Final report OTKA K115698

In this project we aimed to provide a unified description of protein structure organization. In the preceding decades we made significant contribution on the uncovering the structure formation of globular water soluble proteins, trans-membrane proteins and disordered proteins, the three large groups of proteins. In 2015 when we submitted this project proposal, we believed that these three classes cover the whole protein world. It was obvious that proteins inside the membrane are in different environment than water soluble protein or the water soluble part of the integral membrane protein so they need different thermodynamic description. On the other hand it was generally thought that there is a transition from the fully folded proteins through the partially folded ones, to the fully disordered proteins, and these cover all water soluble proteins. However at the beginning of the project we realized that the few disordered proteins, which are able to fold without the help of a stable macromolecule, generally a folded protein, are not exceptions but member of a much larger subclass of proteins, and without their consideration we cannot attempt to develop a general descriptions of water soluble proteins. Therefore we focused our attentions to describe this subset of proteins, called mutual synergestic folding (MSF) proteins.

Our first contribution to this filed was the publication of the MFIB (Mutual Folding Induced by Binding) database. It is a collection of experimentally validated MSF proteins [1]. According to our knowledge it is still the only database of its kind up to our days. The database is the results of a long and exhausting work. After the in silico identification of MSF protein candidates an exhaustive search of the literature was done by searching and evaluating experimental evidences. The results is a database of 205 experimentally validated protein complexes. Almost half of them are homo-dimeric protein structures, and only about one third of them are hetero-oligomeric structures built up from protein chains with different amino acid sequences. It is not a simple catalogue but contains a search motor to find proteins with various feature, like certain biochemical functions. The database was presented in the ECCB2016 (15th European on Conference Computational Biology), Hague(Holland) and a primary report appeared in F1000Research. Our MFIB database publication was recommended by two members of the Professors of Faculty 1000 in the F1000Prime independently. What is more both graded as “very good” (**) papers.

We also created a sister database of MFIB, called DIBS: DIsordered Binding Sites [2]. It is repository of protein complexes that are formed by ordered and disordered proteins. The main difference to MFIB is that in the case of DIBS complexes one interacting partner is ordered, it has a well-defined structure in the absence of its disordered interacting partner. The protein complexed collected in the MFIB database are built from disordered protein subunits, they become folded upon interacting with each other. Both databases were published it in Bioinformatics. Both papers and both databases are publicly available (http://mfib.enzim.ttk.mta.hu ; http://dibs.enzim.ttk.mta.hu). We made significant progress to compare the two dataset to each other and both to a dataset of protein complexes composed from interacting structured proteins.

We made progress in uncovering discriminating features of our recently introduced MSF protein subclass. First we investigated the largest homodimeric subclass of MSF proteins [3]. This is a particularly interesting case, traditional disordered proteins are known to have a special amino acid composition, different from that of globular proteins. In the case of homodimeric proteins the interacting protein chains have identical amino acid sequence, thus the monomeric and homodimeric proteins have identical amino acid compositions. This means, that the amino acid composition cannot be accounted for the change of their structural “orderedness”. MSF proteins are disordered because their ordered protein like

residue composition appear with special residue sequence, resulting in water accessible peptide backbone areas. We have also shown that these energetically unfavorable areas are shaded in the formation of folded homodimers. We published this result in IJMS.

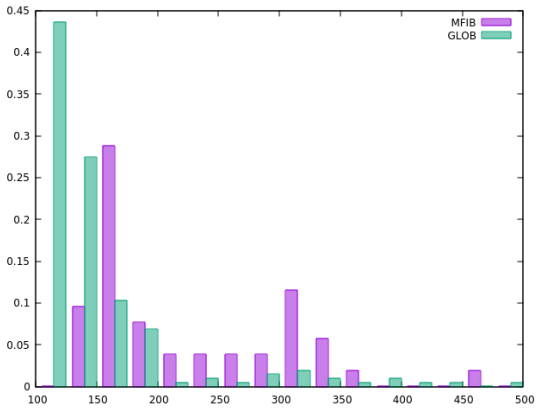
We also made good progress in studying heterodimers [4]. In the case of heterodimers two different proteins chains form the heterodimeric quaternary structure. In this case the amino acid composition of the dimeric for can differ from that of the monomeric protein chains. Still we found that their amino acid composition is globular like and the stabilization of the dimeric structure happens in a similar way to that of the homodimers. We found that the decrease of solvent accessibility is a major driving force of the oligomerization step.

