SMGroup

# Biodata Mining in the Context of Sequence Motif-Based Approaches for the Investigation of the Structural Space of Membrane Proteins

**Steffen Grunert[1]* and Dirk Labudde[1]**

[1]Hochschule Mittweida, University of Applied Sciences, Germany

***Corresponding author:** Steffen Grunert, Hochschule Mittweida, University of Applied Sciences, Technikumplatz 17, 09648 Mittweida, Germany, Email: grunert1@hs-mittweida.de

**Published Date:** December 13, 2015

## ABSTRACT

### Background

Membrane proteins are essential for many cellular processes and are important targets for a wide range of pharmaceuticals. Their sequences provide valuable and partly not yet deciphered information about their three-dimensional structure and functional characteristics. The analysis of membrane proteins has shown to be an important part for the understanding of complex biological processes in the context of proteomics and genomics. In the course of membrane protein investigations, a large number of short, distinct sequence motifs have been revealed. The motifs found so far support the understanding of the folded protein in the membrane environment. In this context, in three different approaches it is shown how the output of sequence motif-based methods can support the understanding of structural and functional properties of membrane proteins.

### Results

In order to increase the understanding of sequence data of membrane proteins, different sequence motif-based methods have been developed. At the beginning of this work, a new approach describes the topology specific separation of motifs by residue specific distributions. Based on

these distributions, the topology structure of the majority motifs in hypothesized membrane proteins can be predicted. The second part consists of the evaluation of different resulting graphs generated from statistical analysis of motif neighborhoods in transmembrane helical structures. The results show motifs with high importance in alpha-helical membrane structure formation and motifs that are important for family-specific functional characteristics. In addition, it can be shown that short interaction patterns of homologous sequence records are membrane protein family-specific signatures and these can be used for structure prediction, protein classification and to observe new evolutionary variations within the signatures. Generally, we introduce simple graphical visualization tools for big data information presentation and to evaluate our results.

## Conclusions

In summary it can be shown how the evolution contributes to realize co-variations within discriminative sequence motifs to maintain structure and function. These points to their general importance for α-helical membrane protein structure formation and interaction mediation. Moreover, from the results of the introduced approaches, it can be better understood how motifs are anchored within protein structures, to realize their role as structural or functional elements. We hypothesize that short sequence motifs can be separated into structure forming motifs on the one hand and on the other hand, into motifs with family-specific functional characteristics. The low cost rapid computational methods postulated, provide valuable information for protein classification, structure comparison and not least to use our *in silico* results for the planning of *in vitro* experiments.

**Keywords:** Membrane proteins; Sequence motifs; Motif; Topology separation; Visualization; Indicators; Structure prediction; Classification; Mutagenesis

# INTRODUCTION

Without them there would not be us - amino acids, they form the basis of all life on our planet. Their emergence about 3.8 billion years ago results from far other environmental conditions in comparison with today's [1]. The opinion of science is that at this time the first primitive life gradually formed. Meanwhile, various theories developed over time, which deal with the question how the life developed on our planet. However, today's scientific community has the certainty that amino acids serve as basic building blocks of proteins. Proteins in turn, are essential components of all cells of single and multi cellular organisms. In the course of this, proteins are essential for almost all fundamental biological processes within an organism and therefore crucial for its survival. They originate from the key relationship between DNA, RNA and the consequent protein synthesis, known as the central dogma of molecular biology. Proteins in general consist of the 20 naturally occurring amino acids that are connected by peptide bonds to form chains. Amino acids have different physic-chemical properties and ultimately they determine the folding of a protein and its structure. The mechanism how a protein folds into its final three-dimensional structure is important for the subsequent correct function of the protein. Therefore, the study of the way

how proteins fold into complex molecular structures and how starting from a protein sequence to predict and better understands the folding of a protein, is an important area in bioinformatics [2].

In principle, proteins can be divided into globular and membrane proteins. Globular proteins freely exist in the cytoplasm of a cell or in other cellular organelles. In contrast, membrane proteins represent a special class. Unlike peripheral membrane proteins, which are single or attached only at one side of the lipid layer, integral membrane proteins traverse the cell membrane completely. With their vital functions such as: photosynthesis, the transport of ions, small molecules and solutes or signal and energy transduction, only some examples are given of cellular processes that are covered by membrane proteins [3]. They get their specific properties through the individual folding and interaction with the hydrophobic environment of the cell membrane, and in the formation of oligomeric complexes and protein-protein interactions [3,4]. The identification of such complexes and interactions is of immense importance and valuable information can be obtained by analyzing interacting proteins with known function. In general, membrane proteins are difficult to solve and cover a wide intracellular concentration range. The inaccessibility of many proteomic methods means that the analysis of membrane proteins is a currently experimental challenging research field [5].

## The Disproportion between Solved and Unsolved Structures

With the X-ray crystal structure analysis and the Nuclear Magnetic Resonance **(NMR)** two experimental methods of investigation for the elucidation of protein structures are given. However, these represent a still challenging field of research when it comes to the elucidation of protein structures with high throughput. Currently, this is still reflected in the small number of solved protein structures. In contrast, the way of providing information on coding DNA sequences and their translation by means of the genetic code is simpler than that of the spatial structure, which is reflected in a variety of manually annotated membrane protein sequences [6]. This disproportion has to be compensated in near future, which represents a major challenge for bioinformatics methods. In addition, the extraction of structural and functional features out of current available unsolved protein structure sequences emerges more and more to a big data problem. For this reason, there is a need for methods, which handle the corresponding information extraction and the prediction of such features in unsolved protein structures. A variety of methods have dealt in history with such tasks. Here in after, a brief review is given to these.

