The aim of this proposal was to bring the functional annotation of the disordered part of the proteome to a completely new level. In particular, on the main goal was to explore the system level occurrences of a specific linear binding motif class involved in signaling pathways of mitogen activated protein kinases (MAPK). Furthermore, taking advantage of emerging data of cancer genomic projects, we also wanted to analyze the role of intrinsically disordered protein regions in cancer. To achieve our goals, we developed and improved various computational tools and approaches. At the same time, we also aimed to significantly improve our understanding of the basic principles governing proteins-linear motif binding. During the project (including publications from the senior participant, Attila Reményi) we published **12 papers** [1-12] and **two book chapters** [13,14]. We also have **two manuscripts** ready that are available **as preprints** (one is under review, the other one is ready to be submitted) [15,16]. An additional manuscript is close to completion (results are ready), while we are still working on two additional projects.

One of the most important outcome of the project was published in Molecular Systems Biology journal [1]. This manuscript describes a complex computational and experimental approach to identify functional D(ocking)-motifs in disordered protein regions that are accessible for MAPK binding in the cell. For this, we developed a computational pipeline to filter likely biologically relevant motif occurrences. Therefore, in addition to sequence similarity to known D-motifs, we also used various sequence features (e.g. IUPRED, and ANCHOR predictions), structural compatibility (by using known MAPK-D-motif peptide crystal structures) and evolutionary conservation based scoring schemes to look for functional sites. This work enabled the identification of many novel MAPK-docking motifs that were elusive for other large-scale protein-protein interaction screens. This work has significantly expanded the inventory of human protein kinase binding sites. It enabled us to examine how kinase-mediated partnerships evolved over time and revealed that most human MAPK-binding motifs are surprisingly new evolutionarily inventions. The novel instances of MAPK docking motifs were incorporated into the central resource for linear motif research, the ELM database [2].

A computational pipeline was also developed to predict novel binding partners for another linear motif binding protein, the dynein light chain LC8. For this work, we implemented a novel decision tree based filtering protocol and combined evolutionary analyses and data from protein-protein interaction database. We have identified novel binding motif instances, discovered novel binding sites within known partners, and found completely novel binding partners which we experimentally verified at the motif level. As one of the main conclusion of our, the novel partners highlighted a likely role of LC8 in the Hippo pathway. This manuscript was published in PLoS Computational Biology [3].

The other main focus of our research was studying how intrinsically disordered proteins and protein regions (IDPs/IDRs) are involved in cancer. As a starting point, we had to developed a method that can identify significantly mutated protein regions that are likely to drive cancer development as opposed to regions that contain only randomly occurring passenger mutations [4]. This method was published in Biology Direct. Using this method, we systematically collected significantly mutated regions located within disordered protein regions. Our analyzis showed that significantly mutation regions located within disordered regions affect similar biological processes compared to globular protein regions involved in cancer, but the details of their molecular function differ and showcase the versatility of protein disorder. While this work is still in preparation (Mészáros B, Zeke A, Hajdu-Soltész B, Reményi A, Dosztányi Z *Intrinsic protein disorder requires new targeting strategies in cancer*, 2018), it has directed our attention to the importance of degron motifs in cancer. Degrons correspond to a special class of short linear motifs (SLIMs) that dictate the regulated degradation of many proteins and are usually located within IDRs. In collaborations with the group of Toby Gibson (EMBL, Heidelberg) we wrote an extensive review that not only provided the first comprehensive overview of degrons motifs but also discussed how their altered functioning can contribute to cancer and how they can be exploited as potential therapeutic targets. This work was published in a special cancer focus issue of Science Signaling in March 2017 [5].

