Uncovering allosteric pathways in caspase-1 with a multiscale community detection method and random-walk network analysis

B. Amor, S. N. Yaliraki, R. Woscholski, M. Barahona

August 2012

Supplementary Information

S.1 Construction of the biophysical graph of the protein: potentials and energy of interactions

Our network is constructed by assigning edges between atoms which interact covalently and non-covalently. Each edge is weighted with weight given by the strength of the interaction between the two atoms it joins. Covalent bond strengths are obtained from tables assuming standard bond lengths. We include three types of non-covalent interactions: hydrophic interactions, hydrogen bonds, and electrostatic interactions.

S.1.1 Hydrophobic tethers

Hydrophobic tethers are assigned between C-C or C-S pairs based on proximity: two atoms have a hydrophobic tether if their Van der Waals' radii are within 2Å. The hydrophobic tethers are identified using FIRST , which does not assign them an energy. The energy of the interaction is then determined based on the double well potential of mean force introduced by Head-Gordon[1], which gives an energy of $\approx$ -0.8kcal/mol for atoms within 2Å.

S.1.2 Hydrogen bonds

The bond strengths were calculated using the same formula used by the program FIRST[2] and is based on the potential introduced by Mayo *et al*[3]. The formula for the calculation of bond strengths is:

$$E_{HB} = V_0 \left\{ 5 \left(\frac{R_0}{R} \right)^{10} - 6 \left(\frac{R_0}{R} \right)^{12} \right\} F(\theta, \phi, \varphi), \tag{S1}$$

where $V_0 = 8$kcal/mol is a constant, $R_0 = 2.80\mathring{A}$ is the equilibrium donor-acceptor distance, and R is their actual distance. $F(\theta, \phi, \varphi)$ is a function of the angles between the donor, acceptor and hydrogen atoms and depends on the type of bond:

- sp^3 donor - sp^3acceptor: $F = \cos^2\theta e^{-(\pi-\theta)^6} \cos^2(\phi - 109.5)$,
- sp^3 donor - sp^2acceptor: $F = \cos^2\theta e^{-(\pi-\theta)^6} \cos^2(\phi)$,
- sp^2 donor - sp^3acceptor: $F = \cos^4\theta e^{-2(\pi-\theta)^6}$,
- sp^2 donor - sp^2acceptor: $F = \cos^2\theta e^{-(\pi-\theta)^6} \cos^2(\max[\phi, \psi])$,

where θ is the donor-hydrogen-acceptor angle, ϕ is the hydrogen-acceptor-base angle, and ψ is the angle between the normals of the planes given by the six atoms attached to the sp^2 centres.

S.1.3 Electrostatics

Electrostatic interactions to be included in the graph are found by a number of different approaches. FIRST implements a geometry based calculation of stacked interactions (for further details see the FIRST documentation). We include electrostatic interactions between the ligands and proteins defined by the LINK entries in the protein's PDB file. Cation-pi interactions are found using the CAPTURE web-server and added to the network manually[4].

S.2 Markov Stability analysis of the caspase graph

Figure S1 shows the Markov Stability analysis of the inactive (1SC1) and active (1ICE) structures of caspase-1 between Markov times 0.001 and 1000. As discussed in the Main text, robust communities in the protein graph are detected as long-lived partitions (i.e., long plateaux in the number of communities) which are robust to the optimisation (i.e., correspond to dips in the variation of information, VI). The drops in VI around Markov times 2×10^{-3} and 10^{-1} correspond, respectively, to chemical groups and amino acids being detected as robust communities. This demonstrates that Markov Stability captures the structure of the protein at smaller scales in the way we would expect. After Markov time 10 we see a drop in the VI, which corresponds to the emergence of the secondary structure elements as communities. At long Markov times we see robust four-way and two-way partitions. For proteins with less than 10,000 atoms we find that 10^3 time steps is long enough for the random-walk to uncover the structure of the protein at the highest scales. For a full explanation of the methodology see Delmotte *et al.*[5].

