Final Paper: Large-Scale Alignment of Encyclopedic Metabolic Networks

Tomer Altman

June 16, 2009

Abstract

Encyclopedic metabolic networks capture the sum of human knowledge of biochemical reactions and substrates as found across a multitude of organisms, and not one particular organism. A method for large-scale alignment of two such encyclopedic metabolic networks, KEGG and MetaCyc, has been designed to allow for a systematic comparison of their contents. A variety of methods for matching reactions and compounds in the two datasets and a method of scoring the fitness of the alignment are proposed.

1 Introduction

Bioinformatics, though in the past exclusively associated with DNA analysis, has grown to cover proteomics, metabolomics, and other such "omics" subfields. Several Model Organism Databases (MODs) focus on not only a community annotation of the MOD genome, but on describing its metabolism as well in what is termed a multidimensional annotation [1].

As the amount of data available on the metabolic reactions and chemical substrates increases, the necessity of central repositories of this data type increases as well. In addition to the inclusion of metabolic data in MODs, several encyclopedic repositories of metabolic network data have been set up over the years, including EBI's Rhea and ChEBI databases [2], KEGG [3], and MetaCyc.

A natural question that one might ask is, "what is the difference between metabolic database X and Y?" In genomics and proteomics, there are established methods and tools for performing such a comparison, such as analyzing the sequence homology using techniques such as Smith-Waterman or BLAST. Unfortunately, there is no current consensus for how to perform a similar large-scale comparative analysis for metabolic networks, let alone standard tools.

The benefit of such a method would be to allow researchers to evaluate the contents of different encyclopedic metabolic network databases, and be able to select the ones that would be most useful for their work. It will also enable the various encyclopedic metabolic network databases to evaluate where they might be lacking information, and thus they can work to complete the gaps in their knowledge. Finally, organizations such as National Center for Biotechnology Information (NCBI) have recently started work on establishing a central repository for such metabolic networks called BioSystems [4]. For such a repository to be of use, a method to map deposited metabolic data to common identifiers will be essential.

Herein we describe a method for comparing two encyclopedic metabolic networks, namely KEGG and MetaCyc. We will discuss the approach taken, compare with existing tools and methods, and outline future work to be undertaken.

2 Comparison of MetaCyc and KEGG

Before describing the methods by which the metabolic networks of KEGG and MetaCyc have been compared, a discussion of the differences of these two datasets is warranted. KEGG and MetaCyc are both part of large database and software projects. All full discussion of the various different software, online, and data file resources of both projects would be beyond the scope of this work. Thus, the discussion is limited to the resources of each project utilized for the purposes of comparison.

MetaCyc is a part of a larger collection of databases known as BioCyc Database Collection [5]. BioCyc consists of 409 databases that combine metabolic and genomic data, in what is termed a Pathway/Genome Database (PGDB). The data represented by MetaCyc is the union of all metabolic compound, reaction, and pathway knowledge represented in BioCyc, and thus serves as a reference database for metabolic networks. Unlike other PGDBs in the BioCyc collection, MetaCyc does not store any genomic (i.e., sequence) information.

The principle objects represented in MetaCyc for the purposes of this study are compounds (i.e., chemicals), reactions, and proteins. Also used was a class of objects known as "enzymatic reactions", which help to depict in a database the many-to-many relationship between reactions and the proteins that catalyze them. While all of these objects have common properties such as a unique identifier, a name, a description, external database links, and links to related objects, each object has specific properties associated with them. For example MetaCyc reaction objects can store an Enzyme Commission (EC) number if it is describing such a reaction.

