Position separation of discriminative sequence motifs with protein domains of unknown functions

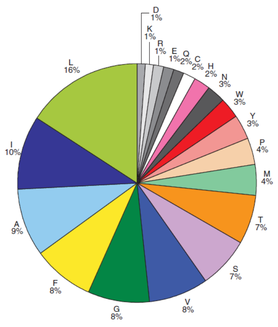
Steffen Grunert, Florian Heinke, Michael Spranger and Dirk Labudde
University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida

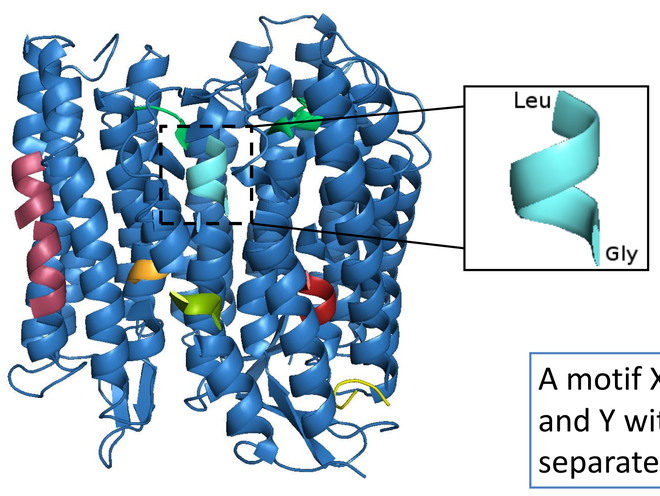
HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES
bigM BioInformatics Group Mittweida
Europa fördert Sachsen. ESF Europäischer Sozialfonds

Background and Motivation

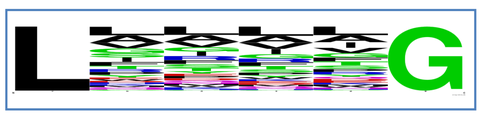
Protein sequence motifs support the understanding of the features that are important for the folded protein in the membrane environment. The protein sequence motif analysis is helpful for target mutagens studies, for investigations of diseases and further studies.

The amino acid composition in the TM-helical regions





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

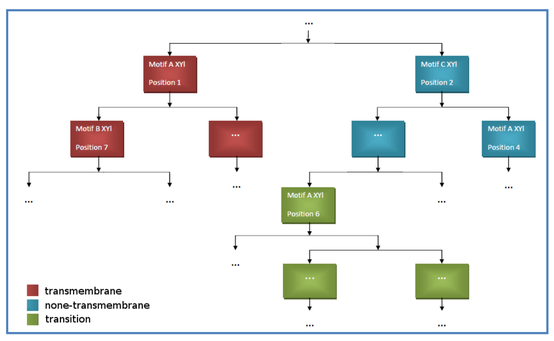
50 protein sequence motifs

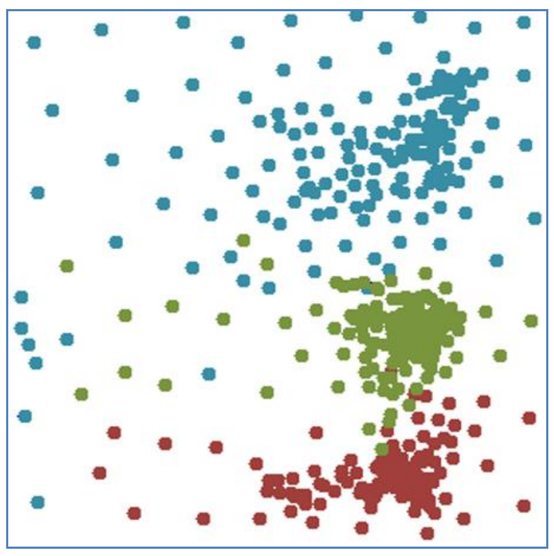
Statistical analysis

$$P(a|pos_i|M) = \frac{\sum_{j=1}^{k} g\,(pos_i)}{k}$$

$$g\,(pos_i, a) = \begin{cases} 1 & pos_i \text{ equals } a \\ 0 & else \end{cases},$$

