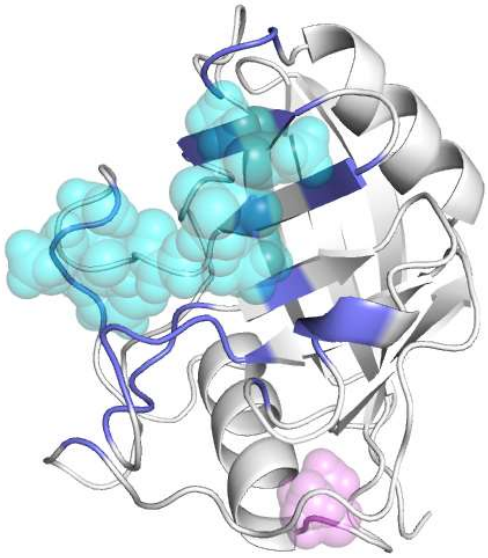

Figure S1 CypA (gray, PDB ID 1M9F) with its substrate peptide Ace-AlaAla-Pro-Phe-Nme (light blue) in the binding pocket (blue). The mutated residue Val29 is shown as a sphere (magenta).

1. Preparing the CypA Structures for MD.

The initial structures for CypA were based on the crystal structure of CypA complex with HIV-1 CA N-terminal domain (PDB ID 1M9F). For theV29L mutant structure, residue 29 was mutated from Val to Leu using Maestro interface of Schrodinger Release 2017-2 (Maestro, Schrödinger, LLC, New York, NY, 2018).

2. Executing MD simulations of CypA.

MD simulations were carried out using Desmond ver. 3.8 package with the OPLS3 force field. The initial model structure was solvated with TIP3P water molecules and 0.15 M NaCl. After minimization and relaxation of the model, a production run was performed for three independent 400 ns simulations in an isothermal-isobaric (NPT) ensemble at 300 K and 1 bar using Langevin dynamics. Trajectory coordinates were recorded every 40 ps. Procedures of MD simulation using Desmond are described elsewhere (Asamitsu *et al.*, 2017).

We used the last 200 ns trajectories of the MD simulations to compare the dynamics of CypA wild-type and mutant.

A MD trajectory should be converted to a file in multi-frame PDB format for input to MDContactCom. Unnecessary elements, such as water molecules and ions, are removed at this stage.

3. Comparing two MD trajectories using MDContactCom.

To run MDContactCom in default mode, the command is as follows:

```
> python mdcontactcom mdtrajectory1 mdtrajectory2
```

where mdtrajectory1 and mdtrajectory2 are input trajectories in multi-frame PDB format.

To Compare the trajectory of CypA mutant V29L (mdtrajectory_V29L) to that of wild-type (mdtrajectory_WT), the command is as follows:

```
> python mdcontactcom mdtrajectory_WT.pdb mdtrajectory_V29L.pdb -t 0 -p 32
```

The optional argument "-t 0" specifies an all-atom mode in which two residues are in contact if the distance between any atoms of these residues is less than threshold (default distance 4.5 Å). The optional argument "-p 32" specifies that MDContactCom processes interresidue contact detections in parallel using 32 CPUs.

4. Results of MDContactCom.

The output files are:

1) a csv format file that provides contact similarities in Tanimoto coefficient and in Euclidean distance for each residue,
2) an image file that plots the contact similarities in Tanimoto coefficient and in Euclidean distance against residue i,
3) PDB files to show the selected residues and their contacts on the protein 3D

structure.

Part of output file 1) is shown in Table S1.

Table S1 Contact similarities in Tanimoto coefficient and in Euclidean distance for each residue between CypA wild-type and mutant V29L.

residue_id	Tanimoto coefficient	Euclidean distance
A0	0.971	0.234
A1	0.725	1.114
A2	0.884	0.722
A3	0.822	1.184
A4	0.912	0.953
A5	0.983	0.423
A6	0.995	0.276
A7	0.999	0.116
A8	1.000	0.057
A9	1.000	0.025
A10	0.999	0.144
A11	0.997	0.200
A12	0.995	0.239
A13	0.978	0.360
A14	0.999	0.078
A15	0.994	0.185
A16	1.000	0.022
A17	0.989	0.345
A18	1.000	0.028
A19	1.000	0.038
A20	0.999	0.115

In file 2), the contact similarity of file 1) is plotted against residue numbers as shown in Figure S2. Both Tanimoto coefficient and Euclidean distance show distinctive peaks

at residues 63, 82-85, 101-109, and 127.

