OTKA project K105415: The functional landscape of proteins
Final report

The sampling problem

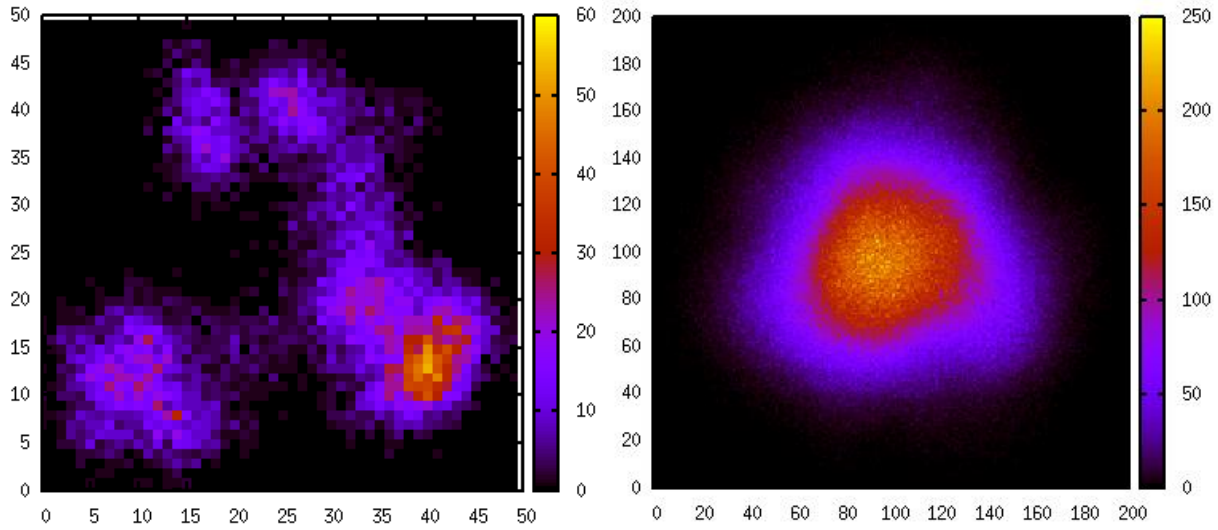
Protein molecules have many degrees of freedom, resulting in a high-dimensional configurational space. Describing the (native) free energy landscape of a protein thus involves sampling a high-dimensional probability distribution. This is a formidable problem, and despite tremendous efforts to develop "enhanced sampling" methods, it still remains a major stumbling block. Unfortunately, measuring sampling quality is often also difficult, and the choice is often between performing some kind of sampling in the hope that it will provide some insight, or giving up on the problem. Therefore, many studies have been published in the literature that perform analyses on various samples, ignoring the possibility that the sample may be insufficient to make any valid conclusion.

Because functional (native) free energy landscapes are much smaller than the full (folding) energy landscape of a protein, it was argued that exploring them is within the reach of current sampling/simulation techniques. In particular, Materese et al. published an influential paper in PNAS (1) presenting what was claimed to be a description of the native energy landscape of the 70-residue protein eglin c, using principle component analysis (PCA) in the space of torsion angles from a pool of molecular dynamics (MD) trajectories of a few hundred nanoseconds. This promising method offered a potentially general way to describe the functional landscapes of proteins. The analysis presented in the paper suggested that the native energy landscape of eglin c is hierarchically organized: clusters of conformations observed in the space of the first two principal components showed up as a collection of smaller clusters in the space of other principal components, and so on.

As this method appeared to be potentially widely applicable, we decided to test it out, and to try to reproduce the results. We performed extensive simulations with eglin c using MD, accelerated MD, and coarse-grained (MARTINI) MD sampling, performing the same type of analyses as in (1). However, we noticed a significant dependence of the results on the sample size, obtaining inconsistent representations of the energy landscape. Examining the trajectories in the space of the first two principal components of the torsion angles, we noticed that the trajectory never actually revisits earlier conformations, indicating poor sampling. Repeating the simulations several times, different results were obtained. These results hinted at a significant sampling quality issue.

As a model experiment, we performed sampling on oversimplified "toy" energy landscapes such as a single quadratic energy well or a completely flat surface (in multidimensional spaces). When these simulations were relatively short, we obtained quite complex, hierarchical "free energy landscapes" which actually only reflected a random walk on a featureless surface. Depending on the dimensionality, only fairly large sample sizes revealed the actual simple shape of the energy landscape.