## State-of-the-Art

Following the introduction of the hydrophobicity scale over 30 years ago by Kyte and Doolittle [7], a variety of approaches and methods have been proven as historic milestones in the prediction and extraction of Transmembrane **(TM)** features. With the hydrophobic moment as a measure of the amphiphilicity of a helix [8] and the Positive-Inside-Rule [9] for the determination of overrepresented positively charged amino acids Arginine **(Arg)** and Lysine **(Lys)** on the cytoplasmic side, first TM helical regions could be predicted. Generally, hydrophobicity-based

methods turned out to be effective in the prediction of TM regions, up to the moment where it has been recognized that globular segments with a highly hydrophobic character can only be imprecisely distinguished from membrane regions. Consequently, methods such as PRED-TMR [10], TMPred [11], Top Pred [12] and SOSUI [13] could improve the prediction accuracy of TM helices by focusing on the identification of potential helical ends or working on the basis of statistical amino acid preferences and hitherto gained measures and rules. With the aid of dynamic programming and grammatical Rules, methods such as MEMSAT [14], TMHMM [15-17], HMM Top [18,19] and BOCTOPUS [20,21] focus specifically on the prediction of structural features such as membrane-spanning α-helices, extra / intracellular domains and membrane-spanning β-Strands of TM β-barrel proteins. With the mapping of energetic interactions in membrane proteins, an important step towards *ab initio* protein structure prediction was made [22-27]. Since 1996 it has been recognized that meaningful structural information can be derived from sequences of homologous protein families [28,29]. In the process methods such as the Direct Coupling Analysis **(DCA)** [30], MiC [31], EVfold [32], PSICOV [33], plmDCA [34,35] or GREMLIN [36-38] benefit from the growth of increasingly large proteomic data like mentioned before. Nowadays, such information-based methods enable the observation and extraction of evolutionary co-variance information to approximate physically interacting amino acid pairs. Homology-based methods are using such valuable information for model generation on the atomar level. Based on these models potential three-dimensional protein structures can be calculated.

Summarized, state-of-the-art methods can show that the spatial structure of proteins has been stronger conserved in evolution than the sequential composition of the folded protein chains. Also, the information of the spatial structure and the folding of proteins to be explored can be evaluated by affinities. There are individual motifs or characteristic sequence parts that expose a certain biochemical function of proteins. Why does the nature pursue the principle of structure and function separation? Residues, which support a stable domain folding, are separated from those that induce a specific function. This procedure is a very efficient strategy of evolution. Two areas were simultaneously optimized [39].

- The stability of the protein backbone in a given folding pattern
- The design of the amino acid sequence according to a specific function.

Due to the conuence of studies on the structural and biochemical understanding of membrane proteins, one begins in principle to understand what mechanisms lead to structural folding of membrane proteins [40-43]. Here, proteome wide investigations have shown that the determining factors of the membrane protein folding can be better understood by the investigation of such individual motifs or characteristic sequence parts in interacting helices [43]. In this context, the geometrical properties and characteristics of interacting helices have been investigated in a variety of studies, and the characteristics in amino acid sequences which are responsible for these

geometries [44]. Generally, based on the recurring sequential elements the structural simplicity of membrane proteins were observed in a large number of investigated TM domains and correlations to helix-helix interactions were confirmed [45-48]. This means that the folding of helical membrane proteins is largely recognizable by interactions between membrane embedded helices which can be seen in recurring helix-helix interaction patterns, the so called packing motifs [43]. As one example, the most famous in literature described Gx3G motif shows a significant presence in TM α-helices [46-51]. Gx3G was observed for the first time in glycophorine proteins [47,52,53], a TM protein of the human erythrocyte that consists of one TM helix. Separated by three variable x-positions (GxxxG) with x ε 20 canonical amino acids, both glycines create a shallow groove that complements the surface of a second helix to form finally a dimer. Previous studies have shown the plurality of such structural and functional motifs, which have to be investigated among other.

## Sequence Motif-Based Approaches, Applied on the Big Data Problem

Related to the mentioned rapid growth of proteomic data, *in silico* approaches can contribute to handle the big data problem. In the scope of this work, meaningful and helpful information about the structural composition and functional involved sequential parts of membrane proteins will be handled by sequence motif-based approaches, which contribute to solve the associated problems. Adapted to this, a variety of discriminative sequence motifs will be examined within the structural space (Figure 1) of α-helical TM proteins in the context of proteomic Data Mining. Based on sequence motif-based methods, valuable structural and functional properties for each investigated motif are to be derived. In the first step, the evolutionary variability in the x-positions will be investigated. Accurately, which physico-chemical properties have to be observed at each x-position to realise the three-dimensional fold within different topologies? From the perspective of sequence motifs, the general understanding about the sequential composition of α-helical structures and the correlated three-dimensional fold is to be extended. The corresponding high-throughput analyses of different membrane protein families lead to different results about motifs and there structural and functional role within the investigated big proteomic datasets. The implicitly of different graph outputs can be used for the evaluation of high-dimensional protein data, to establish a biological relationship and ultimately to provide useful information for structural research. Furthermore, an easy way shows how to generate different kinds of graphical outputs, which can be useful for better understanding of latent information about structural and functional features of membrane proteins out of multidimensional protein data. Also, the derivation of observable structural features represented by sequence motifs can be used as membrane protein family specific indicators. Here, the evolutionary way within helix-helix interacting patterns will be traced and the results provide useful information for structure prediction and classification tasks. Generally, mining big proteom data with sequence motif-based methods support the understanding of the features that are important for establishing stability and functionality of the folded membrane protein in the membrane environment. Moreover, such methods can be useful to investigate the influences of genetic variations within interacting

sequence parts in membrane proteins, which are directly linked to diseases. In summary, this work contributes to support the understanding about the role of motifs within the space of α-helical TM proteins.

