BIOINFORMATICS

Vol. 22 no. 14 2006, pages e290–e297
doi:10.1093/bioinformatics/btl209

A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops

Gorka Lasso[1], John F. Antoniw[2] and Jonathan G.L. Mullins[1,*]

[1]Membrane Proteins Structural Bioinformatics Group, School of Medicine, Swansea University, Singleton Park, Swansea SA2 8PP, Wales, UK and [2]Wheat Pathogenesis Programme, Plant Pathogen Interactions Division, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, England, UK

ABSTRACT

Motivation: Membrane dipping loops are sections of membrane proteins that reside in the membrane but do not traverse from one side to the other, rather they enter and leave the same side of the membrane. We applied a combinatorial pattern discovery approach to sets of sequences containing at least one characterised structure described as possessing a membrane dipping loop. Discovered patterns were found to be composed of residues whose biochemical role is known to be essential for function of the protein, thus validating our approach.

TMLOOP (http://membraneproteins.swan.ac.uk/TMLOOP) was implemented to predict membrane dipping loops in polytopic membrane proteins. TMLOOP applies discovered patterns as weighted predictive rules in a collective motif method (a variation of the single motif method), to avoid inherent limitations of single motif methods in detecting distantly related proteins. The collective motif method applies several, partially overlapping patterns, which pertain to the same sequence region, allowing proteins containing small variations to be detected. The approach achieved 92.4% accuracy in sensitivity and 100% reliability in specificity. TMLOOP was applied to the Swiss-Prot database, identifying 1392 confirmed membrane dipping loops, 75 plausible membrane dipping loops hitherto uncharacterised by topology prediction methods or experimental approaches and 128 false positives (8.0%).

Contact: j.g.l.mullins@swansea.ac.uk

1 INTRODUCTION

Membrane dipping loops

Polytopic membrane proteins are embedded membrane proteins composed of a bundle of α-helices that completely span the membrane. These transmembrane α-helices are generally connected by extramembrane loops of various lengths. However, crystallized structures of membrane proteins such as aquaporins or potassium channels have shown that membrane dipping loops (sometimes called re-entrant loops) can also interconnect α-helical transmembrane regions at the same side of the membrane. These loops are characterised by their particular structure: the N-terminal section of the loop partially transverses the lipid bilayer but with the C-terminal section then returning to the same side as the N-terminal section of the loop. It has been suggested that membrane dipping loops play major roles as selectivity filters in the aquaglyceroporin family (Gonen, *et al*., 2004; Harries, *et al*., 2004; Murata, *et al*., 2000; Ren, *et al*., 2001; Savage, *et al*., 2003; Stroud, *et al*., 2003; Sui, *et al*., 2001), potassium channels (Doyle, *et al*., 1998; Jiang, *et al*., 2002; Jiang, *et al*., 2003; Kuo, *et al*., 2003; Long, *et al*., 2005; Nishida and MacKinnon, 2002; Zhou, *et al*., 2001), chloride channels (Dutzler, *et al*., 2002; Dutzler, *et al*., 2003) and also act as gates of membrane pores, such as in the glutamate homolog transporter (Yernool, *et al*., 2004), and the protein conducting channel (Van den Berg, *et al*., 2004). Prediction of membrane dipping loops from protein sequence has proved difficult as such regions are frequently amphiphilic, containing hydrophobic sections that are too intermittent to be identified as membrane regions. Membrane dipping loops require interactions with adjacent highly hydrophobic helices to become inserted in the membrane and minimise the energy penalty imposed by location of polar or charge residues in a low dielectric environment. In-silico topology prediction approaches often fail to predict membrane dipping loops in polytopic α-helical membrane proteins due to their residue composition differing with that of membrane spanning segments. To date, the bioinformatics approaches of our group, working on the dipping loops of glycerol channels, in collaboration with Stefan Hohmann and colleagues, have relied upon homology modelling (Bill *et al*., 2001), and comparison of test sequences with those of known loops in terms of secondary structure and the propensity scoring of successive residues to reside in α or β conformation (Hedfalk *et al*., 2004; Karlgren *et al*., 2004; Tamas *et al*., 2003), underpinned by extensive laboratory work including measuring channel efflux, mutagenesis and genetic screening. Here we describe the development of a novel and reliable approach to the difficult problem of predicting dipping loops directly from sequence that may be generically applied to membrane proteins.

Pattern discovery

By evolution, conserved nucleotides and residues are often indicative of a common structural or functional role either at the gene or protein level. Sequence similarity detection methods have been successfully applied in fields such as gene discovery, splicing prediction, phylogenesis, protein structure and function prediction or gene expression analysis. Multiple sequence alignment techniques have become the routine approach to measuring sequence similarity and identifying important residues (Altschul, *et al*., 1990; Pearson

*To whom correspondence should be addressed.

and Lipman, 1988). These alignments can be used to develop different motif representation techniques such as single (Falquet, *et al*., 2002) or multiple motif methods (Attwood, *et al*., 1999; Henikoff, *et al*., 1999; Wu and Brutlag, 1995), profiles (Bucher, *et al*., 1996) and hidden markov models (Baldi, *et al*., 1994; Eddy, 1996; Krogh, *et al*., 1994). However, multiple sequence alignment methods have proved to be computationally very expensive (Wang and Jiang, 1994), and the accuracy of the alignment diminishes when distantly related sequences need to be aligned. An alternative approach was based on pattern discovery methods using an unaligned set of sequences. The problem of detecting all possible patterns in a set of sequences has also proven to be computationally expensive but heuristics and restrictions in the architecture of patterns (e.g. maximum length, number of non-wild elements) (Jonassen, *et al*., 1995; Rigoutsos and Floratos, 1998; Sagot, *et al*., 1995) have made it possible to analyse large set of biological sequences and discover structurally and functionally important patterns (Darzentas, *et al*., 2005).

