Hindawi Publishing Corporation
Structural Biology
Volume 2013, Article ID 249234, 10 pages
http://dx.doi.org/10.1155/2013/249234

Research Article

Structure Topology Prediction of Discriminative Sequence Motifs in Membrane Proteins with Domains of Unknown Functions

Steffen Grunert, Florian Heinke, and Dirk Labudde

Hochschule Mittweida, University of Applied Sciences, Technikumplatz 17, 09648 Mittweida, Germany

Correspondence should be addressed to Steffen Grunert; sgrunert@hs-mittweida.de and Dirk Labudde; labudde@hs-mittweida.de

Received 31 October 2012; Accepted 15 January 2013

Academic Editor: Shandar Ahmad

Motivation. Membrane proteins play essential roles in cellular processes of organisms. Photosynthesis, transport of ions and small molecules, signal transduction, and light harvesting are examples of processes which are realised by membrane proteins and contribute to a cell's specificity and functionality. The analysis of membrane proteins has shown to be an important part in the understanding of complex biological processes. Genome-wide investigations of membrane proteins have revealed a large number of short, distinct sequence motifs. *Results*. The in silico analysis of 32 membrane protein families with domains of unknown functions discussed in this study led to a novel approach which describes the separation of motifs by residue-specific distributions. Based on these distributions, the topology structure of the majority of motifs in hypothesised membrane proteins with unknown topology can be predicted. *Conclusion*. We hypothesise that short sequence motifs can be separated into structure-forming motifs on the one hand, as such motifs show high prediction accuracy in all investigated protein families. This points to their general importance in α-helical membrane protein structure formation and interaction mediation. On the other hand, motifs which show high prediction accuracies only in certain families can be classified as functionally important and relevant for family-specific functional characteristics.

1. Introduction

Membrane proteins are essential for many fundamental biological processes within organisms. Active nutrient transport, signal and energy transduction, and ion flow are only a few of the numerous functions enabled by membrane proteins [1]. Membrane proteins obtain their specific functionality by individual folding and interactions with the hydrophobic membrane environment as well as, in many cases, by oligomeric complex formation and protein-protein interactions [1, 2]. The identification of such complexes and interactions is valuable, since, on the one hand, detailed information of the function of an unknown membrane protein can be obtained by analysing its interactions with proteins of known function. On the other hand, biological processes can be comprehended as a dynamically fluctuating system, whereby the biological role of the unknown membrane protein can be defined more precisely [3, 4]. Accordingly, destabilisation of the three-dimensional structure of a membrane protein caused by mutations or ligand interactions are triggers for numerous diseases, for example, diabetes insipidus, cystic fibrosis, hereditary deafness and retinitis pigmentosa [5–7].

Although 20%–30% of all open reading frames of a typical genome are encoding membrane proteins [5, 8, 9] and 60% of all drug targets are membrane proteins [2], membrane proteomics is still an experimentally challenging field due to poor protein solubility, wide intracellular concentration range, and thus, inaccessibility to many proteomics methodologies [10]. Hence, the number of known three-dimensional structures is relatively small, with 394 nonredundant membrane protein chains currently available [11–13]. Therefore, there is a necessity for approaches that allow to predict structural and functional features of unknown membrane proteins. A variety of methods have been developed to predict structural features from sequence, such as α-helical membrane-spanning helices and extra/intracellular domains

(i.e., TMHMM [14], PHDhtm [15], MEMSAT3 [16]) as well as membrane-spanning beta-strands of transmembrane β-barrel proteins (i.e., BOCTOPUS [17]). Furthermore, in genome-wide membrane protein sequence analyses, numerous short conserved sequence motifs were identified [18]. As an example, the most widely discussed GxxxG motif has been shown to be significantly present in transmembrane α-helices. With both glycines resting on one side of the helix as spatially neighbouring residues and by that forming a smooth helix membrane surface, structural studies confirmed that the GxxxG motif plays an important part in mediating helix-helix interactions [18–22]. In general, short conserved membrane protein motifs are considered to be significantly relevant for membrane protein folding and structural stability as well as being involved in defining a protein's function. Hence, sequence motif analyses and resulting insights can support the understanding of protein dynamics. Information can be derived which may contribute to study the dynamics of mutant proteins and the effects of mutagens [23–25]. Additionally, as addressed in [26], the analysis of sequence motifs in proteins with similar function or structure might help to identify essential functional sites and locations which contribute to structural stability.

In this work, we focused on previous studies and results that have been reported by Liu and colleagues [18]. In the process, various integral membrane protein families with polytopic membrane domains had been obtained from Pfam database [27]. As part of their studies, locations of the least conserved residues (glycine, proline, and tyrosine) in α-helical transmembrane regions had been investigated. As a result, short motifs consisting of pairs of small residues (glycine, alanine, and serine) surrounding single or multiple variable positions had been identified in conserved sequences and Pfam-classified families. Based on these results, we have developed a prediction approach to allocate the topological state of a sequence motif in the protein structure based on sequence information. We have used cross-validation to verify the prediction accuracy. However, prediction accuracy has been found to be variable for certain motifs with regard to the investigated protein families. According to this, we hypothesise that short sequence motifs can be separated into structure-forming motifs on the one hand, as such motifs show high prediction accuracy in all investigated protein families. This points to their general importance in α-helical membrane protein structure formation and interaction mediation. On the other hand, motifs which show high prediction accuracies only in certain families can be classified as functionally important and relevant for family-specific functional characteristics.

2. Materials and Methods

2.1. Used Membrane Protein Families. As the first step of our analysis, 32 membrane protein families with domains of unknown functions (DUF) were obtained from the Pfam database [27] using extended keyword searching. All 7051 sequences were retrieved for statistical analysis. The full list of employed membrane protein families is given in Table 1.

Table 1: Thirty-two membrane protein families were derived from Pfam database [28] and employed for statistical analysis.

