Efficient Prediction of Biological Structures from Molecular Formulas

Mai Hamdalla[1], David Grant[2], Sanguthevar Rajasekaran[3], Reda Ammar[3] and Dennis Hill[2]

[1,3]*Computer Science and Engineering Department, University of Connecticut*

[2]*Pharmaceutical Sciences Department, University of Connecticut*

The goal of most metabolomics studies is to identify small-molecule metabolites in tissues and biofluids, and to correlate their levels with physiological and/or toxicological endpoints. It has become clear that small-molecule identification is one of the most problematic challenges in this field, as current chemical databases do not include all possible chemical structures.

We have developed a Biological Structure Predictor (BSP) that predicts all potential biological structures of a given molecular formula. BSP works by enumerating all possible structures corresponding to an input formula into a collection *C.* Only a subset of *C* could possibly be biological. We identify this subset using a database *S* of biological scaffolds. A scaffold in *S* is a substructure prevalent in known biological structures. Each structure *q* in *C* is matched to the scaffolds in *S. q* is ranked based on a similarity score. Only those structures in *C* that score above a threshold are output.

Comparing BSP results with the KEGG LIGAND database showed that 40% of the structures in KEGG were predictable. About 92% of BSP structures were not found in KEGG indicating the potential of this method for identifying unknown biological structures. We believe that increasing the number of biological scaffolds will result in a dramatic improvement.

[1] mai@engr.uconn.edu