Application Optimized Statistical Energy Functions

Laimer J., Lackner P.

University of Salzburg, Faculty of Natural Sciences, Department of Molecular Biology

Background

Statistical energy functions (SEFs) are an important tool in protein structure science. SEFs have a broad range of applications such as protein structure prediction, 3D-model assessment or prediction of stability changes upon point mutations. SEFs are usually evaluated on structure decoy sets as collected in the Decoys 'R' Us database and subsequently applied for the above mentioned tasks. It remains unclear to which extent the numerous SEF parameters of a certain SEF implementation are valid throughout the different applications.

Methods

We enumerate the typical parameters for two SEFs, optimize the SEFs for native fold identification in a decoy set and apply them to change in stability prediction. We then optimize the SEFs for change in stability predictions and apply them to native fold identification.

Datasets

- **SEF compilation**
  - PISCES database[3]
  - resolution cutoff: 1.8Å, R-factor cutoff: 0.25, identity cutoff: 20%, chain length cutoff: 500
  - To prevent undesirable side effects we filter the PISCES data for: (i) chains which are associated with viruses and membranes by PDB keyword search, (ii) chains with an incomplete backbone, (iii) chains found via PDB blast as too similar to the structures in one of the two datasets, and finally (iv) chains with a z-score $> -6.67 - 0.0141x + 2$ (x . . . sequence length) calculated by ProSa2003[4].

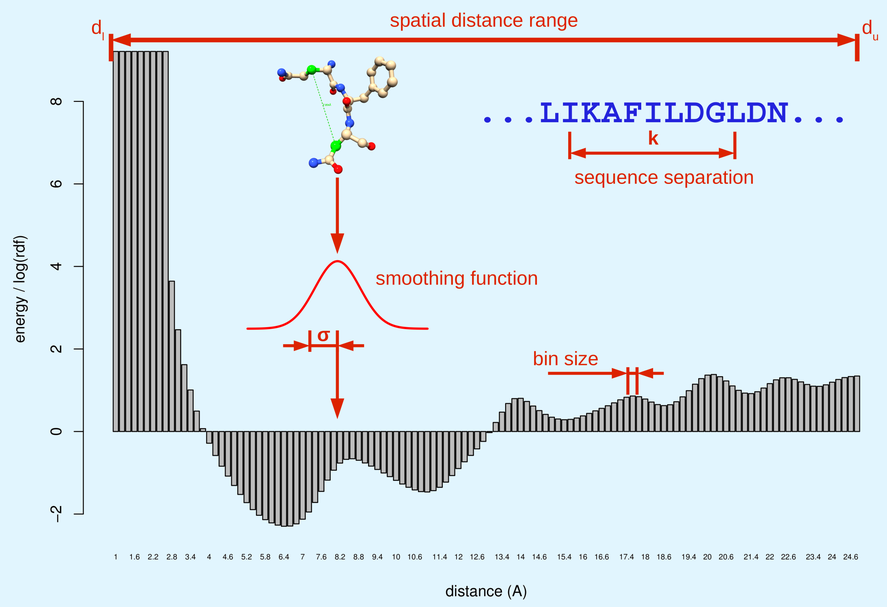
- **Evaluation and optimization**
  - native fold identification: Decoys 'R' Us[1] (multiple decoy sets)
  - stability change prediction: training and validation set published by Dehouck *et al.*[2]