Accession	Family
PF09767	DUF2053
PF09834	DUF2061
PF09842	DUF2069
PF09843	DUF2070
PF09852	DUF2079
PF09858	DUF2085
PF09874	DUF2101
PF09877	DUF2104
PF09878	DUF2105
PF09879	DUF2106
PF09880	DUF2107
PF09881	DUF2108
PF09882	DUF2109
PF09900	DUF2127
PF09913	DUF2142
PF09925	DUF2157
PF09945	DUF2177
PF09946	DUF2178
PF09971	DUF2206
PF09972	DUF2207
PF09973	DUF2208
PF09980	DUF2214
PF09990	DUF2231
PF09991	DUF2232
PF09997	DUF2238
PF10002	DUF2243
PF10011	DUF2254
PF10067	DUF2306
PF10080	DUF2318
PF10081	DUF2319
PF10097	DUF2335
PF10101	DUF2339

Subsequently, 50 sequence motifs, identified by Liu and colleagues [18], were localised in the obtained set of families.

2.2. Programs and Tools. To avoid generating misguiding statistics by including identical or highly similar sequences, a set of nonredundant sequences was generated. Here, we defined the sequence redundancy threshold at 25% sequence identity. In the first step of sequence processing, CD-HIT [29] was applied for first clustering. However, CD-HIT accepts only nonredundancy thresholds of >40%. This limitation is caused by the internal word-length filtering approach and statistical presets. Hence, to ensure clustering sensitivity, a 60% nonredundancy threshold, which corresponds to tetrapeptide word filtering used by the program, was applied. In the second step, sequence clustering using the 25% redundancy threshold was obtained by means of utilising BLASTClust [30]. The representative sequences of all clusters were extracted, leading to a set of 2511 nonredundant sequences.

Subsequently, the determination of membrane and nonmembrane associated sequence regions was derived using by the TMHMM Server v. 2.0 [14]. Basically, TMHMM performs a prediction of intra/extracellular regions and integral membrane helices based on sequence. Additionally, the probability of the prediction is given for each residue as well. According to the obtained results from TMHMM, a topological state was assigned to each residue. A residue was assigned as "TM" if the posterior prediction probability of this residue being a part of a membrane helix has been found to be greater than 90%. If the posterior prediction probability of the residue has been found to be greater 90% for extra/intracellular prediction, the residue was assigned as "nTM."

2.3. Used Motifs. The short sequence motifs analysed in our work have been reported in [18]. In this study, Liu et al. analysed consensus sequences of 168 Pfam-A families to identify significant amino acid pair motifs. By the comparison of their results in earlier published findings (see [20]), a list of 50 significant motifs has been derived which we used in our work (for original data see [18], Table 1, List 3): GG4, GL3, GG7, GL1, AG7, GA7, AG4, PL2, AS4, AL6, LP1, PG9, GA4, FG1, SL1, SG4, PL1, AA7, AG5, LF8, IA1, GV1, AI1, AA2, GL2, AA3, SL2, PG5, PG6, IL4, GS5, VL4, GV2, IG1, PG10, LY6, LF10, SA6, LG5, SA3, PF1, GS4, IV4, LS1, GY8, IG2, LF9, VF8, VG6, GN4.

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. However, in the process of analysis we found that short motifs with a relatively small number of variable positions (more precisely, if n is found to be <3) do not contain enough information to be investigated by our approach. Thus, these motifs have been discarded in the process, which resulted in a final set of 33 sequence motifs. In our nonredundant sequence set, almost 250,000 single motif occurrences were identified. As an example of motifs located in a membrane protein structure, Figure 1 illustrates seven motifs which can be found in the structure of the bacteriorhodopsin (PDB-Id: 1brr).

2.4. Information Extraction and Clustering. In this work, a novel approach is elucidated which predicts the topology state of a short sequence motif in membrane proteins. The following steps were completed to realise this approach.

At first, all single motif occurrences were identified in the nonredundant sequence set. Including TMHMM predictions, each motif occurrence was assigned to a topology state as elucidated in Section 2.2. Additional to the defined topology states "TM" and "nTM," a further state has been defined for this study. Each motif, where the beginning and the end has been located in the different topology states "TM" and "nTM," has been assigned with the "trans" state. Subsequently, all variable positions within each motif occurrence were examined more closely. Ultimately for each variable position, the relative occurrence of each amino acid at the specified position of each motif was calculated.

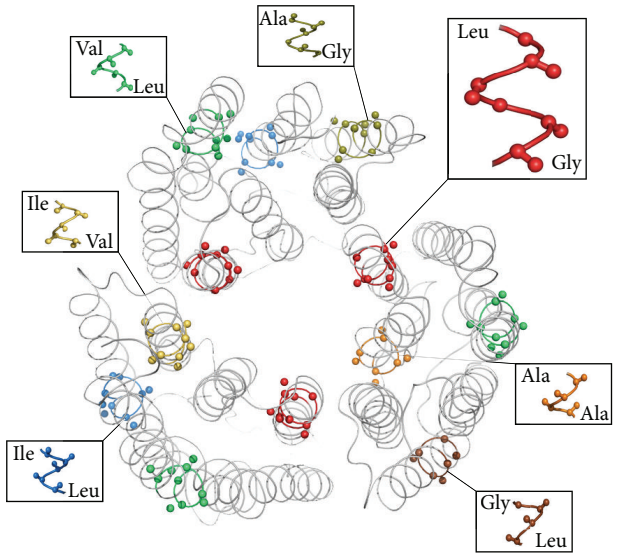

Figure 1: In the bacteriorhodopsin trimer (PDB-Id: 1brr), seven of 33 sequence motifs which were analysed in this study are present. Each motif can be written in a regular expression like *XYn*, where *X* and *Y* are amino acids separated by $n-1$ highly variable positions. For example, the LG5 motif occurrence (highlighted in red) corresponds to a pair of leucine and glycine residues which are separated by four amino acids.

