Graph Analysis of Cellular Reaction Networks

Jocelyn Kishi, Li Gu, Jesse Salomon
Department of Computer Science
California Institute of Technology
Pasadena, California 91125
{jkishi, lggu, jsalomon} @caltech.edu

ABSTRACT

The cellular chemical reaction networks within living organisms are complex network structures that represent different molecular species in living cells and the relationships between them. These relationships are not always well-understood and can potentially hold a large amount of information for biological and medical research.

We investigated several automated approaches to researching these types of networks, drawing upon web graph and text analysis techniques that have proved their worth in other applications. Subsequently, we determined several potentially effective approaches for analyzing such chemical systems and identified next steps for this type of research.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks; J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Theory, Verification

Keywords

Chemical Reaction Network (CRN), Petri Net, PageRank, Hyperlink-Induced Topic Search, Term Frequency-Inverse Document Frequency (tf-idf)

1. INTRODUCTION

As modern scientific fields develop, computational analysis plays an increasingly prominent role in scientific research. Here at Caltech, biological concepts have always inspired novel engineering applications. However, such an interaction can go the other way. Through this project, we extended several graph analysis techniques to study biological processes so that we can better understand biological systems.

We took chemical reaction networks (CRNs) that model biological processes and represented them in different graph formats so that we could apply graph analysis to these networks. By analyzing the PageRank of each molecule and studying the hubs and authorities of the overall graph, we learned which molecules are the most important players in a cellular reaction network and which genes are the most prominent in causing a disease. We also analyzed the clustering relationships between groups and gained more insight in how species interact with one another.

In the last few weeks, we began to compare networks directly against each other and also to investigate the effects of the removal of a single reaction or species from a graph with perturbation analysis. Finally, we looked into incorporating term frequency-inverse document frequency (tf-idf) analysis into our results, in order to better identify the rare but important species in a network that sometimes end up lower on the PageRank and HITS rankings.

2. MOTIVATION / APPLICATIONS

2.1 Insight into the Inner-Workings of a Cell

By applying graph analysis techniques, such as modified PageRank and HITS algorithms, to the CRN graph representations, we can determine which species are the most vital for a cell's function. For example, with the HITS algorithm, a hub species could be one that's produced by a lot of authority species, and authority species could be a species that is a very important reactant (or perhaps enzyme).

We can also use clustering analysis to identify major pathways and shortest paths to identify how many and which reactions need to proceed until a desired species is produced or completely degraded. Finally, incorporating techniques used for text analysis, like tf-idf, we can adjust the rankings results to reflect the rarities of various species to produce a better measure of species importance within a chemical system.

2.2 Medicinal Applications: Disease Research and Potential Treatment

The results from this research can be applicable in many areas of medicine, especially the treatment of diseases. For example, if the CRN graphs turn out to have clustering regions that are generally distinguishable and not overlapping, this modular structure may make the results of an addition of a subnetwork of chemical reactions to treat a particular health condition more predictable.

Furthermore, we can study which genes or molecules are the most important in a disease network. With this information, we can determine which molecules are the biggest factors in the pathology of the disease or which genes are the most likely causes of a disease. This could mean that we can find new potential targets for treatment. This can be extremely useful for devising new ways to deal with diseases that are currently difficult to treat or for learning about diseases that we know little about.

Also, by analyzing components of a CRN, we can predict results of eliminating certain elements or blocking certain receptors. This will allow us to make preemptive guesses towards how some chemical components will react with the overall biological cell cluster. Many diseases are caused by certain subtle changes in cellular structure. By analyzing the CRN of such cells alongside normal cells, we can possibly evaluate the overall impact of small changes in the general reaction network.

Finally, a major benefit of using CRNs to analyze such properties of cells and diseases is that this method of analysis offers a new perspective alongside the traditional biological approaches. The calculation of PageRank, centrality and other graph properties is done via deterministic algorithms by computers. Therefore, this method of analysis is less affected by human perception and bias. Since previously much of the analysis was done by hand, the automation of this process saves a lot of time as well.

3. EXISTING RESEARCH

3.1 Theoretical Prior Research

There has been much theoretical analysis done on chemical reaction networks.

3.1.1 CRN Theory

Fleinberg's Chemical Reaction Network Theory is a classic example of analyzing chemical reaction networks without precise parameters. August and Barahona enhanced that method of analysis by extending the methodology to more quantitative aspects of biochemical reactions networks [5]. August and Barahona were able to show that since many biochemical reactions conserve mass and do not involve inflows or outflows, these graphs can be shown to be not divergent by simply analyzing the structure of the network graph. They achieved this by showing that such biochemical reaction networks have a bounded absorbing set in the case that mass is conserved. This analysis can be applied not only to weakly reversible chemical networks, but also to biological networks because the methodology can be shifted from stationary reactions to oscillatory networks.

3.1.2 Deficiency-Zero Stationary Distributions

Similar to August and Barahona above, Anderson, Craciun, and Kurtz based their research off Fleinberg's Deficiency Zero Chemical Reaction Network Theorem [7]. The theorem states that if a network satisfies the easily-verified deficiency-zero properties, then there is one equilibrium within each compatibility class and that equilibrium is asymptotically stable. This means that if a network satisfies the assumptions of the theorem, then the theorem completely applies to the dynamics of the network. Anderson et al took this theorem further by examining both deterministically and stochastically modeled chemical reaction systems and prove that there is a stationary distribution if the kinetics is general.