The KEGG database houses many kinds of data types. In [3], they describe over 19 datasets. Their data is accessible via their website services, and as flat-file downloads available via FTP. The principle KEGG dataset used in this study is the LIGAND dataset. The LIGAND dataset has evolved over the years to contain not only chemical compound data, but now also enzymatic activity data in their ENZYME file, and a listing of all reactions described in KEGG in the REACTION file. All of KEGG's flat-files are in an idiosyncratic attribute-value format. It should be noted that the ENZYME file in LIGAND is very similar to the ENZYME.dat file available from ExPASy that describes the EC hierarchy of enzyme activities, except in that it has been annotated by KEGG curators with references to corresponding KEGG objects. KEGG also inherited a large database of glycogen molecules known as GLYCAN. Specific care was needed in the processing of KEGG data to exclude GLYCAN compounds and reactions, since MetaCyc does not represent this area of biochemistry to the same extent.

For a number of years KEGG only stored reaction data in the ENZYME file, such that KEGG only was describing EC reactions. As of release 46, KEGG now includes additional reaction data in the REACTION file. A peculiarity in the KEGG representation of their reaction data is that their is some redundant information represented between the ENZYME file and the REACTION file. In our work we had to be sure that not only were we importing distinct reactions, but that we were depicting the correct reactions as being official EC reactions as well.

In contrast to KEGG, the data from BioCyc can be analyzed by a powerful suite of software known as Pathway Tools. Pathway Tools provides not only a desktop application that allows for the visualization, navigation, and analysis of the various types of objects stored in a PGDB, but also provides a rich programming environment based on Common Lisp. The interactive Lisp API provides for a fast method to prototype programmatic queries of MetaCyc.

Though both MetaCyc and KEGG are known primarily as pathway databases, in this study we have limited our comparison to the metabolic network layer concerning reactions and compounds. Our analyses have been agnostic to the representation of pathways in KEGG and MetaCyc, which are differ significantly [6].

2.1 Database Curation

A major difference between KEGG and MetaCyc is that the curation policy of MetaCyc is to not only provide evidence codes for each asserted piece of knowledge, but also to provide clear provenance for each assertion. Gene Ontology (GO) Term-like evidence codes are available for pathways, and there is a built-in system to credit authors that create or update manually-curated information. From the KEGG datasets, it is unclear who might have authored a particular entry. Furthermore, there are no equivalents to evidence codes in their flat-files, which leave data consumers of KEGG to rely on knowing which data types KEGG curates. For example, KEGG manually creates their pathway reference maps, the compound data present in LIGAND, and the REACTION data file. On the other hand, most of the data that associates a particular reaction with a gene in a genome is based on KEGG Orthology (KO) annotations, and is computationally derived. The annotation of a particular KO number to a reaction is an example of a manual curation step.

Another distinction between the two databases is the level of literature citations and descriptions. While MetaCyc provides mini-reviews for certain objects, such as pathways and enzymes, KEGG objects often lack even basic description. KEGG also lacks links to PubMed and other journals, to help verify that their data is in concord with the literature. A more thorough comparison of the datasets can be found at the MetaCyc Literature Curation Guide [7].

3 Methods

3.1 KEGG Loader for the BioWarehouse Relational Bioinformatics Database Integration System

In order to systematically compare the contents of KEGG with MetaCyc, it was first necessary to extract the KEGG metabolic network data into a computable form. For this purpose we used SRI's BioWarehouse database integration system for bioinformatics [8]. The BioWarehouse contains software for parsing KEGG data files (called the KEGG Loader) and mapping the information to a common relational database schema. For the purpose of this work, we have extended the KEGG Loader to parse the REACTION file in the LIGAND dataset, and have imported KO identifiers as well.

As of version 50 of KEGG, the BioWarehouse KEGG Loader imported 15404 compounds and 9164 reactions. MetaCyc currently has 5425 small metabolic compounds and 6798 small-molecule metabolism reactions.

3.2 Compound Match Prediction Methods

The most important step in aligning two graphs is to establish a mapping between the two sets of vertices. The alignment of arcs relies on the quality of the vertex mapping. In this work we have paid especially close attention to this point, and have developed a number of algorithms for mapping compounds between KEGG and MetaCyc.