To define a separation rule for the investigated motifs, an information-based approach was applied. Formally, a motif M, for instance LG5, can be interpreted as a set of variable strings with a length of n. Intuitively, in case of LG5 n equals 4. To include the membership information of the three topology states, we separated M into three motif subsets M_{TM}, M_{nTM} and M_{trans} according to the topology state S in which each single motif occurrence $m \in M$ is located. Furthermore, in each motif M_S each position pos_i with $i \in [1,n]$ can be investigated concerning its amino acid distribution. To this end, interpreting M_S as a set of strings $m_1, m_2, \ldots, m_k$ (all identified motif occurrences found in topology state S) allows formulating the relative probability $P(a \mid \mathrm{pos}_i \mid M_S)$:

$$P\left(a \mid \mathrm{pos}_i \mid M_S\right) = \frac{\sum_{j=1}^{k} g\left(\mathrm{pos}_{i,m_j}, a\right)}{k}, \tag{1}$$

with

$$g\left(\mathrm{pos}_{i,m_j}, a\right) = \begin{cases} 1 & \mathrm{pos}_{i,m_j} \text{ equals } a \\ 0 & \text{else,} \end{cases} \tag{2}$$

where a corresponds to one of the 20 canonical amino acids. To weight the significance of each probability $P(a \mid \mathrm{pos}_i \mid M_S)$, the probability $P(a \mid \mathrm{Nature})$ is applied in a log-odd formula:

$$f\left(a \mid \mathrm{pos}_i \mid M_S\right) = \log\left(\frac{P\left(a \mid \mathrm{pos}_i \mid M_S\right)}{P\left(a \mid \mathrm{Nature}\right)}\right). \tag{3}$$

The amino acid distribution $P(a \mid \mathrm{Nature})$ used to test the significance of the observed relative probability at each

motif position was computed from the NCBI nonredundant protein sequence set [32] (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz).

Using these log-odd values, visualisation, clustering, and information extraction can be performed. To this end, we transformed each position pos_i into a vector consisting of log-odd values which we refer to as log-odd profile $\text{LOP}(\text{pos}_i \mid M_S)$ and which is defined as

$$\text{LOP}(\text{pos}_i \mid M_S) = \begin{pmatrix} f(\text{Ala} \mid \text{pos}_i \mid M_S) \\ f(\text{Arg} \mid \text{pos}_i \mid M_S) \\ \cdots \\ f(\text{Val} \mid \text{pos}_i \mid M_S) \end{pmatrix}. \quad (4)$$

Clustering all resulting $\text{LOP}(\text{pos}_i \mid M_S)$ was finally ensured by implementing the following distance D formula:

$$D\left(\text{LOP}(\text{pos}_i \mid M_S), \text{LOP}\left(\text{pos}_j \mid M_S\right)\right) = 1 - \rho\left(\text{LOP}(\text{pos}_i \mid M_S), \text{LOP}\left(\text{pos}_j \mid M_S\right)\right), \quad (5)$$

where $\rho(\text{LOP}(\text{pos}_i \mid M_S), \text{LOP}(\text{pos}_j \mid M_S))$ corresponds to the Spearman's rank correlation coefficient. Clustering methods were applied to the LOPs to derive characteristics in motifs which determine the protein's structural and functional features.

Furthermore, with these values at hand, the algorithm for predicting the topology state S based on a single motif occurrence m was implemented. At this, the precalculated LOPs of the corresponding motif M are employed as look-up values to compute a straight-forward winner-takes-it-all formula:

$$S = \arg \max_{\forall S \in \{\text{TM,nTM,trans}\}} \sum_{i=1}^{n} f\left(a_{m_i} \mid m_i \mid M_S\right). \quad (6)$$

The assessment of topology state prediction was performed by means of cross-validating and F-measure calculation.

By utilising clustering methods, differences and similarities of all LOPs can be visualised and analysed in detail.

For dimensionality reduction and finally data clustering of the 20-dimensional LOP data, we used the unweighted pair group method with arithmetic mean (UPGMA) [33] and the exploratory observation machine (XOM) [34]. This analysis is helpful to understand the correspondences of physicochemical properties observed in LOPs and topology states. Furthermore, this analysis enforces the found predictability of topology states. We chose the UPGMA as visualisation approach, since it is a widely used bottom-up clustering method that can be understood intuitively.

The XOM algorithm is relatively new for dimensionality reduction. A great advantage lies in its visualisation capabilities, since it can transform neighbourhood or distance relations embedded in multidimensional data into human-intelligible spaces, such as into $\mathbb{R}^2$. In the literature, this property is referred to as topology-preserving mapping. However, the degree of topology-preserving mapping achieved by the XOM depends on the given problem (mainly influenced by the structure of data and applied distance measure), and thus the XOM output can be insufficient for analysis. In application to LOP data, however, it has shown to perform more than satisfying. Further, visualisations were obtained by generating heat maps.

3. Results and Discussion

3.1. Identification of Topology-Discriminative Positions. The identification of topology-discriminative positions in motifs is crucial for drawing meaningful correlations between physicochemical properties plus structural and functional features. A straight-forward approach to address this task is the utilisation of a method to determine the residue conservation at each variable motif position. WebLogo [31], for instance, is a widely used method to address such problems. However, WebLogo does not include any amino-acid-specific background information in deriving residue conservation, since natural amino acid frequencies are not taken into account. To circumvent this problem, we used LOPs for visualisation instead, which, as shown in (4), include natural amino-acid-specific background probabilities. Essentially, this approach is quite similar to the methods recently described in [36]. Single LOPs can be visualised as heat maps [37] (see Figure 3), and amino-acid-specific propensities at each variable position in each motif can be extracted and thus information can be gained.