Here, we present the comparison of a 10-ns vs. a 1000-ns coarse-grained trajectory of eglin c as they appear in the space of the first two principal components of the torsion angles:



Comparison of short-time vs. long time (10 vs 1000 ns) coarse-grained trajectories of eglin c in the space of the principal components of torsion angles.

As the figure shows, the short trajectory indicates a quite complex energy landscape, and further analysis reveals it to be hierarchical. However, as the simulation continues, the complexity vanishes, and we end up with just a single energy basin. (As this was a coarse-grained simulation using the MARTINI forcefield, the simplicity of the final landscape may be due to the lack of detail in the forcefield. However, the fact remains that better sampling results in a simpler energy landscape in this case.)

We conclude that is highly likely that Materese et al's results were due to insufficient sampling. The energy landscape of eglin c is unlikely to be as complex as presented; poor sampling produces an apparent "energy landscape" that is much more complex than it is in reality. There still are several ways to alleviate the problem of poor sampling, but there are various tradeoffs associated with every solution: the number of degrees of freedom, the size of the region to be sampled, the accuracy of the energy function, etc. has to be sacrificed.

Gaussian mixture model of the molecular probability density function

Assuming satisfactory sampling, the molecular probability density function (pdf) can be estimated, which is directly related to the energy landscape when the nature of the ensemble is known (we typically assume a canonical or isobaric-isothermal ensemble). Estimating the density is, however, not trivial. Starting from the idea that a quadratic energy well (which is the simplest form of an energy minimum) gives rise to a (multivariate) Gaussian distribution, we have developed a novel approach to describe energy landscapes: we approximate the molecular pdf with a mixture of multivariate Gaussians:

$$p_k(\mathbf{q}) = \sum_{i=1}^{k} w_i N(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i),$$

where the vector $\mathbf{q}$ represents the variables (in our case, torsion angles), k is the number of Gaussian components, w_i is the weight of the i-th component, and $N(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$ is the multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\sigma}_i$. The advantage of this estimate is that it is analytical, therefore easy to use for various estimations, and that it provides a smooth and realistic density function, informed by the knowledge that harmonic vibrations and therefore Gaussian distributions are ubiquitous in the statistical physics of macromolecules. Also, it has been proven that any function can be arbitrarily closely approximated with Gaussian mixtures.

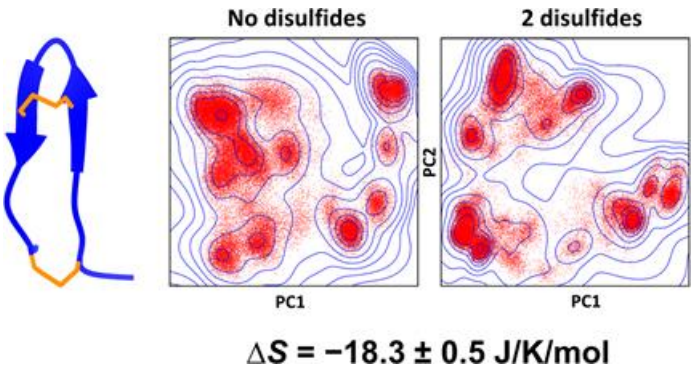
The estimation of a multivariate Gaussian mixture is, however, not trivial. Luckily, an efficient greedy learning method has been developed (2), which we were able to adapt to our purposes.

Combined with a cross-validation based stopping criterion, we have established a robust method that prevents both underfitting and overfitting. (3)

Entropy estimation using Gaussian mixtures

Once the Gaussian mixture model of an energy landscape has been estimated, it can be used for a number of analyses. The entropy of the probability distribution is easy to calculate from the Gaussian mixture, as it is simply a byproduct of the estimation algorithm. This can be converted to physical entropy in a straightforward way. When applied to probability distributions over torsion angles, the resulting absolute entropy is not physical, but the entropy differences between states are. Thus, the Gaussian mixture method gives us a practical way to calculate entropy differences between ensembles.

