Entropy based distance cutoff for protein internal contact networks

Marcin Sobieraj [1,2] and Piotr Setny [1,*]

[1] Centre of New Technologies, University of Warsaw, Warsaw, Poland

[2] Physics Department, University of Warsaw, Warsaw, Poland

*email: *p.setny@cent.uw.edu.pl*

Abstract

Protein structure networks (PSNs) have long been used to provide a coarse yet meaningful representation of protein structure, dynamics, and internal communication pathways. An important question is what criteria should be applied to construct the network so that to include relevant interresidue contacts while avoiding unnecessary connections. To address this issue we systematically considered varying residue distance cutoff length and the probability threshold for contact formation to construct PSNs based on atomistic molecular dynamics in order to assess the amount of mutual information within the resulting representations. We found that the minimum in mutual information is universally achieved at the cutoff length of 5 Å, irrespective of the applied contact formation probability threshold in all considered, distinct proteins. Assuming that the optimal PSNs should be characterised by the least amount of redundancy, which corresponds to the minimum in mutual information, this finding suggests an objective criterion for cutoff distance and supports the existing preference towards its customary selection around 5 Å length, typically based to date on heuristic criteria.

Keywords: protein structure networks, contact cutoff distance, mutual information, computer simulations

1 Introduction

Understanding protein dynamics is the key to understand their function.[1–3] With tremendous advances in computer power and simulation methodologies,[4] we are becoming able to produce atomistic trajectories of protein motions that cover biologically relevant events such as folding, ligand binding, or subtle, allostery-driven shifts of functional properties.[5–7] Intrinsic complexity of such events and their occurrence along pathways that are not *a priori* defined within chaotic, thermally driven motions make it difficult to extract essential information from huge amount of generated data. One possible route is to simplify protein representation from atomistic to that of a network of individual centres (nodes) connected by effective interactions (edges).[8–12] It is driven by the hope that most functional aspects of biological macromolecules, that are essentially folded linear polymers formed by a number of well defined units, are encoded within the topology and dynamics of their inter-residue contacts.

On the one hand, such an idea has long been used to derive simplified potentials to describe protein structure. Indeed, it turned out that so called elastic network models (ENMs),[13] typically consisting of Cα protein atoms connected with near neighbours by simple quadratic potentials with equilibrium lengths taken from native structure geometry, are able to provide surprisingly good reproduction of global macromolecular deformability,[14] as well as local structural fluctuations.[15] Their apparent success indicates that the dynamic properties of protein structure are to a large extent encoded in the topology of close contacts and that harmonic representation of effective potential energy is a reasonable first order approximation of free energy landscape in the vicinity of stable conformational states.

On the other hand, network-based approaches are also adopted to post-proces atomistic trajectories generated by molecular dynamics (MD) simulations. Here, the connectivity of protein residues is determined based on inter-residue distance statistics or average interaction energies,[16,17] evaluated over generated conformational ensembles. The resulting protein structure networks (PSNs),[18–20] typically combined with analysis methods originating from graph theory,[8] have been used to investigate various aspects of structure - function relationships, including allosteric communication pathways, identification of distinct structural motifs or key functional residues, as well as protein folding[21–24]

An important issue related to the development of ENM or PSN models is the choice of macromolecular representation used to determine inter-residue contacts. In the case of protein connectivity graphs based on geometric criteria, possibilities extend from explicitly considering all heavy atoms, defining two or more centres per amino acid, or the selection of a single representative atom (eg. Cα or Cβ) for each residue. Typically, an edge between two residues is established if the closest approach between any pair of their elements falls below the assumed distance cutoff, R_c. The cutoff itself depends on the kind of structure representation. If connectivity is derived based on atomistic structures, cutoff distances are usually assumed in the range between 4.0 to 5.5 Å.[12,24–31] In turn, for coarse grained representation R_c values between 7 – 8.5 Å are adopted.[9–11,14,32,33]