3.2.1 External Database Links in MetaCyc to KEGG

In the process of updating the MetaCyc knowledgebase, curators often add database unification links to connect our data to external resources. As of the writing of this paper, MetaCyc has 3321 database links from compounds to entries in KEGG's LIGAND dataset. These links served as the starting point for the creation of a mapping between MetaCyc and KEGG compounds. It should be noted that even though both MetaCyc and KEGG have database links from their compounds to external databases (many of which are in common), KEGG currently doesn't link to MetaCyc compounds.

3.2.2 Common PubChem Identifiers in MetaCyc and KEGG

A resource that has proven to be very useful for this work has been the PubChem Project from the National Center for Biotechnology Information [9]. PubChem is a centralized resource for chemical compound information as it relates to biological assays. Akin to the open deposition policy of GenBank, any organization can become a depositor of chemical information in PubChem. All compounds deposited in PubChem go though a standardization process that normalizes the chemical structures in terms of protonation, groups, aromatization, and other aspects. After the structure has been standardized, it is entered into the PubChem Structure dataset, and given an identifier unique to that deposition called a SID. After being entered into the PubChem Structure dataset, PubChem attempts to unify each deposited compound with compounds which are already present in PubChem. For each group of compounds from different depositors that are successfully unified, the structure along with the sum of deposited information is stored in the PubChem Compound dataset, and issued a unique identifier known as a CID.

An analysis of PubChem's Compound dataset revealed a number of compounds which were associated with both MetaCyc (described as "BioCyc" via the PubChem interface) and KEGG SIDs. A comparison of these associations to the database links already present in MetaCyc lead our curators to add about 300 additional database links from MetaCyc compounds to KEGG that were hitherto unknown.

3.2.3 Bi-Directional Best Hits of Tanimoto Similarity Scores

The comparison of two chemical structures can be reduced to the subgraph isomorphism or substructure identification problem [10], which in turn has been shown to be an NP-complete problemcitation. While both KEGG and Pathway Tools have software for doing exact chemical sub-structure matching, running a comparison of all unmapped KEGG compounds to MetaCyc compounds would be prohibitively time-consuming. Fortunately there is a resource available from PubChem known as the Score Matrix Service [11]. This web form gives researchers access to the "all versus all" molecular fingerprint comparisons used internally by PubChem for compound matching. A molecular fingerprint in this case is a bit vector of length 880, each representing a binary feature of the molecule. The binary features represent simple properties such as whether there are over a certain number of a specified element in the molecule, or how many aromatic rings are present. This representation allows for a fast determination of the similarity of two molecules, since then a bit-vector similarity measure can be applied. In this case, a Tanimoto coefficient is calculated, multiplied by 100, and is rounded to the nearest integer. It should be noted that two compounds that have a Tanimoto score of 100 are not necessarily isomorphic. What the molecular fingerprint comparison gives up in the way of specificity, it gains in speed.

In analogy to the method used for determining orthologs from “all versus all” BLAST or Smith-Waterman similarity scores, the bi-directional best hits of the unmatched compounds between KEGG and MetaCyc can be calculated. The compounds selected for comparison are by necessity compounds that had an assigned CID, for otherwise they would not be queryable via the PubChem Score Matrix Service. Two sets of CIDs are selected: one from the BioCyc deposition and the other from the KEGG deposition, with no members present in both sets.

Because the Tanimoto coefficient was scaled and rounded to integers between zero and one hundred, often there were several matching compounds with the highest matching value for a given query molecule. For that reason, our definition of a bi-directional best hit excluded KEGG compounds where there was ambiguity as to which compound in MetaCyc was the bi-directional best hit. See Figure 1 for the distribution of “all versus all” Tanimoto scores between unmatched compounds in MetaCyc and KEGG.

In future work, compounds that had ambiguity as to the true bi-directional best hit, and that have a Tanimoto similarity score of 100, will be further compared using an exact chemical sub-structure matching algorithm, to better distinguish between nearly-identical compounds.

