Final report of K104586 application

During the five year of the project, we followed the proposed work plan:

Aim #1: We have developed a new fold recognition algorithm, called TMFoldRec, which based on statistical potentials and a topology filtering approach. TMFoldRec was tested on representatively selected entries of PDBTM containing 116 transmembrane protein chains. In 85% of the cases, the native folds were ranked at the first place, using the topology-filtering algorithm in a 10-fold jackknife benchmark. This accuracy is the highest among the recent state-of-the-art methods. A key feature of this algorithm is the reliability estimation of the obtained model with the lowest energy. Since 80% of the transmembrane folds are currently unknown, a fold recognition approach should make differences between the native, real fold and a non-native fold even if the calculated energy is the lowest. Therefore, this feature makes the TMFoldRec algorithm usable for proteome-wide analyses and can assist in the identification of transmembrane protein sequences with unknown folds as targets for structure determination, to speed up the exploration of the transmembrane structure space and can consequently pave the way for a deeper understanding of the structure- function relationship of transmembrane proteins. We have published the TmFoldRec algorithm in BMC Bioinformatics. Because the algorithm requires high computer capacity, we have developed a web server for the academic users, called TmFoldWeb (http://tmfoldweb.enzim.ttk.mta.hu), which utilizes our high parallel computing cluster system via WSDL protocol. The functionality of the web server is described in a paper published in Biology Direct. After publishing these two manuscripts, we have started to work on a *de novo* 3D structure prediction method for the transmembrane segments (TMSs) of TMPs. The developed method uses the prediction results of TMFoldRec method and generates a model for the TMSs in a modeled double lipid layer using existing molecule dynamic simulation algorithms. We have not finished this new algorithm yet.

Aim #2: We developed a pipeline using the TmFoldRec algorithm (see **Aim #1**), the CCTOP algorithm (see **Aim #7**) and a hidden Markov model based profil-profil alignment method (HHBlits) to create proper sequence alignments for transmembrane proteins. This algorithm was used in **Aim #4** for preparing the target selection database of transmembrane proteins.

Aim #3: We have developed several data mining algorithm to collect topology data from the literature and from the various databases available publicly on the Internet. In addition to collecting the experimentally determined topology data published in the last couple of years, we have gathered topographies defined by the TMDET algorithm using the 3D structures of TMPs from the PDBTM database. Results of global topology analysis of various organisms as well as topology data generated by high throughput techniques, like the sequential positions of N- or O- glycosylations were processed and incorporated into the TOPDB database as well. Moreover, a new algorithm has been developed to integrate scattered topology data from various publicly available databases into one unified piece. We have introduced a new method to measure the reliability of the predicted topologies, and utilized a newly developed topology prediction algorithm to determine the most reliable topology using the results of experiments as constraints. Using these algorithms, we have gathered more than 75.000 topology data for 4.190 proteins by processing more than 4.200 articles. The updated TOPDB database is available at http://topdb.enzim.hu, and it has been published in the Database Issue of the Nucleic Acids Research. With the help of the CCTOP algorithm (see **Aim #7**), we were able to predict the topology of all TMPs in the UniProt database. Moreover, using our super computer, we could detect all sequence motifs and protein domains listed in PFAM, Prosite, Smart and Prints databases for all sequences in the UniProt database. Using the results of these applications, we generated a completely new version of the TOPDOM database, extending the consistently located protein domains and sequence motifs to globular proteins and highly increased the accuracy and the reliability of these domains and motifs in TMPs. The results were published in Bioinformatics.

Aim #4: We solved this aim in two steps. Since the presence of intrinsically disordered regions in protein makes the structure determination almost impossible, first we determine the occurrences of IDRs in TMPs by preparing a benchmark set for prediction of intrinsically disordered regions in TMPs using a stringent definition for disordered regions. The investigation of the resulted, currently the largest experimental dataset reviled there are two different possible roles of intrinsically disordered regions in transmembrane proteins. We found a significant correlation between the spatial distributions of positively charged residues and short disordered regions close to the membrane. This finding suggests a new role of disordered regions in transmembrane proteins by providing structural flexibility for stabilizing interactions with negatively charged head groups of the lipid molecules. The longer disordered regions can be found far to the membrane in space at the N- or C- terminal regions of the polypeptide chains and probably are involved in interactions with other proteins playing roles in signal transduction or cell-cell interactions. We published this work in the BBA Biomembranes.
In the second step we determined the list of transmembrane proteins, which structures' determination may result in the largest modelable transmembrane protein pull. For making a proper sequence alignment we used the combined algorithm developed in specific **Aim #2**. We determined the relationship between our list and the already known structures of transmembrane proteins generated by the various structure genomics consortia consuming millions of dollars. Surprisingly the two lists have not overlapping proteins, because the genome projects choose the save site, i.e. does not applied any bioinformatical study before choosing and determining the structure of a transmembrane protein, but most of the determined structures have similar structures already determined. Two avoid the further spoil of money of genomic consortia; we created a database containing the list of transmembrane proteins, which structure determination can increase the most of the number of modelable transmembrane proteins. This database is publicly available at the https://tstmp.enzim.ttk.mta.hu internet site, and we published this work and the database in the Nucleic Acids Research Database issue. Julia Varga, who works on this project the most; won the first prize on the National Scientific Students' Association in the Bioinformatics section for her contribution to this results.