We have applied a pattern discovery software, TEIRESIAS (Rigoutsos and Floratos, 1998), to various sets of protein subfamilies or families, depending on the residue conservation in the dipping loop region, where at least one of their members has been crystallized and its structure described in the PDB_TM database (Tusnady, *et al*., 2004) and/or in the literature as having at least one membrane dipping loop. The pattern discovery process is carried out using three different types of analysis: i) exact pattern discovery, ii) pattern discovery using a chemical equivalency set and iii) pattern discovery using a structural equivalency set. Our program, TMLOOP, uses the discovered patterns as weighted predictive rules to predict potential membrane dipping loops in polytopic membrane proteins. This software was used to explore the performance of a single motif method compared to a variation of this approach, called the collective motif method approach. Single motif methods require exact pattern matching to find structural or functional relatedness and therefore can miss distant relatives which contain small variations of the pattern (Scordis, *et al*., 1999). The collective method is based on the use of different patterns, partially overlapping, which belong to the same motif and therefore distant relative proteins containing small variations of the most common patterns can be co-detected.

2 METHODS

Data collection

Crystallized membrane proteins containing membrane dipping loops in their structure were identified in the PDB_TM database (Tusnady, *et al*., 2004; update 24/10/05). The predicted membrane dipping loops in each of the membrane proteins listed in the PDB_TM database were cross-referenced to the literature corresponding to the crystallized structures. Although these papers accurately describe the three-dimensional structure of membrane proteins, the boundaries of the lipid bilayer can only be approximated as membrane proteins need to be extracted from the membrane to elucidate their structure. Therefore, most of the loops predicted in the PDB_TM database as membrane dipping loops were found to be described but some loops were not identified in the literature and were considered as potential loops. Some structures contain additional membrane dipping loops that were not listed in the PDB_TM database and so these loops were also considered. In addition, a manual identification of membrane dipping loops in PDB structures of membrane proteins of known 3D structure (the Stephen White laboratory at the University of California, Irvine, http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) was carried out to guarantee that all PDB structures containing a membrane dipping loop had been included. All identified membrane dipping loops were ultimately manually confirmed by being viewed in RasMol (Sayle and Bissel, 1992). In the PDB_TM database 50 structures containing membrane dipping loops and 69 membrane dipping loops were identified. The literature described 5 additional membrane dipping loops in 3 determined structures. No additional structures with membrane dipping loops were manually identified. Of the 50 PDB structures considered, 46 structures were used in this study as membrane proteins containing membrane dipping loops. Members of protein families covered by the crystallized structures containing dipping loops were obtained from the Swiss-Prot database (Boeckmann *et al*., 2003), regardless of their taxonomic group, using the Uniprot/Swiss-Prot family/domain classification. At this stage, the functional and structural annotation of proteins obtained from the Swiss-Prot database was analysed and entries with inappropriate or insufficient functional annotation were discarded from each set. In order to avoid redundancy, protein families were filtered based on the sequence identity of the members composing the set (Hobohm, *et al*., 1992). A bioinformatics tool, Non-Red (Liakopoulos and colleagues, Department of Cell Biology and Biophysics at the University of Athens, http://athina.biol.uoa.gr/bioinformatics/NON-RED/), was used to avoid redundant protein sequences in each set, by removing one of a pair of sequences with homology higher than a user-defined level. Here, Non-Red was used with a setting of the minimum alignment length to 80 and the minimum identity level to 95%. Therefore pairs of sequences sharing a sequence identity of 0.95 or higher were avoided by removing the protein sequence of the given pair more similar to the remaining proteins in the set. The filtered set was defined as the gold standard set for the study. Where the protein containing a dipping loop belonged to a particular subfamily, it was important to ascertain whether the structural motif was conserved only in that particular subfamily or instead was a common feature present in other subfamilies or in the entire protein family. ClustalW (Chenna, *et al*., 2003) was used to analyze the residue conservation in the sequence region pertaining to the dipping loop motifs across the entire protein family set. When no clear differences in residue conservation was observed between subfamilies it was taken that the membrane dipping loop was a structural motif conserved across the entire protein family. By contrast, when there was little or no conservation across the different protein subfamilies, loops were included in the pattern discovery process as members of the particular subfamily only, as there was no evidence that the given membrane dipping loop was conserved throughout the entire protein family.

Isolation of membrane dipping loop regions

For each crystallized protein containing one or more membrane dipping loops a set of similar proteins was assembled as described above. These sets were composed of membrane protein sequences that belonged to the same (sub)family as the crystallized membrane protein found in the PDB_TM database. However, there was no information in the corresponding Swiss-Prot file relating to the location of the membrane dipping loops. The determination of the location of these structural motifs in non-crystallized protein sequences was achieved by aligning the non-crystallized protein sequences, using ClustalW, against the sequence corresponding to the relevant crystallized membrane protein, also known as the reference sequence. The structural motif was then mapped onto the reference sequence and the equivalent motif located in the remaining sequences in the alignment. The beginning and end of each membrane dipping loop were obtained from the PDB_TM database and checked manually. In order to minimise potential errors in identifying the ends of each membrane dipping loop, or possibly missing the appropriate section, 5 residues before the predicted starting position and 5 residues after the predicted ending position were considered. Within each set, all sequences were then reduced to the region corresponding to the particular membrane dipping loop detected in the crystallized membrane protein. At this stage, for each membrane dipping loop detected, a set of partial sequences was assembled.