3.1.3 More Graph Theory

There have been other applications of graph theory to CRNs. For example, in "A Graph-Theoretic Analysis of Chemical Reaction Networks," Othmer laid out a method of producing a representative graph from a set of species and their stoichiometric ratios [28]. This method was then used to set up theoretical proofs of various properties on such graphs including a set of conditions which lead to the existence of a steady state. The proceeding of reactions over time can be viewed as flow through the system, and one can view the steady-state as the system state where the flow is balanced, which is when all time derivatives of concentrations are zero.

3.1.4 Reaction Network Topology

When modeling complicated dynamic chemical models, the molecular changes are not only restricted to basic physical properties. In any chemical reaction network, new molecules can be generated, which is very difficult to properly model. Flamm and Stadler studied several models of theoretical treatments of complex chemical networks and described the topology of such reaction networks using generalized closure functions. Dynamic system models trace the time-dependency of the amounts of different chemicals in a reaction network, which is similar to how the dynamics of genes in organisms are traced. To study such dynamic network topology, Flamm and Stadler studied the dynamic system [12] of Fontana and Buss' work on constructive dynamical systems [13], in which every interaction is an algebraic expression. They also look at computer-aided organic synthesis modeling techniques to study large networks, such as metabolic networks.

3.1.5 Inferring Meaningful Pathways

Since we decided to briefly analyze metabolism as a proof of concept, we needed to be able to find meaning pathways in a small metabolic cycle. Croes and Couche presents a way to use weighted graphs to infer important metabolic pathways [8]. We took a different approach in that we chose to analyze important species instead of inferring important pathways and we did not end up using weighted graphs. We decided that since PageRank of the web graph does not depend on weights and we are trying to model our analysis after that, we should not worry about weights.

3.1.6 Graph Theory and Biological Networks

Finally, Aittokallio and Schiwikowski further developed the actual idea of using actual graph theory to analyze biological networks instead of just chemical reaction networks [3]. Furthermore, they came up with some concrete ideas for how to analyze a biological model. Also, Aerts et al. talked about how to use graph analysis to identify important individual elements in a system using clustering analysis and some form of node ranking [24].

3.2 Prior Work on Actual Analysis

There have also been some prior work done on actual applied analysis of biological systems using graph models.

3.2.1 Medusa: Exploring Clusters

Aerts et al. explored the possibility of using Medusa to perform clustering analysis on biological models [16]. Medusa is a very powerful tool that is used to visualize biological models on a large scale and provides many layouts and means to explore clustering. In our project, we implemented two different measures of clustering analysis. However, we did not use any sort of visualization software to aid our analysis. This is partially due to the fact that none of our network is large enough to really warrant a standalone visualization software.

3.2.2 Centrality Analysis

In their paper, Koschutzki and Schreiber discuss using different centrality analysis methods to learn more about gene regulatory networks [21]. They explored several different centrality measures, such as degree centrality and closeness centrality. While centrality measures can be interesting and informative, we elected to not perform centrality analysis even though we work a lot with gene regulatory networks in diseases. A huge reason for this is the fact that most of databases provide inconsistent data in terms of connections and we would frequently encounter components that are disconnected or nodes that don't link to anything else. This is explained further in Section 5.

3.2.3 Genome Studies

Genome-wide association studies allows biologists to study complex genetic traits and how they are passed down. They also allow the study of mutations and how those manifest as traits. Akula et al. came up with a network-based approach to distinguish what kind of results are the most important or informative from such genome association studies [26]. While we do similar network-based analysis on genes, we focus more on specific networks and less on genomes for a whole species.

4. PROJECT DESIGN AND APPROACHES

4.1 Initial Data Collection

4.1.1 Data Sources

We collected data from a variety of databases and other sources. For human metabolism, we used the BioModels database and the Reactome project. The majority of our data, however, came from the KEGG database. It provided a variety of networks such as metabolism, cell cycle, and disease pathways. Finally, the Microsoft Research tool Visual GEC was used to convert manually inputted CRNs into SBML format.

4.1.2 Data Types

There are two primary data formats that we managed to obtain. The first is simple XML files that contain all the species/genes in a network and the reactions between the species. Many databases have their data in some form of XML. Many other databases have their files in the form of SBML (Systems Biology Markup Language), which is a fairly standardized format for computer modeling of biological processes. Not only does it establish a nice convention, many XML files can also be easily translated into SBML files through tools such as KEGGTranslator, which takes a XML file from KEGG and transforms it into a SBML file.

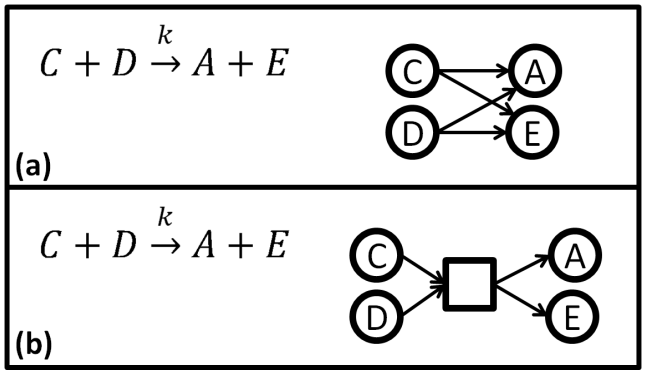

Figure 1: Graph Representations. (a) An example of how a bimolecular reaction is converted to the Reaction-to-Product (RTP) links type of graph format. There is a directed link from every reactant to every product for every reaction in the system. (b) The conversion of the same bimolecular reaction to the simplified Petri Net representation, which uses a special node for each reaction, with the reactants pointing into it and products pointed to from it.

There are many representations, or levels, of SBML, where subsequent levels have more information on reactions compared to the previous. We worked primarily with the level that shows all genes/species and reactions between them in simple graph representation.

