Optimal Use of Protein Structural Data for Knowledge-Based Potentials

Armando D. Solis, PhD
Biological Sciences Department
New York City College of Technology (City Tech)
of The City University of New York (CUNY)
Brooklyn, New York, 11201 U.S.A.
Email: asolis@citytech.cuny.edu

Knowledge-based potentials, used for protein structure analysis and prediction, are routinely derived from high-resolution structures from the PDB. The mechanism of transforming occurrence frequencies in structural data into "energies" or scores requires a well-defined way of assembling and using a comprehensive structural data set. My ongoing work, linking information-theoretic notions to the nature and action of knowledge-based potentials, explores ways in which structural data sets can be constructed and used so that maximal information is extracted. I'm currently working on two directions. First, I propose an improved method to use *all* high-resolution structures in the PDB to construct more informative folding potentials. This is in contrast to the common strategy of using a non-redundant subset of structures, which limits database bias, but also throws out potentially valuable knowledge. Second, I explore "query-specific" potentials, uniquely tailored to the characteristics of the query sequence. This is in contrast to the standard procedure of deriving "one-size-fits-all" potentials that work evenly across various query sequences. These two directions are supported by rigorous information-theoretic formulations, whose goal is to maximize discrimination between native and decoy conformations. I present current results, demonstrating that these two strategies produce potentials that show markedly improved performance in fold recognition tests.