Pattern discovery using TEIRESIAS

The TEIRESIAS algorithm (Rigoutsos and Floratos, 1998) may be used to discover patterns in an unaligned set of nucleotide or amino acid sequences. This software performs unsupervised pattern discovery and reports maximal patterns without enumerating the entire solution. The algorithm restricts the pattern discovery process by limiting the search to patterns with user-defined parameters: the minimum number of literals in any pattern, the maximum extent of an elementary pattern and the minimum support required for a pattern (L, W and K respectively). For the purposes of these analyses L was set to 3 as it has been shown to be the minimum value for which the convolution stage successfully operates during the pattern discovery process (Rigoutsos and Floratos, 1998), W was set to the length of the structural motif to be analyzed in each set (normally between 20-30) in order to detect conserved pairs of residues located in different halves of the structural motif but that may be closely associated in 3D in the membrane, and K was set to the 70% of the sequences contained in each set. The pattern discovery process was carried out using three different types of discovery: i) exact (identical) pattern discovery, ii) pattern discovery using a chemical equivalency set and iii) pattern discovery using a structural equivalency set. Each set was analysed individually using TEIRESIAS, and the dipping loops considered in each set were classified into three different structural categories: helix-in-turn-loop-out, loop-in-turn-helix-out and helix-in-turn-helix-out. These sets were also clustered if sharing structural similarities or assembled from the same protein family and analysed together using TEIRESIAS to find common patterns in structurally related membrane dipping loop motifs and common patterns in membrane dipping loops possibly caused by ancestral gene duplication events.

Pattern validation

The patterns detected by TEIRESIAS were not in themselves guaranteed to be selective as it is not possible to include negative control sets in the pattern discovery process. Therefore, it may be possible to discover patterns from one particular set in other sets of membrane proteins, whose structure does not actually contain a dipping loop, leading to predictive rules with poor specificity. To validate the patterns, an additional tool was implemented, named PATTERNTEST, whose function was to validate the patterns obtained using TEIRESIAS against positive and negative control sets assembled by the user. The patterns discovered for each set were validated against protein sequences belonging to the remaining sets of membrane dipping loop motifs and against the negative control set composed of 363 membrane proteins known not to have membrane dipping loops in their structure. This set was assembled using sequences pertaining to crystallized membrane proteins whose structure was visually checked during the data collection process, and protein families contained in the Swiss-Prot database known not to have membrane dipping loops in their structures (e.g. GPCR family). The subsequent patterns discovered by TEIRESIAS, but found to be present in membrane proteins with a different dipping loop motif and/or membrane proteins without a dipping loop motif and/or in proteins with the corresponding dipping loop motif, but having the pattern outside this motif, were eliminated as candidate predictive rules for TMLOOP.

TMLOOP

A predictive tool was implemented, named TMLOOP, to predict membrane dipping loops in polytopic membrane proteins. TMLOOP uses patterns discovered by TEIRESIAS and validated by PATTERNTEST as weighted predictive rules where the weight was calculated by dividing the number of sequences in the training set containing a particular pattern by the total number of sequences in the training set. The software requires a set of user-defined parameters to run the prediction: i) I is the minimum inter-loop length required between two contiguous loops, where two different patterns would predict the same loop only if the distance of both matches in the sequence is lower than I; ii) S, the minimum pattern support, which restricts the patterns used for the prediction such that only the patterns whose support is equal or higher than S would be used as predictive rules; and iii) C, the minimum prediction confidence, which restricts the report of protein matches to those predictions with a score equal or higher than C.

TMLOOP was evaluated by tenfold cross-validation. During the evaluation process, the single motif approach, using the pattern with the highest support for each set, and the collective motif approach were compared and different values of I, S and C were tested to set up the optimum conditions to maximize the sensitivity and specificity of TMLOOP (Table 2, Figure 1).

Swiss-Prot database prediction

TMLOOP was applied to the Swiss-Prot database using the single motif method and the collective motif method (using values of I, S and C reporting the maximum predictive score during evaluation), a consensus prediction of membrane dipping loops was also undertaken (table 3). Predicted loops were classified as true positives, false positives or possible loops that may merit to be experimentally studied. In order to identify possible hitherto undesignated loops, it was required to identify structural or functional relatedness to the corresponding crystallized protein type known to have a similar membrane dipping loop. This was achieved by: i) searching for structural evidence of the loop or functional relatedness in Swiss-Prot annotation and/or in the IUBMB enzyme nomenclature database and/or in the TCDB transport classification database (Saier *et al*., 2006, http://www.tcdb.org/); ii) looking for distant relationships using BLASTP with an E-value cutoff of 100 (Darzentas, *et al*., 2005); iii) local residue conservation analysis using ClustalW; and iv) relative position of the predicted loop in sequence to the positions of the transmembrane regions in sequence.

3 RESULTS

Pattern discovery and validation of patterns

The 12 sets of partial sequences corresponding to membrane dipping loops found in potassium channels, secY/SEC61 alpha family, aquaglyceroporin family (two loops), sodium/dicarboxylate symporter family, ClC chloride channel family (four loops), psaF family and FecCD subfamily from the binding-protein-dependent family, were analysed using TEIRESIAS individually and combined as described above. Table 1 summarizes the pattern discovery analyses carried out and the subsequent validation of patterns. Only patterns whose support is $\geq$70% were collected.

TMLOOP evaluation by tenfold cross-validation

TMLOOP sensitivity and specificity was tested using different values of I, C and S. TMLOOP was also evaluated using predictive rules for the sole pattern with the highest support found for each training set. Table 2 and Figure 1 summarise the evaluation results (I was set to a default of 30, shown to be the most appropriate minimum inter-loop length, data not shown).

Prediction of membrane dipping loops in Swiss-Prot database

TMLOOP was used to predict membrane dipping loops in polytopic membrane proteins listed in the Swiss-Prot database (version 48.0). The database contained 194,317 protein entries where 29,127 were polytopic membrane proteins (15.0%). TMLOOP was run with two different sets of parameters: i) the single motif approach, using the individual pattern with the highest score for each membrane dipping loop analysed (I=30, while the parameters C and S were not relevant for this prediction; and ii) the collective method approach, using TMLOOP with the most optimal parameters of C and S obtained from the evaluation by tenfold cross-validation (I=30). The results are shown in Tables 3 and 4 (and supplementary information can be found at http://membraneproteins.swan.ac.uk/TMLOOP/Supplementary).