In collaboration with scientists from the Institute of Enzmology, we created a database of structural complexes formed between IDP regions and ordered domains (DIBS). By extensive literature search we focused only on examples for which the disordered status was experimentally verified and collected available Kd values for many of these interactions as well. The corresponding manuscript was published at Bioinformatics [6]. Using this database and the database dedicated to complexes formed between two or more IDPS, we analyzed how the interplay between folding and binding modulates protein sequences, structures, functions and regulation [15]. In addition, we relocated our web servers, IUPred and ANCHOR to a new location provided by ELTE and published an invited manuscript about IUPred in Protein Science [7]. We completely reworked the two web servers and extended with novel functionalities, in our new server, IUPred2A. While only minor bug fixes were implemented for IUPred, the next version of ANCHOR was significantly improved through a new architecture and parameters optimized on novel datasets, including entries of the DIBS database [8]. A completely novel feature was also introduced that can highlight putative redox-sensitive conditionally disordered regions. In a subsequent analysis we found that such regions are surprisingly widespread in various proteomes and play key roles in high-level eukaryotic processes [16].

The PI (Z.D) was also part of an international collaboration through a COST network (BM1405 Non-globular proteins - from sequence to structure, function and application in molecular physiopathology). I (Z.D.) that yielded several published articles in collaboration with Silvio Tosatto (University of Padua, Italy). These included a new version of the DisProt database which collects experimentally verified disordered protein regions. This database was relocated and completely redesigned by the group of Silvio Tosatto, while annotators, including myself, corrected several older entries collected and introduced new entries into the database [9]. We took advantage of the new data collected in the Disprot database to test and compare the performance of disorder prediction methods on an independent dataset [10]. In another publication, a new disorder prediction method was introduced that was developed based on a combination of multiple prediction methods -- including our method IUPred – in order to provide fast and accurate predictions [11]. This method is now used in Interpro to provide annotation for Uniprot sequences. In order to make information related to protein disorder and mobility accessible for the complete UniProt protein set, predictions provided by IUPred and ANCHOR and data collected in the DIBS database were incorporated into the Mobidb3.0 database [12].

We also made good progress in additional projects and we plan to publish them in the near future. Our studies also highlighted the importance of using evolutionary information for linear motif discovery. However, linear motifs are usually located within intrinsically disordered protein regions, which are notoriously difficult to align. To overcome these difficulties, we have been developing a really promising novel approach to align linear motifs located within IDPs (Pajkos et al. A novel motif centric alignment method). We are working on incorporating structural information to the predictions with the aim to predict not only true binding peptides, but also the affinity of their interactions. This approach is currently being tested for the LC8 system (Erdős et al).

We also wrote two book chapters during the project. The first one, entitled "Prediction and analysis of intrinsically disordered proteins", was written in collaboration with Marco Punta (EMBL-EBI) and István Simon (Institute of Enzymology) in the book Methods in Molecular Biology-Structural

Proteomics [13]. I also published a book chapter with Peter Tompa about bioinformatics approaches to the structure and function of IDPs [14].

In addition to the OTKA grant, the project was also supported by the Momentum grant of the Hungarian Academy of Sciences, awarded to the PI in 2014. This caused some delays in the execution of project, especially in the year of 2014 and 2015, which were mostly spent with the relocating and starting a new lab, getting equipment, hiring new people and training them. Thematically, there is significant overlap between the two grants (since I moved to a new host institute, this was allowed). However, the Momentum grant was necessary to equip our newly starting lab to be able to carry out the planned research in the new location. In addition, it also made it possible to widen the scope of our analyses. For example, for our analysis of how intrinsically disordered protein are involved in cancer we could add the analysis of TCGA dataset in addition to the COSMIC database and also study the druggability of our examples of interest. This is likely to significantly increase the impact of this research.

Publications

1: Zeke A, Bastys T, Alexa A, Garai Á, Mészáros B, Kirsch K, **Dosztányi Z**,Kalinina OV, **Reményi A**. Systematic discovery of linear binding motifs targeting an ancient protein interaction surface on MAP kinases. Mol Syst Biol. 2015 Nov 3;11(11):837.