Although relatively few truly systematic studies have been conducted to rationalise cutoff selection,[32,34] a number of arguments has been raised to support particular choices. For atomistic approaches, it has been argued that $R_c = 5$ Å represents the upper range of meaningful attractive dispersion forces.[26] Others have found that a cutoff of 5 Å allows identification of key, multi-connected residues in best agreement with experimental data concerning functional importance of specific amino acids within several protein families.[28] A similar cutoff was shown to produce PSNs that are at the border of percolation transition from multiple, loosely connected residue clusters for $R_c < 5$ Å, to a single large cluster comprising most protein residues for $R_c > 5$ Å, irrespective of protein force field used for MD simulations.[34] In yet another study, a cutoff of 5.5 Å turned out to produce the optimal (smallest) number of distinct, intercorrelated residue clusters based on the community analysis of protein dynamics.[30]

In the current study we consider an approach to cutoff selection based on the observation of linear mutual information (MI) density in the protein contact networks. Such a measure reveals the degree of correlation between inter-residue distances within the network and thereby the degree of unwanted redundancy within the network. We demonstrate that minimum MI is achieved for several unrelated proteins, for distance cutoffs in a very narrow range around 5 Å. Thus, indicating possibly non-redundant protein representation, this observation provides yet another rationale for cutoff length selection.

2 Materials and Methods

2.1 Protein structures

We considered five different, unrelated protein structures, representing: protein kinase A (PKA), lysine-, arginine-, ornithine-binding protein (LAO), lambda repressor (LR), trypsin and SH3 domain. In addition, PKA was present in inactive and active form, the latter of which carried two phosphate groups and a bound ligand (ATP). In turn, LAO protein was simulated in two, distinct conformations corresponding to open and closed protein structure. The summary of structural details is given in the Table 1.

2.2 Molecular dynamics simulations

All simulations were conducted using Gromacs simulation program.[35] Protein crystal structures were taken from the Protein Databank.[36] In each case protonation states of titratable residues were assigned using the PropKa server[37] to reflect pH = 7. The structures were parametrised using Amber99-ILDNP$*$ force field[38] and TIP3P water model[39] was used for aqueous solvent. The systems were simulated with periodic boundary conditions, using rhombic dodecahedron simulation boxes, with 15 Å solvent margin surrounding initial protein structures. The total charge on protein structures was neutralised with Na or Cl counter ions, and excess salt was added to reach the concentration 0.15 mol/l. A cutoff of 10 Å was used for Lennard Jones potential, with additional long range dispersion corrections for energy and pressure. Electrostatic interactions were calculated using the Particle Mesh Ewald method[40] with 1.2 Å grid spacing and 10 Å cutoff in real space. All bonds to hydrogen atoms were fixed using the Lincs algorithm,[41] and the simulation time step was 2 fs. Following energy minimisation and initial thermalisation, the systems were equilibrated for 100 ns in NpT conditions. Temperature of 310 K and pressure of 1 bar were maintained by Nose-Hoover thermostat,[42] and Parrinello-Rahman barostat,[43] respectively. Production runs were conducted in analogous conditions. The summary of collected trajectories is given in the Table 1.

2.3 Determination and analysis of contacts

We assume that a contact between two residues is formed, if the distance between any two of their heavy atoms is shorter than the assumed cutoff length, R_c. Contacts statistics is determined based on the entire MD run, and contacts formed for a fraction $\nu \in [f, 1-f]$ of the simulation time, with considered $f \in [0.1, 0.3]$, are used to determine the elements of the correlation matrix, $\mathbf{r}$, as described in the following.

3 Results and Discussion

3.1 Contact network

A set of inter-residue protein contacts can be established for a given simulation based on two parameters: cutoff length, R_c, and contact probability, ν. The first one determines whether two residues form a contact in a given simulation frame, and the second one narrows the set of contacts to those that are formed for the fraction $\nu \in [f, 1-f]$, with $f \in [0,1]$, of the simulation time. This second criterion allows discarding less informative contacts that are either permanently or only accidentally formed.

We assume that in order to provide the most effective description of protein dynamics, a desired set of contacts should carry possibly most information with possibly few contacts, or, in other words, each contact channel should carry information that is least related to information already contained in the remaining channels. In order to determine the optimal (R_c, f) pair, we thus utilise the concepts based on the information theory and seek to minimise mutual information within contact-based trajectory.