3.2. LOP Visualisation and Classification. The LOP heat map depicted in Figure 3 exemplary shows the apparent amino-acid-specific propensities according to the three topology states. Here, increasing amino acid propensities defined in (3) are illustrated by increasing red colour content. In comparison to the WebLogos (Figure 2), distinct amino acid propensities become obvious. For instance, glycine is observed more frequently in all LG5 motifs which are located in transmembrane regions. In nontransmembrane regions, the propensity of glycine is found to be reduced distinctly. As a second example, the LG5 motif found in transmembrane regions, leucine is observed more frequently at the third variable position as at other positions. This sequence constellation results into two spatially adjacent leucine residues that form a bulky helix surface. In general, relations of topology states and the amino-acid-specific propensities can be derived. This emphasises the predictability of topology states based on single motif occurrences. The full LOP heat map generated by this approach consists of 471 motif positions. To visualise LOP-wide correspondences, we applied UPGMA hierarchical clustering as well as the XOM algorithm. Distance measurement between LOPs was realised by utilising (5). Since 471 variable motif positions were investigated, the UPGMA-tree generated by the first approach consists of 471 nodes. To ease the analysis of the tree, the nodes were coloured according to the topological state in which the corresponding motif is located. Due to the huge number of nodes, we depicted the tree only as a schematically representation which represents the observed general tree topology and identified memberships (see Figure 4). As shown, a distinct clustering, more precisely a formation of

(a) (b) (c)

FIGURE 2: WebLogos [31] of the LG5 motif in the order of three topology states "TM" (a), "nTM" (b), and "trans" (c). However, the symbol height in each logo reflect only the relative occurrence of the corresponding amino acid. Additionally, background amino-acid-specific frequencies are not taken into account which decreases the sensitivity of this method. Compared to the heat map generated from LOPs (see Figure 3), less information can be gained. By applying WebLogo, residue propensities, with regard to the topology states, cannot be derived or identified. For instance, the leucine amino acid in "TM" (WebLogo A) cannot be observed as more frequently at the third variable position as at other positions.

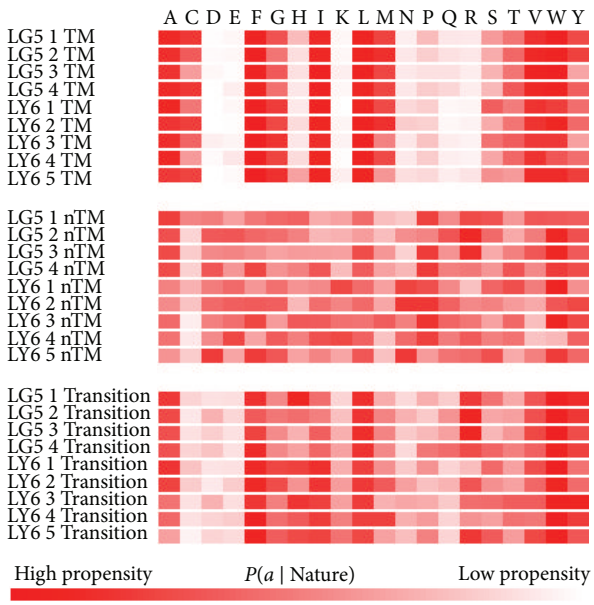

FIGURE 3: LOP heat maps of the LG5 and LY6 motif. LOP heat maps reflect the propensities of each amino acid relative to natural amino-acid-specific frequencies. Increasing amino acid propensities are illustrated by an increased red colour content. The below listed colour scale represents the colour assigning to each amino acid propensities. This visualisation allows a sensitive approach to analyse amino acid propensities of each variable position of a motif according to topology states. Here, the LOP heat map is separated by topology states, so that amino acid propensities become obvious. For example, cysteine can be observed more frequently at the second variable position of transmembrane-located LY6 motif. This results in two spatially adjacent cysteine residues which form a bulky surface in transmembrane helices. Such a bulky helix surface might be important in mediating helix-helix interactions, as knob-to-hole helix packing has been reported as a key folding process in many studies (e.g., [1, 35]).

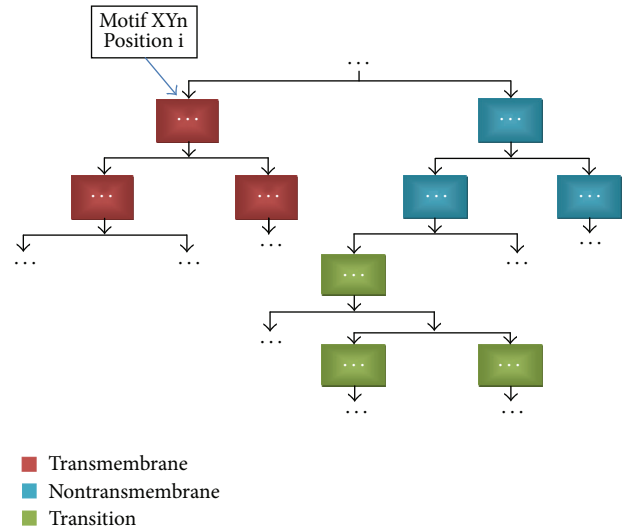

FIGURE 4: Schematic UPGMA-tree derived from LOP clustering: the 471 LOPs of all variable motif positions were clustered using UPGMA hierarchical clustering [33] by utilising the LOP distance measure defined in (5). Due to the original size, the resulting UPGMA-tree is only depicted schematically. However, the tree shows three separated, distinct subtrees which correlate to the topology states in which the corresponding motifs are located. The cluster arrangement corresponds to amino acid propensities and thus to physicochemical properties observed in motifs. This tree proves that the topological location of short sequence motifs are well separable and especially predictable from their amino acid sequence in the variable positions.

three distinct subtrees, according to the topology states is obvious. The cluster arrangement correlates to the physicochemical properties found in membrane and nonmembrane located regions, since greater LOP distances are mainly dictated by the propensities of hydrophobic, hydrophilic, and polar amino acids. The sub-tree mainly consisting of motifs located in "trans" regions is arranged in between, which points to intermediate physicochemical motif compositions and equally distributed amino acid compositions. Similar to these findings, the XOM output (see Figure 5) shows three main clusters which correspond to the topology states too. Additionally, the cluster arrangement is found to be equal to the arrangement observed in the UPGMA-tree, where the causes of cluster formation are analogue as well. The distinct cluster formation observed by the output of both methods points to a good separability of the variable motif positions.