Evaluation

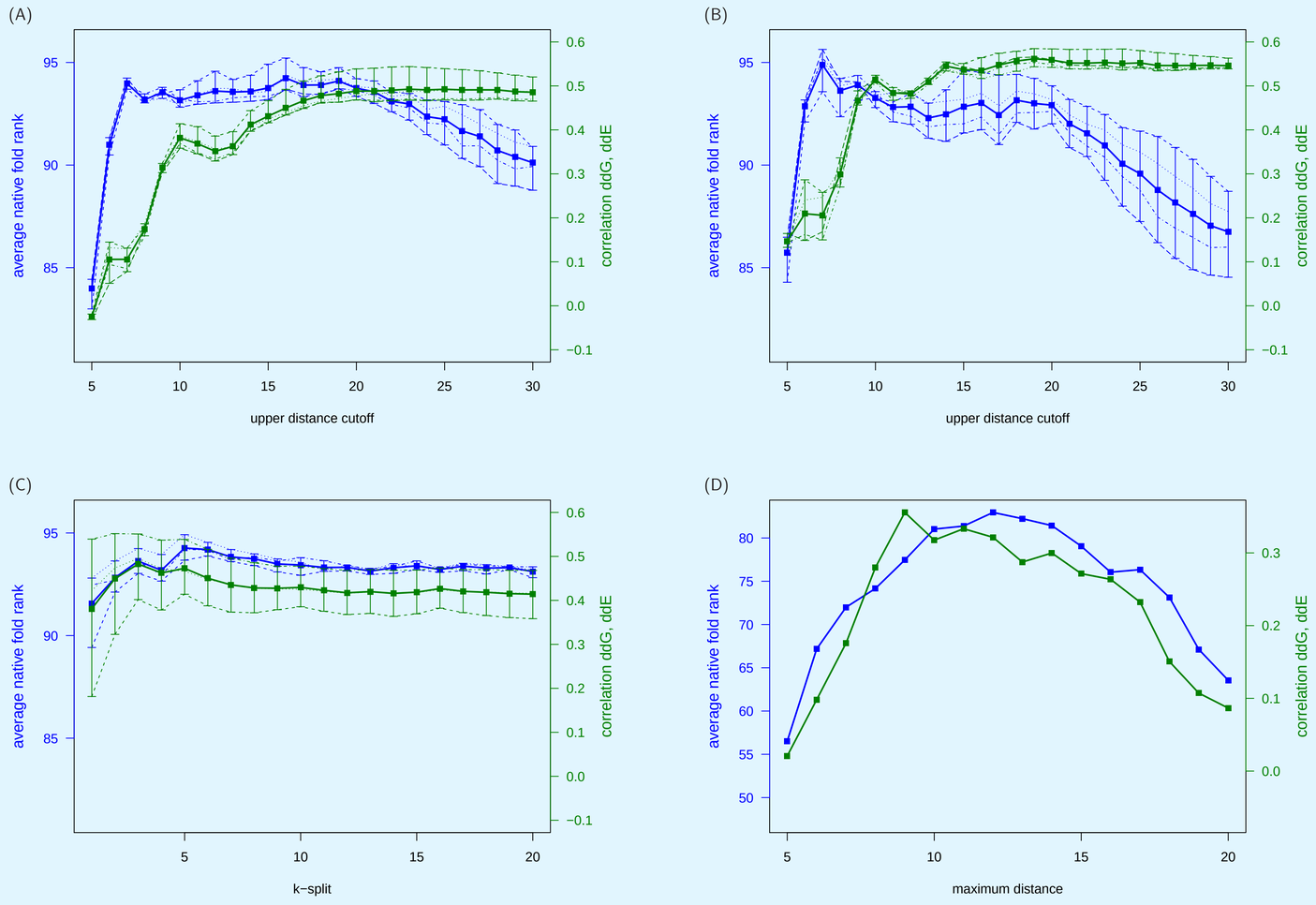
Each parameter set is evaluated by two methods. For the native fold identification we use the average rank of the native folds in percent to the total number of decoys in each set. For stability change prediction we use the Pearson correlation coefficient between the measured ddG values and the delta energy (energy of wildtype - energy of mutant). For this study we only investigate monomeric proteins.

Statistical Energy Functions

- **pair-SEF**
  - preferred distances between two types of atoms/residues
  - parameters: spatial distance range, sequence separation, bin-size, smoothing function
- **contact-SEF**
  - numbers of contacts to other atoms within a defined distance
  - parameters: maximum spatial distance, bin-size