For a set of $N = N(R_c, f)$ contacts between protein residues in a given simulation, we consider respective inter-residue distances, called contact lengths, in subsequent simulation frames as components of N-dimensional time series $\mathbf{x}(t) = (x_0(t), x_1(t), ..., x_N(t))$. Given that we are interested in conformational ensemble representing protein folded into its native, specific tertiary structure, $\mathbf{x}(t)$ samples finite space and is assumed to fluctuate around certain average value. The differential entropy,[44] $H(\mathbf{X})$, of the associated, joint probability distribution $p(\mathbf{x})$, can, in principle, be evaluated as:

$$H(\mathbf{X}) = -\int p(\mathbf{x}) \ln{(p(\mathbf{x}))} d\mathbf{x}. \tag{1}$$

Its direct estimation would require, however, complete sampling of configuration space, and is thus prohibitive. An upper limit[44] of the entropy expressed with Equation 1, corresponds to the entropy of a multivariate normal distribution $\tilde{\mathbf{X}} \sim N(0, \mathbf{C})$ with $\mathbf{C}$ being a covariance matrix of the original signal: $C_{ij} = \mathrm{cov}(x_i, x_j)$,

$$H(\mathbf{X}) \leq H(\tilde{\mathbf{X}}) = \frac{1}{2} \ln{((2\pi e)^N |\mathbf{C}|)}. \tag{2}$$

The model based on normal distribution $\tilde{\mathbf{X}}$ accounts for amplitudes of distance fluctuations and their pairwise linear correlations. The overestimation of entropy with respect to the original distribution, $\mathbf{X}$, comes from the fact that marginal probability distributions, $p(\mathbf{x})$, do not need to be Gaussian (i.e. inter-residue distances may fluctuate in anharmonic effective potentials), and that higher order than pairwise linear correlations are not captured. Nevertheless, in order to assess interdependence between system components (e.g. for the analysis of allosteric correlations) under such harmonic approximation, one may separate the contributions to entropy arising from the amplitude of fluctuations from those that quantify the extent of (linear) correlation within the system. One measure of the latter is mutual information, which under harmonic approximation has a particularly simple form $I(\tilde{\mathbf{X}}) = -\frac{1}{2} \ln |\mathbf{r}|$, with $\mathbf{r}$ being the correlation matrix of $\tilde{\mathbf{X}}$. Indeed, it can be shown (see Supporting Information) that:

$$H(\tilde{\mathbf{X}}) = \sum_{i}^{N} \frac{1}{2} \ln (2\pi e C_{ii}) - I(\tilde{\mathbf{X}}), \tag{3}$$

where the sum represents entropies of i normal distributions, each with its individual variance C_{ii}, and the second term on the right hand side accounts for pairwise linear correlations between them.

3.2 Mutual information density and average correlation

From physical perspective, entropy is an extensive property of a system. If the dimensionality of the signal under study (e.g. the time series of contact lengths) varies depending on different initial assumptions (e.g. the criteria used to define contacts), one may be interested to assess not the total MI content, but rather its amount per single channel in order to determine such a representation that would reveal most internal correlations. Accordingly, an intensive MI (IMI) obtained based on the contact network established for specific distance cutoff and contact probability, denoted in the following as $\iota(R_c, f)$, can be expressed as:

$$\iota(R_c, f) = \frac{I(\tilde{\mathbf{X}}(R_c, f))}{N(R_c, f)}. \tag{4}$$

Being not directly dependent on the size of the considered contact network, such measure allows the comparison of interdependence within networks obtained for different cutoff values, as well as networks characterising different sized proteins.

An undesirable property of IMI is that with increasing correlation between system components it rises to $+\infty$. In order to deal with a more tractable descriptor we consider an average correlation (AC), $\bar{r} \in [0, 1]$, defined as a value of off-diagonal, all-uniform elements of a pseudo-correlation matrix, $\bar{\mathbf{r}}$, constructed such that $\bar{r}_{ij} = \bar{r}$ for $i \neq j$, and 1 otherwise, chosen to produce a determinant equal to the determinant of the original signal correlation matrix: $|\bar{\mathbf{r}}| = |\mathbf{r}|$. Note, that the equality of the two determinants implies that $\iota(\bar{\mathbf{r}}) = \iota(\mathbf{r})$ under the harmonic approximation. Once the correlation matrix $\mathbf{r}$ is determined based on simulation data, $\bar{r}$ can be unambiguously obtained by numericaly solving an analytic equation derived and discussed in the Supporting Information.