3.2.4 Exact Synonym Matching

A method for matching compounds that doesn’t rely on structural information or database links is simply the compound names and synonyms. In this method we build a hash table that relates various different names and synonyms from MetaCyc compounds to the actual compound. We then iterate over all of the compound objects in the KEGG BioWarehouse dataset, and test the name and synonyms of the compounds against the hash table one at a time, collecting any resulting matches to MetaCyc compounds. If the names for the KEGG compound only map to one unique MetaCyc compound, then

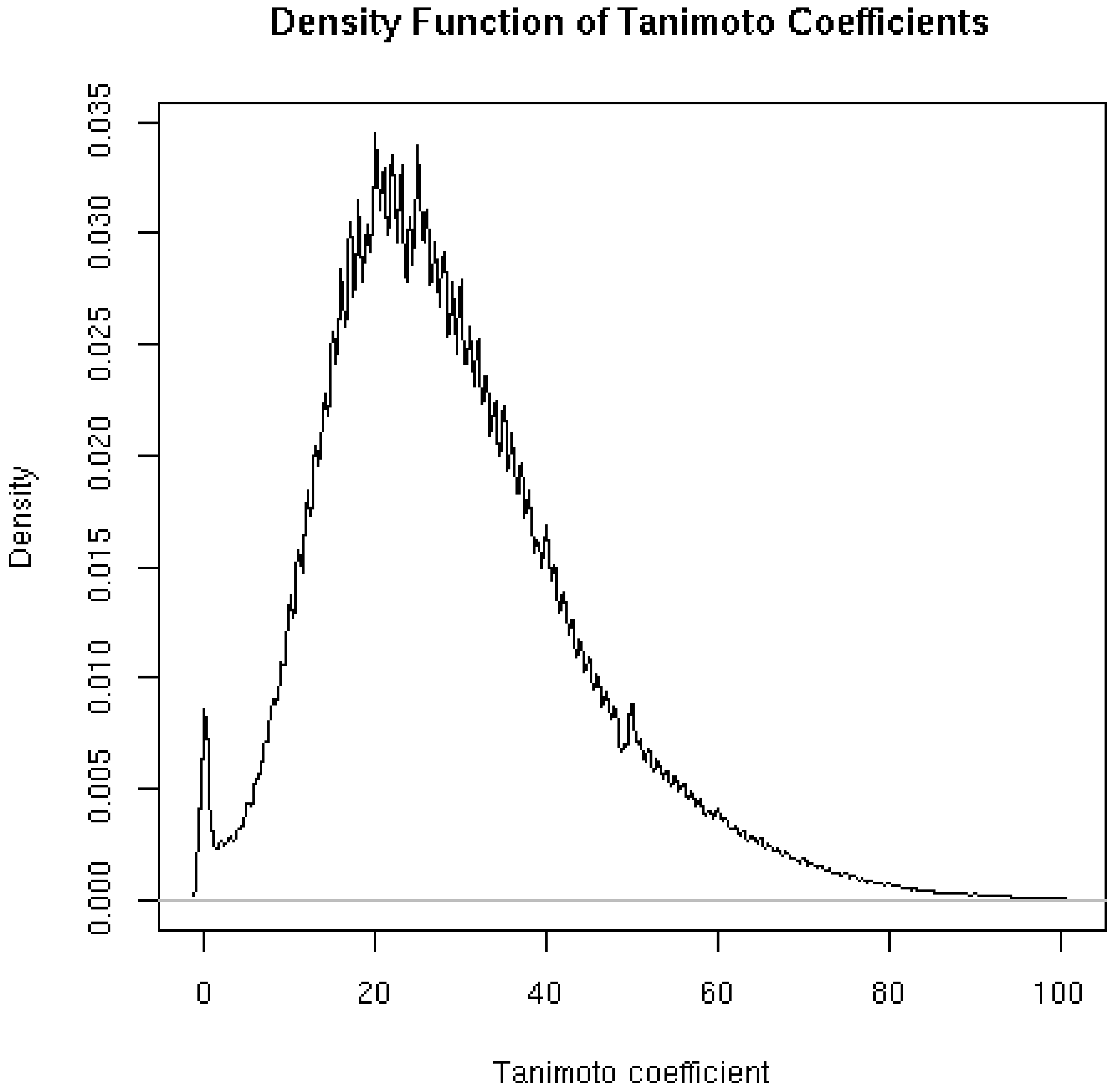

Figure 1: The distribution of Tanimoto similarity scores between KEGG and MetaCyc compounds not mapped via PubChem