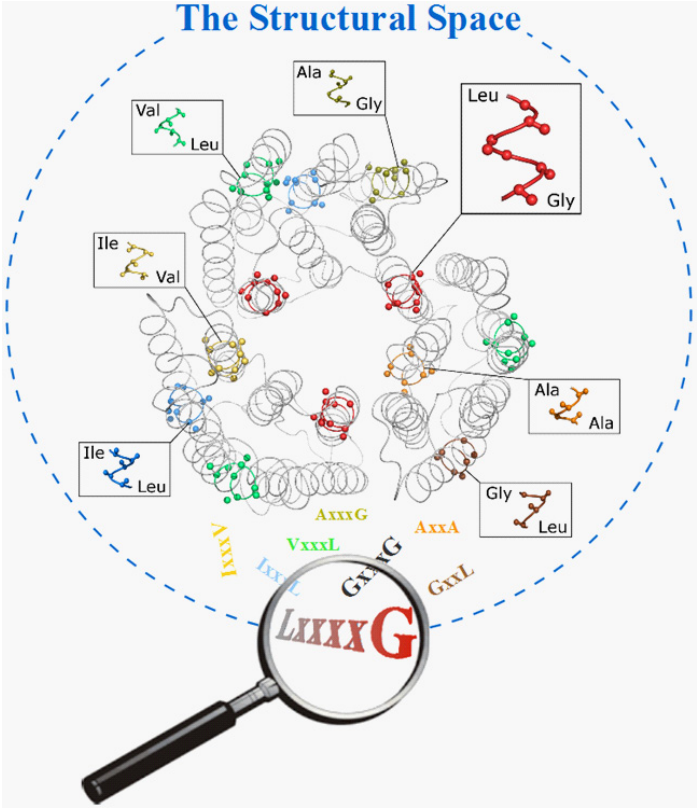


**Figure 1:** The structural space of membrane proteins using the example of bacteriorhodopsin trimer (PDB-Id: 1brr). In general, sequence motifs play key roles within the structural space of membrane proteins. They also play key roles in the general importance of α-helical membrane protein structure formation and interaction mediation just in mediating helix-helix interactions. The investigation of sequence motifs supports the understanding of the features that are important for establishing stability and functionality of the folded membrane protein in the membrane environment. On the other hand, the discovery of common motifs supports the understanding of the characteristics that are important for the stability of membrane spanning structures. Simultaneously, these deliver important structural and functional indications of yet unsolved protein structures. Eventually, sequence motif-based approaches are helpful for a variety of applications e.g. the investigation of mutant proteins, mutagenic effects and protein dynamics. In the present work, different sequence motif-based approaches dealing with the investigation and application of motifs like the illustrated LxxxxG to support the understanding about the three-dimensional fold of sequence parts within different topologies.

# MATERIAL AND METHODS

## Topology Separation of Discriminative Sequence Motifs

As the first step of our analysis, 32 membrane protein families with Domains of Unknown Functions **(DUF)** were derived from the Pfam database [54], using extended keyword searching [83], and finally investigated. To avoid generating misguiding statistics by including identical or highly similar sequences, a set of non-redundant sequences was generated for all in this work investigated membrane protein datasets. The corresponding tools used in this work were CD-HIT [55] and BLASTClust [56]. Furthermore, for the general determination of membrane and non-membrane associated sequence regions, the TMHMM Server v. 2.0 [15,16] was used. Subsequently, a variety of sequence motifs identified in earlier work by Liu et al. [46] were localised in the obtained set of DUF-families. These motifs have been derived from the consensus sequences of 168 Pfam families. By the comparison of their results with findings in [47], a list of 50 significant motifs has been derived that were used in the first project. Each motif shows up to be indispensable for the structure fold. Intuitively, the reported short sequence motifs can be written in a generalised, regular expression-like form of XYn, where X and Y correspond to amino acids separated by n-1 highly variable positions. For example, the LxxxxG motif corresponds to a pair of leucine and glycine residues that are separated by four amino acids. Finally, it can be written as the simple LG5 notation.

In the following, the structural space was examined to investigate one and the same motif e.g. LG5 within three different topologies (**TM**=Transmembrane, **nTM**=None-Transmembrane, **Trans**=Transition). These states come from the obtained TMHMM results, where a topology state was assigned to each residue, if a residue was assigned being a part of a transmembrane helix (TM) or not (nTM). Related to each whole motif, where the beginning and the end has been located in TM and nTM, it has been assigned with the Trans-state. To this, single motif occurrences were identified in the non-redundant sequence set. Including TMHMM predictions, each motif occurrence was assigned to one of the three topology states TM, nTM or Trans. After deriving motif occurrences within different topologies, a novel information-based approach has been described, which was attached to derive latent physico-chemical information, observable in the variable x-positions, in all investigated motifs [57]. Thereby, Log-Odd ProfiLes **(LOP)** describe the observable physico-chemical properties of each x-position in a closer way. General LOP visualization and classification methods were applied to evaluate correlations between physico-chemical properties and corresponding topology states. For that, Multi-Dimensional Scaling **(MDS)** was used to dimensionality reduction and clustering of multi-dimensional log-odd data. At this, the pre-calculated LOPs of the corresponding motifs are employed as look-up values to compute a straight-forward winner-takes-it-all. To assess the resulting topology state prediction, cross-validation and F-measure calculation was performed. Using position-specific LOP information; the scientific community is now able to describe a motif on its observable topology specific physico-chemical properties in detail. Moreover, for each motif in the big space of membrane protein structures, the allocation of its topological state based on sequence information can be done.

## Sequence Motifs within Neighborhoods

Afterwards, to describe the topology specific construction of a sequence motif in detail, we look at the structural space of α-helices from the perspective of motif neighborhoods. In a further approach, the subject of investigation is the sequential composition of α-helices, more precisely, the investigation of constitutive motifs [58]. The investigation of different membrane protein families like e.g. rhodopsin and secretin family including 7 TM receptors and proteins with domains of unknown functions, leads to a better understanding of the construction of α-helical structures and the involved motifs and their structural or functional role. In this work it has been specified that four directly consecutive motifs are to be considered in one statistical frame, so called Motif Architecture **(MA)**. Each of them is represented by a single graph, united in one result graph for each investigated membrane protein family. Basically, such graphs support the understanding about their topological construction and characteristics, which can be investigated using graph-theoretic algorithms. Related to the investigations of organic molecules [59] and proteins [60-62], frequently observable sub-graphs may be important for a variety of applications e.g. for protein classification or prediction of functional regions [63]. Furthermore, highly conserved sub-graphs can be determined by visual representation tools. In this case, statistically over- and under represented neighborhoods emerge to. At long last, using such simple notational conventions in the context of membrane protein analyses, the understanding of the structural space and the involved features for protein folding can be supported.

