CURRENT APPROACHES TO DISENTANGLE THE MYSTERY OF KNOTTED PROTEIN FOLDING

PAWEL DABROWSKI-TUMANSKI[1,2], MATEUSZ SKLODOWSKI[1,2] AND JOANNA I. SULKOWSKA[1,2]

[1] *Faculty of Chemistry, University of Warsaw*
Pasteura 1, 02-093 Warsaw, Poland

[2] *Centre of New Technologies, University of Warsaw*
Banacha 2c, 02-097 Warsaw, Poland

(received: 2 September 2016; revised: 30 September 2016;
accepted: 7 October 2016; published online: 21 October 2016)

Abstract: The folding of knotted proteins remains a mystery both for theoreticians and experimentalists. Despite the development of new models, the driving force for self-tying remains elusive and the principle of minimal frustration cannot be reproduced *in silico*. In this paper we review different models used to understand protein self-knotting and suggest, how to improve the structure based model to observe efficient folding. Our preliminary results show, that including information about some amino acids properties, or reducing the set of physical contacts may be beneficial for modeling of the knotted protein folding.

Keywords: Knotted proteins, minimal contact map, YibK

DOI: https://doi.org/10.17466/tq2016/20.4/f

1. Introduction

Proteins are the most advanced and versatile products of evolution on the macromolecular level. The great diversity of functions they fulfill stems from a vast number of sequences and a large spectrum of physicochemical properties of protein building blocks – amino acids. However, this great diversity, being a blessing for the cell machinery, is also a curse for researchers trying to understand the protein folding mechanism. As the folding is in most cases much faster than the time resolution of any structure-determining technique, there is no other way than to reproduce it within some reasonable model. The most reliable calculations are all atom and explicit-solvent molecular dynamics simulations. These are however highly demanding computationally, and applicable for now only to either very

fast processes or very small systems. Hence, this is the place where the coarse graining (CG) enters.

The CG models differ in the type of pseudoatoms (substituting some specified groups of atoms) and the description and nature of interactions (the force fields). The pseudoatoms may cover the part of the side-group (*e.g.* CABS [1]), whole side-groups (*e.g.* UNRES [2, 3]) whole amino acids (the Cα model *e.g.* [4]), or even groups of residues. The force fields may be derived purely from physical interactions, based on statistical knowledge, or the dynamics may be steered by so-called native contacts (Structure Based Models – SBM). The last approach is justified by the principle of minimal frustration introduced by Bryngelson, Onuchic, Socci and Wolynes [5], according to which the proteins are optimized by evolution in such a way, that the native contacts steer the folding. For more details concerning possible models we recommend a recent review by Kmiecik *et al.* [6].

The large variety of coarse grained approaches and their advantages has led to many discoveries in micro- and milisecond-time processes, such as protein folding, protein-protein interactions, rearrangement upon ligand binding, etc. Finally, multiscale modeling, which includes the CG stage, was appreciated by the Nobel Committee (the Nobel prize in Chemistry, 2013). However, there are some exceptional processes, which although occur in the cell in the timescales available for the CG models, are still difficult to be reproduced *in silico*. One of such processes is efficient folding of knotted and slipknotted proteins. The schematic representation of a knot and a slipknot on an open chain is shown in Figure 1.

Figure 1. Schematic representation of entangled structures discussed in this work: knot (left) and slipknot (right); The green color indicates a knotted core, black – knot tails, and gold a slipknot loop

The existence of knotted proteins was hypothesized in 1994 by Mansfield [7], and in 1996 the first structures were identified by Takusagawa [8]. In 2000, the first protein containing a deep knot, i.e. a knot with at least 15 amino acids (represented by black curves in Figure 1) extending beyond its core, was identified by Taylor [9], whereas the most entangled protein was identified in 2010 by Bölinger [10]. The first slipknotted proteins were identified by Yeates in 2007 [11]. The comprehensive classification of knotted and slipknotted proteins was proposed by Sulkowska in 2012 [12]. Currently, based on the KnotProt database [13] it is known, that knotted and slipknotted structures constitute approx. 1.5%

of all protein structures and they can be categorized into five topologically different structures [12]. However, despite many theoretical and experimental efforts, the knotted protein folding has still much to reveal [14–22] and the proposed mechanisms based on experiments require theoretical reproduction and elucidation of details that are not accessible experimentally.