Table 1. Training sets and pattern discovery of membrane dipping loops

Gold standard sets		Pattern discovery and validation		Single motif method	
Membrane dipping loop set	No. of sequences	No. of patterns	No. of validated patterns	Top scoring pattern	Support
Helix-in-turn-loop-out K^+ channel	134	382, 103, 5	35, 10, 0	[ST]..[ST].G[FY]G	0.89
Helix-in-turn-loop-out secY/SEC61 alpha family	75	12, 0, 0	0, 0, 0	No patterns found	-
L1: Loop-in-turn-helix-out Aquaglycerolporin family	49	863, 73, 22	167, 21, 6	SG.H.N...[ST]	0.96
L2: Loop-in-turn-helix-out Aquaglycerolporin family	49	249, 32, 19	24, 1, 1	[ILMV]NP.R.....[ILMV]	0.94
Helix-in-turn-helix-out Binding protein dependent transport system permease family	25	7506, 479, 31	82, 43, 11	[AG].[ILMV].F[ILMV] [AG]L[IMV].P.[ILMV]	0.96
L1: Helix-in-turn-helix-out Cl^- channel family	35	936, 46, 14	29, 5, 3	[ILMV]G..GP.V	0.86
L2: Helix-in-turn-helix-out Cl^- channel family	35	2419, 97, 63	98, 35, 28	[AG].[AG].G[ILMV]... [FY].....[AG]..F..E	1.0
L3: Helix-in-turn-helix-out Cl^- channel family	35	610, 45, 10	9, 7, 2	P.G...P....G...G	0.91
L4: Helix-in-turn-helix-out Cl^- channel family	35	3751, 137, 26	66, 0, 0	[AG].....[ILMV]... [ILMV][ILMV]..E[ILMV]T	0.91
Helix-in-turn-helix-out psaF family	16	182, 41, 13	27, 16, 9	A.....G..WP..A	1.0
L1: Helix-in-turn-helix-out Na^+: dicarboxylate symporter family	46	3613, 347, 63	134, 19, 11	[ILMV].....T.S[ST]...[ILMV]P	0.89
L2: Helix-in-turn-loop-out Na^+: dicarboxylate symporter family	46	14887, 1327, 103	324, 142, 17	[ILMV].....[ILMV].....S.G..[AG][ILMV].... .[ILMV].[ILMV].....[ILMV]	0.96
L1: Loop-in-turn-helix-out L2: Loop-in-turn-helix-out Aquaglycerolporin family	98	333, 6, 5	27, 0, 0	[ST]G...NP[AG]	0.86
L1: Helix-in-turn-helix-out L3: Helix-in-turn-helix-out Cl^- channel family	70	106, 0, 0	0, 0, 0	No patterns found	-
L2 Helix-in-turn-helix-out L4: Helix-in-turn-helix-out Cl^- channel family	70	244, 0, 0	0, 0, 0	No patterns found	-
L1: Helix-in-turn-helix-out L2: Helix-in-turn-helix-out L3: Helix-in-turn-helix-out L4: Helix-in-turn-helix-out Cl^- channel family	140	0, 0, 0	0, 0, 0	No patterns found	-
L1: Helix-in-turn-helix-out L2: Helix-in-turn-loop-out Na^+: dicarboxylate symporter family	92	124, 2, 0	0, 0, 0	No patterns found	-
Helix-in-turn-loop-out loops	255	0, 1, 0	0, 0, 0	No patterns found	-
Helix-in-turn-helix-out loops	227	0, 0, 0	0, 0, 0	No patterns found	-
All dipping loops	565	0, 0, 0	0, 0, 0	No patterns found	-

Columns one and two describe the different sets of membrane dipping loops assembled. Columns three and four give the number of patterns obtained by chemical ([A,G], [E,D], [F,Y], [K,R], [I,L,M,V], [Q,N], [S,T]), structural equivalence ([C,S], [D,L,N], [E,Q], [F,H,W,Y], [I,T,V], [K,M,R]) and exact discovery respectively. No common patterns were observed between membrane dipping loops across the different structural categories, with the exception of L1 and L2 of the aquaglyceroporin family. Columns five and six describe the single motif method where column five lists those patterns found with the highest support, and column six shows the corresponding pattern support.

4 DISCUSSION

Referencing of discovered patterns

The patterns discovered by TEIRESIAS and validated with PATTERNTEST were considered in the light of the crystallographic structures and literature. These patterns were frequently found to belong to structural motifs, which were described as essential for the function of the protein. Furthermore the biochemical roles of several of the residues described in these patterns have been described in experimental studies, validating our approach. The dipping loop motifs found in potassium channels, aquaglyceroporins and loops 1 and 3 in ClC chloride channels have been described

Table 2. Evaluation of TMLOOP by tenfold cross-validation

				S 70	S 80	S 90	Top score pattern
C	**0.01**	Sensitivity	Av	95.64	92.87	43.24	87.05
			Sd	2.58	2.98	5.57	3.2
		Specificity	Av	95.93	98.18	98.88	100
			Sd	6.93	4.42	0.46	0
	0.1	Sensitivity	Av	90.57	**92.43**	43.09	87.05
			Sd	3.05	**2.48**	5.56	3.2
		Specificity	Av	95.25	**100**	100	100
			Sd	2.36	**0**	0	0
	0.3	Sensitivity	Av	85.93	87.09	41.36	87.05
			Sd	3.85	3.73	5.69	3.2
		Specificity	Av	100	100	100	100
			Sd	0	0	0	0
	0.5	Sensitivity	Av	77.35	84.76	41.07	87.05
			Sd	5.19	4.28	5.04	3.2
		Specificity	Av	100	100	100	100
			Sd	0	0	0	0
	0.7	Sensitivity	Av	63.55	76.78	38.45	87.05
			Sd	5.83	4.16	5.14	3.2
		Specificity	Av	100	100	100	100
			Sd	0	0	0	0

Two different approaches were carried out : i) using the pattern for each membrane dipping loop set with the highest score (top score approach, which is a single motif approach, in orange) using an I value of 30; and ii) using various values of S (minimum pattern support) and C (minimum prediction confidence) with a fixed I value (minimum inter-loop length) of 30 (collective motif approach, in blue). The results shown with the white background are the data relating to the optimal performance of TMLOOP.The top score approach, which proved to be a conservative approach, gave a confidence of 1.0 for each prediction since here TMLOOP uses just one rule per membrane dipping loop considered and therefore the prediction is based upon exact single pattern matching (either yes or no).