## Evolutionary Interacting Patterns as Family-Specific Indicators and their Application

Next, we go back into the structural space, to investigate interacting patterns of TM α-helices in homologous membrane protein families [64]. One of the further aims is the tracing of evolutionary observable variations, within the x-position for each pattern. In combination with derived interaction information from TMPad database [65], the generation of family-specific indicators is aimed. Such indicators having a structural specificity to realise the family-specific fold and the correlated protein function. The derivation of such indicators are to be used for applications like protein classification, structure similarity determination or as general information source for further investigations on mutagen studies. In the further course, this will be evaluated on unknown structures of several membrane protein families. For that, so called pattern alignments leading to the observation of evolutionary influences at the x-positions. In this context, new possible amino acid variations were derived, which could thus not be observed in known structures. Such newly derived variations can be useful for further mutagen investigations and studies. Furthermore, to be an indicator, TMPad interaction information must ensure that two evolutionary influenced pattern stay in spatial contact. This leads to the birth of Evolutionary Influenced Interaction Pattern Pairs **(EIPP)** derived from known structures of major intrinsic proteins, including aquaporins (Pfam-Id: PF00230) and bacteriorhodopsin-like proteins (Pfam-Id: PF01036). Both

families provide homologous protein data where sequential variations can be observed over an evolutionary timescale. In the following, each EIPP can be used to describe the corresponding structural regions within the structure space of the investigated membrane protein families. For this, a mapping task describes the rediscovery of EIPPs in protein sequences of unsolved structures of aquaporine and rhodopsin-like proteins [64]. Additionally, within each EIPP the evolutionary way can be traced and such information can be helpful for the understanding of mutagen effects in membrane protein structures. To illustrate such information in a simple and understandable way, a new graphical output, so called Interaction Block Schema **(IBS),** can close this gap (Figure 4A). Here, an IBS illustrates a two dimensional planar EIPP representation including observable amino acid variations. Such representable information can aid to build a bridge between *in silico* analyses and mutagen investigations. To show this, IBS have been applied on a mutagen investigation task [66]. Here, mutations in Nephrogenic Diabetes Insipidus **(NDI)** involved aquaporine water channel proteins have been investigated. In the following, NDI is described shortly.

## Nephrogenic Diabetes Insipidus

NDI is known as a disorder, which can be acquired as a side effect of surpassing drug taking or which is caused by inherited genetic mutations. Aquaporin-2 water channels and V2R are essential elements in the water re-absorption through the apical cell membrane. Autosomal recessive and dominant inherited NDI are linked to mutations in genes encoding the integral membrane Aquaporin-2 water channel [67,68]. X-linked inheritable NDI is caused by mutations in the gene encoding the AVP type-2 receptor membrane protein (V2R) [69,70]. Aquaporin-2 water channels and V2R are essential elements in the water re-absorption through the apical cell membrane. This water composes the main part of pre-urine; a product that results from ultra-filtration in the kidney. The process of water re-absorption from the pre-urine is essential to ensure the bodys uid balance and is realised by membrane-integrated Aquaporin-2 water channels. The direct inspection of the Aquaporin-2 gene as well as the V2 receptor gene (AVPR2) has become accomplish-able in clinical practice [71] for differential NDI diagnosis and has been substituting dehydration testing over the last years [72].

In the present work, related observable mutational variations causing NDI were derived from UniProt [73,74] and complete the respective IBS. Especially for NDI disease involved Aquaporin-2 water channel proteins, this mixed information can support the structural misfolding of the corresponding interacting sequence parts. For molecular biologist, information is provided for effective drug development.

# RESULTS

One of the main aims was the identification of topology discriminative x-positions, which are crucial for drawing meaningful correlations between physico-chemical properties plus structural and functional features. Thereby, discriminative motifs have been investigated in the structural space of different membrane protein families. A straightforward approach to address this task was the utilisation of a method to determine the residue distribution at each motif x-position.

Different visualisation and classification results showing correlations between LOPs and the three topologies (TM, nTM, trans), where the corresponding motif has been located (Figure 2). This results from LOP distances, which are mainly dictated by the propensities of hydrophobic, hydrophilic and polar amino acids. Summarized, motifs in general can be separated and predicted according to the topology states from its amino acid sequence. An information-based prediction algorithm was implemented and assessed using cross-validation and F-measure evaluation. Motifs showing high F-measures over all or only in certain investigated protein families were identified. From this insight, it can be postulated that short sequence motifs can be divided into structure forming elements, which are present in numerous protein families and highly specific to their topology location. Motifs showing high F-measures only in certain membrane protein families may be important elements in establishing the individual properties which are necessary for the function of a entire protein family.

