The blind leading the blind: how disordered peptides form and ordered complex?

Györffy Dániel

Summary

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics

Supervisors:
Dr. Péter Závodszky
Dr. András Szilágyi

Budapest, 2014

1 Introduction

Proteins usually perform their function through molecular interactions. Since they require some particular interactions for proper working and since aspecific interactions may hamper the function, the proteins must recognize their biological targets accurately. The recognition basically rests on the complementarity of surface patterns of the partners. The complementarity, however, not a static property, in contrast to the assumption of the lock and key hypothesis of Emil Fischer [1]. Because the structure of proteins is intrinsically dynamic, models assuming the conformational transition of proteins via binding have emerged. Two mechanisms exist in the literature for the description of conformational transitions coupled to binding. According to the induced fit mechanism, proposed by Koshland [2], the binding of a partner induces the conformational transition of the protein toward its native conformation. In the conformational selection (fluctuation fit) mechanism, the conformations of the protein preexist in the unbound form and the partner selects the native conformation from the preexisting pool [3, 4]. As the binding shifts the relative populations of protein conformations, it is also called population shift mechanism.

It is very typical of a newly indentified class of proteins, intrinsically disordered proteins (IUPs), that they undergo significant conformational transitions through target binding. IUPs do not have well-determined secondary and tertiary structure in solution, but through binding to a partner, they often adopt well defined tertiary structure, which is called coupled folding and binding [5].

The concept of induced fit and conformational selection is appropriate for the description of binding and folding of an IUP to a partner not possessing significant flexibility. If, however, both chains taking part in the binding process are flexible and undergo a conformational transition, the traditional concept of coupled folding and binding are insufficient [6]. The picture is more complex, if the sequences of the partners are identical, because the issue of symmetry of the complex formation process arises.

According to the thermodynamic hypothesis of Anfinsen, the native state of a protein is the one with the lowest Gibbs free energy [7]. Identification of the native state is an optimization problem which is NP-hard for all two- and three-dimensional protein models [8, 9], which means that the time needed for the search increases very fast with increasing system size. Proteins are systems with a lot of degrees of freedom and a huge state space with complicated transitions between the states. The limits of calculation power of state-of-the-art computers make necessary the use of simplified protein models, which range from the discretization of the state space [10, 11], through the reduction of resolution of the chain representation [12, 10, 11] to the simplification of the energy function [13, 14, 15].

One of the most simplified model is the 2D HP (hydrophobic–polar) lattice model proposed by Lau and Dill [11]. In the HP square lattice model, the protein chain is represented as a self avoiding walk on a square lattice. Only two types of residues, a hydrophobic and a polar one, are discriminated. Only the H-H pairs adjacent in the lattice but not successive along the sequence contribute to the energy with a non-zero term. For HP chains, exhaustive enumeration can be made to discover the full state space up to chains of $l = 25$ bead long. HP models can be used in Monte Carlo simulations, as well, and several move sets are available for this purpose [16, 17]. In my work, I used the move set called "pull moves" proposed by Lesh and coworkers, which contains irreversible moves in the original form, but can be easily made reversible [18].

For the thermodynamic and kinetic description of complex systems with a huge number of degrees of freedom, one traditionally used the free energy landscape perspective. We obtain the free energy landscape to depict the free energy of the system as a function of some predefined reaction coordinates. If we have one reaction coordinate, we obtain a free energy curve, in the case of two reaction coordinates, we obtain a free energy surface, and with three or more reaction coordinates, we obtain a free energy hypersurface. The main bottleneck of the free energy landscape perspective is that, because some kinetically well separated states can coexist at the

same point of the reaction coordinate space, some kinetic barriers may be invisible [19]. This problem can be solved by the methods of transition networks. Mathematically, the transition network is a Markov model, which is a weighted directed graph, the nodes of which are—in the case of a countable state space—the microstates of the system, and the directed edges represent the transitions between them. The edge weights are the probabilities of the transitions. (In the case of an uncountable state space, the nodes of the graph can be obtained by some discretization of the state space [19, 20]). The probability distribution of microstates of a system with constant temperature and pressure is given by the

$$p_i = \frac{\mathrm{e}^{-E_i/k_BT}}{\sum_j \mathrm{e}^{-E_j/k_BT}} \tag{1}$$

Boltzmann distribution, where E_i is the energy of the i state and k_B is the Boltzmann constant. I chose the weights of the transition graph so that they satisfy the Metropolis–Hastings criterion [21].