A possible approach to predict the topology state of a motif from the amino acid sequence alone was implemented as elucidated in Section 2.4. In this calculation, for each motif, the three log-odd sums of all variable positions are computed

Table 2: Statistical analyses of the motifs in the protein families with domains of unknown functions (EDS1). The results are split into three subtables. The "TMHMM prediction," the "Prediction on log-odds," and the "*F*-measures"-table. Thereby the "TMHMM prediction"-table represents the absolute occurrences of a motif in all investigated protein families with domains of unknown functions. The "Prediction on log-odds"-table represents the topology state winners (see (6)) followed by the "*F*-measures"-table which indicates how good or bad a motif can be separated and assigned to a topology state.

Motif	TMHMM prediction			Prediction on log-odds			*F*-measures		
	TM	nTM	Trans	TM	nTM	Trans	TM	nTM	Trans
PG10	430	1556	900	429	1556	901	0.997	1.0	0.998
LF10	2838	1535	2860	2840	1536	2857	0.998	0.998	0.999
PG9	572	1596	896	577	1590	897	0.99	0.998	0.995
LF9	3271	1392	2425	3272	1392	2424	0.998	0.997	0.999
VF8	1936	1065	1116	1933	1065	1119	0.998	0.999	0.996
LF8	3589	1446	2185	3583	1447	2190	0.998	0.998	0.997
GY8	775	863	685	771	860	692	0.995	0.998	0.995
GA7	3035	2907	1943	3047	2889	1949	0.996	0.996	0.993
AG7	3009	2939	2104	3016	2926	2110	0.995	0.997	0.992
AA7	5100	4623	2883	5124	4592	2890	0.993	0.995	0.99
GG7	2380	3171	1463	2373	3175	1466	0.99	0.997	0.987
LY6	1861	1315	1263	1873	1305	1261	0.993	0.995	0.989
VG6	2518	2331	1317	2536	2324	1306	0.987	0.993	0.981
SA6	1747	2683	1269	1757	2674	1268	0.987	0.997	0.985
PG6	566	1756	681	583	1745	675	0.983	0.997	0.982
AL6	6974	3789	2931	7155	3680	2859	0.981	0.98	0.969
PG5	640	1576	696	682	1542	688	0.955	0.989	0.957
GS5	1041	2161	763	1115	2097	753	0.951	0.98	0.959
LG5	4775	3050	1959	5071	2879	1834	0.951	0.952	0.919
AG5	3464	3092	1433	3761	2895	1333	0.942	0.958	0.908
GN4	228	952	271	276	891	284	0.869	0.96	0.919
IV4	3285	1562	723	3568	1339	663	0.905	0.861	0.765
IL4	5700	2244	1282	6209	1889	1128	0.879	0.773	0.699
GS4	1080	2356	651	1381	2063	643	0.807	0.905	0.791
GG4	2302	3822	893	2758	3387	872	0.814	0.89	0.714
SG4	1125	2542	723	1457	2228	705	0.796	0.908	0.789
VL4	6680	3046	1592	7202	2498	1618	0.847	0.737	0.634
AS4	1903	2807	946	2423	2311	922	0.795	0.854	0.769
GA4	3769	3300	1253	4463	2664	1195	0.823	0.807	0.698
AG4	3594	3456	1214	4381	2661	1222	0.813	0.795	0.695
SA3	2005	2965	728	2603	1901	1194	0.65	0.674	0.452
AA3	6719	5358	1327	7855	3199	2350	0.747	0.596	0.386
GL3	5758	3252	1343	6026	2066	2261	0.767	0.597	0.452

with respect to the three topology states. The highest log-odd sum leads to the topology state winner (see (6)). Cross-validation was performed by excluding the evaluation set of motifs from the training motif set, which was used to generate the look-up log-odd values. In the process, each topology state winner has been assessed by *F*-measure. The corresponding *F*-measures for each investigated sequence motif are listed in the given result Tables 1, 2, and 3. It is apparent from these tables that there are motifs with high and rather small *F*-measures. Each representative *F*-measure value indicates how good or bad a motif can be separated and assigned to the respective topology state. For example, the LY6 motif with an *F*-measure >0.8 in all result tables says that this motif is well assignable (by (6)) to each topology state.

3.3. Evaluation of the Prediction Accuracy. To evaluate the prediction accuracy, our new approach has been applied to three datasets. The first dataset (EDS1) consists of DUF-families sequence information described in previous Section 2.1. The second dataset (EDS2) consists of 2254 membrane protein sequences with 55 known structures of the bacteriorhodopsin-like protein (PF01036) family. EDS2 was also obtained from Pfam database [27]. EDS1 and EDS2 include the topology specific recorded statistically occurrence

Table 3: Statistical analyses of the motifs in the bacteriorhodopsin-like protein families (EDS2). The results are split into three subtables. The "TMHMM prediction," the "Prediction on log-odds," and the "*F*-measures"-table. Thereby the "TMHMM prediction"-table represents the absolute occurrences of a motif in all investigated bacteriorhodopsin-like protein families. The "Prediction on log-odds"-table represents the topology state winners (see (6)) followed by the "*F*-measures"-table witch indicates how good or bad a motif can be separated and assigned to a topology state.