S.3 Gaussian Process Regression for identification of statistically significant mutations

To identify residues whose weak interactions are crucial to the community structure of the protein we carry out a full computational mutagenesis by removing from the network all edges corresponding to weak interactions of each residue in turn. Removing bond edges has a greater effect at some scales than at others.

We introduce a novel method for identifying important mutations which not only identifies mutations which have a significant impact on the community structure, but also the scale at which this effect is observed. Using Gaussian Process Regression, we obtain a VI 'trajectory' for each mutated structure and we use this ensemble of trajectories to produce a 'representative' trajectory with statistical bounds associated. Any trajectory falling more than three standard deviations from this representative trajectory for at least one third of the time points in the relevant Markov time plateau is classified as having a significant impact on the robustness of the graph.

In particular, each stability run gives a trajectory $\mathbf{v}_i = [v_1, v_2, \ldots, v_u]$ of VI values at Markov times $\mathbf{t} = [t_1, t_2, \ldots, t_u]$. Carrying out a full mutational analysis of a protein with N residues gives $N+1$ trajectories (including the wild-type data). These trajectories are appended, $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{N+1}]$ and $\mathbf{t} = [\mathbf{t}, \mathbf{t}, \ldots, \mathbf{t}]$, to create a 'training set' $(\mathbf{v}, \mathbf{t})$. The function which maps $\mathbf{v}$ into $\mathbf{t}$ can then be estimated by using a nonparametric fitting method called Gaussian Process Regression[6]. A Gaussian process is an infinite-dimensional extension of a multivariate Gaussian distribution. Any finite subset of the data in a given range has a multivariate Gaussian distribution. The Gaussian process is assumed to have mean zero and is then completely defined by the covariance function $k(x,x')$ which relates each data point to each other data point. We use the following squared exponential covariance function:

$$k(x,x') = \sigma_f^2 \exp\left[-\frac{1}{2l^2}(x-x')^2\right] + \sigma_n^2 \delta_{x,x'}, \tag{S2}$$

where $\delta_{x,x'}$ is the Kronecker delta function and the hyperparameters $\theta = [\sigma_f, \sigma_n, l]$ are estimated by maximising $p(\theta|\mathbf{v},\mathbf{t})$ using the gpml MatLab toolbox*[6]. The output from this toolbox is a mean function which we plot $+/-$ two standard deviations, corresponding to a 95% confidence interval.

Figure S2 shows an example of the results obtained when this process is carried out on active caspase-1 (1ICE). Two trajectories are plotted, one for the mutation Glu151A which does not have a significant impact on the community structure at any scale and one for the mutation Cys136A which has a significant impact at several different scales.

Computational mutagenesis was performed for each residue in active (2HBQ) and inactive (1SC1) caspase-1. The ensemble of VI trajectories obtained for all mutant structures are plotted in grey in Figure S3. Shown in colour are the trajectories of the