In spite of many attempts, the theoretical models which allowed exploring the free energy landscape of the smallest knotted protein, MJ0366, a structure 92 residues long, from *Methanocaldococcus jannaschii* (Figure 2, top-left structure), cannot be adjusted for deeply knotted proteins. Therefore, one may ask, what else can be done to entangle the mystery of knotted protein folding. In this work,

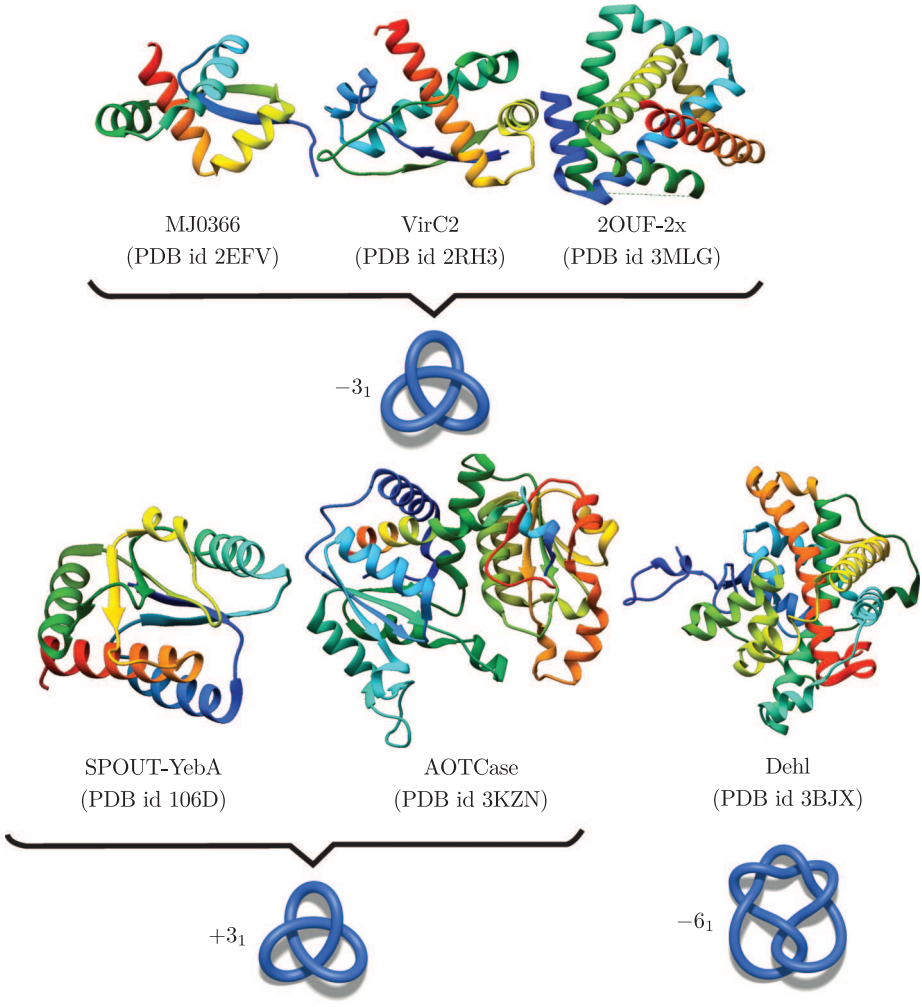

Figure 2. Cartoon representation of knotted protein structures investigated in different *in silico* studies; To our knowledge, the folding mechanism of proteins with 4_1 and 5_2 knot have not been studied *in silico* so far; The dashed line in the structure of 2OUF-2x is a gap in the crystal structure

we try to answer this persistent question. First, we review the successful models of knotted protein folding and show that most of them are simple Cα or all-atom structure based models. Next, we propose some extensions to the Cα structure based model with promising, but still preliminary results in the subject. Finally, we discuss the results, ways of improvement and other possible approaches.