Motif	TMHMM prediction			Prediction on log-odds			*F*-measures		
	TM	nTM	Trans	TM	nTM	Trans	TM	nTM	Trans
PG10	105	17	464	103	17	466	0.942	1.0	0.987
LF10	1900	131	1165	2147	214	835	0.864	0.655	0.782
PG9	187	61	395	223	63	357	0.893	0.952	0.944
LF9	1278	164	565	1170	307	530	0.852	0.586	0.842
VF8	739	118	623	796	104	580	0.945	0.928	0.943
LF8	654	168	362	625	209	350	0.916	0.78	0.966
GY8	715	185	1450	881	186	1283	0.876	0.981	0.928
GA7	1581	2013	1963	1618	1877	2062	0.889	0.95	0.935
AG7	1737	722	1347	1653	782	1371	0.919	0.92	0.924
AA7	1887	1618	1455	1936	1530	1494	0.922	0.923	0.907
GG7	1837	562	1939	1760	506	2072	0.946	0.944	0.955
LY6	1868	189	639	1579	333	784	0.823	0.456	0.704
VG6	1642	199	1011	1562	175	1115	0.956	0.898	0.938
SA6	503	1030	579	614	925	573	0.843	0.926	0.92
PG6	316	56	242	301	54	259	0.94	0.982	0.93
AL6	1969	975	1525	1954	982	1533	0.909	0.908	0.917
PG5	247	39	78	284	37	43	0.904	0.974	0.595
GS5	208	302	574	248	272	564	0.899	0.944	0.965
LG5	2287	949	854	2254	805	1031	0.913	0.796	0.858
AG5	1228	766	746	1222	656	862	0.934	0.878	0.889
GN4	34	222	108	33	179	152	0.925	0.878	0.815
IV4	2612	484	532	2066	821	741	0.811	0.651	0.654
IL4	3586	648	611	2735	1153	957	0.817	0.499	0.681
GS4	136	643	497	193	612	471	0.76	0.969	0.915
GG4	2057	768	945	1972	568	1230	0.95	0.814	0.818
SG4	397	836	365	405	783	410	0.895	0.956	0.88
VL4	2621	619	775	1840	1264	911	0.759	0.579	0.81
AS4	378	1584	943	447	1441	1017	0.815	0.95	0.884
GA4	911	1411	1298	869	1372	1379	0.828	0.934	0.9
AG4	1418	1013	1258	1413	833	1443	0.875	0.813	0.826
SA3	522	1301	359	625	1125	432	0.806	0.893	0.781
AA3	2125	3004	1010	2652	1874	1613	0.695	0.687	0.455
GL3	851	528	435	829	339	646	0.751	0.646	0.614

for each motif generated from TMHMM information. These statistics are listed under the "TMHMM prediction"-table heading and the right of it followed by our predicted (see (6)) information. The prediction quality is determined by the respective *F*-values. The comparison evidence of the number of statistical determined motifs with the predicted ones shows how well our approach for the most motifs works. For all proteins from DUF families and for the bacteriorhodopsin-like protein families, our approach works well and can be stated for the majority motifs. Deviations can be traced back to motifs with different functions. Furthermore, our approach has been transferred to all common known structures. EDS3 as third evaluation dataset consists of all known alpha helical membrane proteins with structures obtained from PDBTM [13]. It is important to note that results from EDS3 only include PDBTM protein information. That means, each found motif has been annotated with one of three given topology states "H," "Side1," and "Side2," in which "H" stands for alpha-helix structure and both Side states refer to the outside or inside of the membrane. Here, "H" can be equated with "TM" because "H" includes only alpha-helical information referring to the interior of the cell membrane. Both Side states can be equated with "nTM." The "trans" state is not included at this point by less membrane information. This means that we have separated a motif M into three motif subsets M_{H}, M_{Side1}, and M_{Side2} according to the topology

Table 4: Statistical analyses of the motifs in all known PDBTM protein structures (EDS3). The results are split into three subtables. The "PDBTM prediction," the "Prediction on log-odds," and the "*F*-measures"-table. Thereby the "PDBTM prediction"-table represents the absolute occurrences of a motif in all investigated PDBTM protein structures. The "Prediction on log-odds"-table represents the topology state winners (see (6)) followed by the "*F*-measures"-table witch indicates how good or bad a motif can be separated and assigned to a topology state.

Motif	PDBTM prediction			Prediction on log-odds			*F*-measures		
	α-helical	Side1	Side2	α-helical	Side1	Side2	α-helical	Side1	Side2
PG10	382	1719	1780	382	1297	2202	1.0	0.86	0.894
LF10	2084	1248	1381	2092	1007	1614	0.998	0.893	0.918
PG9	473	1559	1583	474	1158	1983	0.999	0.852	0.887
LF9	2206	1103	1202	2207	962	1342	0.999	0.93	0.945
VF8	1891	1006	1120	1907	787	1323	0.996	0.878	0.905
LF8	3638	1450	1346	3637	1067	1730	0.998	0.845	0.873
GY8	393	1228	1186	392	930	1485	0.999	0.862	0.888
GA7	2614	2516	2914	2607	1775	3662	0.993	0.817	0.881
AG7	3443	2411	2937	3469	1739	3583	0.995	0.836	0.895
AA7	2870	3288	3650	2870	2280	4658	0.991	0.811	0.873
GG7	2899	2982	3285	2917	2132	4117	0.997	0.834	0.883
LY6	1326	1127	1066	1345	901	1273	0.992	0.888	0.904
VG6	2962	2230	2588	2961	1723	3096	0.996	0.869	0.907
SA6	1499	1984	1947	1497	1551	2382	0.998	0.875	0.899
PG6	347	1697	1558	348	1356	1898	0.999	0.888	0.901
AL6	6110	2672	2947	6140	1951	3638	0.996	0.844	0.889
PG5	971	1651	2095	991	1334	2392	0.985	0.893	0.923
GS5	1101	1609	1708	1154	1131	2133	0.976	0.826	0.874
LG5	5049	3013	3411	5083	2124	4266	0.993	0.826	0.879
AG5	3601	3012	3278	3623	2177	4091	0.986	0.833	0.879
GN4	427	1700	1898	453	1281	2291	0.964	0.857	0.894
IV4	4596	1717	1855	4914	1317	1937	0.947	0.838	0.822
IL4	6972	1827	2344	6956	1299	2888	0.964	0.752	0.842
GS4	1298	1773	1858	1425	1331	2173	0.936	0.854	0.879
GG4	3656	2463	2738	3897	1653	3307	0.93	0.784	0.84
SG4	1493	2141	2419	1629	1349	3075	0.948	0.771	0.86
VL4	6840	2363	3081	7067	1757	3460	0.963	0.821	0.873
AS4	2172	2066	2498	2267	1399	3070	0.934	0.807	0.859
GA4	4397	2954	3845	4685	1883	4628	0.933	0.756	0.85
AG4	3668	3402	3838	3950	2376	4582	0.937	0.807	0.856
SA3	2204	2198	2292	2936	1357	2401	0.773	0.678	0.758
AA3	5085	3463	4144	6342	1865	4485	0.798	0.646	0.733
GL3	5730	3075	3552	6026	2147	4184	0.789	0.591	0.723