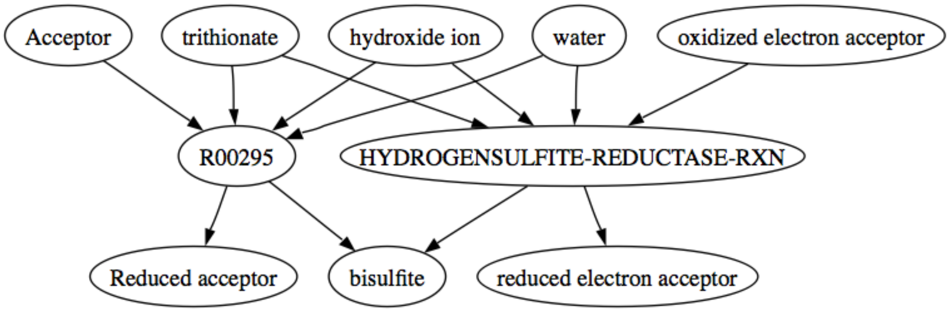
Acceptor
trithionate
hydroxide ion
water
oxidized electron acceptor
R00295
HYDROGENSULFITE-REDUCTASE-RXN
Reduced acceptor
bisulfite
reduced electron acceptor

Figure 2: The KEGG and MetaCyc depictions of EC reaction 1.8.99.3. Note that while many compounds are shared in common, two pairs of similar compounds are not mapped to one another.

it is recorded as a match.

3.2.5 Compound Identification by Reaction Alignment

Once two reactions are aligned, one can compare the two sets of compounds which participate in the reactions. In the case where all but one of the KEGG compounds in a given reaction were mapped to a MetaCyc reaction, and the MetaCyc reaction is also lacking a match for a compound on the same side of the reaction, we are able to use this information to infer a relationship between the two compounds (see Figure 2).

3.3 Reaction Match Prediction Methods

3.3.1 Bi-Directional Best Hit of Stoichiometry Vector Difference

The comparison of chemical reactions is much more complicated than that of a simple graph, since a reaction associates not only several chemical compounds, but does so with a particular orientation. This notion of a reaction can be thought of as a hypergraph arc.

In order to compare reactions to one another, we need a quantitative measure of how similar they are. One way to analyze a set of chemical reactions is to represent them in a stoichiometry matrix. A stoichiometry matrix is a matrix where each row represents a reaction, and each column represents a chemical compound that participates in the reaction. The numbers in the matrix represent the numerical coefficient for the corresponding compound and reaction, with a value of zero given to compounds that don't participate in a particular reaction. The distinction of compounds between the set of "reactants" and "products" is achieved by partitioning the coefficients into positive and negative values.

Once you have reactions represented as vectors, you can use standard techniques from linear algebra to determine the magnitude of the difference vector obtained from the two reaction vectors:

$$v_{diff} = \min \begin{cases} ||v_{f2} - v_{f1}|| \\ ||v_{f2} + v_{f1}|| \end{cases}$$

Where the norm is given by:

$$||v_f|| = \sqrt{\sum_i^n \left(\frac{c_i}{f_i}\right)^2}$$

Where n is the length of the vector, c_i is the stoichiometric coefficient of column i, and f_i is the frequency coefficient for compound c. The inclusion of the frequency of the occurrence of compound i in all of the row vectors of the matrix is used to ensure that the lack of a very common compound, such as water, does not unfairly bias the stoichiometry vector comparison.