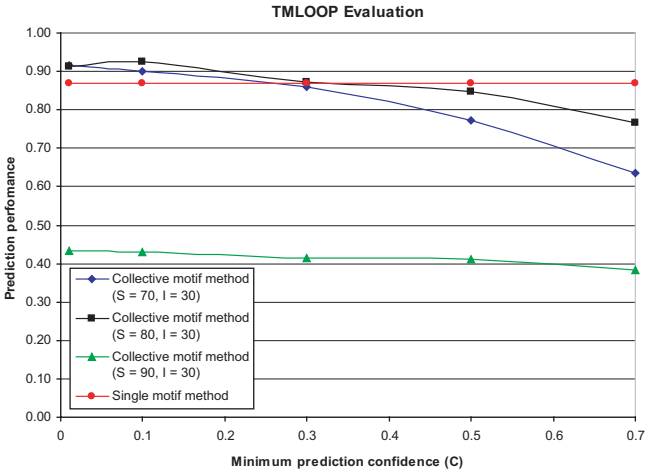

Fig. 1. Comparison of the performance of single and collective motif methods tested by tenfold cross-validation.

This graph shows the prediction performance (considering both the sensitivity and specificity) of each TMLOOP analysis (i. the single motif method in red, ii. the collective method –S (minimum pattern support) = 70, I (minimum inter-loop length) = 30- in blue, iii. the collective motif method –S = 80, I = 30- in black and iv. the collective motif method –S = 90, I = 30 in green) carried out at various levels of minimum prediction confidence (C). The collective method (S = 80, I = 30) showed the highest predictive score at a minimum confidence value C of 0.1. The C value of 0.3 is considered to be the threshold, below which the most accurate prediction method is the collective motif method and above which the single motif method performs better.

Table 3. Prediction of dipping loops in the SwissProt database

		True positives	False positives	Potential loops
Single motif method	Membrane dipping loops	1209	117	32
	Proteins	581	115	32
Collective motif method	Membrane dipping loops	1392	128	75
	Proteins	605	128	75
Consensus prediction	Membrane dipping loops	1204	78	31
	Proteins	576	78	31

The table summarises the analysis of the SwissProt database using TMLOOP (a) when only the pattern with the highest support is used (single motif approach) and (b) when all patterns whose support is $\geq$80 are used and only predictions with score $\geq$0.1 are reported (collective motif approach). The I value (minimum inter-loop length) was set to 30 for both methods. The last two rows (in red) show the consensus prediction considering both approaches.

as selectivity filters. The residues contained in patterns belonging to these selectivity filters have been extensively described and the discovered patterns were found to be refined motifs of those already proposed (e.g. the GYGD motif in potassium channels and the NPA motif in the aquaglyceroporin family). Loops 2 and 4 in ClC chloride channels have been proposed to link the two repeated halves within each monomer and make contacts with each other at the interface between monomers (Estevez and Jentsch, 2002). However, the precise functional relevance of specific residues is not clear. The most common patterns found in our analysis had support of 1.0 and 0.91 for loops 2 and 4 respectively, indicating important roles for particular residues.

In the case of the sodium / dicarboxylate symporter family both loops were proposed to act as gates in the membrane (Yernool, *et al*., 2004). The composition of patterns found in loop 1 were in agreement with the motifs identified in experimental studies, though a proline described previously as being conserved was not included in patterns relating to loop 2. This conserved proline is suggested to act as an anchor together with the serine-rich motif corresponding to loop 1. Further analyses showed that this proline was only conserved in 23 out of 46 sequences.

No patterns corresponding to discrete motifs could be found using TEIRESIAS for the detected dipping loops in the SecY/SEC61 alpha family, which transports soluble proteins across the membrane and passes membrane proteins into the membrane. The dipping loop found in this protein family is also known as the channel plug (Van den Berg, *et al*., 2004) and it has been suggested to block the pore in the closed state, in the open state the channel opens by displacement of the plug which moves away from the pore towards the plug-pocket (Collinson, 2005; Van den Berg, *et al*., 2004). Despite the overall importance of this motif no

