Bayesian Inference of Protein Function Using Homology, Pathway, and Operon Data

Michelle L. Green[1], **Peter D. Karp**[1]

Keywords: pathway/genome database, non-homology, function prediction, sequence analysis

1 Introduction.

The PathoLogic program enables users to create a pathway/genome database (PGDB) predicted from the enzymes present in the genome annotation of an organism of interest. The program maps the genome's enzymes onto a reference set of metabolic pathways and considers a pathway to be present in the organism if any enzymes in the pathway are present. A pathway hole or missing reaction occurs when a genome appears to lack the enzyme needed to catalyze a reaction in the pathway. We have developed a method that combines homology, pathway, and operon-based data to identify and evaluate candidate sequences to fill these pathway holes. Most genome annotation efforts fail to assign function to 40 - 60% of the new sequences. Even when annotated, many functions remain incomplete or nonspecific. Operon- and pathway-based information can provide additional clues about the function of unannotated proteins, and can clarify incomplete or nonspecific annotations.

2 Method.

The activity of each missing reaction is known from the inferred pathway; we use a set of isozyme sequences encoding the required activity in other genomes to search for candidates in the genome of interest. After identifying candidate sequences, our program uses a Bayes classifier to evaluate each candidate. Rather than evaluating the candidates based solely on their similarity to the set of search sequences, we determine the probability that the candidate has the desired function. Our classifier considers evidence from the homology search (e.g., E-values, alignment lengths, and the rank of the candidate in the BLAST output), from the pathway context of the missing reaction (e.g., is the candidate gene adjacent to a gene catalyzing an adjacent reaction in the pathway?), and operon-based data (i.e., is the candidate gene likely to appear in an operon with another gene in the pathway?).

We previously applied the pathway hole filler to three computationally derived PGDBs. Cross-validation studies using the known reactions in these PGDBs resulted in 71% precision at a probability threshold of 0.9. The fact that most of the "true" functional annotations in our studies were derived from sequence analysis complicates assessment of the efficacy of the program. EcoCyc is a PGDB derived from experimental evidence. Hence, we applied our program to the known reactions in EcoCyc to obtain a more accurate indication of the performance of our method.

[1] Bioinformatics Research Group, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, USA. E-mail: `green@ai.sri.com`

3 Results.

Isozyme sequences are available for 494 reactions in EcoCyc's metabolic pathways. A cross-validation study using these known reactions achieved 83% precision and 85% recall at a probability threshold of 0.9. Figure 1 shows the number of true positives versus the number of false positives for known reactions predicted using our model. The chart also includes predictions based only on the E-value of the best alignment between the candidate and the set of query sequences.

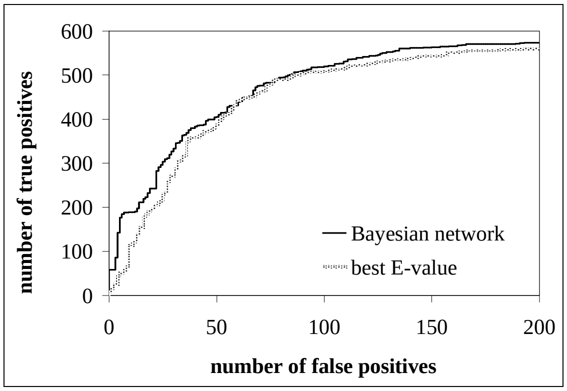

Figure 1: Number of true positives vs. false positives for known reactions in EcoCyc.

As evidenced by the "best E-value" curve, searching an organism's genome with multiple isozyme sequences accurately identifies a significant number of the enzymes assigned to known reactions in EcoCyc. The difficulty in using E-values lies in selecting an appropriate cutoff. At a cutoff of 1e-40 (a reasonably conservative value), precision is 61% -- over 1/3 of the predictions made will be false positives. Identifying enzymes using our Bayesian network allows the selection of a more rational cutoff, based not on a measure of sequence similarity, but on the probability that a sequence has the desired function.