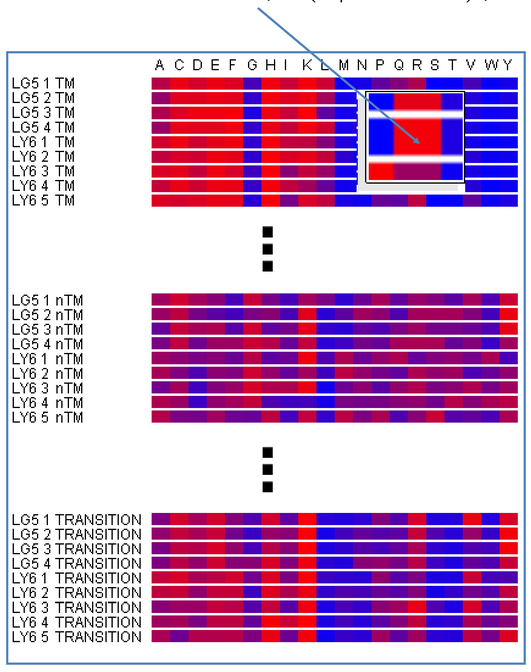
Clustering methods





Naturally occurring

$$f(x) = 1 - \log\left(\frac{P(a|pos_i|M)}{P(a|Nature)}\right)$$

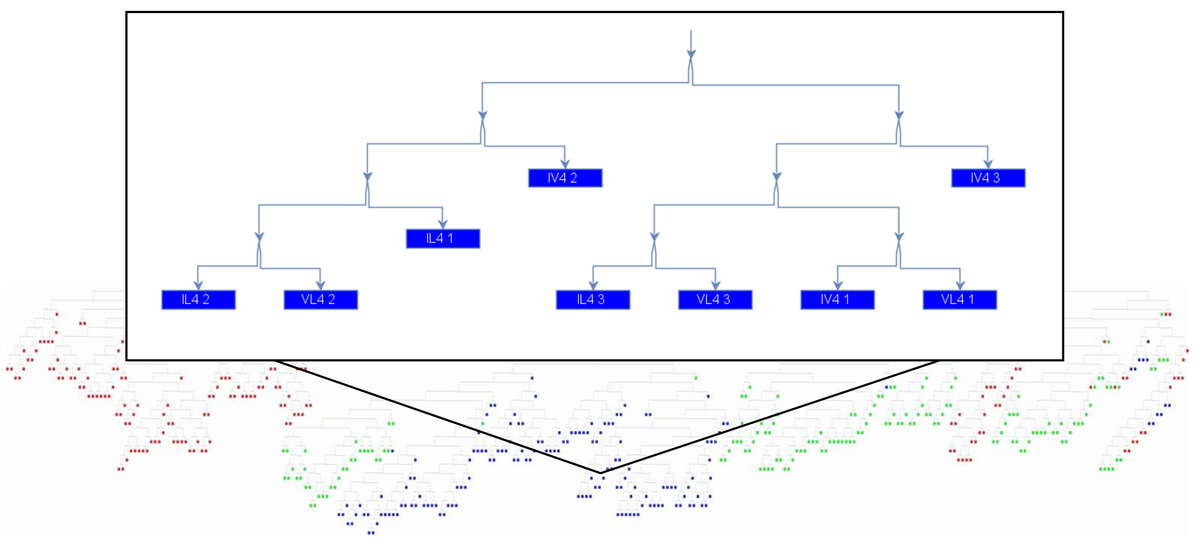


Spearman´s rank to each position

Position specific separation and clustering

Results and Application

- Position specific separation and clustering
- Clustering into transmembrane, non-transmembrane and transition regrions
- Motifpositions with great similarity as a direct leaf neighbors

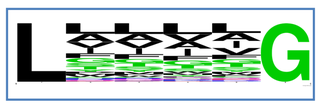


- Motif tendency to region distribution

$$tendency = \sum f(x)$$

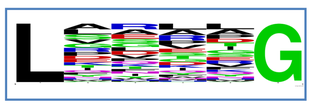
- Separation of motifs by a residue specific distribution.

transmembrane



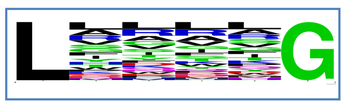
L-[LAY]-[LAY]-[LVA]-[LAI]-G

non-transmembrane



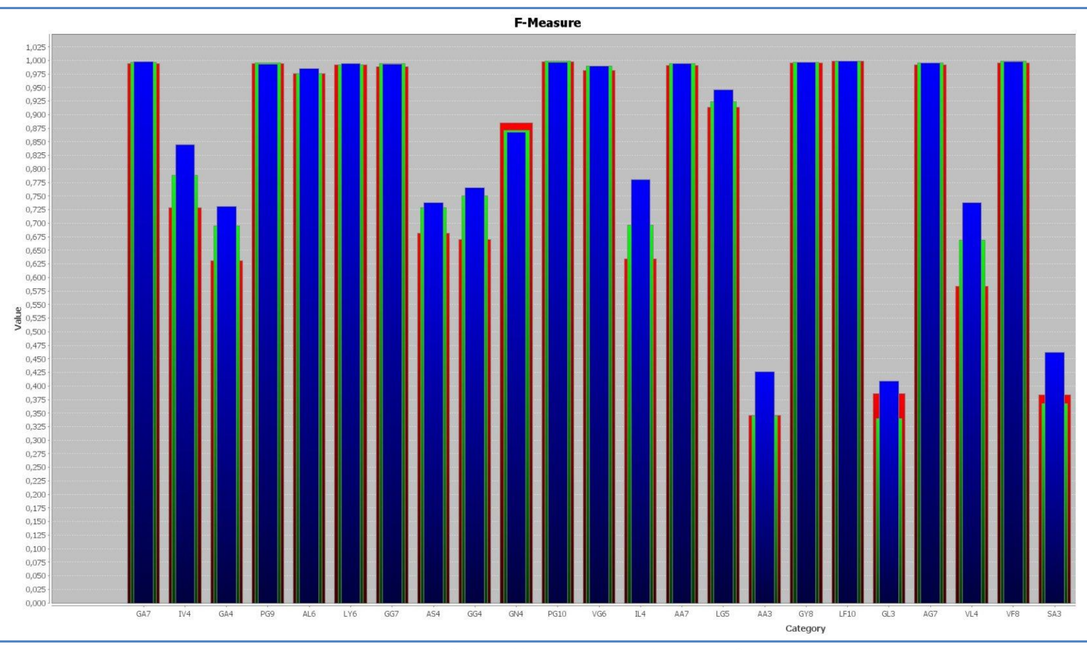
L-[ALS]-[RAE]-[LAR]-[LAP]-G

transition



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

- F-measure up to over 90% 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.



F-measure histogram for all analyzed motifs XYn with n>2

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

contact: steffen.grunert@hs-mittweida.de