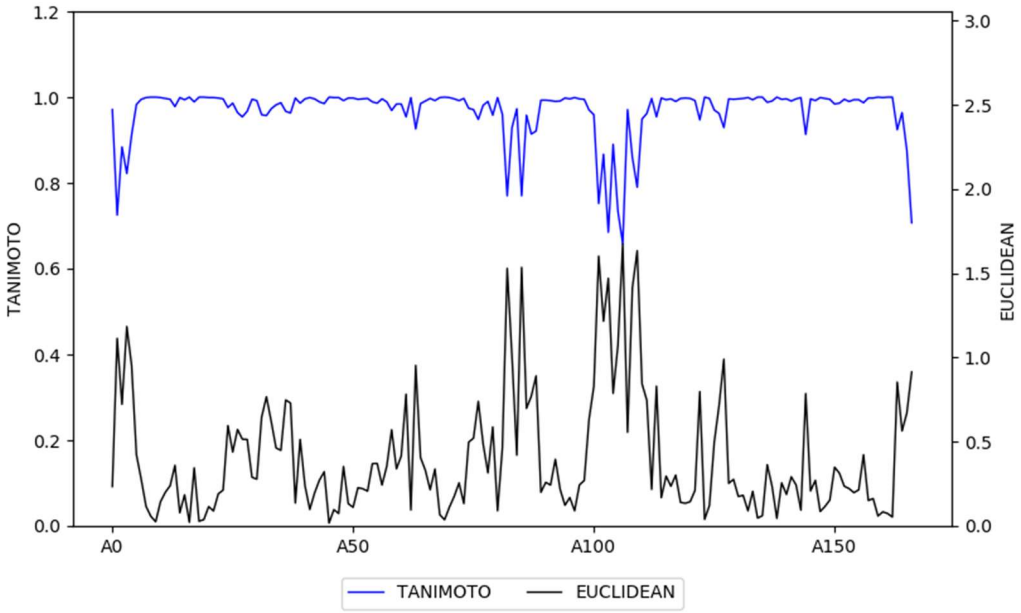

Figure S2 Plot of contact similarity score calculated from MD trajectories of CypA wild-type and variant V29L.

5. Displaying affected residues and their interresidue contacts in PyMOL.

To identify the most affected sites, all residues are sorted by Tanimoto coefficient S_{iAB} and the top 5, 10, 20, 50 and 100% selected (in default mode). PDB files are automatically created to show the selected sites and communication among them. The information is useful to detect the first region of the protein to focus on after MD simulations are performed.

In Figure S3, Residues ranked in the top 5, 10, 20, 50 and 100% of Tanimoto coefficient scores are shown. MDConactCom automatically creates PDB files to show those figures. Contact perturbations including the selected residues are also shown in cylinders. The contacts increased, decreased, or not changed by the mutation V29L are indicated in red, blue, or yellow cylinders, respectively. Top ranked contacts (top 5-20%, panel a-c in Figure S3) are roughly divided into two clusters: a right back side of the molecule and a left front side. We concluded that contact fluctuations in the former

region are due to unstable N-terminal and C-terminal residues and excluded those residues from further analysis.