Results

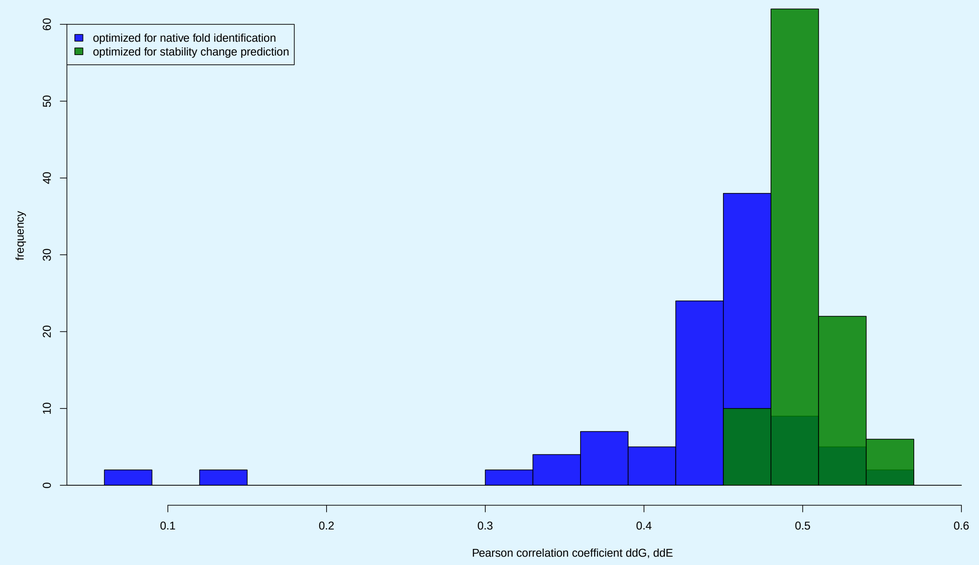


(A) The effect of the **spatial distance range** parameter for pair-SEF compilation on the native fold identification (blue) and the stability change prediction (green). The lower distance cutoff (d_l) is fixed at 0Å. The upper distance cutoff (d_u) varied between 5Å and 30Å. The values of four different k_{split} settings (5,10,15,20) are shown. The parameter k_{split} controls the pooling of distance measurements during SEF compilation, where measurements for $k < k_{split}$ are compiled in distinct SEFs, one per k-level, while the others are collected in a single SEF for $k = [k_{split} \ldots \infty]$. (B) Same setup applied to structures smaller than 100 residues. (C) The $\mathbf{k_{split}}$ parameter varied between 1 and 20. The values for three different upper distance cutoffs (10Å, 15Å, 20Å) are shown. (D) The predictive power of the **contact-SEF for spatial distance** cutoffs between 5Å and 20Å.

In case of pair-SEF, the prediction power for stability changes slightly increases with larger distance ranges, while for native fold identification short distances lead to better results. The native fold identification for short structures even gets worse if long distances are included. In contrast, the maximum distance parameter of the contact-SEF shows nearly the same optima in both applications.

The variation of bin-size and σ affects both evaluation tests comparably (data not shown). However, optimal values for bin-size and σ heavily depend on the size of the SEF compilation data set.

Different parameters have variable effects on the predictive power of the SEFs and there are mutual dependencies. The plot below shows the prediction power for stability changes of the top 100 parameter sets found for native fold identification (blue) and the top 100 parameter sets found for this application (green). The prediction power of the SEFs optimized for stability changes is significantly higher (Wilcoxon rank-sum test p-value$< 2.2e^{-16}$).



Conclusion

SEFs optimized for native fold identification are of limited applicability in change of stability prediction and vice versa. Single SEF parameters may show a small effect but in combination they lead to large differences in the predictive power of the SEFs. In ongoing work we investigate the effect of further parameters, as well as how methods which build on the SEFs are affected by the different optimization approaches.

Contact: josef.laimer@stud.sbg.ac.at

References

[1] **Decoys 'R' Us: A database of incorrect protein conformations to improve protein structure prediction.** Samudrala R and Levitt M. Protein Sci. 2000 Jul;9(7):1399-401. http://dd.compbio.washington.edu/

[2] **Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0.** Dehouck Y *et al.* Bioinformatics. 2009 Oct 1;25(19):2537-43. http://babylone.ulb.ac.be/popmusic/

[3] **PISCES: a protein sequence culling server.** Wang G and Dunbrack RL. Bioinformatics. 2003 Aug 12;19(12):1589-91. http://dunbrack.fccc.edu/PISCES.php

[4] **Recognition of Errors in Three-Dimensional Structures of Proteins.** Sippl MJ. Proteins. 1993 Dec;17(4):355-62. http://www.came.sbg.ac.at/prosa.php