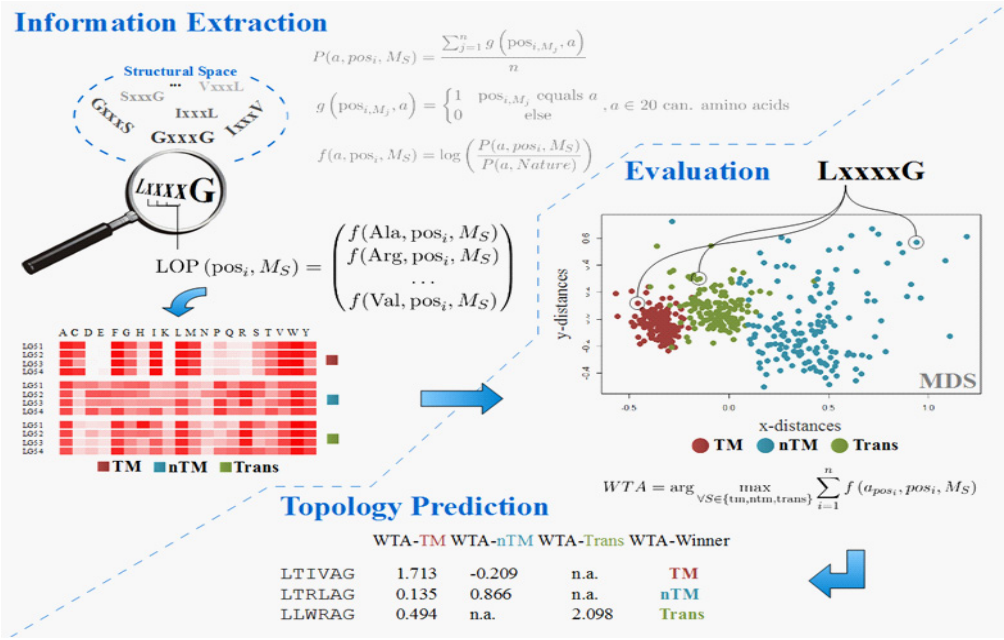


**Figure 2:** Topology separation of discriminative sequence motifs. In the information extraction process, the variable x-positions of discriminative sequence motifs are investigated to register the distribution of topology specific physico-chemical properties. A new information based approach deals with the generation of Log-odd Profiles **(LOP)** for each x-position. Heat map visualization of LOPs makes topology specific distributions apparent. In the evaluation process, LOP data are used for Multidimensional Scaling **(MDS)**. Here, the variable x-positions of all investigated motifs are separable into three topology specific clusters. The arrangement of the clusters correlates to the physico-chemical properties found in membrane and non-membrane located regions, since greater LOP distances are mainly dictated by the propensities of hydrophobic, hydrophilic and polar amino acids. Based on a simple 'winner-takes-it-all' formula, the topology prediction winner for the majority motifs can be determined. In summary it can be said, that it is possible to assign motifs on the physico-chemical properties in the x-positions to well-defined secondary structure elements and topologies.

Related to the high throughput analysis within the structural space of TM α-helices from the perspective of motif neighborhoods', useful information about frequently occurring consecutive motifs has been ascertained for all investigated membrane protein families. The corresponding resulting graphs (Figure 3) provide useful information about highly and less conserved motifs within neighborhoods. Different motifs emerge to structure forming components that are shared amongst to all protein families. Graphs also showing centered motifs, so called hubs (e.g. LL3, LV3, VL3, IL3 and AL3), they often occur together with others within a neighborhood. Such hub-motifs constitute important components within the helical regions. Depending on how TM α-helical structures are constructed, these motifs are required for filling the gaps in the physical and structural context. This hypothesis confirms the results of previous work [75]. The presented results are stating that Ala, Val, Leu or Ilu residues are important members of so called helical wheels, which are related to our representative hub-motifs. Furthermore, available information about TM α-helical non-typical motifs has been revealed, provided by the previous described task for structure topology separation of discriminative sequence motifs. Such motifs indicating to be involved in special protein function. This has been established by their topology non-typical residue distributions that are observable in the x-positions. To evaluate this assumption, networking with known common biological databases like PROSITE [76-79] can be helpful to deliver important information about protein domains, families and functional sites as well associated patterns and Profiles to identify them. As an example, in the present case it can be shown that the α-helical non-typical neighborhood PY3-SN3 seem to be involved in consensus pattern of retinal binding sites within TM receptors of the rhodopsin family (Pfam-Id: Pf00001). Finally, on consideration of all big proteom data processing steps including the final visualization and under networking with biological databases, we are able to build a bridge between graph information and a biological context. Rather, based on sequence motifs one is able to describe the sequential composition in the α-helical space in detail. The provision of such big data in form of graph structures supports further structural comparisons and classification approaches or mutagen investigations. For example, sequential similar regions, detected in sequences of unsolved protein structures, may indicate to be involved in similar structure folding or function related to already solved structures. Generally, in this part it could has been shown how to visualize high-dimensional membrane protein data in form of graph structures and how to fill the lack between high-throughput protein data analyses and evaluation.

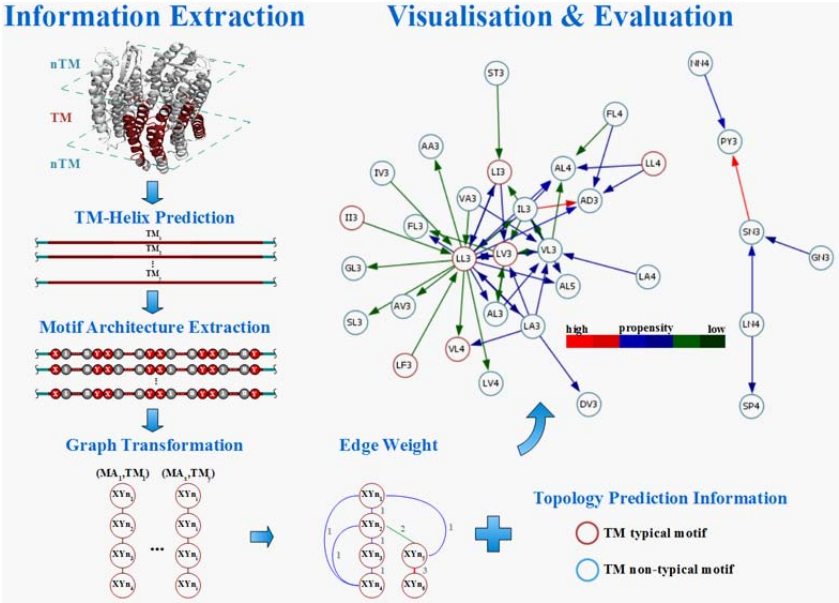


**Figure 3**: Sequence motifs within neighborhoods in the information extraction process, for each investigated membrane protein; all possible Transmembrane **(TM)** α-helices have been predicted using TMHMM [15-17]. Predicted TM sequence parts are colored in red and n TM in blue. After deriving TM region information, all possible motifs with *n-1* highly variable positions by *n = {4-7}* were determined by using a common naive text search algorithm. Further, for each TM sequence part, all possible Motif Architectures **(MA)** consisting of four directly consecutive motifs have been detected. The later application of useful and powerful algorithms which are involved in the statistical information aggregation assumes that each detected MA is considered to be a graph structure. This leads to the transfer of each MA into a graph where each motif can be considered as a node connected by an edge to the following node. The final merge process of all graphs into one resulting main graph includes the registration of co occurring motif neighborhoods. Here, the edge-weight of already existing source and target nodes was updated by an increase by one. Ultimately, a weighted main graph for all investigated membrane protein families exists and its visualization shows over and underrepresented motif neighborhoods connected with red, blue and green colored edges, which have to be examined. Additional information about the predicted topology state for each motif, supplements the visual graph for the detection of TM typical and non-typical motifs.