With the investigation of the transition matrix representation of the transition network, we can identify the metastable states of the system [22]. The methods based on Transition Path Theory are appropriate for the investigation of the kinetics of the system. With the help of Transition Path Theory, we can calculate the reaction rate for any transition between arbitrarily defined sets of microstates. For steady-states of system, the rate constant can be calculated from the reaction rates.

Investigating systems consinsting of more than one chains is particularly problematic because their state space is far larger compared to the state spaces of one chain systems. This stems both from the fact that the chains in the systems explore the same state space as the corresponding monomers and that the chains can be positioned in many ways relative to each other.

The aim of my work is to thermodynamically and kinetically characterize the homodimer formation of IUPs, to elucidate the role of different mechanisms in the dimer formation and to investigate the applicability of concepts originally worked out for conformational transitions coupled to binding of only one flexible participant to the case where more than one partners have significant flexibility.

2 Summary of new scientific results

2.1 Investigating the coupled folding and binding of disordered chains by exact calculations, using an unbiased model

Dimer formation of proteins has been widely investigated with computational methods, but some simplifications had to be applied because of the huge computational requirement. Previously formation of homodimers was investigated by structure based methods (e.g. Gō-model) [23, 24]. Disorder to order transition through binding to an ordered target was investigated by a HP model[25]. However, to my knowledge, this is the first investigation of the mutual folding of two disordered peptide chains coupled to their binding without some information about the native structure built in the energy function.

2.2 Two-layer state network model

The state space of systems consisting of two chains is huge and current computational capacity only allows building the whole transition network up to for $l = 5$ chain length. Applying two simple assumptions, I was able to reduce the size of the state space and establish a two-layer Markov model which made it possible to investigate dimers of chains up to 8 beads of length. In the two-layer model, two states, an associated and a dissociated, belong to each conformation pairs, and they constitute the two layers of the model.

2.3 Method for the calculation of steady-state probabilities in a Markov model

Based on the conservation of material in the steady-state for any microstate, I set up a linear system of equations the solutions of which are the corresponding steady-state probabilities.

2.4 The three mechanism of the coupled folding and binding of IUPs

According to the folding state of the binding partners at the moment of association, three mechanism can be distinguished. In rigid docking, both chains are in the native conformation. If both chains are unfolded, the mechanism is induced folding. If one of the chains is folded and the other is unfolded, the mechanism is conformational selection. Applying Transition Path Theory, I calculated the reactive fluxes through each mechanism in equilibrium and steady-state, together with instantanous fluxes as a function of time. The results show that the dominant mechanism for a particular sequence varies with time and strongly depends on the macrostate the system is in.

2.5 Comparison of equilibrium and steady-state fluxes

Living cells are not equilibrium systems but we can assume for some processes that they are in a steady-state. I compared the behaviour of the system investigated in my work in equilibrium and steady-state. Although it may not be general, for my system the two behaviours did not differ significantly.

2.6 The two chains do not behave fully symetrically in the course of homodimer formation

Theoretical descriptions of homodimer formation, except some works (e.g. [23]), have not dealed with the issue of symmetry of the coupled folding and binding process of identical chains. Moreover, the traditional two- and three-state view of homodimer formation implicitly assumed the

symmetry of the process. I investigated how symmetric the behaviour of the two chains is and I found that some amount of asymmetry arises for every sequence, but for some particular sequences the asymmetry is very expressed at the beginning of complex formation.