Table 4. Newly predicted dipping loops in the SwissProt database

Swiss-Prot accession code	Definition	Predicted membrane dipping loop
Q9H2Y9	Solute carrier organic anion transporter family, member 5A1	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q8KWT2, Q8KWS7, P39642	Putative bacilysin exporter bacE	Loop-in-turn-helix-out Loop 1 & 2 aquaporin like
Q9NRA2, Q8BN82	Sialin (Solute carrier family 17 member 5)	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q58902	Hypothetical protein MJ1507	helix-in-turn-loop-out K^+ channel like
Q64SU9	Hypothetical transport protein BF2680	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q7UH36	Hypothetical transport protein RB4869	helix-in-turn-loop-out K^+ channel like
Q8AAG5	Hypothetical transport protein BT0500	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q57943	Hypothetical protein MJ0523	helix-in-turn-helix-out ClC choride channel loop-4 like loop
Q8NSS8	Hypothetical transport protein Cgl0590/cg0683	helix-in-turn-loop-out K^+ channel like
P74635	Hypothetical protein slr0753	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P0AAC6, P0AAC7	Inner membrane protein yccA	helix-in-turn-helix-out Na^+: dicarboxylate symporter loop-1 like loop
P38745	Hypothetical 61.2 kDa protein in APM2-DUR3 intergenic region precursor	helix-in-turn-loop-out K^+ channel like
P37643	Inner membrane metabolite transport protein yhjE	helix-in-turn-loop-out K^+ channel like
P54181	Hypothetical protein ypnP	helix-in-turn-loop-out K^+ channel like
Q9V7S5	Putative inorganic phosphate cotransporter	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P0A629, P0A628	Phosphate transport system permease protein pstC-1	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P10603, P27182	ATP synthase C chain	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P0A304, P0A305	ATP synthase C chain	helix-in-turn-loop-out K^+ channel like
Q8YGH4, Q8G1E6	Pyrophosphate-energized proton pump	helix-in-turn-loop-out K^+ channel like
P34299, Q8LGN0, Q9C5V5, O81078, Q9ULK0, Q61627, Q62640	Glutamate receptor precursor (glutamate-gated ion channel)	helix-in-turn-loop-out K^+ channel like
Q58671	Probable Na(+)/H(+) antiporter 3 (MjNapA)	helix-in-turn-loop-out K^+ channel like
Q15629, Q01685, Q15629, Q91V04, Q9GKZ4	Translocation associated membrane protein 1	helix-in-turn-loop-out K^+ channel like
Q8XED4, Q8FCT7, P33650, Q57IW8, Q5PLZ1, Q83ST5, P74884, Q57986, P73182	Ferrous iron transport protein B	helix-in-turn-loop-out K^+ channel like
Q97QP7, Q54875, Q59947, Q59986	Immunoglobulin A1 protease precursor	helix-in-turn-loop-out K^+ channel like
Q09917	Hypothetical protein C1F7.03 in chromosome I	helix-in-turn-loop-out K^+ channel like
Q8IZK6, Q8K595	Mucolipin-2	helix-in-turn-loop-out K^+ channel like
P91645, Q13936, Q01815, P15381, P22002, Q24270, Q01668, Q99244, P27732, O60840, Q02789, P07293, Q9JIS7, Q13698, O57483, Q02485, O73700, Q25452, P22316	Voltage-dependent calcium channel alpha-1 subunit	helix-in-turn-loop-out K^+ channel like
O28069 (top score pattern approach)	Hypothetical protein AF2214	helix-in-turn-helix-out ClC choride channel loop-4 like loop

A list of proteins, including the corresponding Swiss-Prot accession codes, containing plausible membrane dipping loops according to TMLOOP. Proteins listed were predicted by using either the single motif approach (I = 30) and/or the collective motif approach (S = 80, C = 0.1 and I = 30).

evidence of residue conservation was found in the multiple sequence alignment or in dipping loops in the pattern discovery process.

No experimental evidence has surfaced to describe the functional role of dipping loops belonging to the FeCD subfamily in the binding-protein-dependent permease family and in the PsaF family. However, the dipping loop region in the FeCD subfamily has been suggested to be important for binding the periplasmic binding protein BtuF (Locher, *et al*., 2002). The highest support found for a pattern corresponding to the membrane dipping loop in the FeCD subfamily (0.96) showed the importance of this motif for the function of the protein and supported the suggestion made by Locher *et al*. On the other hand, members of the PsaF family form part of the photosystem I (PSI). This family has been suggested

to mediate plastocyanin docking and fast electron transport kinetics in the eukaryotic PSI (Haldrup, *et al*., 2000; Hippler, *et al*., 1999). By contrast, in cyanobacteria PsaF proteins have been suggested to contribute to structural features on the surface of PSI and bind carotenoids which serve as a light harvesting and photo-protecting molecule (Jordan, *et al*., 2001). The highest support of patterns found in the dipping loop region (1.0) of PsaF proteins belonging to both cyanobacteria and eukaryote cells showed that this region was universally conserved across the taxa indicating potential residues with an essential and common functional role in both cyanobacteria and eukaryote cells.

TMLOOP evaluation

The main problem of single motif methods, is that prediction of a structural motif or functional category depends upon exact matching with a single pattern. Therefore distantly related proteins containing small variations of the pattern can not be detected. With TMLOOP, a single motif method (using the single pattern with the highest support found for each membrane dipping loop) can be employed to predict a particular membrane dipping loop, or instead, a set of partially overlapping patterns, may be used as weighted predictive rules (a collective motif method). The single motif approach and the collective motif approach using various combinations of C and S (I parameter was set to 30 in both approaches) were evaluated by tenfold cross-validation. The sensitivity and specificity of each method was calculated (table 2) and a single prediction performance score (a product of % sensitivity and % specificity divided by 10,000) was plotted against increasing minimum prediction confidence (C) values (Fig. 1).

Both methods performed well during the evaluation, however the single motif method approach was shown to be more accurate as C parameter increased. This is reflected in figure 1 where the C value of 0.3 is observed to be the threshold at which the accuracy of one method prevails over the other. When C values lower than 0.3 are considered the collective method is found to be the most accurate predictive method whereas when C values higher than 0.3 are considered the single motif method is the most accurate. The reason why the prediction accuracy of TMLOOP dropped significantly when S was set to 90 in the collective motif approach (table 2 and figure 1) was simply because some of the sets of patterns did not have a single pattern whose support was 0.90 or higher and therefore no patterns were considered for the prediction of the given membrane dipping loop. The evaluation showed that the collective approach (S = 80, C = 0.1, I = 30) was the most accurate method where TMLOOP achieved maximum values of sensitivity and specificity of 92.4% and 100% respectively (predictive score = 0.92, table 2 and figure 1). Although the single motif method was found to be a better approach with higher values of C, it also proved to be a conservative prediction. The flexibility of the collective motif approach allowed TMLOOP to detect 91.4% of the dipping loops contained in the two pore domain potassium channel family (in contrast to the 40.3% obtained by the single motif approach, data not shown), where each member of the family has been proposed to have two membrane dipping loops and the second loop showed small variations in sequence compared to the first dipping loop (successfully predicted by the single motif approach).These results reflect the strength of the collective motif method in being able to predict motifs similar but not identical to those used in the gold standard set.