With the addition of spatial information it can be ensured that two motifs stay in contact within the structural space. Though evolutionary changes in the variable x-positions can be registered, the spatial contact is maintained. This leads to family specific Evolutionary Interaction Patterns **(EIPPs)**. In summary, 116,810 (PF01036) and 63,283 (PF00230) EIPPs were extracted from known structures of the corresponding protein families. Here, EIPPs comprise structural tasks and spatial interactions. Specifically, the evolution presents how EIPPs contribute to

emerge different evolutionary mutation types. These types describe the sequence variability in a closer way, which has no significant influence on the protein structure and function. Within a IBS, evolutionary changes (co-variations) can be observed and evaluated in a simple way (Figure 4A) like e.g. the investigation of the folding core, active core uctuations as well as kinetics and the involved residues of in this case bacteriorhodopsin-like proteins [80-82]. Furthermore, a mapping task describes the rediscovery of EIPPs in protein sequences of unsolved structures of aquaporine and rhodopsin-like proteins. Finally, structure similar regions (Figure 4B) and protein family afiliation (Figure 4C) was determined, which leads to 372 of 438 (PF01036) and 5,993 of 6,420 (PF00230) correct assigned proteins to their families. This ultimately makes EIPPs to family specific indicators. However, this is influenced by the Evolutionary Degree Of Freedom **(EDF)** within EIPPs. Here, the EDF describes the number of permitted x-positions, which were ignored in the EIPP-mapping task. That means, the more static occupied x-positions, the more specific is an EIPP for a membrane protein family. And the more specific a EIPP is, the higher the probability of the family specific rediscovery. Furthermore, EIPPs were marked and finally compared to known structures. Here, the EIPP rediscovery contributes also as useful information source for structural similarity determination (Figure 4A). The sum of covered α- helical regions within TMHMM predicted α-helical regions of the investigated unknown structures can be compared with α-helical regions of known structures. This finally leads to a similarity statistic for each investigated protein sequence of unsolved structure, compared to known structures of aquaporine and rhodopsin-like proteins. Consequently we are able to determine the most similar unknown to known structures of a given α-helical membrane protein family. More precise, information are given in the works of Grunert et al. [64,66]. Finally, the present work also resulted in a good agreement with recently published studies that the evolution provides basic building and interacting blocks for maintaining structure and function. Due to sequence homology such blocks are repeated and we have proven structural conservation. The contemplation of a sequence from the perspective of such blocks facilitates the understanding of how the structure space of a membrane protein family is constructed.

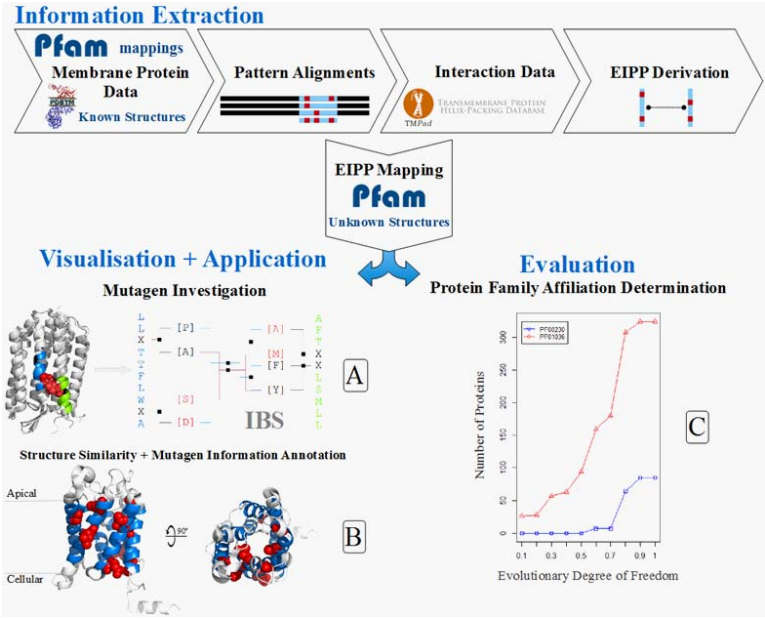


**Figure 4:** Evolutionary Interacting Pattern Pairs **(EIPP)** as family specific indicators and their application in the information extraction process, protein data sets of difierent membrane protein families were obtained from Pfam database [54]. A variety of pattern alignments [64] handle the derivation of observable physico-chemical variants within the variable x-positions. In combination with powerful interaction information from TMPad [65], an EIPP record for each investigated membrane protein family was derived. New visualization tools for the investigation of mutational variants causing diseases were developed. Here, the application of Interaction Block Schemes **(IBS)** can lead to better indicators in laboratory mutagen investigations (A). Using IBS, EIPPs can be evaluated extensively within protein structures. Important information about variable pattern positions that are subjected to a mutation without inuencing attractive pattern interactions can be obtained. Thus, e.g. the folding core and kinetics of bacteriorhodopsin-like proteins can be understood better. This supports the porting of this *in silico* approaches into mutagen investigation studies. Furthermore, a mapping task includes the rediscovery of EIPPs on protein sequences of yet unsolved structures. Under consideration of the Evolutionary Degree Of Freedom **(EDF)**, structural similar regions (B) and the family afiliation (C) of the analyzed protein sequences were determined. Here, the EDF describes the number of permitted x-positions, which were ignored within the EIPP-mapping task.

