Supplementary Figures

Usage of a dataset of NMR resolved protein structures to test aggregation vs. solubility prediction algorithms

Daniel B. Roche[1,2], Etienne Villain[1,2] and Andrey V. Kajava[1,2,*]

[1]Centre de Recherche en Biologie cellulaire de Montpellier, CNRS-UMR 5237, Montpellier, France
[2]Institut de Biologie Computationnelle, Université de Montpellier, Montpellier France

*Corresponding author: Andrey V. Kajava email: andrey.kajava@crbm.cnrs.fr

Analysis of Sequences from NMR-based set of soluble tails

Global statistics

Number of sequences: 361
Total number of residues: 10078
Smallest sequence: 15 residues
Largest: 108 residues
Average sequence length: 27.9 residues

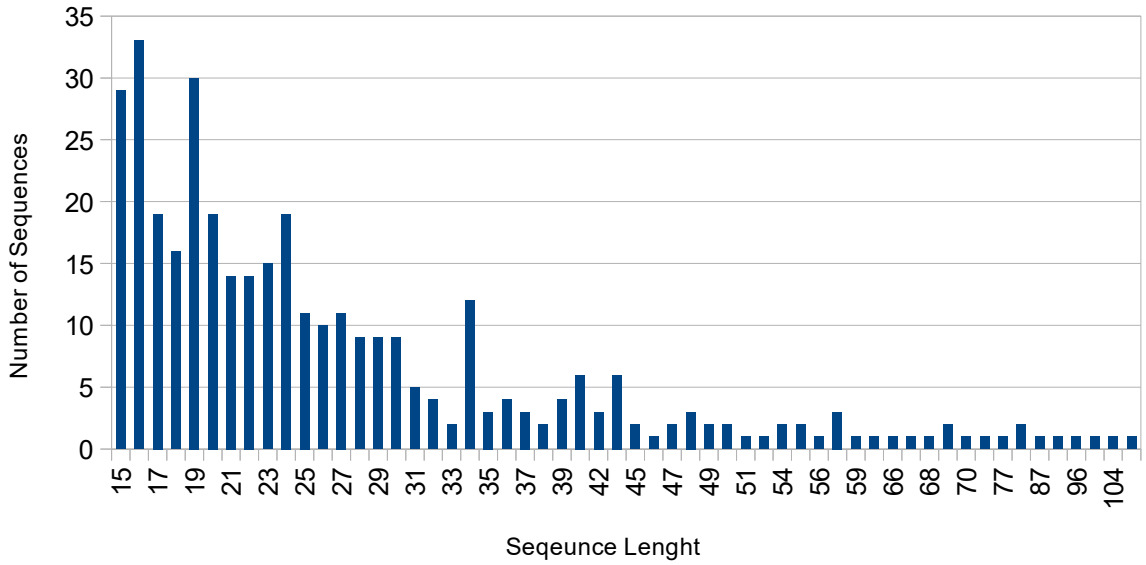

Figure S1: Sequence length distribution for our database.

