Analyzing Gene Relationships for Down Syndrome with Labeled Transition Graphs

Neha Rungta*, Hyrum Carroll*, Eric G Mercer*, Randall J. Roper†, Mark Clement*, Quinn Snell*
*Computer Science Department, Brigham Young University, Provo, UT, USA
Email: {neha, hdc, egm, clement, snell}@cs.byu.edu
†Department of Biology and Indiana University Center for Regenerative Biology and Medicine,
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA
Email: rjroper@iupui.edu

Abstract— **The relationship between changes in gene expression and physical characteristics associated with Down syndrome is not well understood. Chromosome 21 genes interact with non-chromosome 21 genes to produce Down syndrome characteristics. This indirect influence, however, is difficult to empirically define due to the number, size, and complexity of the involved gene regulatory networks. This work links chromosome 21 genes to non-chromosome 21 genes known to interact in a Down syndrome phenotype through a reachability analysis of labeled transition graphs extracted from published gene regulatory network databases. The analysis provides new relations in a recently discovered link between a specific gene and Down syndrome phenotype. This type of formal analysis helps scientists direct empirical studies to unravel chromosome 21 gene interactions with the hope for therapeutic intervention.**

I. Introduction

Researchers currently hypothesize that cancers and other physical ailments are caused by duplication of genes on chromosomes [1]. Chromosomes are made up of genes that encode information about physical characteristics or phenotypes. Through empirical analysis, systems biologists determine how genes are linked to certain phenotypes. Knowing how genes interact to express specific phenotypes is a critical step for therapeutic intervention in many diseases. The large number of genes found in an individual and the possible gene interactions make it difficult to manually discern how gene regulatory networks influence phenotypic characteristics.

Patients with Down syndrome (DS) have an extra copy of chromosome 21 (chr. 21) and have phenotypes like abnormal brain and facial features as well as mental retardation. DS results from three copies of approximately 350 chr. 21 genes and has many well defined phenotypes that make it an excellent model to understand gene regulatory networks. By understanding changes that occur in this model, gene regulatory networks in other diseases that are not as well-defined genetically or phenotypically can be better understood.

Phenotypes are generated by interactions between genes that are described by gene regulatory networks. Each gene regulatory network is a complex feedback circuit that is constructed through large, costly, and time consuming empirical studies. The regulatory networks once understood, however, provide a wealth of information. For example, pharmaceutical researchers use gene regulatory network diagrams to design compounds that inhibit the expression of certain genes. In other words, pharmaceutical researchers are designing drugs to target genes involved in specific medical conditions; the drugs essentially manipulate the gene regulatory networks to produce a desired outcome.

DS researchers and systems biologists believe that changes in expression levels of one or more genes on the extra copy of chr. 21 lead to specific DS phenotypes. Even though DS researchers and bioinformaticians have documented elevated expression levels of chr. 21 genes in DS, they have been unable, for the most part, to directly link a specific chr. 21 gene to a particular phenotype.

Research in human and mouse models shows that there are modifier genes, not found on chr. 21, that contribute to DS phenotypes [2], [3]. For example, the *Sonic hedgehog* gene (SHH) in the Hedgehog signaling pathway has been directly linked to a DS brain phenotype [2]. It is, however, unknown how the SHH gene is linked to chr. 21 genes. The work in this paper helps researchers focus on specific modifier and chr. 21 genes for empirical studies.

Researchers often relate chr. 21 genes to modifier genes by manually examining published gene regulatory network databases. These databases, [4]–[6], are exceptionally complex and regularly evolve as new data is discovered through empirical studies. Manual extraction of indirect connections between chr. 21 and modifier genes is not feasible in such a large, concurrent, and connected system. An automated formal analysis is required to assist researchers in understanding gene regulatory networks important in DS.