4.2 Data Standardization

As mentioned above, our data came from a few different sources, each with their own special way of representing the data. In order for our project to be feasible on a large scale, it was necessary to design one standard format which we can convert all of our data into. The design we settled upon was to create a list of reactions from each graph, using one of two possible formats: RTP and Petri. Each list is stored in a tab-separated data file, where a single line represents an edge of our graph (in particular, each line has the starting vertex, then a tab, then the ending vertex). From there, we can then read in the list of edges from the stored data file into a Python dictionary, which is what all pieces of our analysis code then take as their primary input.

4.2.1 Standard Graph Types

Before converting our given data files into a standard format, we first had to determine what that standard format would be. In the end, we settled upon two different formats, and decided to use both of them, in order to see which would work better with our analysis.

4.2.1.1 RTP. The first form of graph representation we analyzed, reaction-to-product (RTP) links, has a directed edge from every reactant in a reaction to every product. An example is shown in Figure 1a, which shows the nodes and links involved in a single bimolecular reaction. Although it is a rather simplistic model, it does effectively capture the directional cause-effect relationship between a reactant and product. Another advantage of this approach is that each node's ranking can be directly compared to that of

every other, as every node in the graph is a different species in the network. Finally, species that act as catalysts in a reaction (and therefore don't get consumed) will have self-loops, since they will serve as both reactants and products in the same reaction. This makes it very straightforward to remove certain effects of catalysts from network analysis if desired, as all self-loops could just be ignored.

4.2.1.2 Petri Net. The other type of graph representation we investigated was a simplified Petri Net representation. This representation uses a special node for each reaction. All reactants link to the special node, which then points to all of the products. An example of this is in Figure 1b, which shows the Petri net conversion for the same reaction as used in Figure 1a.

In a more quantitative representation of a Petri Net, edges are weighted with the coefficients of the species for each reaction [20]. There are several reasons we chose not to include this additional information or rate constants on the graph. The primary reason is the sparsity of this data, as will be discussed in the Data Sparsity section below. However, we argue that it is not unreasonable to use an unweighted graph - the PageRank algorithm, for example, does not need to take into account number of page visits and link clicks to be effective in its analysis. Our CRN analysis may not need to take into account species concentration and rate constants to be similarly effective.

4.2.2 Parsing Data Files

When writing our parsing code, we had to make sure that we could read any given input file equally well (no matter what data type it was). Thus, when given a file, we would check its data type — when we have an XML file, we can run the XML parser, and when we have an SBML file, we can run the SBML parser.

4.2.2.1 XML. Our primary source of XML files was from the KEGG database, and all such files were formatted in roughly the same way. Each vertex of our graph, representing a compound in our system, was found in an "entry" tag, such as the one shown in Figure 2; the edges of our graph (representing connections between compounds) were found in "relation" tags, such as the one show in Figure 2.

When parsing a given XML file, we then simply looked for entry and relation tags, taking relevant information from each. The format of the XML file places all entry tags first, followed by all relation tags — we take advantage of this later on.

In the case of entry tags, we stored the entry ID value along with the "graphics name" of the vertex (that is, the second string associated with the vertex's name, rather than the first). We had originally used the first value for name, but eventually decided against it, as the second value was more descriptive (whereas the first value simply referred to KEGG's internal name for the compound in question).

In the case of relation tags, we first took the two entry ID numbers associated with each relation and translated them into their associated compound names (since, as mentioned previously, all entry tags will have been parsed before any relation tag is reached). This pair of compounds then represents an edge of our graph, and so we write it to the data file. In particular, when we are writing a data file for RTP graphs, we simply write the edge as is; when writing a data file for Petri Net graphs, we create multiple edges (via the intermediate node that each "true" reaction has) to be written.

4.2.2.2 SBML. As it turns out, parsing SBML files was much easier to do than parsing their XML counterparts. A Python library exists for reading and manipulating SBML files, which we used extensively at this stage of our code. Using the tools from that libary, we were simply able to read the input file as an SBML document, get the list of reactions from said document, and write the reactions to our output data file in our desired tab-separated format (modifying the list of reactions as necessary, depending on whether or not our output type was RTP or Petri).

4.2.3 Conversion to Python Dictionary

Our various methods of analysis all take as an input a Python dictionary, in which each key is a vertex of our graph and each value contains a collection of that vertex's connections (in particular, the value is actually a pair of lists stored as a tuple — the first list contains the vertices which point towards our given vertex, while the second list contains the vertices which our vertex points to). To create this dictionary, we may simply read our data file to get the list of edges which comprise our graph, since each data file is now standardized to actually just be a list of reactions (where each line is a single reaction and each reaction is a tab-separated pair of compounds). As we do so, we can construct entries in the dictionary for the given compounds of a reaction (if they don't exist already), and add to the dictionary the relevant connections which arise from each given edge.

4.3 Data Analysis

4.3.1 Analysis of One Graph

The first few types of analysis we perform are meant to be done on a single graph, and yield useful information about how some important compounds in a particular system as compared to other compounds in that system.