We also presented a scalable clustering-based classification scheme, built on redundancy filtered features that describe the sequence and structure properties of MSF complexes and the role of the interaction, which is directly responsible for structure formation [5]. Using this approach, we defined six major types of MSF complexes, corresponding to biologically meaningful groups.

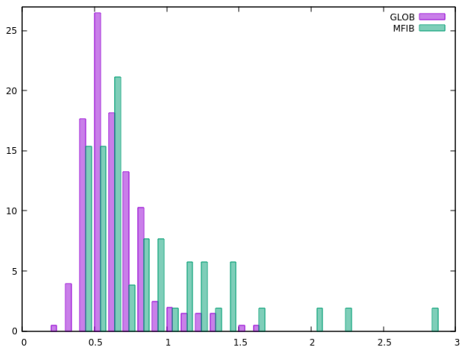
We defined three categories of protein complexes depending on the unbound structural state of the interactors and analyzed them in detail [6]. We found that the properties of interactors are defined not only by the intrinsic structural state of the protein itself but also by the structural state of the binding partner. The different types of interactions are also regulated through divergent molecular tactics of post-translational modifications. This presents a distinct molecular mechanisms compatible with specific biological processes. About these and other result in the fields of disordered proteins we published a book chapter [7] and an editorial [8] in the Special Issue "Functionally Relevant Macromolecular Interactions of Disordered Proteins" in the International Journal of Molecular Sciences.

Our latest result are just being submitted to a new Special Issue "Frontiers in Protein Structure Research" in IJMS [9]. We are trying to find common properties among MSF proteins, which distinguish them from globular oligomeric proteins. Next to the basic research relevance this would enable us to develop prediction methods. In order to be able to develop a sequence based prediction method first we plan to increase the size of our MFIB database. The first step in this direction is the development of a structure based method which is able to identify MSF proteins based on their oligomeric structure. We found several properties which can be the basis of such a prediction method. First we found that the most prominent change between MSF and globular proteins can be seen in the increase in solvent accessible surface area (SASA) during the oligomerization step. We define a residue buried if its relative SASA accessibility is below 20%, and accessible if it is over 20%. We calculate the ratio of buried/accessible (B/A) residues for our MSF and globular homodimeric datasets in both monomeric and dimeric forms, then we calculated a value representing the increase of B/A ration upon dimerization by dividing the dimeric B/A ratio by the monomeric B/A ratio. The following figure shoes the ratio of the proteins with a given increase in the B/A ration upon dimerization. We can see that in more than 40% of the globular homodimers this ratio is in the [100%,125%[ interval, thus the increase is less than 25% (green bars). While in MSF proteins the increase much higher (purple bars).

There are very few globular homodimers with a higher than 200% B/A dimeric value. We suspect that most of this hits are actually MSF proteins. The question arose as to what is in the background of the surface accessibility change? Sequence and structural studies of protein interactions have revealed that sequential and spatial neighboring residues often play important roles in the environmental hydrophobicity and long-term binding site interactions,

