4-11-2014

Proposing Genes for Gap Reactions in Metabolic Pathways

Carl Deeg

Shinnosuke Kondo

Hope COLLEGE
Dr. Matthew DeJongh
Department of Computer Science
dejongh@hope.edu

Proposing Genes for Gap Reactions in Metabolic Pathways

Carl Deeg, Shinnosuke Kondo
Dr. Matthew DeJongh

Abstract

A metabolic model is a map of the biochemical reactions that take place in an organism. These reactions are catalyzed by enzymes, which are encoded by genes in the organism's genome. However, there are reactions that are known to exist and needed to complete the metabolic model, but are not associated with any genes. These are called "gap reactions" (see Figure 1). Our goal is to find the genes that encode the enzymes that catalyze these gap reactions. We have researched two approaches: a knowledge-driven approach that focuses on finding a small set of good candidates, and a data-driven approach that focuses on scoring all candidates to rank their plausibility. Identifying the genes that are associated with gap reactions produces better predictive models and directs laboratory experimentation.

Background

Data Used in Previous Approaches:

- Gene-Reaction Association: shows how gene activity is related to reaction activity
- Chromosomal Contiguity: genes that are found on the same or similar chromosomes
- Metabolic Paths: follows the transformation of compounds through a series of reactions

Data in Our Approaches:

We used the following online databases: The SEED (http://www.theseed.org/) and KBase (http://kbase.us/). These databases contain the types of data previously used for this kind of analysis, as well as two other types not previously used:

Functional Coupling indicates how frequently a pair of genes is conserved in different organisms. For example, in Figure 2, a gene with function *purF* and another gene with function *purD* co-occur in many organisms, including ST, PN, BS, CA, and EF. Thus, the functional coupling score for these two genes is high.

Gene expression correlation measures how closely a pair of genes are active in different experimental conditions. Assuming the amount of RNA represents the activity of a gene, expression levels are shown by the brightness of cDNA, which is the complement of the RNA (Figure 3). The Pearson correlation for each gene pair is calculated by comparing expression levels across many experiments.

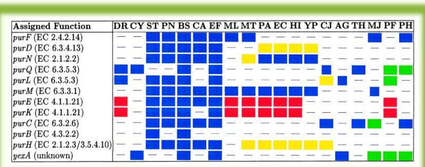

Figure 2. Occurrence of genes with the same function over different organisms

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA. 1999;96:2896–901.

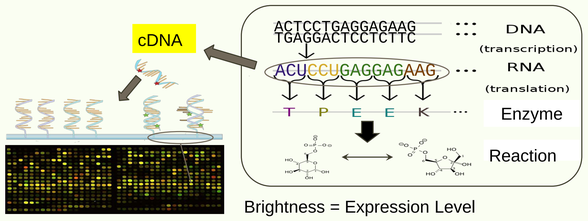

Figure 3. Measure of gene expression by capturing RNA.

Adapted from:
https://en.wikipedia.org/wiki/File:Genetic_code.svg, https://en.wikipedia.org/wiki/File:NA_hybrid.svg

Data-Driven Approach: Focus on Scoring Candidates

Method:

Genes are compared to the neighbor genes associated with reactions surrounding the orphan reaction. If the data suggest that a gene is within a close metabolic distance of the neighbors, the gene is proposed as a candidate.

Finding Candidates:

All non-neighbor genes are considered as candidates.

Scoring Candidates:

Gene pair data is compiled into a table containing:

- Expression correlation
- Functional coupling
- Metabolic distance

Using Bayes' Theorem, the known data is used to estimate the probability that a given pair of genes is within a specified metabolic distance of one another

Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

A = metabolic distance < x
B = within expression correlation cutoff

Gene neighbors are then compared individually to every gene in the organism, and if they are predicted to be within a certain metabolic distance, the gene is proposed as a possible candidate.
Candidates that are close to multiple neighbors with high Bayes probabilities are scored as the best candidates.

Results:

In initial testing, our Data-Driven approach produced limited results.
Treating known reactions as though they were orphan reactions, this approach proposed the correct gene if the gene is indeed functionally coupled and has expression correlations with its neighbors.
If the correct gene is neither functionally coupled nor has expression correlations with its neighbors, the gene was not proposed.