2.7 The size of the overlapping region between regions occupied by ordered and disordered sequences on the average hidrophobicity–absolute average net charge plane decreases with the chain length

It was observed several times that for disorder-prediction methods, relying on the characteristic differences between the amino acid compositions of ordered and disordered sequences, the accuracy of prediction increases with increasing length of the disordered regions to be predicted [26, 27, 28]. One of the main reasons may be the fact that although the ordered and disordered sequences occupy distinct regions of the amino acid composition space, these regions overlap. The separation and the overlap can be observed if the amino acid composition space is mapped onto the average hidrophobicity–absolute average net charge plane [29]. In our work, we show that the overlapping region, called *twilight zone*, becomes narrower and narrower as the length of the sequence increases. The position of the twilight zone on the average hidrophobicity–absolute average net charge plane does not vary significantly with chain length. HP and HPN model proteins also show the decrease in the size of the twilight zone with increasing chain length. In addition, if chain length dependent contact energies are used then the position of the twilight zone does not vary with chain length [30].

2.8 The "pull moves" Monte Carlo move set is not fully reversible

Lesh and coworkers introduced an ergodic, local and—according to the statement of the authors—reversible Monte Carlo move set, called "pull moves" [17]. I proved that the move set originally proposed contains irreversible moves. Because of the irreversible moves, the system does not satisfy the *detailed balance*, which can lead to improper sampling when the move set is used in sampling methods. I suggested a modification of the move set which makes it fully reversible. I showed that in some conditions, the flawed move set does not lead to improper sampling. I showed by Wang–Landau sampling that for short chains, irreversibility can cause severe deviations from the exact values [18].

3 Outlook, the applicability of results of the work

In my work, a methodological and a theoretical result are worthy of further consideration. Based on two simple assumptions, I defined a two-layer Markov model to significantly reduce the state space of the dimeric system. This two-layer Markov model can be used for analysis of results obtained from calculations with protein representations less simplified than that used by me. In the case of systems consisting of more than one protein chains, the contribution of the huge number of relative positions enlarges the state space of the system and so the size of the corresponding Markov model. With the application of the two-layer model, the significant decrease of system size can be achieved which makes possible more detailed investigation of the conformational space.

For the description of complex chemical reactions, one often uses kinetic schemes. In experimental studies, the states of the kinetic scheme are defined based on some measurable parameter. It is not guaranteed, however, that these states are true metastable states of the system. The rate constants in the kinetic schemes are true kinetic constants, that is their values do not depend on the instantanous state of the system, only if the defined states are true metastable states. If not, the measured values are rate constants only under special conditions such as equilibrium or steady-state.

Related publications

1. Szilágyi A, Györffy D és Závodszky P. "The twilight zone between protein order and disorder." *Biophys J*, **2008**;95:1612–26.

2. Györffy D, Závodszky P és Szilágyi A. ""Pull moves" for rectangular lattice polymer models are not fully reversible." *IEEE/ACM Trans Comput Biol Bioinform*, **2012**;9:1847–9.

Poster presented in international conference

Györffy D, Závodszky P és Szilágyi A.: "The blind leading the blind: how disordered peptides form an ordered complex". 3rd Prague Protein Spring meeting, Prága, 2014

References

[1] Fischer E. "Einfluss der configuration auf die Wirkung der Enzyme". *Ber Dtsch Chem Ges*, **1894**;27:2985–93.

[2] Koshland DE. "Application of a theory of enzyme specificity to protein synthesis." *Proc Natl Acad Sci U S A*, **1958**;44:98–104.

[3] Straub FB. "Formation of the secondary and tertiary structure of enzymes." *Adv Enzymol Relat Areas Mol Biol*, **1964**;26:89–114.

[4] Boehr DD, Nussinov R és Wright PE. "The role of dynamic conformational ensembles in biomolecular recognition." *Nat Chem Biol*, **2009**;5:789–96.

[5] Dyson HJ és Wright PE. "Coupling of folding and binding for unstructured proteins." *Curr Opin Struct Biol*, **2002**;12:54–60.

[6] Piana S, Lindorff-Larsen K és Shaw DE. "Atomistic description of the folding of a dimeric protein". *The Journal of Physical Chemistry B*, **2013**;117(42):12935–12942.

[7] Anfinsen CB. "Principles that govern the folding of protein chains." *Science*, **1973**; 181:223–30.