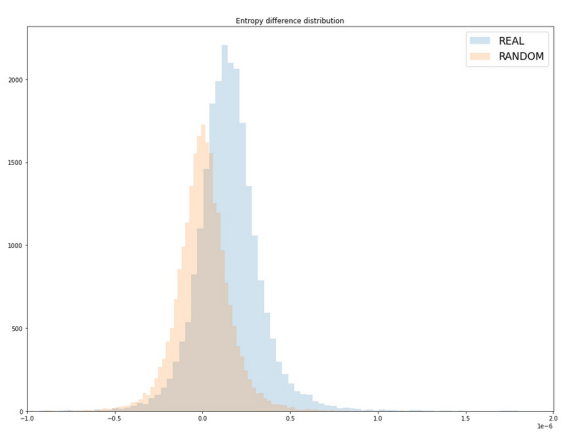


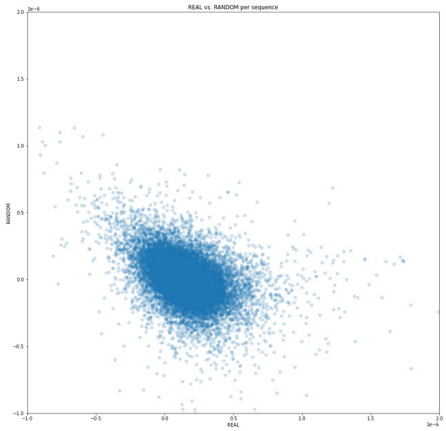

thus, determining the structural and functional behavior of proteins. Based on this fact, the shielding effect (reducing hydration of peptide group) can be divided into local and non-local terms. The local contribution is provided by the side-chain atoms of the amino acid residues, which are connected by the peptide bond, while the non-local contribution is provided by the shielding effect of other sequentially distant residues. We studied the solvent accessibility of peptide bonds in MSF and globular dimer proteins in light of the relationship between dipeptide frequencies, the ordered/disordered nature of the monomeric protein form. Furthermore, we investigated the entropy calculated from local and spatial neighboring residues of MFS-like and globular proteins, which may indicate sequential differences between the two groups of proteins. We calculated Shannon information entropy values based on amino acid pair occurrences. We calculated the Qij pair frequencies in a non-redundant subset of the PDB database and Pij frequencies from our MSF/globular homodimeric datasets and calculated Shannon information entropy values using the S = SUM {P * log(P/Q)} equation over all possible residue pairs. We obtained the following entropy distribution for out two reference datasets (purple: globular, green) MSF). We can see that there is a tendency for globular proteins to have lower values.




When we combine this property with the above described increase of buried residues we hope to be able to create a structure based MSF prediction method.

Our investigations of the „sequence entropy” resulted in the development of a novel alignment free method to estimate the information content of protein sequences. This method is in the final development phase and is expected to be published in 2021. The non-random nature of protein sequences, like our own amino acid pair preference is known for a long time. We developed a new method based on the Shannon information (entropy) theory. The method check amino acid pairs along the protein sequence up to the 10th neighbor. The pair preference of a given residue pair is described with a continuous value, where 0 means a random occurrence, positive values are assigned for preferred residue pair and negative values mean that a given pair has a lower than expected occurrence. These continuous scale is mapped to a three state -1,0,+,1 discrete value using appropriate cutoffs. These values denote unflavored neutral and preferred pairs. The whole UniRef90 database is scored using this 3 state functions. The proteins sequences are scored using 13 residue long windows with +/-6 residues from the residue in question. Thus we obtain a 6+6 element vector for all residues containing -1 / 0 /+*1* values. This procedure map a protein sequence to a 729x729 matrix (729=3^6). By scoring the UniRef90 database we obtained a 729x729 frequency table. The procedure was repeated with randomized proteins sequences, leading to a different 729x729 matrix. We can calculate the Shannon information entropy of a sequence using the following equation: S = P * log(P/Q), where P is the frequency calculated from real proteins sequences, and Q is value calculated from randomized sequences, respectively. The following figures show the distribution of the entropy values for real and random protein sequences, and the entropy values calculated for individual real protein sequences plotted versus values calculated from a randomized version of the same sequence.




These figures demonstrate the power of our new method in differentiating between real and randomized sequences with an efficiency of about 80%. We plan to expand our investigations to differentiate between MSF and globular sequences, when our database will be expanded.

Parallel with our effort on understanding folding of MSF proteins we were involved in other basic research projects. We finished our basic research project on stabilization center elements [10]. These are special residues involved in higher than average long range interactions, which we believed, have an effect on the thermal stability of proteins. Stabilization centers were thought to have a role to prevent unfolding of the stable 3D structure of proteins by due to their cooperative long range interactions. We found that in the case of proteins with moderate thermal stability (up to 85 C melting temperature) higher thermal stability correlates with a higher number of stabilization center elements and mutations, which change the thermal stability are more frequently found at stabilization center residues.

We pursued a research project related with cancer diseases. We published a paper presenting our results on somatic mutations driving cancer [11] and a review paper was published in *Science Signaling* presenting the current knowledge of about degrons in cancer [12], suggesting that research should be focused on the "dark degrome" to enhance progress in cancer research.