We have implemented the method as a program named GMENTROPY, and made it publically available (http://gmentropy.szialab.org). Careful testing on small peptide test systems have shown that the method yields more accurate entropy differences on smaller samples than other, competing methods. It also scales well to larger molecules: we have calculated the entropy difference between the disulfide-bonded and disulfide-less states of a 17-residue antimicrobial peptide (see figure), and the entropy difference between two sub-states of the native state of the 58-residue bovine pancreatic trypsin inhibitor (BPTI).



Tachyplesin is a 17-residue antimicrobial peptide with 2 disulfide bridges. We estimated the molecular probability density of two variants (shown projected on the first two principal components in torsion angle space), and estimated the entropy difference.

Application of entropy calculation on relative domain motion

Uroporphyrinogen-3 synthase (UP3S) is a segment-swapped protein, i.e. it consists of two domains that evolutionary formed from a domain-swapped dimer (4). This results in the presence of two linkers between the domains, rather than one. We hypothesized that this helps reduce the entropy cost of ligand binding as the ligand binds between the domains and fixes the relative position of the domains. We generated conformational ensembles representing relative domain motions in several ways, and applied our entropy calculation method to the angles describing domain orientation. Comparisons were made with in silico variants of the protein with only one linker between the domains. The results confirmed the favorable effect of the two linkers on the free energy of ligand binding (5).

Energy landscape analysis using Gaussian mixtures

In addition to calculating entropies, the Gaussian mixture based probability density model also offers a novel way to analyze the free energy landscape of biomolecules. As there is a strict correspondence between the peaks of the molecular probability density function and the local energy minima of the energy landscape, analyzing the landscape is equivalent to analyzing the probability density.

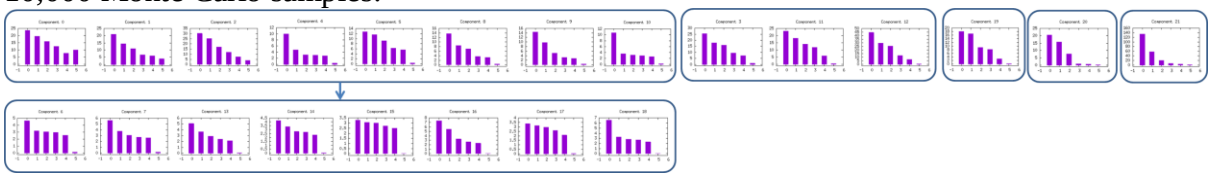
We have developed a set of algorithms to identify the various features of the energy landscape using the individual components of the Gaussian mixture. This also allows us to answer the difficult question whether an energy landscape is hierarchical.

The assumption behind this analysis is that each Gaussian component represents an underlying quasi-quadratic energy minimum on the energy landscape, even though there may not be an actual minimum on the landscape at that particular location due to the other components in the mixture which may suppress or shift the minimum. The first step of our analysis method is to associate a multidimensional ellipsoid with each Gaussian component. This "principal ellipsoid" is calculated from the covariance matrix of the Gaussian component, with its axes corresponding to the principal components and its border defined to be $n\sigma$ from the mean where σ is the standard deviation whose value depends on the direction with regard to the mean, calculated from the eigenvalues of the covariance matrix, and n is a predefined number (we used n=4.0 in our calculations). In the second step, we determine the overlaps of the ellipsoids with each other. Ellipsoids that have no overlap with any other ellipsoid correspond to separate energy minima on the landscape, with their locations corresponding to the mean of the Gaussian component. Then, we define energy basins as connected sets of overlapping ellipsoids. A basin is considered to be a wider area that may include several energy minima.

A further step of the analysis examines the relationships between individual ellipsoids and basins. An ellipsoid may lie completely inside another ellipsoid or a basin, thereby defining a lower level in the organization of the landscape. Nested sets of basins define a hierarchical landscape.

Are native energy landscapes hierarchical?

The hierarchical nature of free energy landscapes is a recurring theme in the literature (1, 6, 7), in the context of both folding landscapes and functional landscapes. However, different authors use different definitions of hierarchicity, and the algorithms used to define the hierarchy are often such that they will always yield a hierarchy regardless of how the landscape actually looks like (e.g. hierarchical clustering will always produce a hierarchical arrangement of clusters). In order to avoid tautology, and meaningfully ask the question whether a particular landscape is hierarchical, we need a stricter and consistent definition of hierarchicity. Our definition is based upon well-defined energy basins (as identified by our Gaussian mixture based algorithms), and the criterion that an energy basin is assigned to a lower level of hierarchy if it lies completely inside another (necessarily wider) energy basin. As an example, we show here the organization of the 5-dimensional free energy landscape of the Ala-Val-Ala peptide as obtained from a sample of 10,000 Monte Carlo samples:



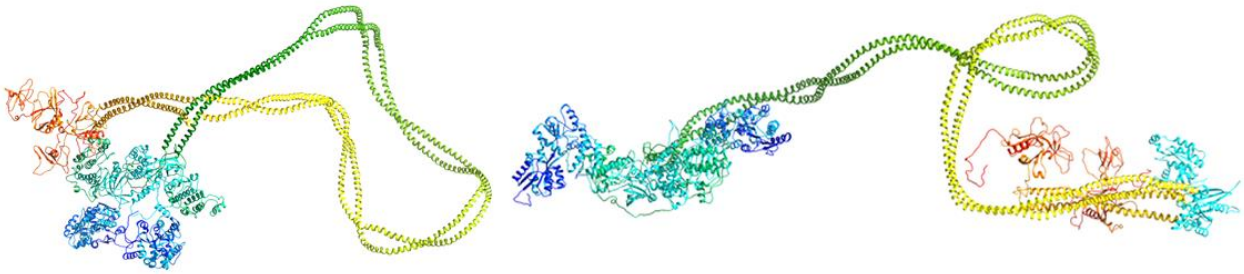
The organization of the free energy landscape of the Ala-Val-Ala peptide, as gleaned from Gaussian mixture fitting on 10,000 Monte Carlo samples. Rounded rectangles represent energy basins; the arrow indicates a step down in the hierarchy. Bar graphs represent individual Gaussian components, with the bars proportional to the lengths of the semi-axes of the principal ellipsoids; the last bar in each graph represents the weight of the component.

As the figure shows, this landscape has 6 energy basins organized into 2 levels of hierarchy; 3 basins are only composed of a single Gaussian component each. However, we found that when we increase the sample size, the hierarchy disappears, again indicating that some of the complexity observed here is only a result of insufficient sampling.

We have applied our algorithm to conformational ensembles of several systems, from small peptides with only 3 degrees of freedom up to proteins with over 200 degrees of freedom (i.e. torsion angles). In general, we found little hierarchicity, if any. Running the algorithm on a 51-component Gaussian mixture in 221-dimensional space obtained from a 1-millisecond trajectory of bovine pancreatic trypsin inhibitor (BPTI) (8) revealed 3 distinct energy basins but no hierarchy. When applied to ensembles obtained from coarse-grained simulations of eglin c (a 70-residue protein), two-level hierarchies were seen on short sub-trajectories but no hierarchy on the full trajectory. From extensive simulations, the (non-)disulfide-bonded 17-residue tachyplesin was found to have 7 (10) energy basins but no hierarchical organization. Although it is difficult to generalize based on this small number of systems, the results suggest that native free energy landscapes tend to have little or no hierarchy (apart from spurious hierarchy resulting from insufficient sampling). Hierarchicity may be more prominent in folding energy landscapes.

Using SAXS data to estimate the conformations of ROCK kinase

Although the energy landscapes of large protein molecules are expected to be very complex and inaccessible due to the high number of degrees of freedom, there are cases when a significant part of the landscape can be assumed to be nearly flat. This is the case when the molecule contains a long flexible part. The ROCK2 kinase molecule is a homodimer of 2 chains of 1388 residues. Much of the molecule is a coiled coil, which appears to be relatively flexible. The kinase domain is at the N-terminus but is usually inactive as it is assumed to be inhibited by the C-terminal domain binding to it. The RhoA protein binds to a region on the coiled coil, and there is some contradictory experimental evidence that it may activate the kinase by allowing the C-terminal domain dissociate from it. Thus, we were interested in comparing the energy landscape of ROCK in its RhoA-free and RhoA-bound state. Due to the flexibility of the coiled coil, we expect a relatively flat landscape. Thus, we were able to use geometric simulation (the FRODAN program (9)) to generate very large ensembles of conformations and filter them using available small-angle X-ray scattering data. The results confirmed that the RhoA-free form is more compact, with its termini close to each other, while the RhoA-bound form is more extended (see figure). We have also developed a protocol using geometric simulation to help interpret SAXS data and estimate structures. A publication presenting the results has been submitted.



