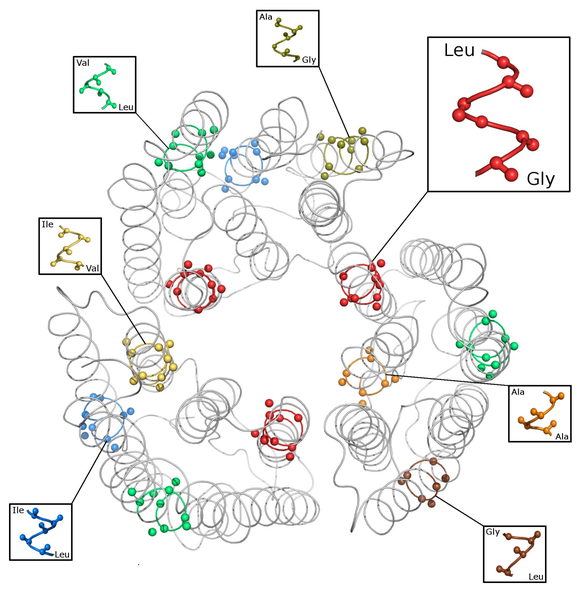
Topology separation of discriminative sequence motifs located in membrane proteins with domains of unknown functions

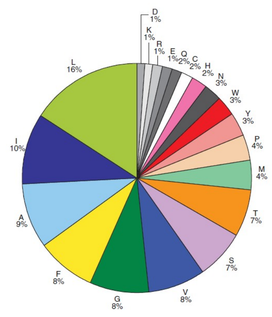
Steffen Grunert, Florian Heinke, Dirk Labudde
Hochschule Mittweida, University of Applied Sciences, Germany, Technikumplatz 17, 09648 Mittweida

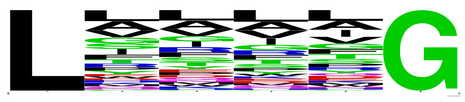
Background and Motivation

Protein sequence motifs support the understanding of the features that are important for the folded protein in the membrane environment. The protein sequence motif analysis is helpful for target mutagens studies, for investigations of diseases and further studies.



e.g. LG5 = LXXXXG





A motif XYn corresponds to amino acid X and Y with X,Y ϵ 20 canonical amino acids separated by (n-1) variable residues.

Methods

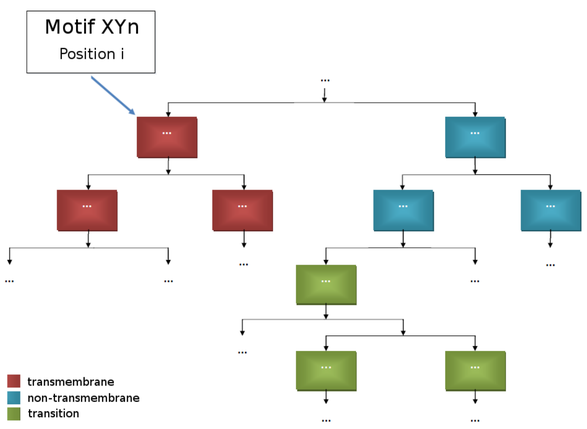
Pfam

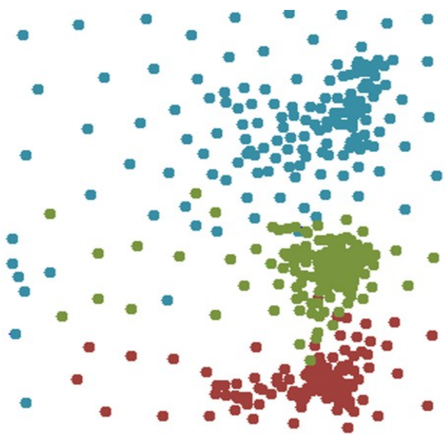
32 membrane protein families with domains of unknown functions