Conclusion:

In the current state of the script, we can achieve limited success by reaching the "low-hanging fruit," finding the genes that are easily identified. In order to reach for the higher, novel fruit we must expand this approach beyond the two variables that were used. More data is available to be incorporated into the Bayes estimation and will hopefully increase its predictive power.

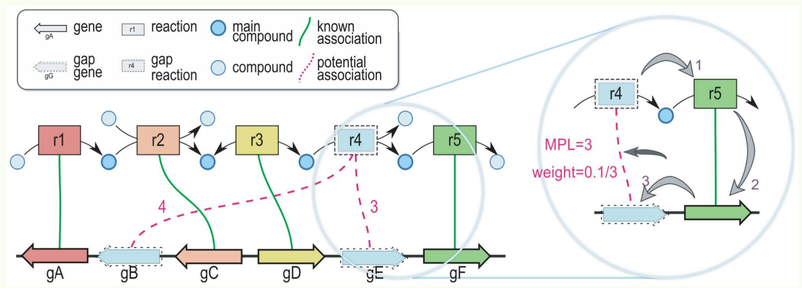

Figure 1. Reaction paths and cluster genes connected by known associations.

Source: Smith AAT, Belda E, Viari A, Medigue C, Vallenet D (2012) The CanOE Strategy: Integrating Genomic and Metabolic Contexts across Multiple Prokaryote Genomes to Find Candidate Genes for Orphan Enzymes. PLoS Comput Biol 8(5): e1002540.doi:10.1371/ journal.pcbi.1002540

Knowledge-Driven Approach: Focus on Finding Candidates

Method:

As a starting point, we adapted a method called CanOE (Candidate genes for Orphan Enzymes) developed by Smith et. al.(2012).

Finding candidates:

In CanOE, subsets of a metabolic model, called genomic *metabolons*, are created first. In a metabolon, genes are connected in a graph representation based on their positions on chromosomes, and reactions are connected by common compounds. Genes and reactions are connected by known gene-reaction associations (Figure 1). Frequently two genes are close on chromosome when the reactions associated with them are close in a metabolic map, so every possible combination of gap reactions and gap genes within a metabolon are proposed as potential association candidates. In our method, genes in a metabolon are clustered using gene expression correlation and functional coupling as well as positions in chromosomes, because we know that when reactions are neighbors on a metabolic map, associated genes frequently have high values of gene expression correlation and function coupling. At the start of our clustering process, a distance threshold is set. Then, if the value for distance between a pair of genes is within the threshold, that pair of genes is put in the same cluster (Figure 4). After clustering, candidate associations are proposed in the same manner as CanOE.

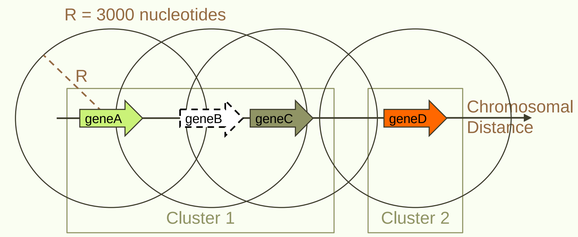

Figure 4. Single linkage with chromosomal distance

Scoring candidates:

Gene expression correlation is used to score candidates. Candidates that are highly correlated with other genes in the metabolon receive the highest scores.

Results:

There were no predictions which were obviously wrong, and our method was able to rediscover several known or newly found associations. However, we did not succeed in limiting the number of candidates to an extent necessary for biological validation. For example, our method proposed 4679 candidates along with one recently found association for *Shewanella oneidensis* MR-1.
Nevertheless, working with some biologists, we determined that gene expression correlations are useful for evaluating candidates.

Conclusion:

Though there is limited success at constraining the number of candidates, we can make use of gene expression correlation for scoring candidates when other evidence is available to constrain the number of candidates.

Acknowledgement

- Hope College Computer Science Summer REU Program
- Argonne National Lab
- Dr. Nathan Tintle (Dordt College)
- Dr. Ross Overbeek (Fellowship for the Interpretation of Genomes)
- Dr. Aaron Best (Hope College Biology)
- National Science Foundation
- Department of Energy Systems Biology Knowledgebase