Which distance measure is best for training and testing protein pair potentials?

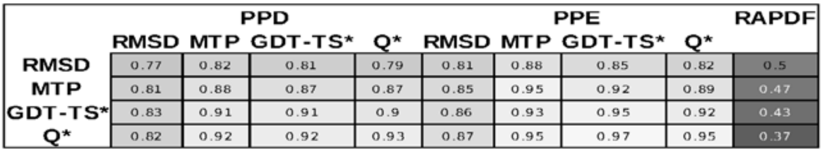
Martin Carlsen[a] , Peter Røgen[a], and Patrice Koehl[b]
a, Department of Applied Math. and Computer Science, Technical University of Denmark.
b, Department of Computer Science and Genome Center, University of California, Davis*.*

Predicting the structure of a protein based on its sequence usually involves two steps, sampling the conformational space and selecting native-like conformations. We study metric training and testing of knowledge-based potentials designed for the second step.

Method: In previous work[1], a high performing, b-spline based, C-α potential that includes a local term matching the sequence and structures of 7-mer fragments, a pair-wise C-α distance potential, and a solvation term was trained on the Titan high resolution decoy set (Titan-HRD)[2] using the so called metric training such that the entire potential (F) is imposed to form a linear funnel shape around a set of native structures given by F(sequence, structure)=F(sequence, native structure)+Dist(structure, native structure).

In this work, we trained two C-α pair potentials following the same methodology, one based on all inter-residue distances (PPD), while the other had the set of all these distances filtered to reflect consistency in an ensemble of decoys (PPE). To investigate the importance of which notion of distance is used to quantify near-native conformations, we tested four different distance measures, two based on extrinsic geometry (RMSD and GTD-TS), and two based on intrinsic geometry (Q*=1-fraction of native contacts and MTP an anharmonic normal node potential).

Results: We found a striking improvement of energy distance correlation when using the intrinsic distances. Both PPD and PPE perform extremely well on high resolution decoy sets, with correlation coefficients between energy and distance usually well above 0.8. PPE always performs better than PPD on this set, emphasizing the benefits of capturing consistent ensemble information. The performance of the statistical potential RAPDF[3] is shown for comparison. For lower resolution decoys the performance of PPE decays only slowly, mimicking the correlations between the distance measures it was trained on, while PPD requires additional local energy terms to sustain performance.



	PPD				PPE				RAPDF
	RMSD	MTP	GDT-TS*	Q*	RMSD	MTP	GDT-TS*	Q*	
RMSD	0.77	0.82	0.81	0.79	0.81	0.88	0.85	0.82	0.5
MTP	0.81	0.88	0.87	0.87	0.85	0.95	0.92	0.89	0.47
GDT-TS*	0.83	0.91	0.91	0.9	0.86	0.93	0.95	0.92	0.43
Q*	0.82	0.92	0.92	0.93	0.87	0.95	0.97	0.95	0.37

TABLE 1: *Showing the average correlation between energy and distance on Titan-HRD*

References

1. P. Røgen, and P. Koehl, Proteins:Struct. Func. Bioinfo. , 2013, **81**,841–851.
2. R. Rajgaria, S. McAllister, and A.C. Floudas, Proteins: Struct. Func. Bioinfo. 2006, **65**, 726–741.
3 R. Samudralam and J. Moult, *Journal of Molecular Biology 1998,* **275**, 895-916.