4.3.1.1 PageRank. PageRank is a link analysis algorithm created by Larry Page. Basically, PageRank assigns a number to a set of elements where each number represents how "important" the element is in the set. Although it is primarily used by the Google search engine, PageRank can also be used on any collection of elements that form a graph. In this case, we apply the algorithm to cellular reaction networks and calculate the numerical weights, or "rank" for individual elements in the CRNs and measure their relative importance. [31]

4.3.1.2 HITS. Another popular link analysis algorithm is HITS, or hyperlink-induced topic search. This algorithm

```
<entry id="28" name="hsa:10379" type="gene"
    link="http://www.kegg.jp/dbget-bin/www_bget?hsa:10379">
    <graphics name="IRF9, IRF-9, ISGF3, ISGF3G, p48" fgcolor="#000000" bgcolor="#BFFFBF"
        type="rectangle" x="676" y="844" width="46" height="17"/>
</entry>

<relation entry1="27" entry2="28" type="PPrel">
    <subtype name="binding/association" value="---"/>
</relation>
```

Figure 2: Example code from a KEGG database XML file.

was developed by Jon Kleinberg before the invention of PageRank. This algorithm focuses on two types of pages: hubs and authority pages. Hubs may not have accurate or authoritative information, but hosts (points to) a large set of pages which do have authoritative information. Authorities, as the name implies, are pages that may not be hubs, but hold valid and accurate information. This algorithm ranks web pages by their hub values and authority values. [31, 19]

4.3.2 Analysis of Many Graphs

While comparing compounds within a single system is useful, there is also much to be gained by comparing multiple graphs to one another, and determining the ways in which they differ.

4.3.2.1 Perturbation. Given two different graphs, we make a list of all vertices from the two, along with their PageRanks and the order of their PageRanks in their respective graphs. Note that if a node from one graph is not present in the other, we treat it as having PageRank 0 in that second graph. Then, we sort the list of vertices by the absolute value of the change in PageRank from one to the other and also list the percentile change of ranked-order.

4.3.2.2 TF-IDF. When rating text documents given a text query, the rarest words are often considered the most important, since they can give more insight into what differentiates one document or query from another. On the other hand, more common stop words, like "the" and "of", are not that useful. One sort of "rarity" measure of a word is the tf-idf, or term frequency — inverse document frequency, which is computed based on the frequency of a word in a given document and the frequency of that word in all documents. [29]

Extending this to our project, we use the analogy of individual species as specific words and different CRNs as separate documents. Thus, tf-idf measures could indicate the importance of a species in a particular CRN, based on how many reactions it occurs in in that CRN and how many times it appears in reactions in other CRNs in a set. It was our hope that the equivalent of "stop words" (i.e. species that are common in most of the networks), would no longer dominate the tops of the rankings once tf-idf analysis got incorporated, and differences between diseases could be more effectively highlighted. One thing to note about this analysis is that it doesn't utilize the specifics of the document's word structure (i.e. word order), so in our case the network structure of the CRN is not taken into account when evaluating a species in it. It will be interesting to compare the results of this analysis to the results of those that do utilize network structure.

Incorporating this structure-independent technique into our structurally-focused network analysis can allow us to negate the effects of species which may dominate the PageRank results simply by being involved in many reactions. Our results from this analysis are discussed in Section 6.

5. CHALLENGES

5.1 Data Standardization

Possibly the biggest challenge that we encountered had to do with getting data in different formats to work together and be able to analyze them side-by-side. There are an obscene number of databases online that provide networks and graphs in some form or another. However, there is zero guarantee that the different databases have the same data formats. Therefore, it took a lot of work just to get our analysis code to work with data of different formats.

Even within data of the same formats, the naming convention and the labels may be completely different. One database might store genes and molecules with a specific number that only matches to a name in their own local database. Obviously this would cause problems with perturbation analysis. Molecule KP-463 could be the exact same thing as molecule BG-568 in another database, but we wouldn't know that just from the data format that they provide. Therefore, we had to do some cross-referencing between databases and settle on a universal convention of names and standard molecule/gene database to extract the species names from.

As we previously mentioned, there are many levels of SBML representation. Therefore, we would have the problem where some database stores SBML in a level with less information than desired but display the graphs with no associated datafile. We even ran into this problem with some KEGG files where the actual XML or SBML file does not maintain a list of the connections between nodes, only the nodes themselves. However, they would have a picture of the graph in their database, just no file to reconstruct that graph. This was one of the more frustrating things to deal with and usually when that happened we just looked for a similar network

elsewhere. Of course, converting from a higher-level SBML to a lower-level one is not extremely tough, KEGGTranslator has that capability. The real problem was when we found SBML files in a lower level than we would have liked.

5.2 Pathway Integration

Often in our search for newer and better data, we would encounter a large network that we would like to analyze. Upon closer examination, we would then find that the large network incorporates several smaller sub-networks. However, different databases treat these sub-networks differently. Sometimes they are treated as a single node in the graph, while other times they are simply referred to but not actually connected to anything. Usually we managed to find the specific graph representations of those smaller networks as well and were able to analyze them separately. In the rare cases where they are unavailable, we simply ignored them, which might mean that some of the important but smaller changes were overshadowed in the process.

5.3 Data Sparsity

Another issue we encountered with the data was the inconsistency in information provided. Some data sources included information about species' concentrations and which compartments they were in, whereas others included neither. Furthermore, only a few of the sources had any information about rate constants and stoichiometries. This is one of the reasons we opted to not include this information in our graphs (e.g. as weighted edges in Petri Nets). Since species didn't always have a consistent set of quantifiable data about them available, it seemed more reasonable to only use metrics that were applicable to all of the networks we were analyzing. This meant that we could easily compare our results for different networks and not worry about the most appropriate way to incorporate this information (which usually wasn't available) into our graphs.