*Available from http://www.gaussianprocess.org/gpml/code/

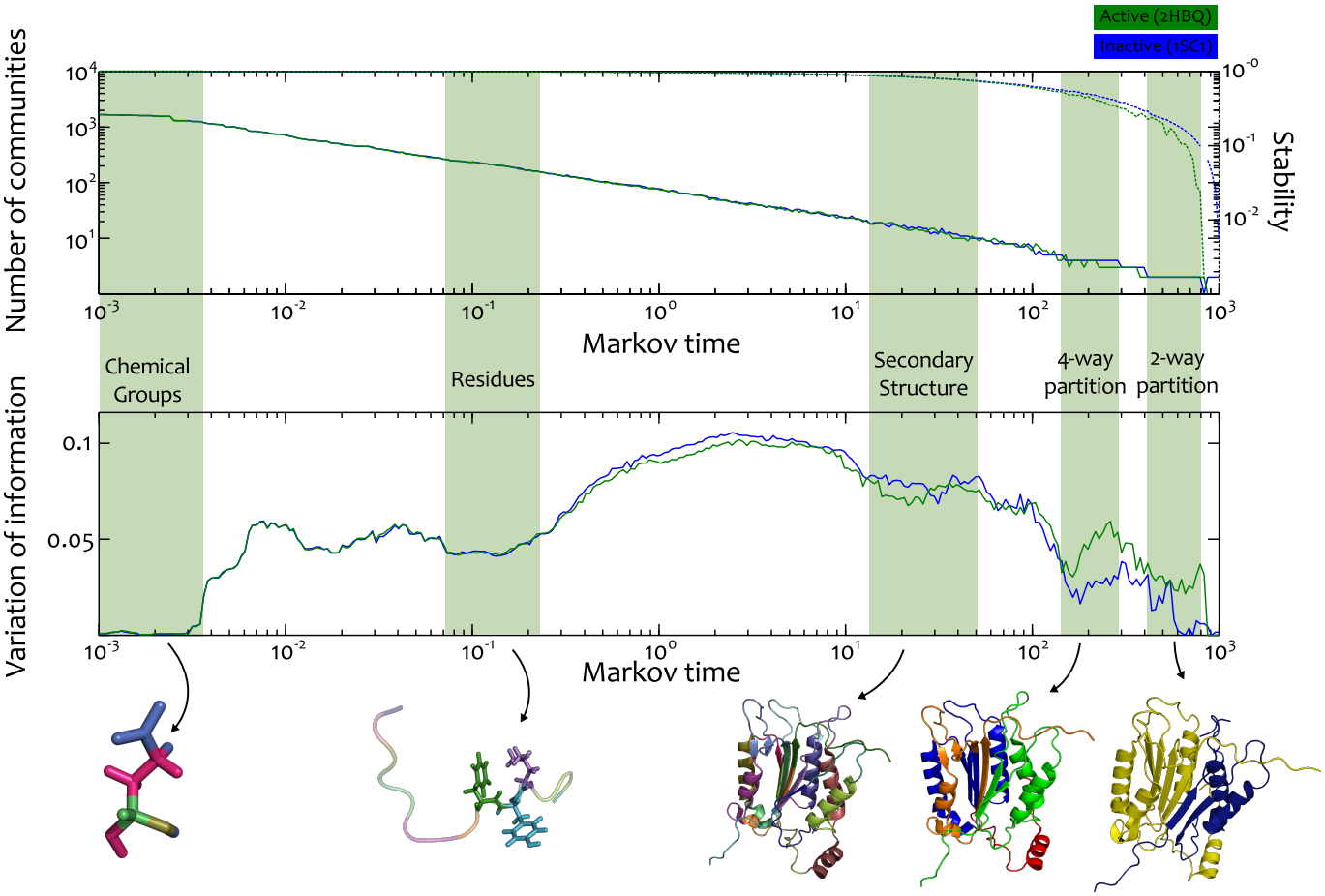

Figure S1 (Top panel) Markov Stability (dashed line) and number of communities (solid line) and (bottom panel) variation of information of the partitions found for the inactive (blue) and active (green) caspase-1 between Markov times 0.001 and 1000. We find communities of chemical groups and amino acids (at small Markov times), secondary structure (at intermediate Markov times), and domains/sub-units (at long Markov times).

residues identified as statistically significant (see also Fig. 4). The location of these significant residues and the weak interactions they form are shown in Figure S4.

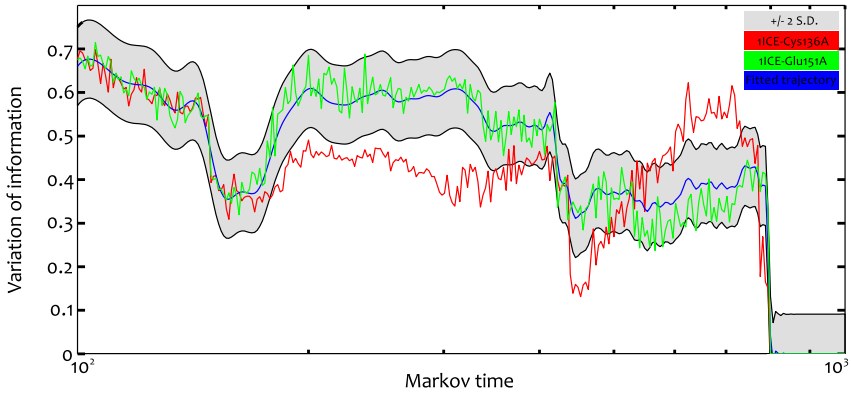

Figure S2 Example output from Gaussian Process Regression. The blue line and the grey shaded area are the fitted trajectory and error for the entire set of VI trajectories obtained by mutating every residue in the active (1ICE) conformation of caspase-1. The green line is the trajectory obtained by mutating Glu151 and represents a non-significant mutation. The red line is the trajectory obtained by mutating Cys136 and falls outside the bounds over the periods 170-400, 440-462, and 600-730.