3.3 Optimal distance cutoff

Plots of AC as a function of R_c and f, obtained based on long MD trajectories of five proteins with varying sizes and secondary strcucture types (Table 1) are shown in Fig. 1. All curves share a similar form, revealing generally increasing AC for increasing cutoff distance. This is an expected behaviour that reflects gradual transition from relatively uncorrelated, disjoined short range contacts to dense contact networks rich in multiple, redundant long range connections. Strikingly, however, there is a shallow local minimum at $R_c^* \simeq 5$ Å, uniformly present in most AC curves. It indicates, that new, non-permanent contacts included into the network as R_c approaches R_c^* from below tend to introduce relatively non-redundant information about the system. Overall, the minimum becomes more apparent for contacts whose probability of formation, f, remains close to 0.5. Such "maximum entropy" contacts, neither preferentially formed nor broken, are indeed expected to be least correlated with the rest of the network, and thus are possibly most interesting. Accordingly, $R_c^* \simeq 5$ Å appears to encode protein structure networks with highest proportion of such contacts providing for the optimally efficient representation. The above observations made for proteins represented with the Amber force field, seem to generally hold also for CHARMM force field parametrisation. An analalysis of CHARMM-based test run for λ-repressor, reveals AC minimum at $R_c^* \simeq 5$ Å for $f \in [0.1, 0.2]$, albeit with a slight shift towards $R_c^* \simeq 5.4$ Å for $f \in [0.2, 0.3]$ (see Supporting Information).

3.4 Structural interpretation

The fact that the minimum in average correlation between inter-residue distances in PSNs occurs at $R_c^* \simeq 5$ Å for rather diverse proteins seems to indicate that it stems from the local properties of protein structures rather than their global architecture. An interpretation of structural determinants underlying the $\bar{r}(R_c)$ behaviour can be attempted based on the analysis of $N(R_c, f)$. Somewhat counterintuitively, the total number of potentially interesting contacts that are formed only for a fraction of the simulation time (eg. $\nu \in [0.1, 0.9]$) initially drops while R_c increases till $\lesssim 5$

Å, uniformely in all considered cases (Fig. 3). We attribute this observation to the fact that the most of short range interactions correspond to permanent contacts formed owing to residue packing in protein structure: with increasing cutoff length they become excluded from the $\nu \in [0.1, 0.9]$ category (and shift to the $\nu \in (0.9, 1.0]$ category) at a faster rate than new, longer contacts enter it. This is reflected by the radial distribution of "permanent" contacts (Fig. 3) which achieves its maximum around 5 Å, indicating the location of the first neighbour shell around an average amino acid. The contacts that are formed only transiently for $R_c = 5$ Å have the ability to reversibly penetrate and abandon the first neighbour shell of nearby amino acids (Fig. 4), which apparently corresponds to events that are relatively little correlated with the rest of system dynamics. Such picture complements well the observation of percolation threshold within PSN $R_c = 5$ Å, reported before[34] and indicates high allosteric potential of transient contacts defined by this cutoff radius.

4 Conclusions

We considered dynamic protein structure networks in which nodes (amino acids) are connected by edges whose existence depends on the probability that given inter-residue distance falls below certain cutoff, so that to exclude permanently formed or broken, potentially less interesting contacts. Being interested in network representation that conveys possibly most essential information per degree of freedom we evaluated harmonic approximation of intensive mutual information between contact distances fluctuating in unconstrained molecular dynamics simulation as a function of cutoff distance and the probability of formation. To this end we introduced an average correlation parameter that quantifies the amount of mutual information per contact. We found that the minimum in such correlation occurs at a cutoff distance of 5 Å in all considered proteins, in spite of differences in their dominant secondary structure type. We attribute this observation to the structure of local protein packing. In this light, non-permanent contacts falling within 5 Å cutoff length should have the ability to transiently penetrate the first neighbour residue shells of the nearby amino acids, thus providing connection between local residue clusters. This interpretation is in agreement with previous PSN analyses indicating a percolation threshold at 5 Å, and provides another objective rationale for this cutoff selection.

Acknowledgement This work was supported by EMBO IG 3051/2015 to PS.

References

[1] Frauenfelder H, Sligar S., Wolynes Peter G. The energy landscapes and motions of proteins *Science (80-.)..* 1991;254:1598–1603.