Left: the most probable RhoA-free; Right: the most probable RhoA-bound structure of the ROCK2 kinase as selected using SAXS data from ensembles of conformations generated by geometric simulations

Solving the sampling problem by reducing the state space: coupled folding-binding of homodimers

As large molecules and molecular complexes have an astronomical number of possible states, sampling all of them to estimate the probability density is not feasible. But even the real molecules themselves are unable to sample all their states, which suggests that a full sampling should not be necessary to understand their function. In fact, large groups of microscopic states

are essentially equivalent. This suggests that sampling can be simplified if we merge those states that are functionally and structurally equivalent.

To test this idea, we have chosen the coupled folding and binding of homodimers of intrinsically disordered peptides. The source of intractability here is the huge size of the part of the state space reflecting the relative positions and orientations of the chains.

To reveal the whole dynamics of a two-protein system, all the microstates and the transitions between them must be known. They constitute a graph (kinetic network) the nodes of which are the microstates and the edges are the transitions. Analysis of the kinetic network provides insight into the mechanisms of dimer formation. Based on the foldedness state of the monomers at the moment of association, we defined three possible mechanisms of dimer formation: induced folding (association between two unfolded monomers); conformational selection (association between a folded and an unfolded monomer); and rigid docking (association between two folded monomers). This is an extension of the traditional binary classification of mechanisms of folding of an unstructured peptide coupled to its binding to a partner.

Two-layer kinetic network model

To perform exact calculations for the dimer formation, we reduced the dimensions of the state space belonging to the relative positions and orientations of the chains by compressing the relative positional and orientational subspace belonging to a particular conformational state into a unique associated state, i.e. a state where at least one interchain contact is present, and a unique dissociated state, i.e. a state where there are no contacts between the chains. Thus, we defined a network where each conformational state was represented by two nodes representing an associated and a dissociated state, respectively, with the associated and dissociated states constituting the two-layers of the graph.

The two assumptions underlying the definition of our two-layer kinetic network model are: *i)* the translational movements and the binding dissociation events are fast compared to the conformational transitions of the chains and *ii)* among microstates belonging to the same pair of monomer states, only the associated states of minimal energy are populated significantly. The second assumption corresponds to the fact that the partial energy landscape of binding between the particular conformations is funneled.

We combined our two-layer kinetic network model with both the two-dimensional HP lattice model (10) and the Wako-Saito-Muñoz-Eaton (WSME) ensemble-based model (11). Both studies concluded that all three dimer formation mechanisms (as defined above) are simultaneously present, but with widely different probabilities for different proteins; their relative importance also depends on external criteria such as concentration, time, or the type of process investigated (i.e. equilibrium, steady-state, transient, etc.).

WSME

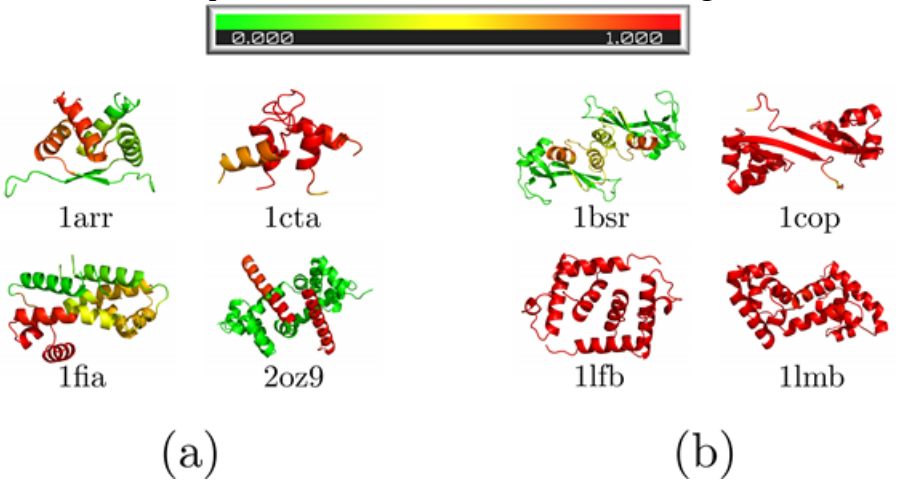
We applied our two-layer approximation in combination with the WSME model to eight homodimers, four of which are two-state and the other four are three-state homodimers.