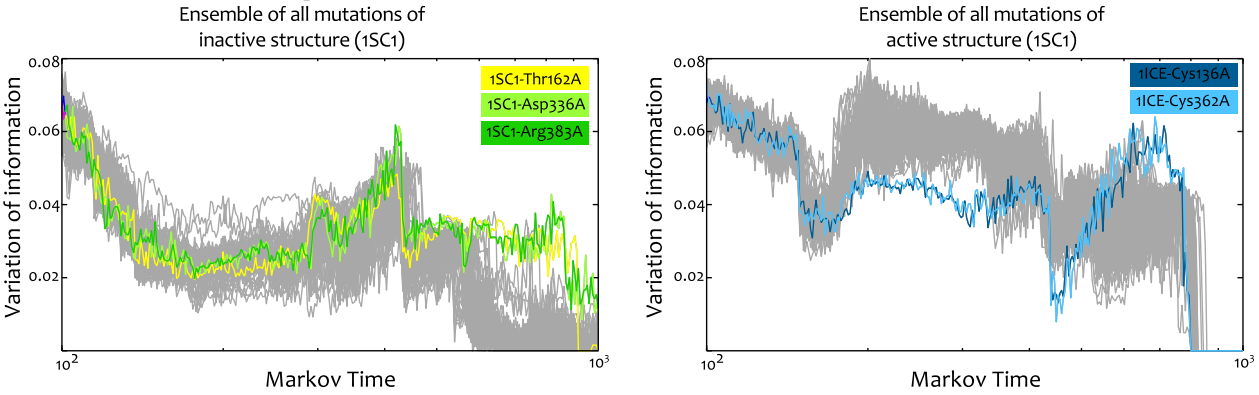

Figure S3 Ensemble of VI trajectories. The VI trajectories from all mutated structures of inactive (a) and active (b) caspase-1. The significant mutations are shown in colour.