[2] Vendruscolo Michele, Dobson Christopher M.. Dynamic visions of enzymatic reactions *Science (80-.)..* 2006;313:1586–1587.

[3] Henzler-Wildman Katherine, Kern Dorothee. Dynamic personalities of proteins. *Nature.* 2007;450:964–72.

[4] Maximova Tatiana, Moffatt Ryan, Ma Buyong, Nussinov Ruth, Shehu Amarda. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics *PLoS Comput. Biol..* 2016;12:1–70.

[5] Lindorff-Larsen Kresten, Piana Stefano, Dror Ron O., Shaw David E.. How fast-folding proteins fold. *Science (80-.)..* 2011;334:517–520.

[6] Dickson Alex, Tiwary Pratyush, Vashisth Harish. Kinetics of Ligand Binding Through Advanced Computational Approaches: A Review *Curr. Top. Med. Chem..* 2017;17:2626–2641.

[7] Wodak Shoshana J., Paci Emanuele, Dokholyan Nikolay V., et al. Allostery in Its Many Disguises: From Theory to Applications *Structure.* 2019;27:566–578.

[8] Kannan N., Vishveshwara S.. Identification of side-chain clusters in protein structures by a graph spectral method *J. Mol. Biol..* 1999;292:441–464.

[9] Vendruscolo Michele, Dokholyan Nikolay V., Paci E., Karplus Martin. Small-world view of the amino acids that play a key role in protein folding *Phys. Rev. E.* 2002;65:061910.

[10] Dokholyan Nikolay V., Li L., Ding F., Shakhnovich E. I.. Topological determinants of protein folding *Proc. Natl. Acad. Sci..* 2002;99:8637–8641.

[11] Atilgan Ali Rana, Akan Pelin, Baysal Canan. Small-World Communication of Residues and Significance for Protein Dynamics *Biophys. J..* 2004;86:85–91.

[12] Brinda K V, Vishveshwara Saraswathi. A Network Representation of Protein Structures : Implications for Protein Stability *Biophys. J..* 2005;89:4159–4170.

[13] Bahar Ivet, Lezon Timothy R, Yang Lee-wei, Eyal Eran. Global Dynamics of Proteins : Bridging Between Structure and Function *Annu. Rev. Biophys..* 2010;39:23–42.

[14] Hinsen Konrad. Analysis of domain motions by approximate normal mode calculations. *Proteins.* 1998;33:417–29.

[15] Bahar Ivet, Atilgan Ali Rana, Erman Burak. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential *Fold. Des..* 1997;2:173–181.

[16] Vijayabaskar M. S., Vishveshwara Saraswathi. Interaction energy based protein structure networks *Biophys. J..* 2010;99:3704–3715.

[17] Ribeiro Andre A.S.T., Ortiz Vanessa. Determination of signaling pathways in proteins through network theory: Importance of the topology *J. Chem. Theory Comput..* 2014;10:1762–1769.

[18] Paola L Di, Ruvo M De, Paci P, Santoni D, Giuliani A. Protein Contact Networks : An Emerging Paradigm in Chemistry *Chem. Rev..* 2013;113:1598–1613.

[19] Grewal Rajdeep, Roy Soumen. Modeling proteins as residue interaction networks *Protein Pept. Lett..* 2015;22:923–933.

[20] Bhattacharyya M, Ghosh S, Vishveshwara Saraswathi. Protein Structure and Function: Looking through the Network of Side-Chain Interactions *Curr Protein Pept Sci.* 2016;17:4–25.

[21] Amitai Gil, Shemesh Arye, Sitbon Einat, et al. Network Analysis of Protein Structures Identifies Functional Residues *J. Mol. Biol..* 2004;344:1135–1146.

[22] Bagler Ganesh, Sinha Somdatta. Assortative mixing in Protein Contact Networks and protein folding kinetics *Bioinformatics.* 2007;23:1760–1767.

[23] Ghosh Amit, Vishveshwara Saraswathi. A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis *Proc. Natl. Acad. Sci..* 2007;104:15711–15716.

[24] Doshi Urmi, Holliday Michael J, Eisenmesser Elan Z, Hamelberg Donald. Dynamical network of residue - residue contacts reveals coupled allosteric effects in recognition , catalysis , and mutation *PNAS.* 2016;113:4735–4740.