This approach, where the dipping loop is specifically targeted, has distinct advantages over the baseline approach of identifying proteins that possess membrane dipping loops by ''association'' through global sequence similarity searching, where large portions of sequences may be common, but not the loop region, and *vice versa*. A thorough comparison of the targeted pattern approach with similarity search approaches is underway. This new approach has a further advantage in that it also predicts the specific residues composing the dipping loop, and the loop type. It is envisaged that the full value of TMLOOP will be realised through its use in conjunction with transmembrane region topology prediction programs.

Prediction of membrane dipping loops in the Swiss-Prot database

TMLOOP was applied to the Swiss-Prot database to predict membrane dipping loops in polytopic membrane proteins. Prediction was carried out by the two different approaches mentioned above: the single motif method approach using only the pattern with the highest support for each membrane dipping loop analyzed (I = 30); and the collective motif approach using TMLOOP with S, C and I set to 80, 0.1 and 30 respectively which maximized the predictive score during the tenfold cross-validation. The single motif method was shown to be a more conservative method whereas the collective motif method detected more potential membrane dipping loops not tested yet by experimental approaches (table 3). A good example of these highly plausible membrane dipping loops was found in the voltage-dependent calcium channel α-1 subunits (table 4) where a potassium-like membrane dipping loop (helix-in-turn-loop-out) was predicted (prediction score was 0.138). The low prediction score may be indicative of a distantly related structural motif that while not necessarily acting as a selectivity filter for potassium ions, may work for calcium ions in a similar fashion.

In conclusion, we have undertaken a full characterisation of all membrane dipping loops known to date. We have detected conserved patterns, with both high sensitivity and specificity, for most of these membrane dipping loop types. The corresponding literature highlighted some of the residues contained in these patterns as essential for the function of the protein, thus supporting our pattern discovery approach. We have implemented a tool to predict membrane dipping loops using a variation of the single motif method approach, named the collective motif approach, which was shown to be capable of detecting distantly related membrane dipping loops. Evaluation of TMLOOP by tenfold cross-validation showed impressive levels of both sensitivity and specificity. TMLOOP was successfully applied to the Swiss-Prot database predicting 75 plausible membrane dipping loops not detected previously by other methods. The program is available for use at http://membraneproteins.swan.ac.uk/TMLOOP (supplementary information can be found at http://membraneproteins.swan.ac.uk/TMLOOP/Supplementary).

ACKNOWLEDGEMENTS

We thank Vasilis Promponas for his valuable suggestions for identification of membrane dipping loops in crystallized membrane proteins. The work of G.L. was supported by a ''Beca de formación de investigadores'' grant from the Basque Government. Rothamsted

Research receives grant aided support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the UK.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, *J Mol Biol*, **215**, 403–410.

Attwood, T.K., Flower, D.R., Lewis, A.P., Mabey, J.E., Morgan, S.R., Scordis, P., Selley, J.N. and Wright, W. (1999) PRINTS prepares for the new millennium, *Nucleic Acids Res*, **27**, 220–225.

Baldi, P., Chauvin, Y., Hunkapiller, T. and McClure, M.A. (1994) Hidden Markov models of biological primary sequence information, *Proc Natl Acad Sci U S A*, **91**, 1059–1063.

Bill R.M., Hedfalk R., Karlgren S., Mullins J.G.L., Rydstrom J., Hohmann S. (2001) Analysis of the pore of the unusual MIP channel, yeast Fps1p. J. Biol. Chem. **276**, (39), 36543–36549.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res*. **31**:365–370.

Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles, *Comput Chem*, **20**, 3–23.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res*, **31**, 3497–3500.

Collinson, I. (2005) The structure of the bacterial protein translocation complex SecYEG, *Biochem Soc Trans*, **33**, 1225–1230.

Darzentas, N., Rigoutsos, I. and Ouzounis, C.A. (2005) Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families, *Proteins*, **61**, 926–937.

Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T. and MacKinnon, R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity, *Science*, **280**, 69–77.

Dutzler, R., Campbell, E.B., Cadene, M., Chait, B.T. and MacKinnon, R. (2002) X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity, *Nature*, **415**, 287–294.

Dutzler, R., Campbell, E.B. and MacKinnon, R. (2003) Gating the selectivity filter in ClC chloride channels, *Science*, **300**, 108–112.

Eddy, S.R. (1996) Hidden Markov models, *Curr Opin Struct Biol*, **6**, 361–365.

Estevez, R. and Jentsch, T.J. (2002) CLC chloride channels: correlating structure with function, *Curr Opin Struct Biol*, **12**, 531–539.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K. and Bairoch, A. (2002) The PROSITE database, its status in 2002, *Nucleic Acids Res*, **30**, 235–238.

Gonen, T., Sliz, P., Kistler, J., Cheng, Y. and Walz, T. (2004) Aquaporin-0 membrane junctions reveal the structure of a closed water pore, *Nature*, **429**, 193–197.

Haldrup, A., Simpson, D.J. and Scheller, H.V. (2000) Down-regulation of the PSI-F subunit of photosystem I (PSI) in Arabidopsis thaliana. The PSI-F subunit is essential for photoautotrophic growth and contributes to antenna function, *J Biol Chem*, **275**, 31211–31218.

Harries, W.E., Akhavan, D., Miercke, L.J., Khademi, S. and Stroud, R.M. (2004) The channel architecture of aquaporin 0 at a 2.2-A resolution, *Proc Natl Acad Sci U S A*, **101**, 14045–14050.