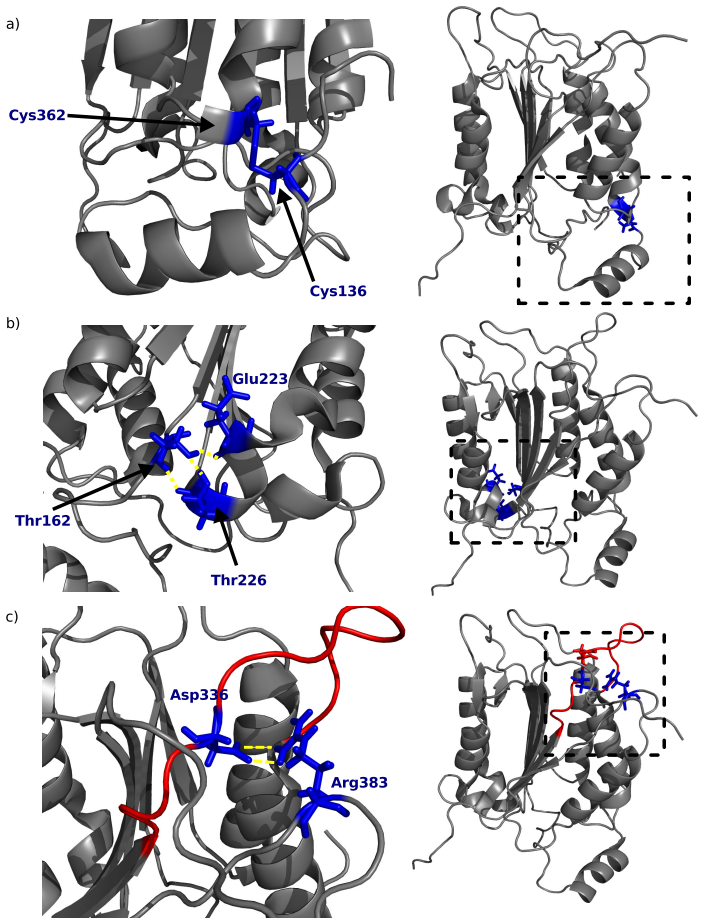

Figure S4 Statistically significant mutations and their structural location. a) Asp336/Arg383 form salt-bridge which stabilises the binding loop (coloured red) in the inactive conformation. Removal of this bond causes the two subunits to be more closely associated at long Markov times. b) Thr162 forms hydrogen bonds with residues Glu223 and Thr226 in the inactive conformation. Removal of this bond also causes the two subunits to be more closely associated at long Markov times. c) Cys136/Cys362 disulfide bond provides a bridge between the two subunits in the active conformation. Removal of this causes an increase in the quality of the partition in the active conformation due to the decreased association of the two subunits.

S.4 Evolutionary conservation of residues with high conformational ratios Δ_{CF}

We use the popular ConSurf[7] package to estimate the degree of evolutionary conservation of the residues in caspase-1. Figure S5 shows the evolution score of each residue in chain A (residues 120-297) and chain B (residues 317-404). Each score is calculated relative to the other residues in the same chain and negative scores indicate a high level of conservation. Residues identified as having large conformational $t_{1/2}$ ratios in Table 1 and Table 2 are highlighted in red. Residues with high $t_{1/2}$ ratios generally are well conserved, particularly in chain B.

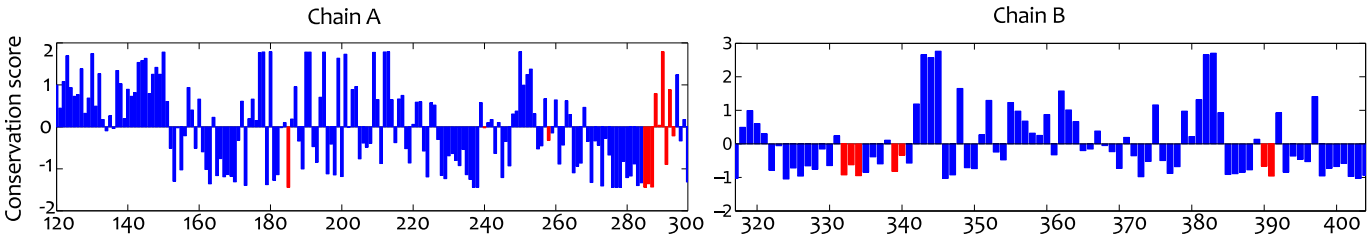

Figure S5 The relative evolutionary conservation scores for residues in chain A (left) and chain B (right) as calculated using the ConSurf server[7]. A more negative score indicates a high degree of conservation. The residues identified as having largest conformational, bond removal, and mutational $t_{1/2}$ ratios (see Tables 1-4) are coloured red.

References

[S1] M. S. Lin, N. L. Fawzi and T. Head-Gordon, *Structure*, 2007, **15**, 727–740.

[S2] D. Jacobs, A. Rader, L. Kuhn and M. Thorpe, *Proteins: Structure, Function, and Bioinformatics*, 2001, **44**, 150–165.

[S3] B. Dahiyat, D. Benjamin Gordon and S. Mayo, *Protein Science*, 1997, **6**, 1333–1337.

[S4] J. P. Gallivan and D. A. Dougherty, *Proceedings of the National Academy of Sciences*, 1999, **96**, 9459–9464.

[S5] A. Delmotte, E. Tate, S. Yaliraki and M. Barahona, *Physical Biology*, 2011, **8**, 055010.

[S6] C. Rasmussen, *Advanced Lectures on Machine Learning*, 2004, 63–71.

[S7] H. Ashkenazy, E. Erez, E. Martz, T. Pupko and N. Ben-Tal, *Nucleic acids research*, 2010, **38**, W529–W533.