3.3.2 Enzyme Commission Number Matching

The Enzyme Commission of IUBMB (EC) defines a hierarchy of enzymatic activities. For each number in the EC hierarchy, a unique identifier known as an EC number is assigned. EC numbers are in extensive use for characterizing the catalytic properties of proteins (and their associated genes). Along with giving a description of the enzymatic activity, the EC number in the EC hierarchy often includes a reaction equation. For this reason, many of the encyclopedic metabolic network databases include so-called EC reactions. Currently there are 4533 active EC numbers in the EC hierarchy, which provides a large common subset of reactions from which to align encyclopedic metabolic networks.

Both KEGG and MetaCyc have a large sub-graph in common which consists of the EC reactions. Alignment via EC numbers is straight-forward in the case where there is only one reaction in each database for a given EC number. In both databases there are EC numbers with more than one associated reaction.

3.3.3 Enzyme Mapping by KEGG Orthology and UniRef

Both KEGG and MetaCyc contain information about the enzymes that catalyze their reactions. More specifically, KEGG often has UniProt accession number database links from their Gene objects, and MetaCyc does as well from its protein objects. Finding a common UniProt accession number between KEGG and MetaCyc can aid in mapping together reactions that are catalyzed by the same protein sequence.

One problem with this approach is that MetaCyc only puts in a UniProt link that is representative of the space of sequences that are known to perform a particular catalytic function. Put another way, if there are twenty UniProt entries for a particular EC number, and they are all 100% identical, then MetaCyc will only contain a database link to one of the twenty sequence accession numbers.

A solution to this mapping problem can be achieved by making use of UniProt's UniRef database. This database contains sequence clusters at 50%, 90%, and 100% identity. By mapping a KEGG UniProt accession number and a MetaCyc UniProt accession number to the same UniRef cluster, especially in the case of UniRef 100, one can conclude that the two accession numbers are referring to the same sequence.

We justify the use of TrEMBL in addition to SwissProt sequence data in that we are using the UniProt interface to UniRef mainly to achieve the sequence redundancy reduction and mapping via UniRef clusters. We are relying on the fidelity of the KEGG and BioCyc external database links to UniProt in order to provide meaningful results.

3.3.4 Exact Synonym Matching

As with the compound data, an exact name matching approach was taken for mapping KEGG reactions to MetaCyc reactions. In contrast to the approach taken with the compound mapping, we gather name and synonyms not only for the reaction objects in the KEGG BioWarehouse dataset, but also from the associated enzymatic reaction objects and enzyme objects as well.

3.4 Initial Alignment

For the first step in the iteration, we wish to be more conservative with our predictions, for subsequent predictions will only amplify any errors made in