2: Gouw M, Michael S, Sámano-Sánchez H, Kumar M, Zeke A, Lang B, Bely B, Chemes LB, Davey NE, Deng Z, Diella F, Gürth CM, Huber AK, Kleinsorg S, Schlegel LS, Palopoli N, Roey KV, Altenberg B, **Reményi A**, Dinkel H, Gibson TJ. The eukaryotic linear motif resource - 2018 update. Nucleic Acids Res. 2018 Jan 4;46(D1):D428-D434.

3: **Erdős G**, Szaniszló T, **Pajkos M**, Hajdu-Soltész B, Kiss B, Pál G, Nyitray L, **Dosztányi Z**. Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway. PLoS Comput Biol. 2017 Dec 14;13(12):e1005885.

4: Mészáros B, Zeke A, **Reményi A**, Simon I, **Dosztányi Z**. Systematic analysis of somatic mutations driving cancer: uncovering functional protein regions in disease development. Biol Direct. 2016 May 5;11:23.

5: Mészáros B, Kumar M, Gibson TJ, Uyar B, **Dosztányi Z**. Degrons in cancer. Sci Signal. 2017 Mar 14;10(470). pii: eaak9982.

6: 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. 2018 Feb 1;34(3):535-537.

7: **Dosztányi Z**. Prediction of protein disorder based on IUPred. Protein Sci. 2018 Jan;27(1):331-340. doi: 10.1002/pro.3334.

8: **Mészáros B, Erdos G, Dosztányi Z**. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018 Jul 2;46(W1):W329-W337.

9: Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, **Dosztányi Z**, Elofsson A, Gasparini A, Hatos A, Kajava AV, Kalmar L, Leonardi E, Lazar T, Macedo-Ribeiro S, Macossay-Castillo M,Meszaros A, Minervini G, Murvai N, Pujols J, Roche DB, Salladini E, Schad E, Schramm A, Szabo B, Tantos A, Tonello F, Tsirigos KD, Veljković N, Ventura S,Vranken W, Warholm P, Uversky VN, Dunker AK, Longhi S, Tompa P, Tosatto SC.DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res. 2017 Jan 4;45(D1):D219-D227. doi: 10.1093/nar/gkw1056.

10: Necci M, Piovesan D, **Dosztányi Z**, Tompa P, Tosatto SCE. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics. 2018 Feb 1;34(3):445-452.

11: Necci M, Piovesan D, **Dosztányi Z**, Tosatto SCE. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics.
2017 May 1;33(9):1402-1404.

12: Piovesan D, Tabaro F, Paladin L, Necci M, Micetic I, Camilloni C, Davey N, **Dosztányi Z**, Mészáros B, Monzon AM, Parisi G, Schad E, Sormanni P, Tompa P, Vendruscolo M, Vranken WF, Tosatto SCE. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res. 2018 Jan 4;46(D1):D471-D476.

Book chapters

13. Bioinformatics Approaches to the Structure and Function of Intrinsically Disordered Proteins
Dosztányi Z and Tompa, P.
From Protein Structure to Function with Bioinformatics pp 167-203

14. Prediction and Analysis of Intrinsically Disordered Proteins
Marco Punta , István Simon , and **Zsuzsanna Dosztányi**
Raymond J. Owens (ed.), Structural Proteomics: High-Throughput Methods, Methods in Molecular Biology, vol. 1261,

Under review

15. Interplay between folding and binding modulates protein sequences, structures, functions and regulation
Bálint Mészáros, László Dobson, Erzsébet Fichó, Gábor E. Tusnády, **Zsuzsanna Dosztányi,** István Simon
doi: https://doi.org/10.1101/211524

16. Large-scale analysis of redox-sensitive conditionally disordered protein regions reveal their widespread nature and key roles in high-level eukaryotic processes
Gábor Erdős, Bálint Mészáros, Dana Reichmann, **Zsuzsanna Dosztányi**
doi: https://doi.org/10.1101/412692