5.4 Graph Structure

In close relation to the non-standardized data issue, many of the datafiles that we can obtain have nodes that are not connected to anything. For example, some KEGG pathways would have nodes that appear to be connected on their graph representation, but the connections would not be there in the XML files. Therefore, we would have some nodes that are considered by KEGG to be not particularly important, but completely disappear when we reconstruct the graph using the XML files. However, we may want to still look at them for analyses, such as tf-idf. Furthermore, since closeness is only viable on connected graphs, we could not really incorporate closeness analysis into most of our research.

5.5 Coding Challenges

One member of our group was working primarily on Linux throughout the project while the other two members were using Windows. Since we used a version control program (Git), this should not have been a problem. However, the SBML library does not work well with Linux, which meant that the member using Linux could not actually run any of the code that he wrote. This meant that we spent a lot more time than we would have liked on debugging. The Windows vs Linux problem also came into focus when we needed Python to read and write filenames. Since file paths are formatted differently by Windows and Linux, we ran into some more issues with that discrepancy. Eventually we managed to fix all the bugs and get everything working, so this was a relatively minor setback.

5.6 Database Unavailable

Another issue that we encountered was that sometimes, the databases that we have come to rely on would just become unavailable. For a long time, the Reactome database would tell us that the Caltech mirror was down and then redirect us to a British mirror, which would then also fail to work. Obviously, there was nothing we could do about that, so we just had to wait it out and hope that it was not down forever. Fortunately, that was not the case and it went up in a week.

At one point, the KEGG database underwent some sort of renovation or database cleanup and the button that previously allowed us to download the files disappeared. We could still view the graph representations, but we had no way to download the files. We figured out a slightly hacky way eventually that allowed the KEGGTranslator software to not only save the SBML that it converted to, but also the XML file itself for us to parse. Two weeks later, the button that allowed us to download files came back and we had no further problems with that.

5.7 General Lack of Expertise

Clearly, our project integrates computer science with a lot of biology. This is somewhat problematic as our entire group consists of computer science majors. Therefore, it was difficult to decide on what kind of data would be most conclusive for us to analyze. Furthermore, even after we had the analysis complete, it was not the easiest for us to make sense of our results and determine what is reasonable. Fortunately, we were not completely unschooled in biology and most of us had some degree of experience with genes and molecules in a cell reaction. We also had a lot of outside help in the form of professors and other students. Finally, we made up any further lack of knowledge through extensive research and literature reading.

6. RESULT AND EVALUATION

6.1 Perturbation Analysis: Gene Regulation, a Toy Model

We considered several different metrics for perturbation analysis, and in the end we narrowed this to two. Consider species X with a PageRank value of PR_1 and rank R_1 in network 1 and PR_2 and rank R_2 in network 2. Then, the metrics we investigated were:

- Direct comparison: $\mathrm{PR}_1 - \mathrm{PR}_2$
- Proportional: PR_1 / PR_2
- Given some weighted distribution function $f(x)$ over the domain $[0, 1]$, calculate $|f(\mathrm{PR}_1)f(\mathrm{PR}_2)|$
- Jump in ranks: $|\mathrm{R}_1 - \mathrm{R}_2|$
- Change in percentile $|\%\mathrm{R}_2 - \%\mathrm{R}_1|$

As a proof of concept and as a tool for narrowing down our metrics, we used a toy model of gene repression. This model has four molecular players involved — G, M, P, and GP. Species G represents a gene, M the mRNA (which gets transcribed from the gene), P the protein (which is translated from the mRNA), and GP the protein bound to the gene (i.e. repression). In a sort of feedback mechanism, the produced protein can go back and repress the gene that originally produced it. This prevents the gene from continuing to transcribe more mRNA and therefore prevents more of the protein from being created until the protein then unbinds from the sequence. The transcription, translation, binding, and unbinding steps can be represented in the following chemical reaction network:

$$\begin{aligned} G &\rightarrow G + M && \text{(Transcription)} \\ M &\rightarrow M + P && \text{(Translation)} \\ M &\rightarrow \emptyset && \text{(mRNA degradation)} \\ P &\rightarrow \emptyset && \text{(Protein degradation)} \\ G + P &\rightarrow GP && \text{(Binding)} \\ GP &\rightarrow G + P && \text{(Unbinding)} \end{aligned}$$

The mRNA strands and protein will also degrade naturally at some rate in a cell, which is represented by M and P being converted to nothing in the two degradation reactions.(The reactions for this network come from [30]).

As a perturbation to this network, we decided to look at how PageRank results changed if the effects of repression (i.e. the binding and unbinding reactions) were removed. The removal of this medium of interaction dramatically changed the rankings - mRNA went from having the lowest PageRank value to the highest when repression effects were removed, and GP went from having the top PageRank value to the lowest (due to its non-existence) in the non-repression model. On a high level this makes sense, as one would expect the mRNA to play a more important role in the system when it is continually being produced and producing the protein, whereas its role would be more limited in a system where the protein it produces goes back and prevents more mRNA from being transcribed.

Through this process, we determined that the most useful metrics would be either a direct difference in PageRanks or the same difference with some distribution function applied. The difference in ranks between two networks is highly volatile and not very informative, and the proportional analysis results in a lot of 0's and ∞'s (which is not very useful), since not all species are likely to be present in both networks. In the end, we opted for two metrics: direct difference in PageRanks and difference in percentile, each of which reflects its own wealth of information.

6.2 Strong Hubs as Major Initiators

We found the top-scoring hubs returned by the HITS algorithm were often the species that initiate the process modeled by their networks, either by sensing the environment (as in the case of chemotaxis) or by being activated (as in apoptosis). Because these are such strong hubs, the next level of species (i.e. those triggered by the hubs) are generally highly ranked authorities. We investigated this behavior with three networks — bacterial chemotaxis, the human cell cycle, and programmed cell death.