2. Methods

Proteins In simulations, we used the crystals structure of YibK (PDB id 1J85).

The model We used a coarse-grained structure based model, where each single amino acid was represented by a Cα atom implemented in a SMOG server [23] with a Gaussian potential for attracting interaction between beads in native contacts. The approaches used differed only in the contact maps definition. We performed molecular dynamics simulation using the Gromacs 4.5.4 package, as described before [24]. The starting conformation was obtained as a random frame in a high-temperature unfolding simulation, for which the fraction of native contacts (Q) was lower than 0.2.

The maps Two different maps were used in the simulations. In the minimal map approach, the contacts forming the β-strands and those in the vicinity of the knot were chosen out of the SMOG-based map (see Figure 3, below the diagonal). In what we refer to as the Miyazawa-Jernigan approach, the contacts obtained from the SMOG server were scaled according to the quasi-chemical energies determined by Miyazawa and Jernigan [25]. The energies were scaled linearly with the requirement that the lowest (most negative) energy should correspond to the highest strength of the contact, while the strength of the contact corresponding to the most positive (most destabilized) energy was equal to 0. To retain the balance between the contribution of native contacts and other parts of the model, the strength was scaled with the restriction that the average strength should be equal to 1 (as in the original SMOG model).

Knot detection and visualization The presence of the knot was determined using the method described in detail in [12, 26, 27]. Molecular graphics and analyses were performed with the UCSF Chimera package [28].

3. Existing knotting models

As stated in the introduction, theoretical analysis of folding of knotted proteins concentrates mainly on the self-tying of the smallest knotted MJ0366 protein, most often using the truncated structure at both termini with PDB id 2EFV (Figure 2, top-left structure). The MJ0366 protein model was used in the all atom explicit solvent simulation on the Anton supercomputer to test if the protein could self-tie via a slipknot configuration [29], threading the protein terminus in the bent conformation through a twisted native loop, as it was suggested also for other proteins [30, 31]. Qualitatively the same results were also obtained using the Cα and all atom [24] Structure Based Models [24, 32–34], showing

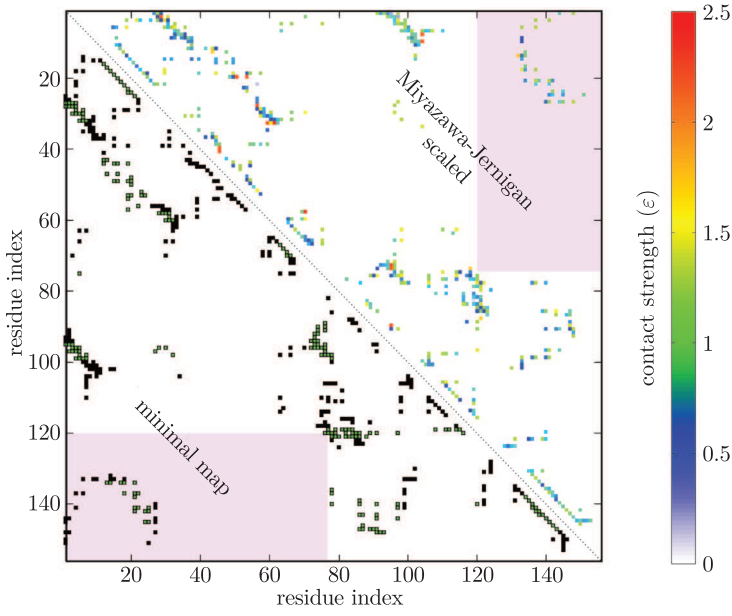

Figure 3. Comparison of used contact maps. Below the diagonal, the original SMOG-based map (black squares) and the minimal map – green dots; The strength of each contact was equal to 1; Above the diagonal, the Miyazawa-Jernigan scaled SMOG-based contact map – the strength of each contact was scaled according to the quasi-chemical potential determined by Miyazawa and Jernigan (details in the text); The strength of each contact is described by different color, explained by the bar in the right; The light pink rectangles denote the position of the knotted core, determined by the KnotProt database

that the mechanism to overcome topological barrier (threading through a twisted loop) via the slipknot or the plugin (straight) conformation, depended on the number of threading residues. The Cα coarse graining was also used including the ribosome effect during cotranslational protein folding [35], or with the steric confinement [36]. An interesting approach was used in [37], where the force field used to steer the dynamics of Cα beads was developed based on the rigidity of the chain in the MC procedure. The atom-level approach was also used with the algorithm efficiently choosing the path leading to the native structure [38], however, such approach cannot sample all the available configurational space.