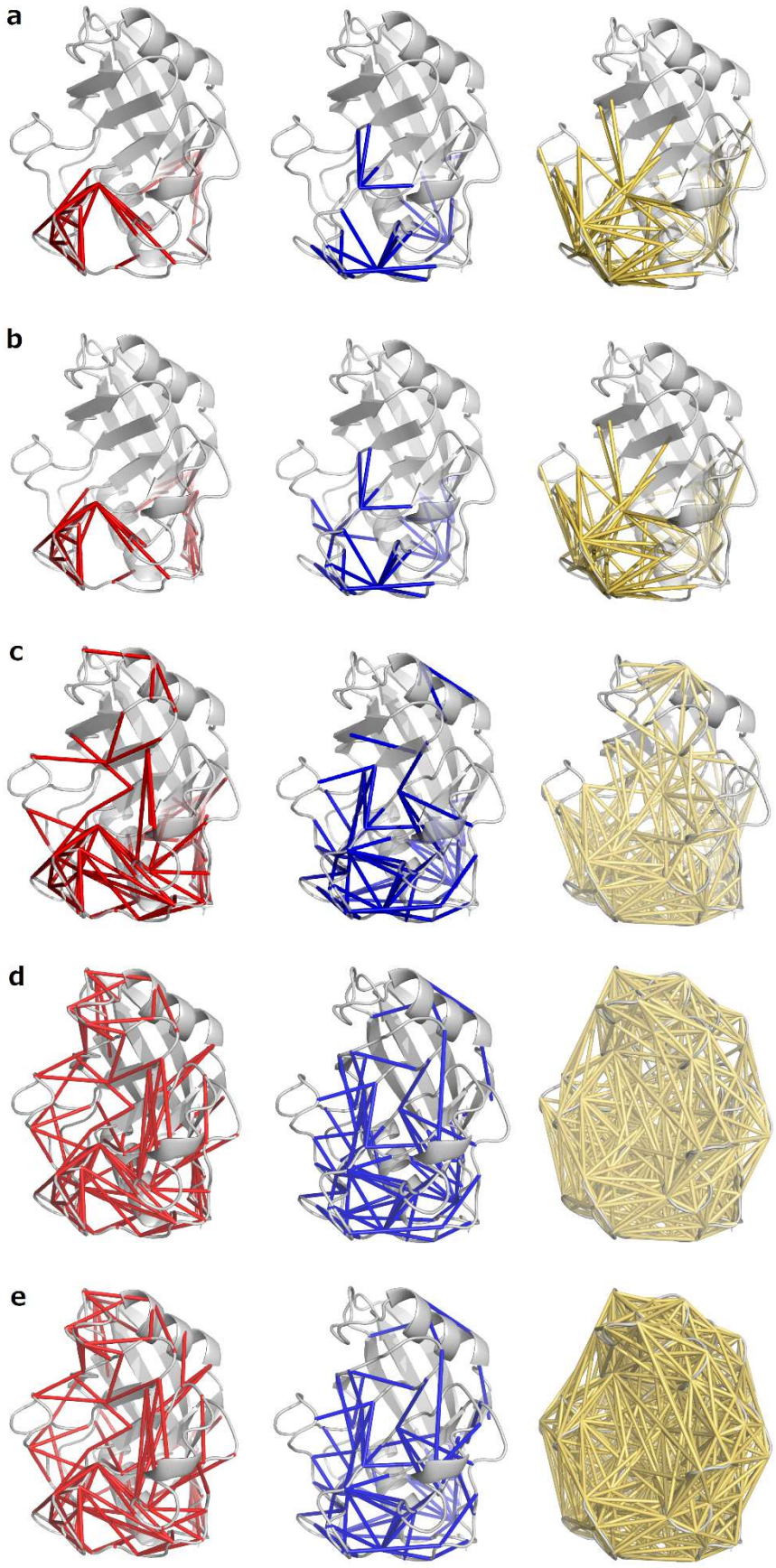
a
b
c
d
e

Figure S3 Overview of the perturbation in residue-residue contacts in CypA domain caused by the mutation V29L. Residues ranked in the top 5, 10, 20, 50, and 100% of contact similarity scores (Tanimoto coefficient) are shown in panel a, b, c, d and e. (left) Red line: contact frequency fij increased ≥0.1 in the mutated CypA domain. (middle) Blue line: contact frequency fij decreased ≥0.1 in the mutated CypA domain. (right) Yellow line: contact frequency fij increased <0.1 or decreased <0.1 in the mutated CypA domain.

NMR experiments and chemical-shift-restrained MD simulations has shown the putative communication pathways response 1 (residues 91, 92, 96, 97, 113, 114, 118, 119, 120, 122, 123 and 125) and response 2 (residues 48, 55, 56, 61, 62, 63, 65, 77, 78, 83, 87, 102, 108, 109, 112, 115) through which mutation to Val29 affects the active site (Holliday *et al.*, 2017). The residues in Figure S4 are clustered in the left-hand half of the CypA molecule, covering the pathway response 2. Among the top 20% of residues from MDContactCom analysis, half (7/16) of pathway response 2 were detected. However, only two of 12 residues were detected in pathway response 1. When contacts including top 20% of residues are counted in, 10 of 16 residues in response 2 and five of 12 residues in response 1 were detected. In Figure S4 the changes in contact dynamics caused by mutation V29L reaches a binding pocket through a left side of the molecule.

Many contacts shown in Figure S4 are formed over a rim of the binding pocket. A higher affinity for the substrate of V29L mutation could be a consequence of the binding site strengthened by modulation of contacts dynamics. Though further analysis and quantification is required to the conclusion, NMR experiments or MD simulations has revealed the same pathways (Holliday *et al.*, 2017; Rodriguez-Bussey *et al.*, 2018). These results demonstrate a representative mechanism to propagate allosteric information from a mutated site towards a binding region. Comparison of two trajectories using MDContactCom is helpful to detect the first region of CypA molecule to focus on after MD simulations are performed.