state in which each single motif occurrence was located. Further calculations are described in Section 2.4 based on these motif subsets. All results from Table 4 show that our approach can be applied on known structures. The topology specific recorded statistically motif occurrence is listed in the "PDBTM prediction"-table heading and the right of it followed by our predicted information.

4. Conclusion

In this work, 33 short sequence motifs reported in [18] were investigated in 32 polytopic membrane protein families with domains of unknown functions. Transmembrane and nontransmembrane sequence regions were predicted using the TMHMM method [38] and topology states were annotated to all detected sequence motif occurrences. These amino acid propensities were derived and employed to define log-odd profiles (LOP) of all variable sequence positions in the investigated motifs. Propensity tendencies according to the topology states were identified using UPGMA and XOM clustering. Both methods pointed to good separability and predictability of the topology state of a motif 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

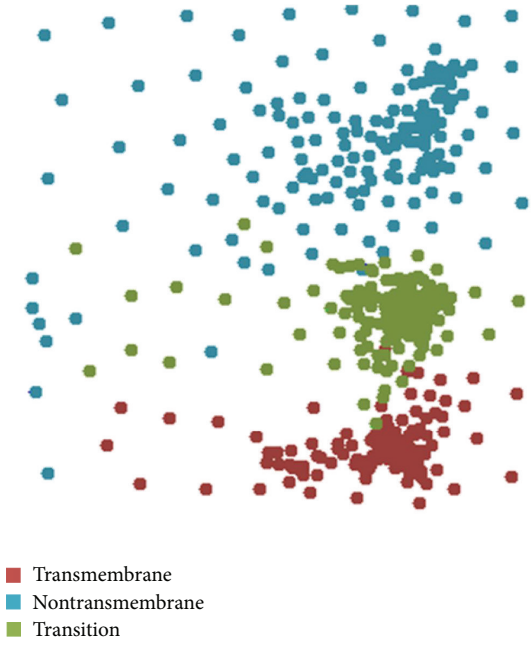

Figure 5: Output of the XOM clustering: XOM [34] is a relatively new approach for dimensionality reduction and clustering of multidimensional data. We used this approach to visualise the distance relations of the 471 investigated variable motif positions by employing the distance measure defined in (5). Here, XOM delivers a two-dimensional mapping of the distance relations of all LOPs. Coloured according to the topology state in which the corresponding motif is located, three well separable clusters can be seen. The LOP distances which contribute to the cluster formation are mainly dictated by the propensities of hydrophilic, hydrophobic, and polar residues. Thus, the XOM output reflects physicochemical correspondences which also applies for the general cluster arrangement, with the cluster of LOPs mainly observed in "trans" topology states (which corresponds basically to helix caps) located between the other two clusters. Similar to the UPGMA-tree depicted in Figure 4, the XOM output points to a good separability and predictability of topology states of short sequence motifs from their amino acid sequence in variable motif positions.

all or only in certain investigated protein families were identified. From this insight, we postulate that short sequence motifs can be divided in general, structure-forming elements, which are present in numerous protein families and highly specific to their topology location. But they are probably less important for functional properties. Finally, 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 an entire protein family.

Also, the information of the spatial structure and the folding of proteins to be explored can be evaluated by affinities, because the spatial structure of proteins has been stronger conserved in evolution than the sequential composition of the folded protein chains. These are individual motifs or characteristic sequence parts which 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 optimised [39]:

(i) the stability of the protein backbone in a given folding pattern,

(ii) the design of the amino acid sequence according to a specific function.

Based on this information, further work will discuss and deal with how the evolution has spawned motifs in their function as structure building blocks. In addition, motifs originated by evolution and spatially interacting with other should be determined as structure stabilizing.

Acknowledgment

The authors would like to thank the Free State of Saxony and the European Social Fond (ESF) for the financial support.

References

[1] M. Luckey, *Membrane Structural Biology*, Cambrigde University Press, 2008.

[2] M. H. Y. Lam and I. Stagljar, "Strategies for membrane interaction proteomics: proteomics: no mass spectrometry required," *Proteomics*, vol. 12, no. 10, pp. 1519–1526, 2012.

[3] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, "Protein function the post-genomic era," *Nature*, vol. 405, no. 6788, pp. 823–826, 2000.

[4] N. Lan, G. T. Montelione, and M. Gerstein, "Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level," *Current Opinion in Chemical Biology*, vol. 7, no. 1, pp. 44–54, 2003.

[5] A. Marsico, D. Labudde, T. Sapra, D. J. Muller, and M. Schroeder, "A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy," *Bioinformatics*, vol. 23, no. 2, pp. e231–e236, 2007.

[6] M. Childers, G. Eckel, A. Himmel, and J. Caldwell, "A new model of cystic fibrosis pathology: lack of transport of glutathione and its thiocyanate conjugates," *Medical Hypotheses*, vol. 68, no. 1, pp. 101–112, 2006.