Aim #5: Using the data generated in Aim #3, we could accurately map interacted protein domains in TMPs. Unfortunately, the results showed that data in the various interaction databases are highly biased towards the TMPs, therefore they cannot be used for reliable analyses. As a side project of this module, we characterized disease-associated mutations in human TMPs and found characteristic differences in the spectrum of amino acid changes within TMSs and showed that the majority of disease associated mutations result in glycine to arginine and leucine to proline substitutions. Our analysis contributes to the better understanding of the effect of disease associated mutations in TMPs, which can help prioritize genetic variations in personal genomic investigations. This work was published in PLOS One.

Aim #6: We slightly modified the plan; because we were able to develop a new high-throughput topology assay using the native form of proteins in intact cells. The planned and most of the currently existing topology assays, (such as glycosylation site identification methods, epitope mapping of proteins, protease protection assays, etc), could generate topology data only for a single protein within months, while the newly developed method, can produce topology data for hundreds of proteins at the same time. We optimized the parameters of the wet lab experiments, and finally we were able to generated topology data for about two hundred TMPs and showed that the generated topology data can help to model more than 2500 TMPs in the UniProt database. We published this work in Nature Scientific Research. Besides these works, we can successfully determine the topology and membrane localization and routing of ABCB6 protein in the cooperation with Gergely Szakacs' research group, and published the results in Biochemical Journal.

Aim #7: We focused our efforts to human transmembrane proteins. In order to distinguish transmembrane from not transmembrane proteins in the genome as well as for topology prediction, we used a newly developed consensus prediction method (CCTOP), which is a consensus and constrained prediction method based on our HMMTOP algorithm. The CCTOP method combines the results of i, 10 different topology or topography prediction methods; ii, the deposited experimental topology data of the homologous proteins in TOPDB; iii, the sequential information of any homologous transmembrane protein in the UniProt database; and iv, conservatively localized protein domains and sequence motifs form TOPDOM database. All of these data were integrated into the probabilistic framework of the hidden Markov model. The per protein topology prediction accuracy of the CCTOP algorithm was shown to be the highest among the other currently available state-of-the-art methods (98.5% for filtering and 84% for per protein topology prediction). We published CCTOP in Nucleic Acids Research Web server issue, and made it publicly available at the web (http://cctop.enzim.ttk.mta.hu). Using the CCTOP algorithm we can accurately predict that the human genome contains 4998 (26%) transmembrane proteins. The predicted and the collected experimental topology data can be downloaded from and can be visualized at the homepage of the human transmembrane proteome (http://htp.enzim.hu). We have published the paper about the HTP database in Biology Direct.

Besides these works, we cooperated in characterizing the genome of the DT40 chicken cell line using whole genome shotgun sequencing and single nucleotide polymorphism array hybridization. The results of this work was published in G3: Genes, Genomes, Genetics. Moreover, we cooperated in identifying viruses in various plants by NGS with the Diagnostic Group of Agricultural Biotechnology Center, Gödöllő led by Éva Várallyai and published our results in the Journal of Plant Biology and presented them in several international scientific conferences. We were involved in investigation of mutation in cancer and developed a method to detect mutations from patients NGS data responsible for certain cancer types as well as revealed the mutagenic impact of common cancer cytotoxics. These works were published in Appl Immunohistochem Mol Morphol and Genome Biology, respectively.

During this OTKA proposal we have launched 7 new or renewed web servers and publicly available databases, and published our work in 19 scientific publications or book chapters, which cumulative impact factor is almost 90 and were cited more than 150 times.