Modified single sequence approximation. The problem of solving the WSME model (12) exactly is computationally intractable. On the other hand, exact solution is not capable of generating all the conformational states. Instead, often the single- or double-sequence approximation is used where only one or two stretches of folded residues are allowed. The single- and double-sequence approximations, however, provide very inaccurate estimations for the free energies. We modified the single sequence approximation by considering the entropic effect of non-specific residue folding out of the single sequence.

Calculations with the modified single sequence approximation for acylphosphatase (1aps) provides a free energy curve very similar to that provided by exact calculation in opposed to the original form of single-sequence approximation or the double sequence approximation.

Relative importance of the dimer formation mechanisms. The probabilities of the possible mechanisms were calculated for the eight homodimers by the Transition Path Theory (13). We found that for almost all of the eight proteins, all three mechanisms are present but often with very different probabilities.

Preformed structural elements. We investigated how folded the monomers are before they bind together, and found that even for the two-state sequences, some (occasionally even quite significant) residual structure is present in the monomer form, as the figure below shows.



The foldedness of particular residues at the moment of association for (a) the two-state and (b) the three-state dimers. Two-state dimers are rather unfolded when the two chains bind together, while three-state dimers are essentially fully folded with the exception of BS-RNase

Effect of concentration on the dimer formation mechanism. The effect of the protein concentration on the dominant mechanism was investigated. We found that as the concentration increased, the importance of rigid docking relative to induced folding also increased. For several proteins, the probability of conformational selection had a maximum at medium concentrations. Also, the proportion of preformed segments increased with the concentration.

HP lattice models

We applied our two-layer network model to two-dimensional HP (hydrophobic-polar) lattice model sequences to investigate the coupled folding and binding of intrinsically disordered peptides to an ordered complex. Sequences of 4-8 beads having a degenerated ground state as monomers and a unique native state conformation in the dimer were selected for investigation, representing sequences for which the monomers are disordered but the dimer is ordered.

Defining an energy function reproducing two-state folding behavior. As the adjacency-based energy function proposed by Lau and Dill (10) does not reproduce two-state folding behavior, we tested three additional energy functions: a distance-based energy function, a cluster-based energy function, and a “squared diagonal” energy function. In the distance-based energy function, the energy of a conformation depends on some negative power of the Euclidean distances between each pair of hydrophobic beads. In the cluster energy function, clusters of adjacent hydrophobic beads are defined and the distance-based energy between pairs of beads belonging to the same cluster are calculated. In the squared diagonal energy function, the energy of a conformation is the squared sum of distance-based energy terms calculated for pairs of beads for which the squared Euclidean distance less than 2.

The four energy functions were compared, and it was concluded that only the squared diagonal energy function exhibits two-state folding as shown by the ratio of *van't Hoff* and calorimetric enthalpies. Thus, the squared diagonal energy function was used in the further calculations.

Transition matrix. A transition matrix was built with the microstates of a two-chain system as nodes and transitions defined by the 'pull moves' move set as edges. A fixed version of the pull moves move set was used (14). Transition probabilities were calculated according to the Metropolis-Hastings criterion (15, 16).

Metastable states. We applied Perron Cluster Cluster Analysis (PCCA) methods (17, 18) to reveal the metastable states in the state space. We found that our sequences do not have a few metastable states containing many microstates but rather many metastable states each with a few microstates and having similar escape times.

Relative importance of dimer formation mechanisms. Calculations based on Transition Path Theory, and time-dependent flux calculations showed that each of the three mechanisms was present for all the sequences. The relative importance of the mechanisms was also dependent on whether a non-stationary, a steady-state of an equilibrium process was investigated. The dominant mechanism may also change in time as the process advances.

Symmetry of dimer formation. We investigated whether the immanent symmetry being present in the native structure of the studied homodimers manifests itself in the process of dimer formation. An asymmetry parameter for the two chains was defined and calculated as a function of time. For some sequences, the maximal asymmetry during the dimer formation significantly exceeded the equilibrium value of the asymmetry parameter indicating that asymmetry can be essential in the coupled folding and binding process.

Our results related to dimer formation mechanisms have been presented at several conferences, and a publication has been submitted (expected to be published early 2018).