[25] Xu Ying, Xu Dong, Gabow Harold N.. Protein domain decomposition using a graph–theoretic approach *Bioinformatics.* 2000;16:1091–1104.

[26] Greene Lesley H., Higman Victoria A.. Uncovering Network Systems Within Protein Structures *J. Mol. Biol..* 2003;334:781–791.

[27] Kamagata Kiyoto, Kuwajima Kunihiro. Surprisingly high correlation between early and late stages in non-two-state protein folding *J. Mol. Biol..* 2006;357:1647–1654.

[28] Del Sol Antonio, Fujihashi Hirotomo, Amoros Dolors, Nussinov Ruth. Residues crucial for maintaining short paths in network communication mediate signaling in proteins *Mol. Syst. Biol..* 2006;2:1–12.

[29] Vanwart Adam T., Eargle John, Luthey-Schulten Zaida, Amaro Rommie E.. Exploring residue component contributions to dynamical network models of allostery *J. Chem. Theory Comput..* 2012;8:2949–2961.

[30] Bowerman S., Wereszczynski Jeff. Detecting Allosteric Networks Using Molecular Dynamics Simulation *Methods Enzymol..* 2016;578:429–447.

[31] Yao X.-Q., Momin M.F., Hamelberg Donald. Elucidating Allosteric Communications in Proteins with Difference Contact Network Analysis *J. Chem. Inf. Model..* 2018;58:1325–1330.

[32] Silveira Carlos H., Pires Douglas E V, Minardi Raquel C, et al. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins *Proteins Struct. Funct. Bioinforma..* 2009;74:727–743.

[33] Wriggers Willy, Stafford Kate a., Shan Yibing, et al. Automated Event Detection and Activity Monitoring in Long Molecular Dynamics Simulations *J. Chem. Theory Comput..* 2009;5:2595–2605.

[34] Viloria Juan Salamanca, Allega Maria Francesca, Lambrughi Matteo, Papaleo Elena. An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass *Sci. Rep.*. 2017;7:1–11.

[35] Abraham Mark James, Murtola Teemu, Schulz Roland, et al. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers *SoftwareX.* 2015;1-2:19–25.

[36] Berman H. M., Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res.*. 2000;28:235–242.

[37] Olsson Mats H M, SØndergaard Chresten R., Rostkowski Michal, Jensen Jan H.. PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions *J. Chem. Theory Comput.*. 2011;7:525–537.

[38] Lindorff-Larsen Kresten, Piana Stefano, Palmo Kim, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field *Proteins Struct. Funct. Bioinforma.*. 2010;78:1950–1958.

[39] Jorgensen William L., Chandrasekhar Jayaraman, Madura Jeffry D., Impey Roger W., Klein Michael L.. Comparison of simple potential functions for simulating liquid water *J. Chem. Phys.*. 1983;79:926 – 935.

[40] Essmann Ulrich, Perera Lalith, Berkowitz Max L, Darden Tom, Lee Hsing, Pedersen Lee G. A smooth particle mesh Ewald method *J. Chem. Phys.*. 1995;103:8577 – 8593.

[41] Hess Berk, Bekker Henk, Berendsen Herman J. C., Fraaije Johannes G. E. M.. LINCS: A linear constraint solver for molecular simulations *J. Comput. Chem.*. 1997;18:1463–1472.

[42] Nosé Shichi. A molecular dynamics method for simulations in the canonical ensemble *Mol. Phys.*. 1984;52:255–268.

[43] Parrinello Michele, Rahman A.. Polymorphic transitions in single crystals: A new molecular dynamics method *J. Appl. Phys.*. 1981;52:7182–7190.

[44] Cover Thomas M., Thomas Joy A.. *Elements of Information Theory*ch. Differential Entropy, :224–238. Hoboken, New Jersey: Wiley-Interscience 1991.

Figure legends

Fig. 1 Upper row: sample $\bar{r}(R_c, f)$ curves for three proteins with different dominant secondary structure types. Lower row: points representing the occurrence of local minima in $\bar{r}(R_c, f)$ curves, and histograms gathering the numbers of minima at each R_c.