We also participated in a couple of applied research projects, like the “Identification of potential glutamine cyclase inhibitors from lead-like libraries by in silico and in vitro fragment-based screening” [13] project, which work was published in Molecular Diversity. Together with the group of Gergely Szakács we submitted a paper to Journal of Chemical Information and Modeling entitled “ Identifying new topoisomerase II poison scaffolds by combining publicly available toxicity data and 2D/3D-based virtual screening” [14]. With the group of Julianna Kardos we published the paper “Peptide Binding Sites of Connexin Proteins” in Chemistry [15]. We collaborated with the group of Csilla Özvegy-Lacka in the project “Synergistic transport of a fluorescent coumarin probe marks coumarins as pharmacological modulators of Organic anion-transporting polypeptide, OATP3A1”, which was published in Biochemical Pharmacology [16]. In these projects our main contribution was the in silico screening of the small molecule and peptide libraries using the Schrödinger Small Molecule Drug Discovery Suite.

**Publication list**


1. Fichó E, Reményi I, Simon I, Mészáros B: MFIB: a repository of protein complexes with mutual folding induced by binding, BIOINFORMATICS 33: (22) 3682-3684, 2017

2. Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B: DIBS: a repository of disordered binding sites mediating interactions with ordered proteins, BIOINFORMATICS 34: (3) 535-537, 2018

3. Magyar, C; Mentes, A; Fichó, E; Cserző, M; Simon, I: Physical Background of the Disordered Nature of “Mutual Synergetic Folding” Proteins, INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES 19 : 11 Paper: 3340 , 12 p., 2018

4. Mentes, A; Magyar, C; Fichó, E; Simon, I: Analysis of Heterodimeric “Mutual Synergistic Folding”-Complexes, INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES 20 : 20 p. 5136, 2019

5. Mészáros B, Dobson L, Fichó E, Simon I: Sequence and Structure Properties Uncover the Natural Classification of Protein Complexes Formed by Intrinsically Disordered Proteins via Mutual Synergistic Folding., Int. J. Mol. Sci. 20: 5460., 2019

6. Mészáros B, Dobson L, Fichó E, Tusnády GE, Dosztányi Z, Simon I: Sequential, Structural and Functional Properties of Protein Complexes Are Defined by How Folding and Binding Intertwine, J. Mol. Biol. 43: 4408–4428., 2019

7. Mészáros, B; Dosztányi, Z; Fichó, E; Magyar, C; Simon, I: Bioinformatical Approaches to Unstructured/Disordered Proteins and Their Complexes, In: Liwo, Adam - Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes Springer International Publishing, (2019) 561-596, 2019

8. Simon, I: Macromolecular Interactions of Disordered Proteins, INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES 21 : 2 p. 504, 2020

9. Magyar C, Mentes A, Cserző M, Simon I: Effect of peptide bond solvent accessibility on the stability of Mutual Synergestic Folding proteins, *submitted to the INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES.*

10.Csaba Magyar, M Michael Gromiha, Zoltán Sávoly, István Simon: The role of stabilization centers in protein thermal stability, BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS 471: (1) pp. 57-62, 2016

11. Mészáros B, Zeke A, Reményi A, Simon I, Dosztányi Z: Systematic analysis of somatic mutations driving cancer. BIOLOGY DIRECT 11:(1) Paper 23, 2016

12. Meszaros B, Kumar M, Gibson TJ, Uyar B, Dosztanyi Z: Degrons in cancer, SCIENCE SIGNALING 10:(470) eaak9982, 2017

13. Szaszkó M, Hajdú I, Flachner B, Dobi K, Magyar C, Simon I, Lőrincz Z, Kapui Z, Pázmány T, Cseh S, Dormán G: Identification of potential glutaminyl cyclase inhibitors from lead-like libraries by in silico and in vitro fragment-based screening, MOLECULAR DIVERSITY 21 (1) 175-186, 2017

14. Lovrics, A; Pape, VFS; Szisz, D; Kalaszi, A; Heffeter, P; Magyar, C; Szakacs, G: Identifying new topoisomerase II poison scaffolds by combining publicly available toxicity data and 2D/3D-based virtual screening, JOURNAL OF CHEMINFORMATICS 11 : 1 Paper: 67 , 14 p., 2019

15. Simon Á, Magyar C, Héja, L, Kardos J: Peptide Binding Sites of Connexin Proteins, CHEMISTRY 2 : 3 pp. 662-673., 2020

16. Bakos, É; Tusnády, GE; Német, O; Patik, I; Magyar, C; Németh, K; Kele, P; Özvegy-Laczka, C: Synergistic transport of a fluorescent coumarin probe marks coumarins as pharmacological modulators of Organic anion-transporting polypeptide, OATP3A1, BIOCHEMICAL PHARMACOLOGY 182 Paper: 114250, 2020.