## CONCLUSION

Generally, the analyses of big membrane protein data based on sequence motif-based approaches can show that membrane protein structures can be studied and understood in more detail, not least to better understand functional changes within cellular processes. The investigation of short sequence motifs supports the understanding about their intended and

unintended changes, observable over the evolutionary time-scale, not least their basic construction within different topologies. Sequence motif-based approaches in general are able to generate meaningful information, which may contribute to study the dynamics of mutant proteins and the effects of mutagens. Accordingly, destabilisation of the three-dimensional structure of membrane proteins caused by mutations or ligand interactions is triggers for numerous diseases. In this context, mutagen effects can lead to disease patterns, which are normally of immense importance to realise family specific structural or functional tasks. For this reason, such patterns need to be analysed in more detail. Besides, mutations are used in the diagnosis of biomarkers. However, the introduced *in silico* approaches have shown, how meaningful structural and functional information can be derived based on sequence motif-based approaches. The final union in visualization tools based on e.g. graphs or Interaction Block Schemes **(IBS)** ultimately leads to better indicators for further laboratory mutagen studies and thus the investigation of disease patterns. Here, derived evolutionary background information within IBS can contribute to support the understanding about the corresponding incorrectly folded interacting structure building blocks. Furthermore, a new introduced information based approach contributes to understand observable amino acid distributions in the variable positions of discriminative sequence motifs, which are required for the topology specific fold within a investigated membrane protein dataset. The high-throughput analysis of different membrane protein families revealed over and under representative motifs within neighborhoods. Summarized, our introduced sequence motif-based approaches in general, support the understanding of anchoring motifs within membrane protein structures and their structural and functional role.

## ACKNOWLEDGEMENT

SG and DL participated in the design of this chapter. SG designed all methods and performed the implementation. SG evaluated the results. DL provided valuable consultation on structural biology and procedural steps. SG and DL wrote the chapter. All authors read and approved the final manuscript. In addition, SG and DL thank to the bioinformatics group members at University of Applied Sciences in Mittweida, which have revised the chapter.

### References

1. Miller SL. A production of amino acids under possible primitive earth conditions. Science. 1953; 117: 528-529.
2. Zvelebil M, Baum JO. Understanding Bioinformatics. New York: Garland Science-Taylor and Fancis Group. 2008.
3. Luckey M. Membrane Structural Biology. Cambridge: Cambridge University Press. 2008.
4. Lam MHY, Stagljar I. Strategies for membrane interaction proteomics: no mass spectrometry required. Proteomics. 12: 1519-1526.
5. Sadowski PG, Groen AJ, Dupree P, Lilley KS. Sub-cellular localization of membrane proteins. Proteomics. 2008; 8: 3991-4011.
6. Holtzhauer M, Behlke J. Methoden in der Proteinanalytik. Berlin: Springer. 1996.
7. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982; 157: 105-132.
8. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature. 1982; 299: 371-374.
9. Heijne G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J. 1986; 5: 3021-3027.
10. Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ. A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng. 1999; 12: 381-385.

11. Hofmann K, Stoffel W. TMbase - A database of membrane spanning proteins segments. Biol. Chem. Hoppe-Seyler. 1993; 374:166.

12. von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol. 1992; 225: 487-494.

13. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics. 1998; 14: 378-379.

14. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all helical membrane protein structure and topology. Biochemistry. 1994; 33: 3038-3049.

15. Sonnhammer EL, von Heijne G, Krogh A. A hidden markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998; 6: 175-182.

16. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001; 305: 567-580.

17. Möller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics. 2001; 17: 646-653.

18. Tusnády GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol. 1998; 283: 489-506.

19. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001; 17: 849-850.

20. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics. 2008; 24: 1662-1668.

21. Hayat S, Elofsson A. BOCTOPUS: improved topology prediction of transmembrane Î² barrel proteins. Bioinformatics. 2012; 28: 516-522.

22. Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics. 2004; 20: 1565-1572.

23. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 1999; 286: 295-299.

24. Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol. 2003; 10: 59-69.

25. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH. Evolutionary information for specifying a protein fold. Nature. 2005; 437: 512-518.

26. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in arti cial ww domains. Nature. 2005; 437: 579-583.

27. Bartlett GJ, Taylor WR. Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction. Proteins. 2008; 71: 950-959.

28. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins. 1994; 19: 55-72.

29. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci. 1995; 4: 521-533.

30. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identi cation of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A. 2009; 106: 67-72.

31. Gomes M, Hamer R, Reinert G, Deane CM. Mutual information and variants for protein domain-domain contact prediction. BMC research notes. 2012; 5: 472.

32. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, et al. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012; 149: 1607-1621.

33. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28: 184-190.

34. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys. 2013; 87: 012707.

35. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. Journal of Computational Physics. 2014; 276: 341-356.

36. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics. 2011; 79: 1061-1078.

37. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolutionbased residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013; 110: 15674-15679.

38. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife. 2014; 3: e02030.

39. Carl Ivar Branden, John Tooze. Introduction to protein structure. New York: Garland Science-Taylor and Fancis Group. 1999.

40. DeGrado WF, Gratkowski H, Lear JD. How do helix-helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles. Protein Sci. 2003; 12: 647-665.

41. Von Heijne G. Membrane protein assembly in vivo. Adv Protein Chem. 2003; 63: 1-18.

42. Bowie JU. Solving the membrane protein folding problem. Nature. 2005; 438: 581-589.

43. Walters RF, DeGrado WF. Helix-packing motifs in membrane proteins. Proc Natl Acad Sci U S A. 2006; 103: 13658-13663.

44. Chamberlain AK, Faham S, Yohannan S, Bowie JU. Construction of helix-bundle membrane proteins. Adv Protein Chem. 2003; 63: 19-46.