50 protein sequence motifs

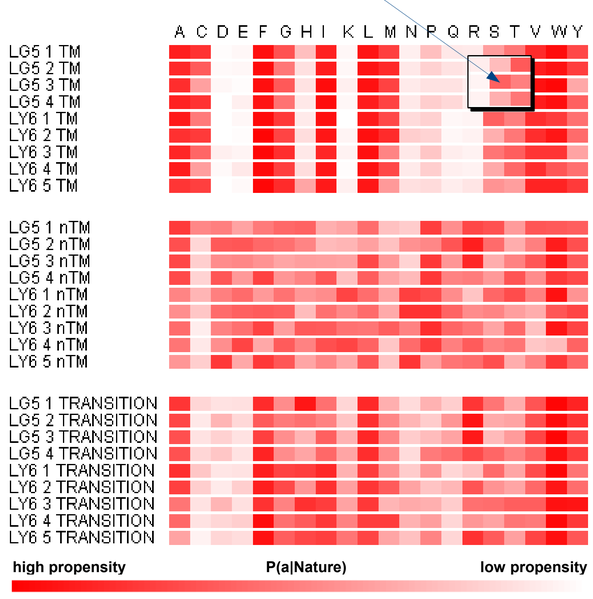
$$P(a|\text{pos}_i|M_S) = \frac{\sum_{j=1}^{k} g\left(\text{pos}_{i,m_j}, a\right)}{k}$$

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





$$f(a|\text{pos}_i|M_S) = \log\left(\frac{P(a|\text{pos}_i|M_S)}{P(a|\text{Nature})}\right)$$



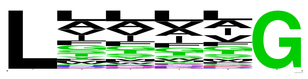
This heatmap visualization allows a sensitive approach to analyse amino acid propensities of each variable position of a motif according to topology states. Here, the heatmap is separated by topology states, so that amino acid propensities become obvious. For example, cystein can be observed more frequently at the fourt variable position of transmembrane-located LG5 motifs.

Results and Application

- Separation of motifs by a residue specific distribution.
- Position specific separation and clustering
- Clustering into transmembrane, non-transmembrane and transition regrions
- Motif positions with great similarity as a direct leaf neighbors
- Propensities of each amino acid relative to natural amino acid-specific frequencies
- F-measure up to over 0.9 within all derived protein families
- Prediction of location inside of the protein's topology for the majority of identified motifs.
- Functional relevance and diverse specificity of motifs with rather small F-measure.

Motif	TM	nTM	trans
LF10	0,9900	0,9972	0,9984
LF9	0,9981	0,9966	0,9986
VF8	0,9980	0,9980	0,9957
LF8	0,9979	0,9978	0,9967
GA7	0,9972	0,9969	0,9940
PG10	0,9963	0,9996	0,9977
VL4	0,7379	0,7276	0,5842
AS4	0,7379	0,8331	0,6810
GA4	0,7305	0,7865	0,6306
AG4	0,7283	0,8013	0,6466
SA3	0,4615	0,4433	0,3831
AA3	0,4258	0,4049	0,3456
GL3	0,4086	0,3783	0,3854

A



L-[LAY]-[LAY]-[LVA]-[LAI]-G

B

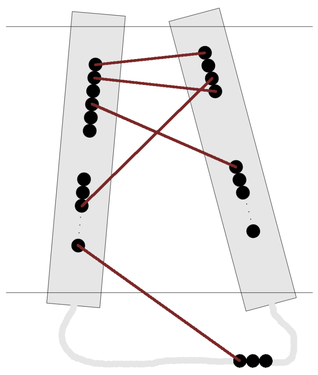
L-[ALS]-[RAE]-[LAR]-[LAP]-G

C

L-[LAR]-[LRA]-[LRA]-[LAR]-G

WebLogos of the LG5 motif in the order of three topology states 'TM' (WebLogo A), 'nTM' (WebLogo B) and 'trans' (WebLogo 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.

Further Work



- Investigation of motifs based on evolution conserved residue interactions
- Detection of structure forming and protein stabilizing motifs
- Motifs in relationship to other motifs (co-*-co-occurrence).

References

1. Pearson WR, Lipman DJ:Improved tools for biological sequence comparison., Proc Natl Acad Sci USA 1988, 85:2444-48.
2. Krogh A, Larsson B, von Heijne G, Sonnhammer EL:Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes., J Mol Biol 2001, 305: 567-580.
3. Yang Liu, Donald M Engelman and Mark Gerstein,Genomic analysis of membrane protein families: abundance and conserved motifs, , Sept. 2002
4. Senes A, Gerstein M, and Engelman,Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions.J Mol Biol 2000, 296:921-936.
5. Wismueller Axel,A Computational Framework for Nonlinear Dimensionality Reduction and Clustering, Lecture Notes in Computer Science, 2009, Volume 5629/2009, 334-343, DOI: 10.1007/978-3-642-02397-238
6. G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. Genome Res., 14(6):1188–1190, Jun 2004.

Europa fördert Sachsen.
ESF
Europäischer Sozialfonds

contact: sgrunert@hs-mittweida.de