There are only a few examples of folding simulations of proteins with a deep knot, showing that the techniques used for MJ0366 cannot be easily generalizable. The most thoroughly studied experimentally deeply knotted proteins are members of the SPOUT family: proteins YibK and YebA (Figure 2, bottom-left structure). Therefore, they were also tested several times in simulations: pure native centric Cα model revealed that those proteins could self-tie [30], however, only a few successful trajectories were observed [30, 39, 40]. An addition of the cavity simulating the ribosome in cotranslational folding enhances the folding rate by an order of magnitude [16]. The highest fraction of the folded structure is obtained upon addition of non-native contacts [31], however, it results in irreversible folding.

All other studies of knotted proteins were done in Cα Structure Based Models. These models allowed to uncover the free energy landscape of artificially designed proteins [39]. For the Vir2C protein (Figure 2, top-center panel), structurally similar to MJ0366, folding kinetics was explained in [32] and in the Cα model driven by the stiffness of the main chain [37]. The N-acetyl-L-ornithine transcarbamylase (AOTCase, Figure 2, bottom-center panel) was analyzed in Cα without [41] or with addition of non-native contacts [42]. Addition of non-native contacts allows to achieve 1–2% higher folding rates. The most complex protein folded in simulations to date is the haloacid dehalogenase (DehI) with -6_1 knot (Figure 2, bottom-right panel), analyzed also with a Cα native-centric model [10].

4. Possible developments and promising results

The models used to fold and tie knotted proteins fall into two categories. These are either explicit models which are too demanding computationally even for the smallest knotted protein or Cα Structure Based Models with possible addition of some external factors (like the cavity modeling the ribosome). There are only singular exceptions, like the rigidity based force field of Najafi *et al.* [37]. In terms of the folding funnel theory which is the basis and justification of SBM, the failure to fold the knotted protein with high efficiency translates in a rough energy landscape, forcing the protein to explore local minima instead of reaching the global minimum corresponding to the native fold. Let us recall here that experimental studies indeed suggest that proteins can self-tie, albeit slowly [43]. From this perspective, any changes in the model should aim at smoothing the energy landscape. How can one smooth the folding funnel? Basically, one can either remove local minima by increasing their energy, or decrease the barrier between the minima. Both approaches are schematically depicted in Figure 4.

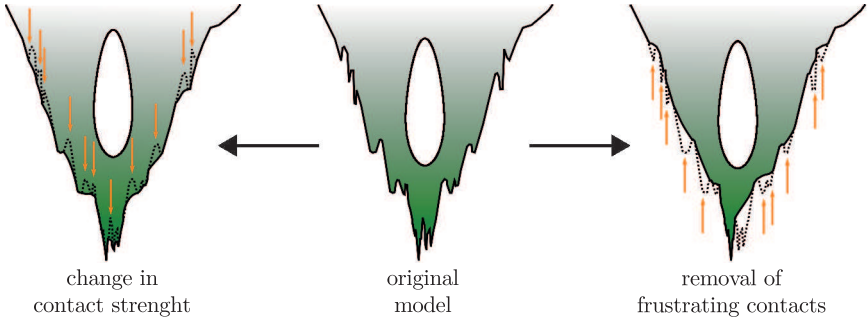

Figure 4. Schematic depiction of changes in a funneled landscape upon manipulation with the contact map; Changing the contact strength – introducing heterogeneity – may decrease the free energy barrier between different minima (left panel); Eliminating contacts creating energetic traps removes the minima (right panel); The empty space emphasizes the bifurcation in the landscape in the case of protein with a non-trivial topology; Tying the knot requires threading one terminus across a twisted loop, in consequence, not all routes are allowed