We use a formal approach to analyze gene regulatory networks that are mined from different biological databases. We build a single labeled transition graph from these databases by defining a way of connecting the gene regulatory networks. We perform an exhaustive reachability analysis on the graph using a randomized breadth-first search (BFS). Randomized BFS gives a sample of shortest path connections between modifier genes and chr. 21 genes. The random paths are tabulated and graphed to show researchers the names and frequencies of genes found in these paths. We hypothesize that genes frequently found in paths connecting modifier and chr. 21 genes are more likely to be involved in a specific phenotype. Our results demonstrate a possible relationship between a nonchr. 21 modifier gene and chr. 21 genes. These relationships help DS researchers to direct resources for future empirical studies.

The principle contributions of this work are: (1) a biologically feasible technique connecting different gene regulatory networks into a single labeled transition graph suitable for formal analysis; (2) a reachability analysis using randomized BFS to generate different traces between chr. 21 genes and modifier genes; and (3) tabulated results showing potential interactions between specific chr. 21 genes and the SHH modifier gene.

II. Related Work

Several databases store pictorial representations of empirically curated gene regulatory networks [4]–[6]. Among all the databases, KEGG [4], currently has the largest amount of data in the most comprehensible and accessible format. It provides 175 pathways with over 12,000 genes from the human genome. Many of its pathways include both a pictorial and XML representation; although, the XML descriptors often have fewer defined interactions than those defined in the pictorial descriptors. Another source of gene interactions is found in a PubMed abstracts database [7]. PubMed is a premier journal index for published bio-medical articles. Existing research, [7], extracts gene relationships from PubMed using natural language processing algorithms. The work in this paper uses the KEGG database and the PubMed extraction to build the labeled transition graph of the gene regulatory networks.

State of the art pathway analyses tools such as `Cytoscape` [8], `Reactome` [9], and `Bind` [10] visualize gene regulatory networks. They provide some basic query mechanisms to probe structure in the regulatory networks. The queries, however, are simple and do not provide an ability to derive indirect relationships between modifier and chr. 21 genes in a fully automated manner.

Other formal verification approaches that analyze regulatory networks consider only a single network or pathway [11]–[14]. A single network is modeled in isolation to predict pathway behavior based on the gene interaction rates. The work in this paper abstracts many details of individual regulatory networks to consider the interactions of the complete system of regulatory networks. Where existing research, [11]–[14], tries to understand intra-network interactions, we extend the analysis to inter-network interactions in expressing a specific phenotype.

III. Graph Construction

We build a labeled transition graph from the KEGG and PubMed databases. The process follows three steps: first, we abstract the reactions and compounds in gene regulatory networks in the KEGG database to create intra-pathway gene interactions; second, we use gene interactions to create inter-pathway connections; and third, we add PubMed interactions to the inter-pathway and intra-pathway connections. Adding the gene-to-gene connections between individual pathways essentially flattens the KEGG database in step two. The final labeled transition graph is an over approximation of the actual biological system. The process is illustrated in more detail with a simple example.

Fig. 1(a) shows parts of gene regulatory networks and metabolic pathways: the Hedgehog signaling pathway (`HSP`), Basal cell carcinoma (`BCC`), and Alzheimer's disease (`AD`) regulatory networks. The rectangle boxes in each pathway represent the genes in the regulatory networks. For example, the Hedgehog signaling pathway shown in Fig. 1(a) contains the genes SHH, PTCH, GLI and WNT1. An edge between any two genes is a direct or indirect interaction between the genes. We maintain scalability by abstracting away most of the details in the actual biological system to focus purely on gene interactions—direct and indirect. Results suggest that the abstraction retains enough information to be meaningful. Fig. 1(b) is the final labeled transition graph for the pathways and PubMed relations shown in Fig. 1(a).