Collaborations with other groups

During the project, we have also utilized our methods in several collaborations including the modeling of the effect of mutations on the Sleeping Beauty transposon (19), modeling of transposon proteins (20), kinetics of proteinase networks (21), allo-network drugs (22), template-based prediction of protein-protein interactions (23). Several projects with collaborators have produced further manuscripts that are submitted.

References

1. Materese, C.K., C.C. Goldmon, and G.A. Papoian. 2008. Hierarchical organization of eglin c native state dynamics is shaped by competing direct and water-mediated interactions. Proc. Natl. Acad. Sci. U. S. A. 105: 10659–10664.

2. Verbeek, J., N. Vlassis, and B. Kröse. 2003. Efficient greedy learning of Gaussian mixture models. Neural Comput. 15: 469–485.

3. Gyimesi, G., P. Závodszky, and A. Szilágyi. 2017. Calculation of Configurational Entropy Differences from Conformational Ensembles Using Gaussian Mixtures. J. Chem. Theory Comput. 13: 29–41.

4. Szilágyi, A., Y. Zhang, and P. Závodszky. 2012. Intra-chain 3D segment swapping spawns the evolution of new multidomain protein architectures. J. Mol. Biol. 415: 221–235.

5. Szilágyi, A., D. Györffy, and P. Závodszky. 2017. Segment swapping aided the evolution of enzyme function: The case of uroporphyrinogen III synthase. Proteins. 85: 46–53.

6. Noé, F., I. Horenko, C. Schütte, and J.C. Smith. 2007. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. J. Chem. Phys. 126: 155102.

7. Jain, A., and G. Stock. 2014. Hierarchical folding free energy landscape of HP35 revealed by most probable path clustering. J. Phys. Chem. B. 118: 7750–7760.

8. Shaw, D.E., P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, and W. Wriggers. 2010. Atomic-level characterization of the structural dynamics of proteins. Science. 330: 341–346.

9. Wells, S.A. 2014. Geometric simulation of flexible motion in proteins. Methods Mol. Biol. Clifton NJ. 1084: 173–192.

10. Lau, K.F., and K.A. Dill. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules. 22: 3986–3997.

11. Zamparo, M., and A. Pelizzola. 2006. Kinetics of the Wako-Saitô-Muñoz-Eaton model of protein folding. Phys Rev Lett. 97: 68106.

12. Bruscolini, P., and A. Pelizzola. 2002. Exact solution of the Muñoz-Eaton model for protein folding. Phys Rev Lett. 88: 258101.

13. Metzner, P., C. Schütte, and E. Vanden-Eijnden. 2009. Transition Path Theory for Markov Jump Processes. Multiscale Model Simul. 7: 1192–1219.

14. Györffy, D., P. Závodszky, and A. Szilágyi. 2012. “Pull moves” for rectangular lattice polymer models are not fully reversible. IEEEACM Trans Comput Biol Bioinform. 9: 1847–9.

15. Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21: 1087–1092.

16. Hastings, W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 57(1): 97–109.

17. Deuflhard, P. 2000. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. Linear Algebra Its Appl. 315: 39–59.

18. Deuflhard, P., and M. Weber. 2005. Robust Perron cluster analysis in conformation dynamics. Linear Algebra Its Appl. 398: 161–184.

19. Abrusan, G., Y. Zhang, and A. Szilagyi. 2013. Structure prediction and analysis of DNA transposon and LINE retrotransposon proteins. J Biol Chem. 288: 16127–38.

20. Abrusan, G., A. Szilagyi, Y. Zhang, and B. Papp. 2013. Turning gold into “junk”: transposable elements utilize central proteins of cellular networks. Nucleic Acids Res. 41: 3190–200.

21. Oroszlán, G., R. Dani, A. Szilágyi, P. Závodszky, S. Thiel, P. Gál, and J. Dobó. 2017. Extensive Basal Level Activation of Complement Mannose-Binding Lectin-Associated Serine Protease-3: Kinetic Modeling of Lectin Pathway Activation Provides Possible Mechanism. Front. Immunol. 8: 1821.

22. Szilágyi, A., R. Nussinov, and P. Csermely. 2013. Allo-network drugs: extension of the allosteric drug concept to protein- protein interaction and signaling networks. Curr. Top. Med. Chem. 13: 64–77.

23. Szilagyi, A., and Y. Zhang. 2014. Template-based structure modeling of protein-protein interactions. Curr. Opin. Struct. Biol. 24: 10–23.