Hedfalk K, Bill R.M., Mullins J.G. L., Karlgren S, Filipsson C, Bergstrom J, Tamas M.J., Rydstrom J., Hohmann S. (2004) A regulatory domain in the C-terminal extension of the yeast glycerol channel Fps1p. J Biol Chem.**279** (15):14954–60.

Henikoff, J.G., Henikoff, S. and Pietrokovski, S. (1999) New features of the Blocks Database servers, *Nucleic Acids Res*, **27**, 226–228.

Hippler, M., Drepper, F., Rochaix, J.D. and Muhlenhoff, U. (1999) Insertion of the N-terminal part of PsaF from Chlamydomonas reinhardtii into photosystem I from Synechococcus elongatus enables efficient binding of algal plastocyanin and cytochrome c6, *J Biol Chem*, **274**, 4180–4188.

Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) Selection of representative protein data sets, *Protein Sci*, **1**, 409–417.

Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B.T. and MacKinnon, R. (2002) Crystal structure and mechanism of a calcium-gated potassium channel, *Nature*, **417**, 515–522.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B.T. and MacKinnon, R. (2003) X-ray structure of a voltage-dependent K+ channel, *Nature*, **423**, 33–41.

Jonassen, I., Collins, J.F. and Higgins, D.G. (1995) Finding flexible patterns in unaligned protein sequences, *Protein Sci*, **4**, 1587–1595.

Jordan, P., Fromme, P., Witt, H.T., Klukas, O., Saenger, W. and Krauss, N. (2001) Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution, *Nature*, **411**, 909–917.

Karlgren S., Philipson C., Mullins J.G.L., Bill R.M., Tamas M.J., Hohmann S. (2004) Identification of residues controlling transport through the yeast aquaglyceroporin Fps1 using a genetic screen. Eur J. Biochem **271**, 771–779.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling, *J Mol Biol*, **235**, 1501–1531.

Kuo, A., Gulbis, J.M., Antcliff, J.F., Rahman, T., Lowe, E.D., Zimmer, J., Cuthbertson, J., Ashcroft, F.M., Ezaki, T. and Doyle, D.A. (2003) Crystal structure of the potassium channel KirBac1.1 in the closed state, *Science*, **300**, 1922–1926.

Locher, K.P., Lee, A.T. and Rees, D.C. (2002) The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism, *Science*, **296**, 1091–1098.

Long, S.B., Campbell, E.B. and Mackinnon, R. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K+ channel, *Science*, **309**, 897–903.

Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J.B., Engel, A. and Fujiyoshi, Y. (2000) Structural determinants of water permeation through aquaporin-1, *Nature*, **407**, 599–605.

Nishida, M. and MacKinnon, R. (2002) Structural basis of inward rectification: cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8 A resolution, *Cell*, **111**, 957–965.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A*, **85**, 2444–2448.

Ren, G., Reddy, V.S., Cheng, A., Melnyk, P. and Mitra, A.K. (2001) Visualization of a water-selective pore by electron crystallography in vitreous ice, *Proc Natl Acad Sci U S A*, **98**, 1398–1403.

Rigoutsos, I. and Floratos, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm, *Bioinformatics*, **14**, 55–67.

Sagot, M.F., Viari, A. and Soldano, H. (1995) A distance-based block searching algorithm, *Proc Int Conf Intell Syst Mol Biol*, **3**, 322–331.

Saier, M.H. Jr, Tran, C.V., Barabote, R.D. (2006) TCDB: The transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res*.34:D181–6.

Savage, D.F., Egea, P.F., Robles-Colmenares, Y., O'Connell, J.D., 3rd and Stroud, R.M. (2003) Architecture and selectivity in aquaporins: 2.5 a X-ray structure of aquaporin Z, *PLoS Biol*, **1**, E72.

Sayle, R. and Bissel, A. (1992) RasMol: A Program for Fast Realistic Rendering of Molecular Structures with Shadows., *Proccedings of the 10th Eurographics UK 1992*. University of Edinburgh, UK.

Scordis, P., Flower, D.R. and Attwood, T.K. (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database, *Bioinformatics*, **15**, 799–806.

Stroud, R.M., Miercke, L.J., O'Connell, J., Khademi, S., Lee, J.K., Remis, J., Harries, W., Robles, Y. and Akhavan, D. (2003) Glycerol facilitator GlpF and the associated aquaporin family of channels, *Curr Opin Struct Biol*, **13**, 424–431.

Sui, H., Han, B.G., Lee, J.K., Walian, P. and Jap, B.K. (2001) Structural basis of water-specific transport through the AQP1 water channel, *Nature*, **414**, 872–878.

Tamas M.J., Karlgren S., Bill R.M., Hedfalk K., Allegri L., Ferreira M., Thevelein J.M., Rydstrom J., Mullins J.G.L., Hohmann S. (2003) A short regulatory domain restricts glycerol transport through yeast Fps1p. J Biol Chem. **278** (8), 6337–45.

Tusnady, G.E., Dosztanyi, Z. and Simon, I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification, *Bioinformatics*, **20**, 2964–2972.

Van den Berg, B., Clemons, W.M., Jr., Collinson, I., Modis, Y., Hartmann, E., Harrison, S.C. and Rapoport, T.A. (2004) X-ray structure of a protein-conducting channel, *Nature*, **427**, 36–44.

Wang, L. and Jiang, T. (1994) On the complexity of multiple sequence alignment, *J Comput Biol*, **1**, 337–348.

Wu, T.D. and Brutlag, D.L. (1995) Identification of protein motifs using conserved amino acid properties and partitioning techniques, *Proc Int Conf Intell Syst Mol Biol*, **3**, 402–410.

Yernool, D., Boudker, O., Jin, Y. and Gouaux, E. (2004) Structure of a glutamate transporter homologue from Pyrococcus horikoshii, *Nature*, **431**, 811–818.

Zhou, Y., Morais-Cabral, J.H., Kaufman, A. and MacKinnon, R. (2001) Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 A resolution, *Nature*, **414**, 43–48.