The addition of non-native contacts or interaction with the cavity on one hand may stabilize the native state, on the other, it lowers the energy barrier between the minima. The same effect may be achieved by manipulating with the contact strength. Increasing the interaction stabilizes while decreasing it destabilizes a state in which the particular contact is formed. In general, constructing a force field by seeking the optimal contact strength to tie a knotted protein is an exciting, although very challenging task. Instead of adjusting each contact strength separately, one may use the knowledge-based approach, utilizing *e.g.* the energies of quasi-chemical interactions of pairs of amino acid residues calculated by Miyazawa and Jernigan [25, 44]. To introduce heterogeneity in the strength of interactions we used Miyazawa and Jernigan potentials. We performed its simplest implementation with linear scaling of the contact strength (described in details in the Methods section). As our aim was to observe if the knotting probability would increase in such a model, we conducted 100 constant-temperature simulations at different temperatures, below T_m, starting from an unfolded conformation, obtained first in high-temperature unfolding (see the Methods section). In several temperatures we observed correct folding of the YibK protein following the same slipknot mechanism observed in earlier works [30, 39] (Figure 5).

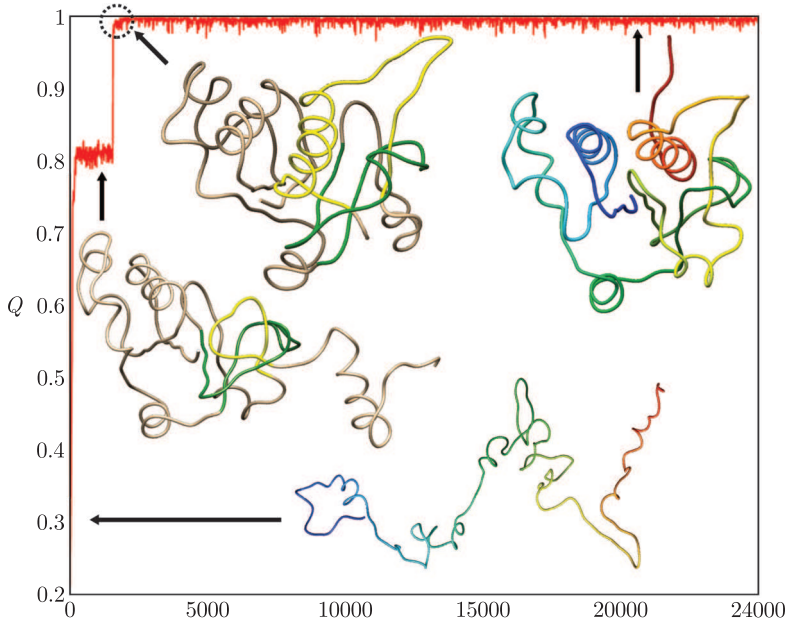

Figure 5. The fraction of native contacts (Q) during typical successful folding trajectory of YibK methyltransferase (PDB id 1J85) with the $+3_1$ knot (the protein discussed in this work); In the plot the snapshots of the trajectory corresponding to each part of the plot are shown: an unfolded protein (bottom-right panel), a slipknot conformation (left panels) and a folded native structure of YibK (top-right panel); In the slipknot conformations the green color indicates a twisted loop and yellow a threading loop; Note, that the top-left panel corresponds to the short-living metastable state (denoted as the dashed circle), which turns into a folded state upon threading the very C-terminus through the loop

