NMR SELECTION OF PROTEINS FOR STRUCTURE DETERMINATION WITH NMR SPECTROSCOPY

Wojciech Augustyniak[1], Łukasz Słabiński[2], Łukasz Jaroszewski[3]

[1] *The Scripps Research Institute, La Jolla, CA,* [2] Silesian University of Technology, Gliwice, Poland,

[3] *The Burnham Institute for Medical Research, La Jolla, CA*

An appropriate target selection is used for increasing the success rate of structure determination in structural genomics. The initial selection must be based on bioinformatics criteria and these criteria are studied here in detail to establish a reliable filter for selection of proteins suitable for structure determination using NMR spectroscopy. It was assumed that the targets suitable for structure determination shall be similar to the proteins solved by NMR spectroscopy. To avoid the competition with X-ray crystallography, the convenient targets shall also differ from proteins solved by crystallographers. The appropriate protein characteristics were identified in order to select targets fulfilling these two major requirements. The procedure consisted of the following steps: 1) Variety of protein parameters were calculated starting from amino acid sequences. The compared parameters includes for example: occurrence of amino acids residues and predicted physicochemical parameters. 2) Distribution of proteins versus each calculated parameter was computed. 3) The obtained distributions were compared between proteins solved by NMR spectroscopy and annotated proteins - to identify the properties favored by NMR spectroscopists; and proteins solved by X-ray crystallography - to find the parameters disfavored by crystallographers. 4) The selected protein parameters will be used to construct a filtering function for target selection, which will facilitate target selection for NMR spectroscopy in the future.

Data Preparation

Solved Proteins

Sequences from Protein Data Bank[1]
46051 entries

Protein Selection

Protein Sequences
44116 entries

His-Tag Elimination

Protein Sequences without His_6

NMR Solved Proteins
4164 entries

Clustering to 70% of homology CD-HIT[3]

Final Set:
NMR Solved Non-Homologous Proteins
3347 entries

Genomic Proteins

Annotated Protein Sequences from GenBank[2]
1321613 entries

Clustering to 70% of homology CD-HIT[3]

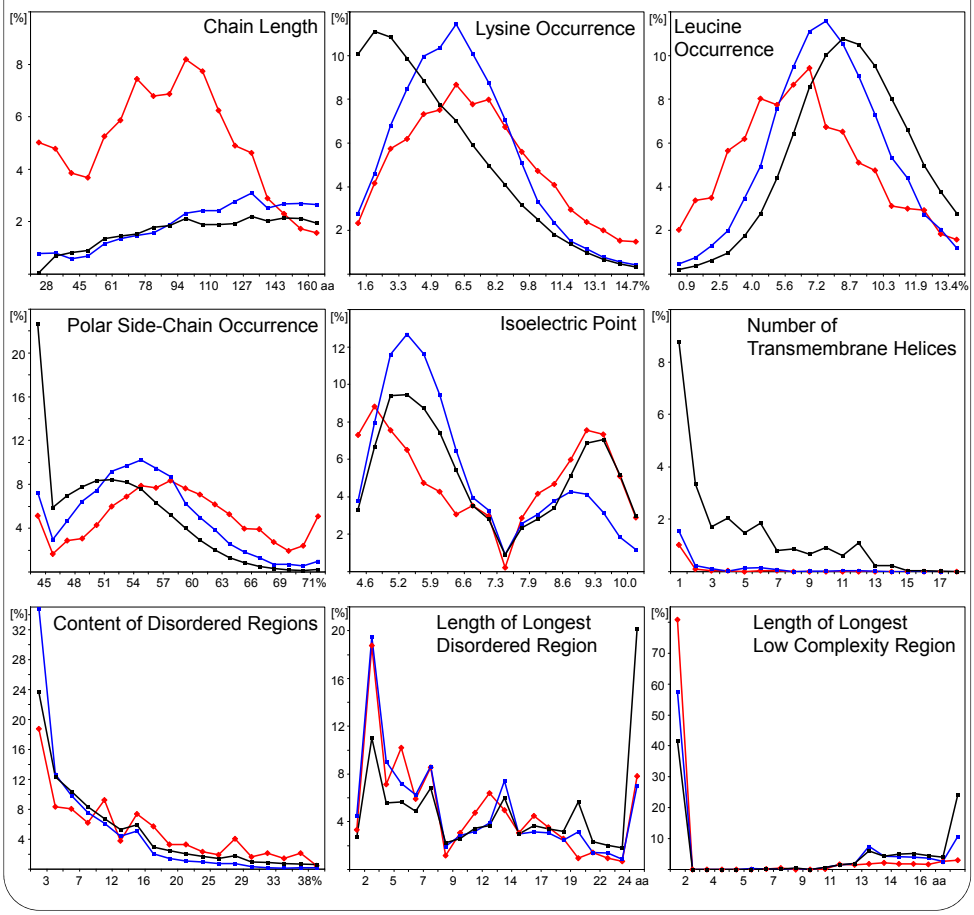
Final Set:
Non-Homologous Proteins
813895 entries

Crystallography Solved Proteins
20257 entries

Clustering to 70% of homology CD-HIT[3]

Final Set:
Crystallography Solved Non-Homologous Proteins
8918 entries

Distribution Comparison: NMR-, Crystallography-Solved and Genomic Proteins



Analyzed Parameters

- occurrence of each amino acid
- amino acid chain length
- occurrence of functional groups and chemical moieties
- occurrence of charges and acidic/basic/polar groups
- physical parameters: GRAVY[4], aliphatic index[4], instability index[5], extiction coefficient[6]
- transmembrane helices (TMHMM[7])
- secondary structure content (PSIPRED[8])
- disoredered region content (DISOPRED2[9])
- low complexity region content (SEG[10])

Future Outlook: Selection Filter

- fit distributions of selected parameters to appropriate functions for 3 data sets: NMR-solved, crystallography-solved and genomic proteins
- compute occurence probability for each selected parameter of the protein of interest using the above mentioned distribution functions
- compute overall score favoring differences between 1) proteins solved by NMR and crystallography, and 2) NMR-solved and genomic proteins
- prepare website with selection filter

Acknowledgments

Grant Sponsor: National Institute of General Medical Sciences, Protein Structure Initiative; Grant Number: U54 GM074898. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.
W.A. is a Marie Curie Outgoing International Fellow sponsored by the European Community.

References

1. Protein Data Bank, www.pdb.org
2. Benson D. A., et al, *Nuc. Acid Res.*, 2007, 35, D21
3. Li W., Godzik A., *Bioinformatics*, 2006, 22, 1658
4. Kyte J., Doolittle R. F., *J. Mol. Biol.*, 1982, 157, 105
5. Guruprasad K., et al, *Protein Eng.*, 1990, 4,155
6. Gill S. C., von Hippel P. H., *Anal. Biochem.*, 1989, 1182, 319
7. Krogh A., Larsson B., von Heijne G., Sonnhammer E. L., *J. Mol. Biol.*, 2001, 305, 567
8. Jones D. T., *J. Mol. Biol*,, 1999, 292, 95
9. Ward J. J., Sodhi J. S., McGuffin L. J., Buxton B. F., Jones D. T., *J. Mol. Biol,*, 2004, 337, 635
10. Wootton J. C., *Comput. Chem.*, 1994, 18, 269