BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1

Automating Ontological Function Annotation: Towards a Common Methodological Framework

Cliff Joslyn[*]

[a] *Computer and Computational Science, Los Alamos National Laboratory,*
[b] *Biosciences, Los Alamos National Laboratory*

A new paradigm for functional protein annotation is the use of automated knowledge discovery algorithms mapping sequence, structure, literature, and/or pathway information about proteins whose functions are unknown into a functional ontology, typically (a portion of) the Gene Ontology (GO). Our own work in this area has resulted in the development of the POSet Ontology Laboratory Environment (POSOLE), which has been deployed for CASP.

In doing this work, and in looking at the literature, we have uncovered a variety of methodological issues which we believe could be valuable for the community to focus on. In this paper, we first introduce the POSOLE architecture as applied to the CASP functional annotaiton task, and use it as a point of deprature to discuss the following issues:

*) First, there's the need to settle on a set of annotations. Might it be preferable if this was regularized across the community? Annotation mappings from sets of proteins to sets of GO nodes, for example, Uniprot, GOA, IEAs, etc.

*) The overall paradigm is to compare the results of one's automated method against a set of known annotations (call it a "gold standard"). Developing a coherent test set, which can potentially be shared within the community, is critical. Published, within the community. Trusted to be true, and selected for evaluation. In our case, structures. Refer to others here. NR gold standard (J).

*) Evaluation methods vary greatly, with different methods (ROCs, F-scores, sensitivity and specificity measures) measuring ratios of true and false positives and negatives against each other in different ways. This can be standardized to some extent, although particular architectures can force particular choices here. For example, the results produced by POSOLE do not form a simple set, but rather a ranked list of effectively indefinite length. This has forced us to create a non-standard measure of precision in the context of an F-score balancing precision and recall.

*) Beyond that, however, the question is begged of what a "true positive" actually is, and in particular, what a "near miss" can be. The structure of the GO is such that measuring distance between nodes is a non-obvious but central question, one that has occupied our research recently. An example of the kinds of motivating questions present is how do we compare an automated annotation which is a parent or child of a correct annoation with one which is a sibling? This leads to a more general set of questions about comparing collections of annotations against each other, for example the collection of correct annotations, considered as a set of GO nodes, to another set of automated annotations. Different concepts of depth and location in the GO are available, but also still in development, and these need to be appreciated and internalized by the bio-ontology community better.

*) Central to all these questions is the need for a much better analytical infrastructure for analyzing portions of the GO. For example, consider collections of GO nodes which might be generated as known annotations of either test or target sequences (or their BLAST neighbors), putative automated annotations, or combinations of these, or just arbitrary GO portions like the CC branch. We might wish to understand such things as what the "average depth" of that set is, of the "size" of the region they circumscribe, and the relative amounts of "back-branching" present.

[1]

REFERENCES

[1]Pal, Debnath and Eisenberg, David: (2005) "Inference of Protein Function from Protein Structure", *Structure*, v. **13**, pp. 121-130

[*]Knowledge Systems and Computational Biology Team, Computer and Computational Sciences, Mail Stop B265, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, joslyn@lanl.gov, http://www.c3.lanl.gov/˜joslyn, (505) 667-9096.