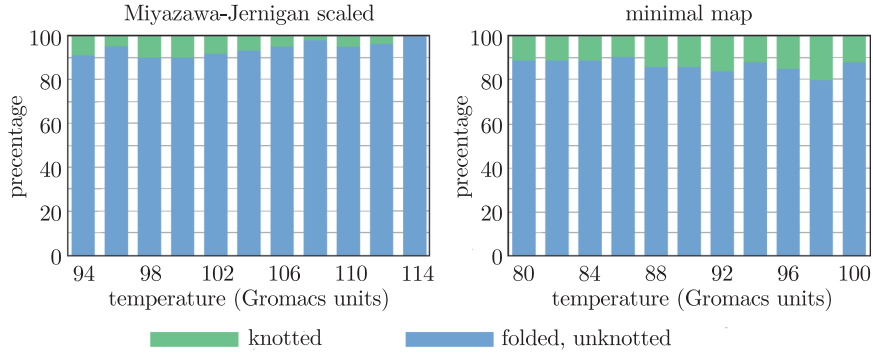

Figure 6. Histograms of final states of proteins depending on the temperature observed in two versions of a structure based model; Left panel: Implementation of Miyazawa-Jernigan quasi-chemical interaction energy; Right panel: The minimal map approach

The percentage of successful folding trajectories (finishing in the native, knotted state) depends on temperature. In one temperature (94 Gromacs units) the folding was observed in 9% of trajectories – see Figure 6, left panel. It shows, that even the unoptimized, crude implementation of the residue properties in the native contact map smoothes the free energy landscape and facilitates the folding of the knotted protein.

On the other hand, one can smooth the folding landscape by removing the local minima. These may be introduced by addition of artificial contacts, stabilizing some non-native states. For example, in the case of knotted proteins, topological constrains apart from local contacts also stabilize the fold of the protein. In consequence, uniform treatments of all interactions in protein can lead to too dense network of contacts. Therefore, to eliminate such minima, one should remove the unnecessary contacts. In practice, determining which groups of contacts are not necessary for folding is again a tedious and not-well defined task. Still, some native contacts are clearly indispensable, like the β-sheet forming contacts. In the case of self-tying, one may also guess, that the contacts in the vicinity of the knot may be at least a part of the self-tying driving force. We refer to the selection of those contacts from the original contact map as a minimal map. The minimal map approach has been utilized so far in the case of the smallest knotted protein [33]. The results have shown that in such an approach, the free energy barrier between the unfolded and folded states decreases, facilitating and speeding up the folding with retention of the folding mechanism compared to the explicit model [29]. Using the same approach in the case of YibK methyltransferase led to the observation of knotting in each series of constant-temperature simulations, reaching up to 20 properly folded structures for some specific temperatures (Figure 6, right panel). This shows that, on one hand, the minimal map approach may be a generalizable solution for folding of knotted proteins. On the other hand, it may suggest that the standard distance-based contact definition may omit some important factors (like the nature of residues), introducing many false-positive contacts and hindering folding.

5. Discussion

In this work, we have discussed the models used to explore knotted proteins folding. We have shown that they are mainly structure based models (with some exceptions), however the model rarely is able to recreate the folding of a deeply knotted protein with sufficient yield. Nevertheless, the $+3_1$ deeply knotted studied proteins (YibK, YebA) can fold *in vivo* without the need of chaperone machinery [43] (although addition of chaperones speeds up the process by at least an order of magnitude). Rooted in theory of a folding funnel and the principle of minimal frustration we have proposed two extensions of the Cα structure based model, differing in the definition of contact maps.

In one approach, we tested the so-called minimal map used before in the folding of the smallest knotted protein. This approach led to fairly good results with regular observation of successful folding of the YibK protein at some range of temperatures. The success of this approach needs comparison with a previous result of Wallin *et al.* [31] who added non-native contacts achieving 100% efficiency in irreversible knotting. The added non-native contacts were precisely between the C-terminal helix and the knotting loop, *i.e.* in one of the regions included in the minimal map (although over 60% of the original contacts were removed in the construction of a minimal map). This shows, that possibly the same effect, crucial for the knot formation, is present in both approaches.