45. Lehnert U, Xia Y, Royce TE, Goh CS, Liu Y, et al. Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions. Q Rev Biophys. 2004; 37: 121-146.

46. Liu Y, Engelman DM, Gerstein M. Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biol. 2002; 3: research0054.

47. Senes A, Gerstein M, Engelman DM. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J Mol Biol. 2000; 296: 921-936.

48. Senes A, Engel DE, DeGrado WF. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr Opin Struct Biol. 2004; 14: 465-479.

49. Arkin IT, Brunger AT. Statistical analysis of predicted transmembrane alpha-helices. Biochim Biophys Acta. 1998; 1429: 113-128.

50. Russ WP, Engelman DM. The GxxxG motif: a framework for transmembrane helix-helix association. J Mol Biol. 2000; 296: 911-919.

51. Mueller BK, Subramaniam S, Senes A. A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical c-h hydrogen bonds. Proc Natl Acad Sci U S A. 2014; 111: E888-E895.

52. Lemmon MA, Flanagan JM, Treutlein HR, Zhang J, Engelman DM. Sequence speci city in the dimerization of transmembrane. alpha-helixes. Biochemistry. 1992; 31: 12719-12725.

53. MacKenzie KR, Prestegard JH, Engelman DM. A transmembrane helix dimer: structure and implications. Science. 1997; 276: 131-133.

54. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J. The Pfam protein families database. Nucleic Acids Res. 2012; 40: D290-301.

55. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22: 1658-1659.

56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403-410.

57. Grunert S, Heinke F, Labudde D. Structure topology prediction of discriminative sequence motifs in membrane proteins with domains of unknown functions. Structural Biology. 2013.

58. Grunert S, Labudde D. Graph representation of high-dimensional alpha-helical membrane protein data. BioData Min. 2013; 6: 21.

59. Borgelt C, Berthold MR. Mining molecular fragments: Finding relevant substructures of molecules. Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. 2002; 51-58.

60. Milik M, Szalma S, Olszewski KA. Common Structural Cliques: a tool for protein structure and function analysis. Protein Eng. 2003; 16: 543-552.

61. Huan J, Bandyopadhyay D, Wang W, Snoeyink J, Prins J, et al. Comparing graph representations of protein structure for mining family-speci c residuebased packing motifs. Journal of Computational Biology. 2005; 12: 657-667.

62. Dhifli W, Saidi R, Nguifo EM. Smoothing 3d protein structure motifs through graph mining and amino acid similarities. Journal of Computational Biology. 2014; 21: 162-172.

63. Dhifli W, Diallo AB. PGR: A novel graph repository of protein 3d-structures. J Data Mining Genomics Proteomics. 2015; 6: 2153-0602.

64. Grunert S, Labudde D. The observation of evolutionary interaction pattern pairs in membrane proteins. BMC structural biology. 2015; 15: 6.

65. Lo A, Cheng CW, Chiu YY, Sung TY, Hsu WL. TMPad: an integrated structural database for helix-packing folds in transmembrane proteins. Nucleic Acids Res. 2011; 39: D347-355.

66. Grunert S, Labudde D. Evolutionary influenced interaction pattern as indicator for the investigation of natural variants causing nephrogenic diabetes insipidus. Computational and Mathematical Methods in Medicine. 2015.

67. Deen PM, Verdijk MA, Knoers NV, Wieringa B, Monnens LA. Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine. Science. 1994; 264: 92-95.

68. Mulders SM, Bichet DG, Rijss JP, Kamsteeg EJ, Arthus MF. An aquaporin-2 water channel mutant which causes autosomal dominant nephrogenic diabetes insipidus is retained in the Golgi complex. J Clin Invest. 1998; 102: 57-66.

69. van den Ouweland AM, Dreesen JC, Verdijk M, Knoers NV, Monnens LA. Mutations in the vasopressin type 2 receptor gene (AVPR2) associated with nephrogenic diabetes insipidus. Nat Genet. 1992; 2: 99-102.

70. Rosenthal W, Seibold A, Antaramian A, Lonergan M, Arthus MF, et al. Molecular identi cation of the gene responsible for congenital nephrogenic diabetes insipidus. Nature. 1992; 359: 233-235.

71. Fujiwara TM, Bichet DG. Molecular biology of hereditary diabetes insipidus. J Am Soc Nephrol. 2005; 16: 2836-2846.

72. Los EL, Deen PMT, Robben JH. Potential of nonpeptide (ant)agonists to rescue vasopressin v2 receptor mutants for the treatment of x-linked nephrogenic diabetes insipidus. J Neuroendocrinol. 2010; 22: 393-399.

73. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford). 2011; 2011: bar009.

74. Jungo F, Bougueleret L, Xenarios I, Poux S. The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. Toxicon. 2012; 60: 551-557.

75. Schiffer M, Edmundson AB. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. Biophys J. 1967; 7: 121-135.

76. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, et al. New and continuing developments at prosite. Nucleic Acids Res. 2013; 41: 344-347.

77. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, et al. Prosite: a documented database using patterns and pro les as motif descriptors. Brief Bioinform. 2002; 3: 265-274.

78. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006; 34: W362-365.

79. Sigrist CJ, De Castro E, Langendijk-Genevaux PS, Le Saux V, Bairoch A, et al. Prorule: a new database containing functional and structural information on prosite pro les. Bioinformatics. 2005; 21: 4060-4066.

80. Wood K, Lehnert U, Kessler B, Zaccai G, Oesterhelt D. Hydration dependence of active core fluctuations in bacteriorhodopsin. Biophys J. 2008; 95: 194-202.

81. Curnow P, Di Bartolo ND, Moreton KM, Ajoje OO, Saggese NP. Stable folding core in the folding transition state of an alpha-helical integral membrane protein. Proc Natl Acad Sci U S A. 2011; 108: 14133-14138.

82. Schlebach JP, Cao Z, Bowie JU, Park C. Revisiting the folding kinetics of bacteriorhodopsin. Protein Sci. 2012; 21: 97-106.

83. http://pfam.xfam.org/search/keyword?query=DUF