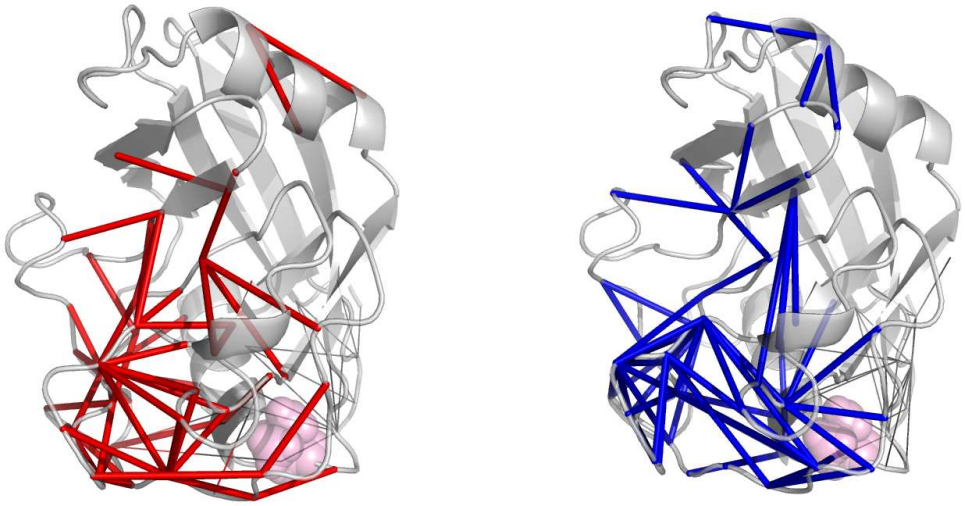

Figure S4 Overview of the perturbation in residue-residue contacts in CypA domain caused by the mutation V29L. Residues ranked in the top 20% of contact similarity scores are shown. (left) Red line: contact frequency fij increased ≥0.1 in the mutated CypA domain. (right) Blue line: contact frequency fij decreased ≥0.1 in the mutated CypA domain. Contacts including N-terminal and C-terminal residues are shown in thin gray lines. Mutated residue V29 is represented as a pink sphere.

References

Abraham,M.J. *et al.* (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1**–**2**, 19–25.

Asamitsu,K. *et al.* (2017) MD simulation of the Tat/Cyclin T1/CDK9 complex revealing the hidden catalytic cavity within the CDK9 molecule upon Tat binding. *PLoS One*, **12**, 1–14.

Brooks,B.R. *et al.* (2009) CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.

Case, D.A., Aktulga, H.M., Belfon, K., Ben-Shalom, I.Y., Brozell, S.R., Cerutti, D.S., Cheatham, T.E.III, Cruzeiro, V.W.D., Darden,T.A., Duke,R.E., Giambasu,G., Gilson,M.K., Gohlke,H., Goetz,A.W., Harris,R., Izadi,S., Izmailov,S.A., Jin,C., Kasavajhala,K., Kaymak,M.C., King,E., Kovalenko,A., Kurtzman,T., Lee,T.S., LeGrand,S., Li,P., Lin,C., Liu,J., Luchko,T., Luo,R., Machado,M., Man,V., Manathunga,M., Merz,K.M.,Miao,Y., Mikhailovskii,O., Monard,G., Nguyen,H., O'Hearn,K.A., Onufriev, A., Pan,F., Pantano,S., Qi,R., Rahnamoun,A., Roe,D.R., Roitberg,A., Sagui,C., Schott-Verdugo,S., Shen,J., Simmerling,

C.L., Skrynnikov,N.R., Smith,J., Swails,J., Walker,R.C., Wang,J., Wei,H., Wolf,R.M., Wu,X., Xue,Y., York,D.M., Zhao,S., and Kollman, P.A. (2021), Amber 2021, University of California, San Francisco.

Holliday,M.J. *et al.* (2017) Networks of Dynamic Allostery Regulate Enzyme Function. *Structure*, **25**, 276–286.

Kumawat,A. and Chakrabarty,S. (2017) Hidden electrostatic basis of dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci.*, **114**, E5825–E5834.

Liu,J. and Nussinov,R. (2017) Energetic redistribution in allostery to execute protein function. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 7480–7482.

Pettersen,E.F. *et al.* (2004) UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

Phillips,J.C. *et al.* (2020) Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.*, **153**.

Rodriguez-Bussey,I. *et al.* (2018) Decoding Allosteric Communication Pathways in Cyclophilin A with a Comparative Analysis of Perturbed Conformational Ensembles. *J. Phys. Chem. B*, **122**, 6528–6535.

Schrödinger Release 2021-2: Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, 2021. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2021.

Stacklies,W. *et al.* (2009) Dynamic allostery in the methionine repressor revealed by force distribution analysis. *PLoS Comput. Biol.*, **5**.

Tsai,C.J. *et al.* (2008) Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play. *J. Mol. Biol.*, **378**, 1–11.

Tzeng,S.R. and Kalodimos,C.G. (2009) Dynamic activation of an allosteric regulatory protein. *Nature*, **462**, 368–372.