6.2.1 *Bacterial Chemotaxis*

Bacteria sense chemical gradients of attractants and repellants and move their flagella (and therefore the entire organism) based on this sensing. This process, known as bacterial chemotaxis, is quite effective. Bacteria move in a sort of random walk motion, turning in random directions with their "tumbling" motions. If a larger amount of attractant is being detected by a bacterium than was detected in the past, then its tumbling frequency gets lowered, which means it is more likely to continue moving in the direction of the sensed increasing temporal gradient. Conversely, if more repellant is being seen, the tumbling frequency gets ramped up, increasing the likelihood that the organism will move in a different direction than the high concentration of repellant. Thus, the sensing of temporal gradients of chemicals can help to guide the organism in a more advantageous direction. [4]

The top-scoring hubs in the chemotaxis reference pathway in the Kegg database are the Aerotaxis receptor in first place, followed by the Methyl-accepting chemotaxis proteins (MCPs). These are the species involved in sensing the environment and initiating the process which will end in cell motility. The Aerotaxis receptor, referred to as Aer, senses inputs from within the cell. The four MCPs, MCPI, MCPII, MCPIII, and MCPIV are tied for second place hubs, and they are sensor receptors for extracellular serine, aspartate, ribose and galactose, and peptide, respectively. [15] The top-ranked authorities in the system are CheA (a protein kinase) and CheW (a protein), the two species directly linked to by the Aerotaxis receptor and MCPs. Thus, the species crucial in initiating and carrying out the first steps of flagellar movement based on temporal chemical gradients are those that are high-ranking hubs and authorities, which provides good evidence for the hypothesis.

6.2.2 *Human Cell Cycle*

In the case of the human cell cycle reference pathway in the Kegg database, the subunits of the origin recognition complex (ORC) all came in first place for top hubs, tied with the cell division control protein 45 (CDC45) and the S phase kinase activator (DBF4). The only species with any significant authorities values were MCM2, MCM3, MCM4, MCM5, MCM6, and MCM7, which are all components of the mini-chromosome maintenance complex. These results are in line with those expected - for DNA replication to be initiated, the first thing that must happen is the binding of the ORC to a strand; then the MCM complex can form around it. DBF4 and CDC45 also act upon the MCM components, which is why they also rank highly as hubs. The ORC and MCM are crucial pieces of the pre-replication complex, which, once fully formed, can initiate DNA replication through a series of other interactions. [22]

6.2.3 *Programmed Cell Death (Apoptosis)*

In the programmed cell death reference pathway for humans, the only species with a significant hub value is FADD, or the Fas-associated protein with death domain. This protein will be discussed in more depth in the cancer results below, but it is relevant to note here that it initiates the process of

programmed cell death [9], which matches the hypothesis regarding hub values proposed above. FADD bridges death receptors on cells with caspase-8 and that process starts cell death. The top authorities, caspase-8 and caspase-10, are cysteine proteases that are very important in cell apoptosis. [9].

6.3 Applying Text Analysis Techniques: TF-IDF Results

As we processed many of the networks, we noted that sometimes the PageRank results were dominated by species like water and proton. Although these species are clearly very important in many chemical systems, we realized that having just these very common species recognized by our analysis could be detrimental to the rarer, but crucial species in a network. This is why we turned to the tf-idf analysis — to find some way to reduce the elevated effects of extremely common species on rankings and identify the species that really differentiate one CRN from others.

The first group of networks we investigated with tf-idf analysis was a set of metabolic pathways taken from Kegg for photosynthesis: the citrate cycle, carbon fixation, oxidative phosphorylation, and glycolysis, among several others. Our focus was on species with low tf-idf values, which we would expect to be very common in the set of all networks analyzed. We found that several of these networks had aldehyde dehydrogenase and pyruvate synthase with low tf-idf values, which are both very common in many metabolic pathways. Although these species can play important roles in the processes they are involved in, their presence in a single network given their prevalence in others should be considered less important for comparative network analysis than, for example, a species that is only present in a single network. [23, 6] It is important to note that these are just initial results, and we would need to do more work to determine a suitable manner of incorporating tf-idf analysis into our network structure-based approaches. Nevertheless, it does seem to be a reasonable avenue to venture down.

6.4 Bacterial vs. Viral Infections

6.4.1 Interferon Regulatory Factors

Interferon regulatory factors (IRFs), as their name suggests, regulate the production and release of interferons in the human body. Interferons are primarily released by the host body when an immune response is triggered. There are many interferons, some of them can trigger macrophages and some of them can activate cytotoxic T-cells to kill off tumors [1], but all of them are antiviral agents [11]. In fact, they are called interferons because they interfere with virus replication by inhibiting the replication process or inducing infected cells to die through apoptosis. Therefore, it would be reasonable to assume that IRFs would have higher rank in viral pathways than in bacterial pathways.

Our analysis show that the PageRank's of IRF3 and IRF9 are consistently very high in viral infections such as measles, influenza, and hepatitis. At the same time, IRF ranks are not very high in bacterial infections.

6.4.2 JAK-STAT Signaling Pathway

The JAK-STAT (Janus kinase and Signal Transducer and Activator of Transcription) signaling pathway is a pathway that relays messages from the outside of the cell into gene promoters within the nucleus, which in turn causes DNA transcription [2]. DNA replication is a vital point in both healthy cell regulation and proliferation of viruses after they have infected cells in the host body. Bacteria, on the other hand, can replicate themselves and do not need to alter host replication processes.

