Poster I-82

Protein Structure Prediction Aided Text Mining for Functional Inference

Authors:
Andreas Rechtsteiner (*CGB, Indiana University*)
Jeremy Luinstra (*B-Div, Los Alamos National Lab*)
Judith D. Cohn (*B-Div, Los Alamos National Lab*)
Charlie E. M. Strauss (*B-Div, Los Alamos National Lab*)

Short Abstract: Ab-initio structure prediction with Rosetta and Mammoth can be highly effective on structural variations beyond the limits of homology modeling and threading approaches. We report here that putative annotation rankings based on literature keyword similarity dramatically improve the annotation accuracy over either sequence or structure based annotation alone.

Long Abstract:

Each newly sequenced genome is, in due course, principally annotated by comparison of its sequences to previously annotated genomes.
Typically 40 to 60% of a new genome can be reliably annotated in this fashion. However, this method is most successful for the genes we often care least about, placing a premium on methods that can
annotate unusual or highly diverged sequences. That is, if an organism was chosen for sequencing based on it's unique characteristics, those special features are likely to have elements that are highly specialized and hence highly diverged from other genomes and only possessing weak, ambiguous, sequence similarity.

In this twilight recognition realm, structure based annotation can be useful. By prediction of a protein's approximate structure we can compare it's structure to proteins of known function. It turns out that the process is highly insensitive to the accuracy of structure prediction allowing for recognition even for sequences with large deletions and variations in their sequences. It has been shown previously that ab initio modeling and comparison methods like Rosetta and Mammoth are highly effective on structural variations
beyond the limits of homology modeling and threading approaches.

Because it is less specific than sequence based annotation, it is useful to confirm putative transitive annotations by other means. We have been developing methods to screen genome-scale predictions in an
automated fashion to selected candidates for human curation effort.

To approach this we compare sequence and structure based similarity measures for common features. The comparison measure we describe here is text based. We generate a body of text from sequence-based
comparison to the non-redundant sequence database by selecting the literature associated with the top BLAST hits and extracting keywords. Then we do the same with structure comparison to SCOP families. We then apply a text key word similarity measure between

each SCOP match and the ensemble of sequence-based key-words.

We report that putative annotation rankings based on this approach dramatically improve the annotation accuracy over either sequence or structure based annotation alone. This is demonstrated on large
benchmark sets that are carefully screened for homolog removal to simulate highly diverged sequences.