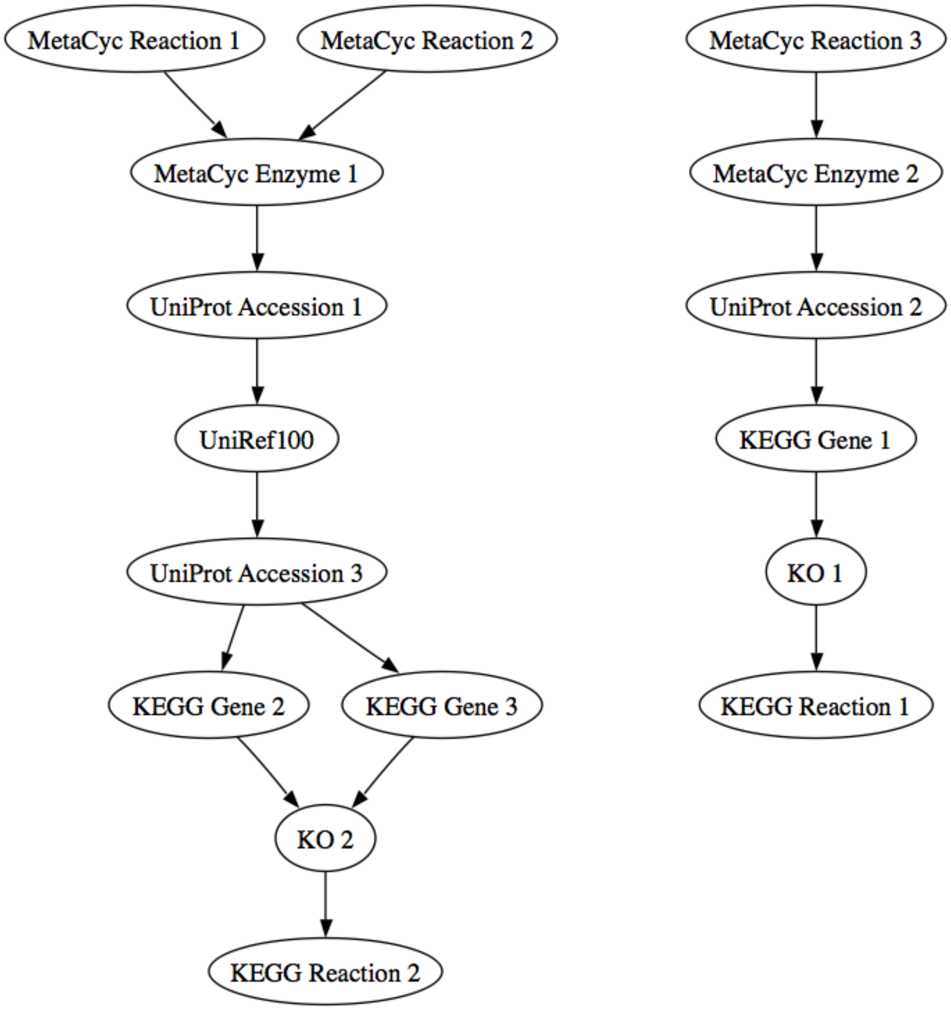

Figure 3: Mapping enzymes between MetaCyc and KEGG. The mapping on the right shows the case where there is a shared UniProt accession number between the two datasets. The mapping on the left shows how different UniProt accession numbers can be related to one another by common membership in a UniRef cluster.

the initial mapping. For each non-structural mapping approach, there should be corroborating evidence from a structural mapping approach. For example, if a compound has an exact string match, but the structures are completely different, then something is obviously wrong, and the match shouldn't be made. Similarly, if a map between reactions is made via EC numbers or UniRef, it should be validated using a stoichiometry vector comparison.

3.5 Iterative Alignment Refinement

Once the initial set of predicted compounds and reactions have been added to the metabolic network, we can iterate between compound predictors and reaction predictors until no new associations are predicted. Compound predictors such as the reaction alignment method described previously rely on reaction data, and reaction predictors such as the bi-directional best hit stoichiometry vector difference, rely on compound data. This inter-dependence allows for the method to iterate until no new predictions are made.

It must be stressed that due to known duplicate compounds and reactions, that any alignment method must be flexible enough to allow for such occurrences.

3.6 Alignment Score

The aspects of this comparison guide the selection of an appropriate set of metrics:

1. There is a large aligned sub-graph present in both the KEGG and BioCyc databases, as they both represent the full set of reactions described by the Enzyme Commission within their networks.

2. Due to differences in manual curation approaches and areas of interest, there is no assumption that both datasets have covered the breadth of human knowledge on metabolism to the same degree (beyond the common EC reaction core) in similar parts of the network. For that reason, we do not wish to penalize non-overlapping sub-networks.

3. Because the KEGG notion of a pathway is significantly different than the pathways in BioCyc, the comparison is done at the level of reactions and molecules.

4. To take stoichiometry into account, reactions are modeled as vectors in a stoichiometry matrix.

5. All of the molecules in the KEGG and BioCyc network have been compared using the molecular fingerprint Tanimoto similarity score available from PubChem.