Fig. 2 Points representing the occurrence of local minimum in $\bar{r}(R_c, f)$ curves for different values of f, for all 5 considered proteins. Colour denotes the number of independent counts.

Fig. 3 Left: normalised (for comparison) absolute number of transient contacts as a function of R_c. Right: radial distribution functions of permanent contacts.

Fig. 4 Schematic representation of interconnecting residues that form transient contacts at $R_c =$ 5 Å (green circle), as opposed to residues engaged in permanent neighbour shells (red circles).

Table 1: The summary of considered protein structures and conducted simulations. N_{res} - the number of amino acids, α / β / o – relative fractions of secondary structure elements (classified as: α helices, β sheets, and other).

protein	PDB id	runs [μs]	N_{res}	α / β / o [%]	R_c^* [Å]
PKA (active)	3fjq	2.7, 2.5, 2.8, 2.7, 1.2	343	34/14/52	0.50
PKA (inactive)	4dfy	1.1, 1,1, 1.3, 0.5, 1.4	337	34/12/54	0.50
LAO (open)	2lao	1.0	238	37/23/40	0.50
LAO (closed)	1lst	1.0	238	37/21/42	0.50
λ-repressor	1lmb	1.0, 1.0, 1.0	92	67/0/33	0.51
trypsin	4i8h	1.0, 1.0	223	10/34/56	0.50
SH3 domain	1shg	1.0, 1.0	60	4/46/50	0.50

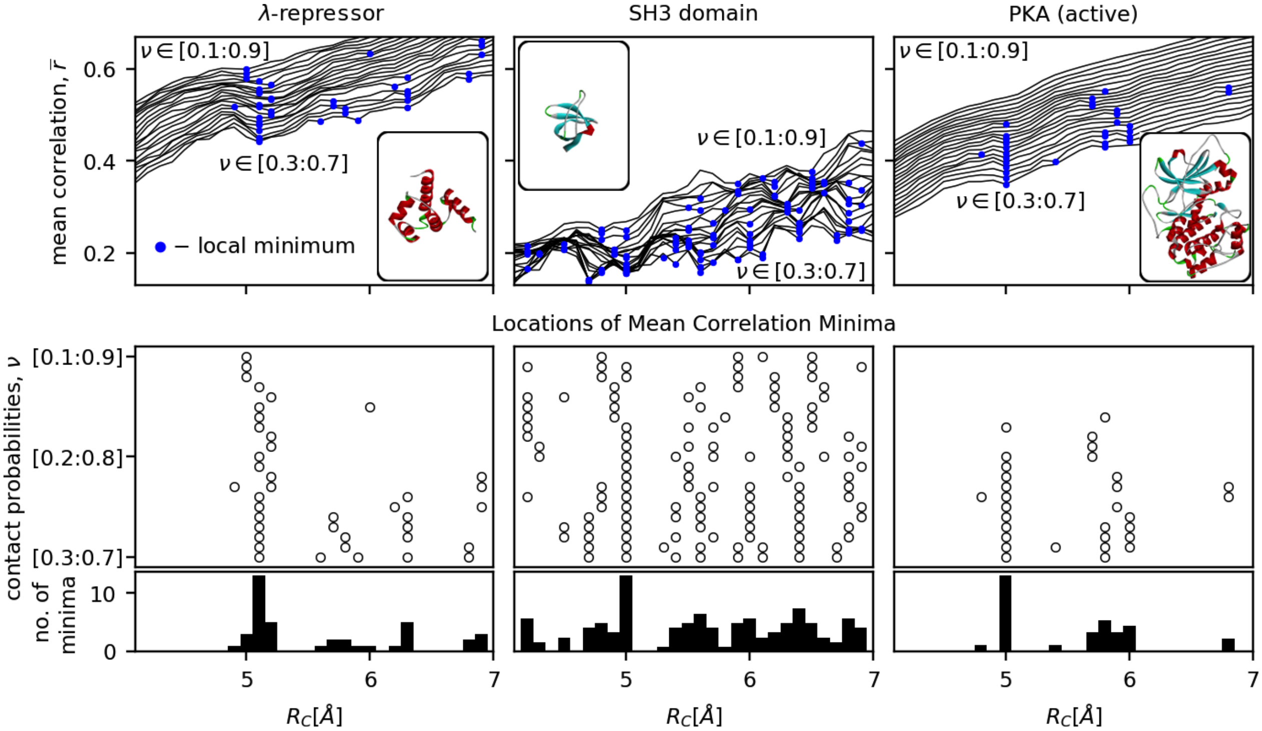

Figure 1: Upper row: sample $\bar{r}(R_c, f)$ curves for three proteins with different dominant secondary structure types. Lower row: points representing the occurrence of local minima in $\bar{r}(R_c, f)$ curves, and histograms gathering the numbers of minima at each R_c.