[7] S. M. Rowe, S. Miller, and E. J. Sorscher, "Cystic fibrosis," *The New England Journal of Medicine*, vol. 352, no. 19, pp. 1992–2001, 2005.

[8] S. Tan, T. T. Hwee, and M. C. M. Chung, "Membrane proteins and membrane proteomics," *Proteomics*, vol. 8, no. 19, pp. 3924–3932, 2008.

[9] G. C. Brito and D. W. Andrews, "Removing bias against membrane proteins in interaction networks," *BMC Systems Biology*, vol. 5, article 169, 2011.

[10] P. G. Sadowski, A. J. Groen, P. Dupree, and K. S. Lilley, "Subcellular localization of membrane proteins," *Proteomics*, vol. 8, no. 19, pp. 3991–4011, 2008.

[11] J. U. Bowie, "Solving the membrane protein folding problem," *Nature*, vol. 438, no. 7068, pp. 581–589, 2005.

[12] G. E. Tusnady, Z. Dosztanyi, and I. Simon, "Transmembrane proteins in the protein data bank: identification and classification," *Bioinformatics*, vol. 20, no. 17, pp. 2964–2972, 2004.

[13] G. E. Tusnady, Z. Dosztanyi, and I. Simon, "Pdbtm: selection and membrane localization of transmembrane proteins in the protein data bank," *Nucleic Acids Research*, vol. 33, pp. D275–D278, 2005.

[14] A. Krogh, B. Larsson, G. Von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.

[15] B. Rost, R. Casadio, P. Fariselli, and C. Sander, "Transmembrane helices predicted at 95% accuracy," *Protein Science*, vol. 4, no. 3, pp. 521–533, 1995.

[16] D. T. Jones, "Improving the accuracy of transmembrane protein topology prediction using evolutionary information," *Bioinformatics*, vol. 23, no. 5, pp. 538–544, 2007.

[17] S. Hayat and A. Elofsson, "Boctopus: improved topology prediction of transmembrane β-barrel proteins," *Bioinformatics*, vol. 28, no. 4, pp. 516–522, 2012.

[18] Y. Liu, D. M. Engelman, and M. Gerstein, "Genomic analysis of membrane protein families: abundance and conserved motifs," *Genome Biology*, vol. 3, no. 10, research0054, 2002.

[19] I. T. Arkin and A. T. Brunger, "Statistical analysis of predicted transmembrane alpha-helices," *Biochim Biophys Acta*, vol. 1429, no. 1, pp. 113–128, 1998.

[20] A. Senes, M. Gerstein, and D. M. Engelman, "Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and association with β-branched residues at neighboring positions," *Journal of Molecular Biology*, vol. 296, no. 3, pp. 921–936, 2000.

[21] W. P. Russ and D. M. Engelman, "The GxxxG motif: a framework for transmembrane helix-helix association," *Journal of Molecular Biology*, vol. 296, no. 3, pp. 911–919, 2000.

[22] A. Senes, D. E. Engel, and W. F. DeGrado, "Folding of helical membrane proteins: the role of polar, gxxxg-like and proline motifs," *Current Opinion in Structural Biology*, vol. 14, no. 4, pp. 465–479, 2004.

[23] R. A. Melnyk, S. Kim, A. R. Curran, D. M. Engelman, J. U. Bowie, and C. M. Deber, "The Affinity of GXXXG Motifs in Transmembrane Helix-Helix Interactions Is Modulated by Long-range Communication," *Journal of Biological Chemistry*, vol. 279, no. 16, pp. 16591–16597, 2004.

[24] D. Schneider, C. Finger, A. Prodöhl, and T. Volkmer, "From interactions of single transmembrane helices to folding of α-helical membrane proteins: analyzing transmembrane helix-helix interactions in bacteria," *Current Protein and Peptide Science*, vol. 8, no. 1, pp. 45–61, 2007.

[25] D. Schneider and D. M. Engelman, "Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions," *Journal of Molecular Biology*, vol. 343, no. 4, pp. 799–804, 2004.

[26] R. Jackups and J. Liang, "Combinatorial model for sequence and spatial motif discovery in short sequence fragments: examples from beta-barrel membrane proteins," *IEEE Engineering in Medicine and Biology Society*, vol. 1, pp. 3470–3473, 2006.

[27] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The pfam protein families database," *Nucleic Acids Research*, vol. 40, pp. D290–D301, 2012.

[28] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research Acids Res*, vol. 40, pp. 290–301, 2012.

[29] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[30] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[31] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[32] E. W. Sayers et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 40, pp. 13–25, 2012.

[33] R. R. Sokal and C. D. Michener, *A Statistical Method for Evaluating Systematic Relationships*, vol. 38 of *University of Kansas Science Bulletin*, 1958.

[34] A. Wismueller, "A computational framework for nonlinear dimensionality reduction and clustering," *Lecture Notes in Computer Science*, vol. 5629, pp. 334–343, 2009.

[35] D. Langosch and J. Heringa, "Interaction of transmembrane helices by a knobsinto-holes packing characteristic of soluble coiled coils," *Proteins*, vol. 31, no. 2, pp. 150–159, 1998.

[36] M. C. Thomsen and M. Nielsen, "Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion," *Nucleic Acids Research*, vol. 40, pp. W281–W287, 2012.

[37] L. Wilkinson and M. Friendly, "The History of the Cluster Heat Map," 2008, www.cs.uic.edu/~wilkinson/Publications/heatmap.pdf.

[38] E. L. Sonnhammer, G. von Heijne, and A. Krogh, "A hidden markov model for predicting transmembrane helices in protein sequences," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 6, pp. 175–182, 1998.

[39] B. Carl and T. John, *Introduction to Protein Structure*, Taylor and Francis, Auflage, 2nd edition, 1998.