The minimal map despite increasing the folding efficiency has a drawback, because it is not based on solid physical considerations concerning the interaction of residues. This can be included in the second approach – via heterogeneity in the contact strength. Actually, the minimal map is a map in which the strength of some contacts is retained, while the strength of others is negligible, therefore it is a special case of introduction of heterogeneity between the contact strength. In our approach, we applied the quasi-chemical potentials derived by Miyazawa and Jernigan [25] to differentiate the strength of native contact pairs. This has an effect of taking into account not only the native structure but also the nature of interacting residues. In fact, such approach should reduce the influence of false-positive contacts in the model, *e.g.* the contact between repelling residues, which were brought close in space due to the interaction of the neighboring residues. In our simplest approach, this method resulted in a small but considerable fraction of knotting events, which shows that averaging the residue properties in the model (treating each residue as an identical Cα bead) may be too "coarse". This result does not disprove the principle of minimal frustration, as possibly in most cases this effect is negligible, but self-tying of proteins is a sophisticated and delicate process, for which the driving force may be lost at the Cα graining level. From this viewpoint, it would be desirable to test the protein self-tying in models taking into account also the structure of the side-group, such as [4] or models with the force field constructed based on the mean force derived from a statistical analysis of the structural correlations seen in the known protein structures: CABS [45] or UNRES [46]

The model with re-scaled contact strength can be developed in many ways. First of all, in our model, no contact was repelling. However, in the so-called Miyazawa-Jernigan approach, some contacts have positive (*i.e.* destabilizing) energy. Therefore, one may introduce other scaling functions. More generally, one may construct one's own statistics of the residue-residue distance for each contact (similarly to the procedure described in [47]).

Despite the unquestionable progress in the area of knotted proteins, their free energy landscape still holds many puzzles for researchers. Especially, that we can only guess how the processes inside the cell actually occur. Life is a great mystery.

Acknowledgements

This work was supported by the National Science Centre [#2012/07/E/NZ1/01900 to JIS and PDT] and [#2016/21/N/NZ1/02848 to PDT]. We would like to thank Karol Pikul for helpful discussions about contact maps construction.

References

[1] Kolinski A *et al.* 2004 *Acta Biochimica Polonica* **51**
[2] Liwo A, Baranowski M, Czaplewski C *et al.* 2014 *Journal of molecular modeling* **20** (8) 1
[3] Liwo A, Arłukowicz P, Czaplewski C, Ołdziej S, Pillardy J, Scheraga H A 2002 *Proceedings of the National Academy of Sciences* **99** (4) 1937
[4] Sulkowska J I, Cieplak M 2008 *Biophysical Journal* **95** (7) 3174
[5] Bryngelson J D, Onuchic J N, Socci N D, Wolynes P G 1995 *Proteins: Structure, Function, and Bioinformatics* **21** (3) 167
[6] Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid A E, Kolinski A 2016 *Chemical Reviews* **116** (14) 7898
[7] Mansfield M L 1994 *Nature Structural & Molecular Biology* **1** (4) 213
[8] Takusagawa F, Kamitori S 1996 *Journal of the American Chemical Society* **118** (37) 8945
[9] Taylor W R 2000 *Nature* **406** (6798) 916
[10] Bölinger D, Sułkowska J I, Hsu H-P, Mirny L A, Kardar M, Onuchic J N, Virnau P 2010 *PLoS Comput Biol* **6** (4), e1000731
[11] Yeates T O, Norcross T S, King N P 2008 *Curr Opin Chem Biol* **11** (6) 595
[12] Sułkowska J I, Rawdon E J, Millett K C, Onuchic J N, Stasiak A 2012 *Proceedings of the National Academy of Sciences* **109** (26), E1715
[13] Jamroz M, Niemyska W, Rawdon E J, Stasiak A, Millett K C, Sułkowski P, Sulkowska J I 2014 *Nucl. Acids Res.* **43** (D1), D306
[14] Soler M A, Faísca P F N 2013 *PloS ONE* **8** (9), e74755
[15] Faísca P F N, Travasso R D M, Charters T, Nunes A, Cieplak M 2010 *Physical biology* **7** (1) 16009
[16] Chwastyk M, Cieplak M 2015 *Journal of Physics: Condensed Matter* **27** (35) 354105
[17] Lou Sh-Ch, Wetzel S, Zhang H, Crone E W, Lee Y-T, Jackson S E, Hsu Sh-T D 2016 *Journal of molecular biology* **428** (11) 2507
[18] Wang I, Chen Sz-Y, Hsu Sh-T D 2015 *The Journal of Physical Chemistry B* **119** (12) 4359
[19] Wang I, Chen Sz-Y, Hsu Sh-T D 2016 *Scientific Reports* **6**
[20] Hsu Sh-T D *et al.* 2016 *Understanding Enzymes: Function, Design, Engineering, and Analysis* 167
[21] Virnau P, Mallam A, Jackson S 2010 *Journal of Physics: Condensed Matter* **23** (3) 33101
[22] Mallam A L, Jackson S E 2007 *Journal of molecular biology* **366** (2) 650