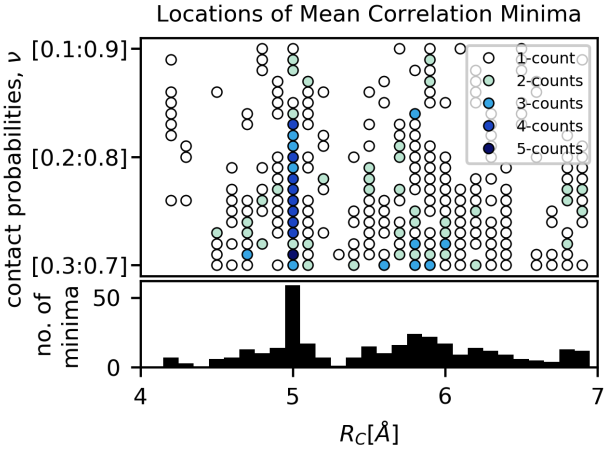

Figure 2: Points representing the occurrence of local minimum in $\bar{r}(R_c, f)$ curves for different values of f, for all 5 considered proteins. Colour denotes the number of independent counts.

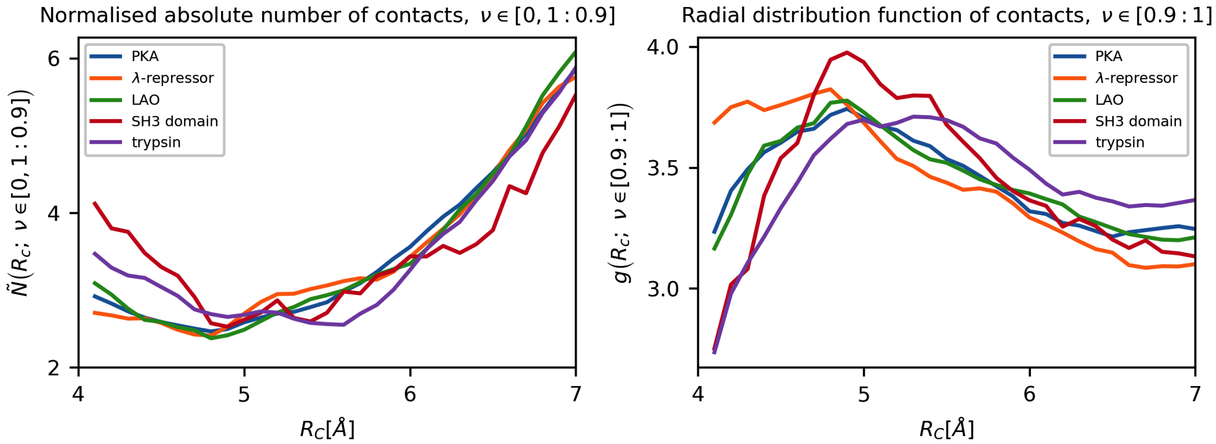

Figure 3: Left: normalised (for comparison) absolute number of transient contacts as a function of R_c. Right: radial distribution functions of permanent contacts.

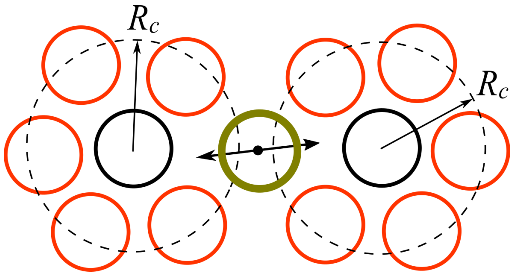

Figure 4: Schematic representation of interconnecting residues that form transient contacts at $R_c = 5$ Å (green circle), as opposed to residues engaged in permanent neighbour shells (red circles).