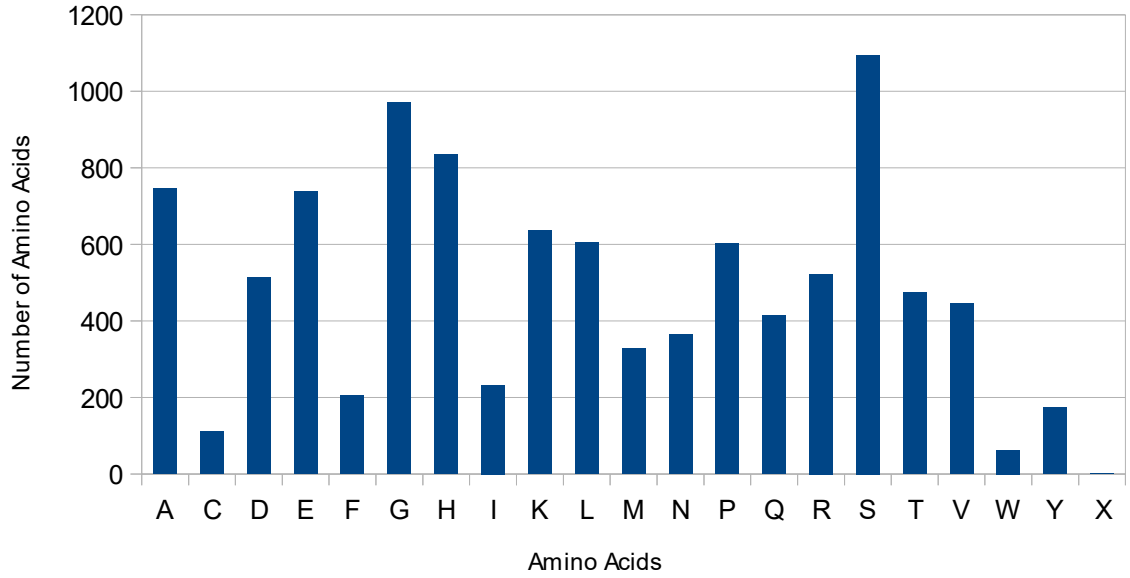

Figure S2: Amino acids composition for all the sequences contained in our database.

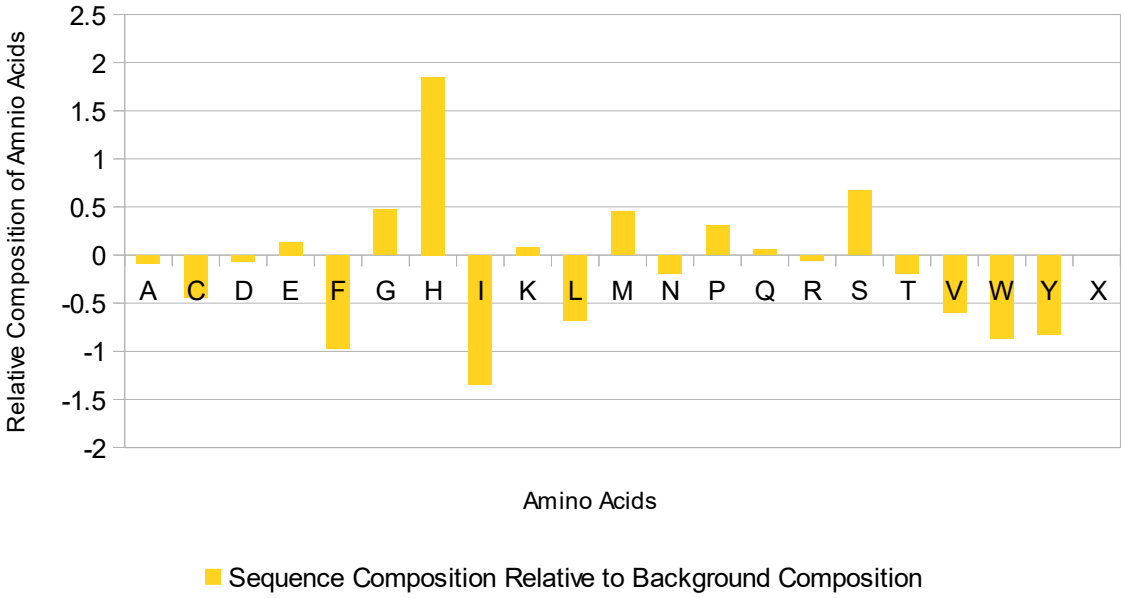

Figure S3: Relative amino acids composition for all the sequences contained in our database relative to background composition, generated using esl-seqstat from HMMER (1,2).

1. Eddy SR. Profile Hidden Markov Models. Bioinformatics, 14:755-763, 1998.
2. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions. Nucleic Acids Research, 41:e121, 2013.