Our results show that in viral infections, JAK and STAT are very much disturbed while they are relatively uneventful for bacterial infections. This also closely matches the results of the IRF analysis as IRF3 and IRF9 are the two primary regulatory factors that interact with the JAK-STAT pathway.

6.5 Generic Cancer Results

6.5.1 Fas-Associated Protein with Death Domain

Fas-associated protein with death domain (FADD) is referred to as an adaptor molecule. Its primary function is bridging death receptors to caspase-8. This linking forms a death domain, which in turn forms a death-inducing signaling complex and leads to cell death [9]. Apoptosis is key because cancer cells do not go through apoptosis and divide infinitely, which is what causes them to be dangerous. At the same time, caspase proteins, especially caspase-8, play a central role in cell apoptosis. Therefore, it stands to reason that caspase and FADD will have high ranks in analysis of cancer pathways.

Our analysis shows that caspase-8 and FADD are prominent in every cancer pathway that we managed to get our hands on.

6.5.2 MAPK

Mitogen-activated protein kinases (MAPK) is a family of kinases that respond to a wide range of stimuli to the cell such as osmotic changes, heat changes, and other proteins. The MAPK family governs very important cell functions, such as cell mitosis, proliferation, and apoptosis [14]. Again, the reason that tumors are so dangerous is because tumorous cells proliferate uncontrollably and do not go through apoptosis.

Our analysis shows that all cancer pathways that we found have multiple MAPK's near the top of the PageRanks, if not at the top. They are also very strong authorities for the HITS algorithm, which is reasonable because they respond to so much stimuli.

6.5.3 HRAS

The HRAS gene encodes for the production of transforming protein p21, which regulates cell division. They do this by activating in response to growth factors binding to cell membrane [27]. Irregular growth factor production is one of the effects of cancer and p21 activity is consequently raised in cancer pathways.

Our analysis shows that the HRAS gene plays a prominent role in several types of cancer, such as melanoma, glioma, and bladder cancer.

6.5.4 AKT3

AKT3 is a gene that encodes for the production of RAC-gamma serine/threonine-protein kinase. They respond to growth factors (similar to p21) and insulin production. These kinases are involved in cell proliferation and tumorigenesis [18]. Therefore, one would expect AKT3 to be very prominent in PageRanks.

Although our analysis did show that AKT3 is towards the top in many cancers, it is not the most prominent in that it never has top rank. It shows up as fairly high in endometrial cancer, glioma, and leukemia, but never at the top. However, they are very high hubs on the HITS algorithm. We were unable to find appropriate literature explaining this discrepancy. Therefore, it can either be just a side-effect of our analysis algorithm coupled with improperly represented data or it can be a new direction for future research.

As we mentioned, AKT3 also responds to insulin production. Therefore, we expect that it would have high ranks or be a very prominent hub for pancreatic cancer. We can only say that this is partially confirmed because it is near the top, but not quite there.

6.6 Specific Cancer Results and Possible Discoveries

6.6.1 RB1

Retinoblastoma protein 1 is a tumor suppression gene [25]. Obviously, it is dysfunctional in many cancers. Since it is a blastoma protein, we expect it to be prominent in skin cancers. Through analysis, we show that it is indeed prominent in melanoma, basal cell carcinoma, and interestingly, small cell lung cancer. The reason that a blastoma protein interacts so much with lung cancer is unclear to us, so it could be a direction for future research.

6.6.2 CCND1

Cyclin D1 is a protein encoded by the CCND1 gene. Through our analysis, we found that it is prominent in thyroid cancer and prostate cancer, but not in most other cancers. Through further research, we found that CCND1 is indeed associated with thyroid cancer [10]. In fact, CCND1 staining is an effective technique in detecting thyroid cancer. As to prostate cancer, we also found that CCND1 frequently interacts with androgen receptors, which are prominent in prostate cancer. Finally, research has shown that CCND1 is a target for colorectal cancer studies [10]. In correspondence with this, we discovered that CCND1 has very high PageRank in colorectal cancer.

6.6.3 VEGF

Vascular endothelial growth factor (VEGF) is an important growth factor for the vascular system and it is especially important for the development for the pancreas [17]. Since pancreatic cancer involves uncontrollable division of pancreatic cells, VEGF must play a big role in pancreatic cancer. We confirmed this through both PageRank and HITS analysis.

6.6.4 Renal-cell Carcinoma and EPAS1

One particularly interesting discovery we made is that EPAS1 is extremely high for both PageRank and hub score in renal-cell carcinoma. EPAS1 is a gene that encodes for transcription factor in oxygen-regulated genes. We were unable to locate any existing literature that discuss the role of EPAS1 in renal-cell carcinoma. This could mean that EPAS1 is simply a dead-end for research or that it has not been the target for research yet and could potentially be a crucial component to treating or researching renal-cell carcinoma.

7. FUTURE WORK

There are several things we would like to do if we have more time to work on this project that would improve our research. One important direction to take would be to try to find more genes and species that are important according to our analysis but have not been thoroughly researched, such as EPAS1 for renal-cell carcinoma. These kinds of new findings are what promotes our work from being a proof of concept to actually contributing. Of course, if we can get those results corroborated by actual biology research teams, it would be ideal.

Another direction we want to go in our analysis is to better deal with sub-networks or compartments. Right now we analyze small compartments individually or ignore them altogether. However, there can be really subtle changes in sub-networks that induce large changes in the overall network if integrated properly. Right now our analysis does not consider that possibility. Also, instead of looking at nodes on a graph, we can look at individual paths and add in considerations for edge weights. We also have not truly incorporated self-perturbation, which is when we compare a network to itself once we add or remove one or more genes/molecules. This can be especially important when determining the effects of changes such as inhibitors and enzymes being added to a reaction. Moreover, we have basic tf-idf analysis, which is not incorporated into network analysis. Given more time in the future, we would probably work that into our current analysis methods.