[8] Unger R és Moult J. "Finding the lowest free energy conformation of a protein is an np-hard problem: proof and implications." *Bull Math Biol*, **1993**;55:1183–98.

[9] Fraenkel AS. "Complexity of protein folding." *Bull Math Biol*, **1993**;55:1199–210.

[10] Skolnick J, Kolinski A és Yaris R. "Monte Carlo simulations of the folding of β-barrel globular proteins." *Proc Natl Acad Sci U S A*, **1988**;85:5057–61.

[11] Lau KF és Dill KA. "A lattice statistical mechanics model of the conformational and sequence spaces of proteins". *Macromolecules*, **1989**;22:3986–3997.

[12] Clementi C, Nymeyer H és Onuchic JN. "Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins." *J Mol Biol*, **2000**;298:937–53.

[13] Krantz AT. “Analysis of an efficient algorithm for the hard-sphere problem”. *ACM Trans Model Comput Simul*, **1996**;6(3):185–209.

[14] Dokholyan NV, Buldyrev SV, Stanley HE és Shakhnovich EI. “Discrete molecular dynamics studies of the folding of a protein-like model.” *Fold Des*, **1998**;3:577–87.

[15] Taketomi H, Ueda Y és Gō N. “Studies on protein folding, unfolding and fluctuations by computer simulation. i. the effect of specific amino acid sequence represented by specific inter-unit interactions.” *Int J Pept Protein Res*, **1975**;7:445–59.

[16] Chain HS és Dill KA. “Energy landscape and the collapse dynamics of homopolymers”. *J Chem Phys*, **1993**;99:2116–2127.

[17] Lesh N, Mitzenmacher M és Whitesides S. “A complete and effective move set for simplified protein folding”. In “Proceedings of the seventh annual international conference on Research in computational molecular biology”, RECOMB ’03. ACM, New York, NY, USA, 188–195. URL `http://doi.acm.org/10.1145/640075.640099`.

[18] Györffy D, Závodszky P és Szilágyi A. “"Pull moves" for rectangular lattice polymer models are not fully reversible.” *IEEE/ACM Trans Comput Biol Bioinform*, **2012**;9:1847–9.

[19] Noé F és Fischer S. “Transition networks for modeling the kinetics of conformational change in macromolecules.” *Curr Opin Struct Biol*, **2008**;18:154–62.

[20] Chodera JD, Singhal N, Pande VS, Dill KA és Swope WC. “Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics.” *J Chem Phys*, **2007**;126:155101.

[21] Hastings WK. “Monte Carlo sampling methods using Markov chains and their applications”. *Biometrika*, **1970**;57(1):97–109.

[22] Deuflhard P. “Identification of almost invariant aggregates in reversible nearly uncoupled markov chains”. *Linear Algebra and its Applications*, **2000**;315(1-3):39–59.

[23] Levy Y, Cho SS, Onuchic JN és Wolynes PG. “A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes.” *J Mol Biol*, **2005**;346:1121–45.

[24] Levy Y, Wolynes PG és Onuchic JN. “Protein topology determines binding mechanism.” *Proc Natl Acad Sci U S A*, **2004**;101:511–6.

[25] Gupta N és Irbäck A. “Coupled folding-binding versus docking: a lattice model study.” *J Chem Phys*, **2004**;120:3983–9.

[26] Li X, Romero P, Rani M, Dunker AK és Obradovic Z. “Predicting protein disorder for n-, c-, and internal regions.” *Genome Inform Ser Workshop Genome Inform*, **1999**;10:30–40.

[27] Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ és Dunker AK. “Predicting intrinsic disorder from amino acid sequence.” *Proteins*, **2003**;53 Suppl 6:566–72.

[28] Melamud E és Moult J. “Evaluation of disorder predictions in casp5.” *Proteins*, **2003**;53 Suppl 6:561–5.

[29] Uversky VN, Gillespie JR és Fink AL. “Why are "natively unfolded" proteins unstructured under physiologic conditions?” *Proteins*, **2000**;41:415–27.

[30] Szilágyi A, Györffy D és Závodszky P. “The twilight zone between protein order and disorder.” *Biophys J*, **2008**;95:1612–26.