With that said, the following scoring scheme was determined:

$$score = \sum_{i}^{C} t_i - \sum_{j}^{R} v_{diff}(j)$$

... where a compound match score t_i is 1 for exactly matched compounds, and otherwise the Tanimoto coefficient of the two molecular fingerprints as obtained from PubChem. $v_{diff}(j)$ is the vector difference between the matched reactions in KEGG and MetaCyc.

4 Conclusion

4.1 Comparison with Other Methods

Most of the literature on aligning graphs in a biological context is concerned with finding conserved motifs, whether for sequence data or for those of biochemical networks. In Berg et al. [12], they discuss a method they term "local graph alignment" for the detection of local motifs in gene regulatory and protein-protein interaction networks. Their method builds a statistical model of the motif from aligning the interaction networks from several experiments. In Fratkin et al. [13], they develop a method termed MotifCut for the detection of novel sequence motifs, which relies on determining the maximum density subgraph among all k-mers of a sequence.

These approaches differ from the work described above in that they are seeking to detect subtle patterns based on data analysis. For the alignment of large encyclopedic metabolic networks, the entity types are distinct as are the instances of both reactions and compounds. The essence of the problem is to disambiguate two differing representations of the same system.

4.2 Future Work

The work described here is currently in progress at SRI's Bioinformatics Research group under the direction of Dr. Peter D. Karp.

In addition to implementing all of the predictors described herein, there are further steps that can be taken to improve the quality of the alignment. Instead of strict exact string matching for names and synonyms of compounds, one might use the techniques from text-mining to cast the problem in the framework of Information Retrieval. Also, many of these predictors discard ambiguous matches, such as when the bi-directional best hit of the Tanimoto score for a compound matches more than one compound in the target database. A more refined algorithm would handle these situations as a special case of the general mapping problem. In the specific case of compounds with ambiguous bi-directional best hit Tanimoto scores of 100, a formal sub-structure matching algorithm can be applied.

4.3 Acknowledgements

This work was performed jointly for BioChemistry 218 and at SRI International in the Bioinformatics Research Group. Thank you to Dr. Peter D. Karp for supervising this work and providing guidance. Tom J. Lee extended the BioWarehouse KEGG Loader to parse the KEGG LIGAND REACTION data file. Thanks to Joseph M. Dale for insightful comments and feedback.

References

[1] P.D. Karp, I.M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S.M. Paley, I.T. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M.I. Penaloza-Spinola, C. Bonavides-Martinez, and J. Ingraham. *Multidimensional annotation of the Escherichia coli K-12 genome*. Nuc Acids Res, 35:757790, 2007.

[2] Rhea: EBI Annotated Reactions Database.
`http://www.ebi.ac.uk/rhea/`

[3] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yaman-

ishi, Y.; KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480-D484 (2008).

[4] NCBI BioSystems.
http://www.ncbi.nlm.nih.gov/biosystems/

[5] The BioCyc Database Collection.
http://www.biocyc.org

[6] MetaCyc User's Guide.
http://metacyc.org/MetaCycUserGuide.shtml

[7] MetaCyc Literature Curation Guidelines.
http://metacyc.org/MetaCycLiteratureCuration.shtml

[8] T.J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D.W.J Stringer-Calvert, J.D. Tenenbaum, and P.D. Karp. BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics, 7:170, 2006.

[9] PubChem. http://pubchem.ncbi.nlm.nih.gov

[10] Garey, M. R. and Johnson, D.S. Computers and Intractability: A Guide to the Theory of NP-Completeness. (1979) W.H. Freeman. ISBN 0-7167-1045-5. A1.4: GT48, pg.202.

[11] Pubchem Score Matrix service.
http://pubchem.ncbi.nlm.nih.gov/score_matrix/score_matrix.cgi

[12] Berg, J. and Lassig, M. Local graph alignment and motif search in biological networks. (2004) *Pro. Natl. Acad. Sci. USA* **101** 14689 - 14694.

[13] Fratkin, E., Naughton, B.T., Brutlag, D.L., and Batzoglu, S. MotifCut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics.* 22 (14) e150-e157.