Finally, we can improve our code's readability and usability by creating a graphical user interface and better formatting our output. If our research is proven to be useful in aiding future research, we can create a distributable code package so other people can use our work.

8. ACKNOWLEDGEMENTS

We would like to thank Dr. Michael Hucka for providing us with invaluable information about prior work that has been done in the field, introducing us to the SBML data format, and assisting us in data acquisition; Professor Dianne Newman for suggesting some very useful databases for CRN data; and Professor Adam Wierman for his mentorship, constant encouragement, stimulating feedback, and ongoing support.

9. REFERENCES

[1] S. H. A. Takaoka and H. Yanai. Integration of interferon-alpha/beta signalling to p53 responses in tumour suppression and antiviral defence. *Nature*, 424(6948):516–523, 2003.

[2] D. Aaronson and C. Horvath. A road map for those

who don't know jak-stat. *Science*, 296(5573):1653–1655, May 2002.

[3] T. Aittokallio and B. Schwikowski. Graph-based methods for analysing networks in cell biology. *Briefing In Bioinformatics*, 7(3):243–255, May 2006.

[4] U. Alon. *An introduction to systems biology: design principles of biological circuits*, volume 10. Chapman & Hall/CRC, 2007.

[5] E. August and M. Barahona. Network analysis of biochemical reactions in complex environments. 2006.

[6] E. Chabriere, C. Cavazza, C. Contreras-Martel, and J. C. Fontecilla-Camps. *Pyruvate-Ferredoxin Oxidoreductase*. John Wiley & Sons, Ltd, 2011.

[7] G. C. David F. Anderson and T. G. Kurtz. Product-form stationary distributions for deficiency zero chemical reaction networks. *Annals of Applied Probability*, 2008.

[8] S. W. Didier Croes, Fabian Couche and J. van Helden. Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356:222–236, 2006.

[9] M. Eberstadt. Nmr structure and mutagenesis of the fadd (mort1) death-effector domain. *Nature*, 392(6679):941–945, 1998.

[10] J. B.-A. S. Elizabeth Musgrove, Elizabeth Caldon and R. Sutherland. Cyclin d as a therapeutic target in cancer. *Nature Reviews Cancer*, 11:558–572, August 2011.

[11] V. Fensterl. Interferons and viral infections. *BioFactors*, 35(1):14–20, 2009.

[12] C. Flamm and P. Stadler. Topology of chemical reaction networks. June 2003.

[13] W. Fontana and L. Buss. Arrival of the fittest. *Bulletin of Mathematical Biology*, 56(1):1–64.

[14] T. G.-B. X. K. B. G. Pearson, F. Robinson and M. Cobb. Mitogen-activated protein (map) kinase pathways: regulation and physiological functions. *Endocr. Rev.*, 22(3):153–183, April 2001.

[15] T. W. Grebe and J. Stock. Bacterial chemotaxis: the five sensors of a bacterium. *Current biology*, 8(5):R154–R157, 1998.

[16] A. S.-R. S. Jan Aerts, Sean Hooper and G. Pavlopoulos. Medusa: A tool for exploring and clustering biological networks. *BMC Research Notes*, 384(4), 2011.

[17] W. C.-M. G. B. J. C. John Ebos, Christina Lee and R. Kerbel. Accelerated metastasis after short-term treatment with a potent inhibitor of tumor angiogenesis. *Cancer Cell*, 15.

[18] D. T.-R. W. R. R. K. Nakatani, H. Sakaue. Identification of a human akt3 which contains the regulatory serine phosphorylation site. *Biochem Biophys Res Commun*, 257(3):906–910, June 1999.

[19] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[20] I. Koch. Petri nets–a mathematical formalism to analyze chemical reaction networks. *Molecular Informatics*, 29(12):838–843, 2010.

[21] D. Koschutzki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*.

[22] M. Lei and B. K. Tye. Initiating dna synthesis: from recruiting to activating the mcm complex. *Journal of Cell Science*, 114(8):1447–1454, 2001.

[23] S. A. Marchitti, C. Brocker, D. Stagos, and V. Vasiliou. Non-p450 aldehyde oxidizing enzymes: the aldehyde dehydrogenase superfamily. 2008.

[24] T. S.-S. K. J. A. R. S. G. P. Maria Secrier, Charalampos Moschpoulous and P. Bagos. Using graph theory to analyze biological networks. November 2010.

[25] A. Murphree and W. Bendict. Retinoblastoma: clues to human oncogenesis. *Science*, 223(4640):1028–1033, March 1984.

[26] D. S.-J. S. M. A. N. A. S. L. F. T. T. S. B. Y. S. C.-Y. J. K. J.-Y. L. B.-G. H. F. J. M. Nirmala Ancha, Akula Baranova. A network-based approach to prioritize results from genome-wide association studies. January 2011.

[27] T. T. S. T.-F. T. J. L. K. B. I. Y. E. H. M. T. P. L. K. K.-S. M. F. V. A. F. O. Ohta, Y. and S. J. K. H-ras ribozyme-mediated, alteration of the human melanoma phenotype. *Annals of the New York Academy of Sciences*, 716:242–256, 1994.

[28] Othmer. A graph-theoretic analysis of chemical reaction networks.

[29] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[30] M. Ullah and O. Wolkenhauer. Biochemical reaction networks. In *Stochastic Approaches for Systems Biology*, pages 23–52. Springer, 2011.

[31] W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE, 2004.