[23] Noel J K, Levi M, Raghunathan M, Lammert H, Hayes R L, Onuchic J N, Whitford P C 2016 *PLoS Comput Biol* **12** (3), e1004794
[24] Noel J K, Sułkowska J I and Onuchic J N 2010 *Proceedings of the National Academy of Sciences* **107** (35) 15403
[25] Miyazawa S, Jernigan R L 1996 *Journal of molecular biology* **256** (3) 623
[26] Millett K C, Rawdon E J, Stasiak A, Sułkowska J I 2013 *Biochemical Society Transactions* **41** (2) 533
[27] Rawdon E J, Millett K C, Sułkowska J I, Stasiak A 2013 *Biochemical Society Transactions* **41** (2) 538
[28] Pettersen E F, Goddard T D, Huang C C, Couch G S, Greenblatt D M, Meng E C, Ferrin T E 2004 *Journal of computational chemistry* **25** (13) 1605
[29] Noel J K, Onuchic J N, Sulkowska J I 2013 *The Journal of Physical Chemistry Letters* **4** (21) 3570
[30] Sułkowska J I, Sułkowski P, Onuchic J 2009 *Proceedings of the National Academy of Sciences* **106** (9) 3119
[31] Wallin S, Zeldovich K B, Shakhnovich E I 2007 *Journal of molecular biology* **368** (3) 884
[32] Sulkowska J I, Noel J K, Ramírez-Sarmiento C A, Rawdon E, Millett K C, Onuchic J N 2013 *Biochemical Society Transactions* **41** (2) 523
[33] Dabrowski-Tumanski P, Jarmolinska A I, Sulkowska J I 2015 *Journal of Physics: Condensed Matter* **27** (35) 354109
[34] Dabrowski-Tumanski P, Niewieczerzal S, Sulkowska J I 2014 *TASK Quarterly* **18** (3)
[35] Chwastyk M, Cieplak M 2015 *The Journal of chemical physics* **143** (4) 45101
[36] Soler M A, Rey A, Faísca P F N 2016 *Physical Chemistry Chemical Physics* **18** (38) 26391
[37] Najafi S, Potestio R 2015 *The Journal of chemical physics* **143** (24) 243121
[38] Beccara S, Škrbić T, Covino R, Micheletti C, Faccioli P 2013 *PLoS Comput Biol* **9** (3), e1003002
[39] Li W, Terakawa T, Wang W, Takada S 2012 *Proceedings of the National Academy of Sciences* **109** (44) 17789
[40] Prentiss M C, Wales D J, Wolynes P G 2010 *PLoS Comput Biol* **6** (7), e1000835
[41] Sułkowska J I, Sułkowski P, Szymczak P, Cieplak M 2008 *Proceedings of the National Academy of Sciences* **105** (50) 19714
[42] Škrbić T, Micheletti C, Faccioli P 2012 *PLoS Comput Biol* **8** (6), e1002504
[43] Mallam A L, Jackson S E 2012 *Nature chemical biology* **8** (2) 147
[44] Miyazawa S, Jernigan R L 1985 *Macromolecules* **18** (3) 534
[45] Kmiecik S, Kolinski A 2008 *Biophysical Journal* **94** (3) 726
[46] Maisuradze G G, Senet P, Czaplewski C, Liwo A, Scheraga H A 2010 *The Journal of Physical Chemistry A* **114** (13) 4471
[47] Sippl M J 1990 *Journal of molecular biology* **213** (4) 859