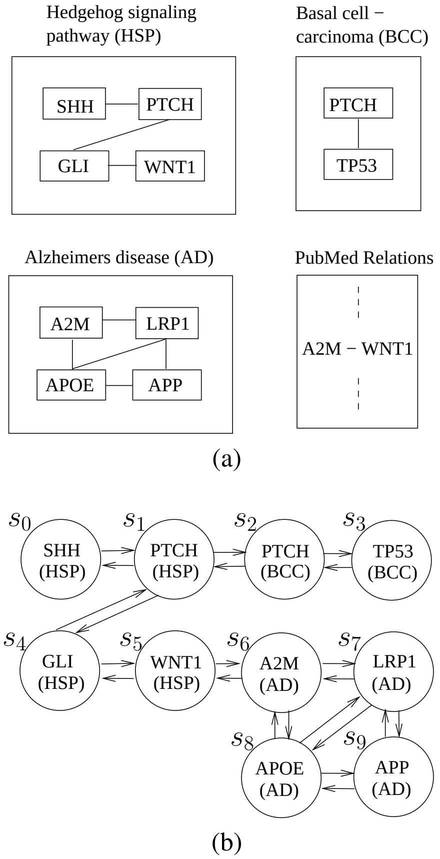
The abstract pathways form a set of intra-network gene interactions. Each graph is a separate abstracted regulatory network. For example, the abstract graph of the Hedgehog signaling pathway contributes s_0, s_1, s_4, and s_5 nodes to the transition graph in Fig. 1(b). The intra-network connections between the genes in the Hedgehog signaling pathway are maintained in the transformation. The next step connects the abstract networks to form an inter-network graph. The separate abstract networks are connected by creating a set of nodes labeled by both the gene and the network owning the gene. Each node contains the gene label and the owning pathway label. Once the set of nodes is known, then edges are created between nodes that contain common gene labels. For example, Fig. 1(a) shows that both HSP and BCC contain the PTCH gene. As such, Fig. 1(b) connects state s_1 to state s_2 showing a relation between the HSP and BCC pathways through the common PTCH gene. The two different nodes labeled with PTCH gene for the HSP and BCC networks enable us to easily detect that PTCH is the gene which connects the HSP and BCC networks. Connecting regulatory networks in this way essentially flattens the KEGG database through gene interactions.

The final step in building the inter-network graph augments the XML data in the KEGG database with known interactions published in the PubMed database using the work in [7]. In essence, any gene pair related in PubMed is also related in the inter-network graph. For example, the bottom right member of Fig. 1(a) shows a relation defined in the PubMed database; it connects the gene A2M to WNTI. The relation is expressed in Fig. 1(b) by the edge between states s_5 and s_6.

IV. Analysis

DS researchers are interested in finding the gene interactions between modifier genes (e.g., SHH gene) and chr. 21 genes. We define a randomized BFS to find shortest path traces between modifier genes and chr. 21 genes.

A regular BFS enumerates all nodes reachable in one-step from the initial node before enumerating nodes reachable in two-steps. This feature guarantees that the BFS finds the



(a)

(b)

Fig. 1. Converting a set of gene regulatory pathways into a labeled transition graph. (a) KEGG pathways and PubMed Relations. (b) The corresponding labeled transition graph.

shortest distance between the start node and another node of interest. The search, however, deterministically generates the same shortest trace between the two nodes in every single trial. For example, the shortest distance between nodes s_6 and s_9 is two steps in Fig. 1(b). The successors of node s_6 are sorted in some default order. If the successors are lexicographically sorted where node s_7 is inserted in the queue before s_8, the BFS generates the trace $s_6 \rightarrow s_7 \rightarrow s_9$ and does not generate the trace $s_6 \rightarrow s_8 \rightarrow s_9$.

The labeled transition graph derived from the gene regulatory networks database is a highly connected graph. It has approximately 12,000 nodes and one million edges. About 90% of the connections are a result of the abstraction applied to the pathways while creating the inter-network connections. The high degree of connectivity creates a large number of unique shortest paths between two nodes in the graph. There are different algorithms to find the k shortest paths between two nodes in a graph such as the one presented by Eppstein in [15]; however, to randomly sample different shortest traces we use a randomized BFS for its algorithmic simplicity and low complexity. Generating $s_6 \rightarrow s_7 \rightarrow s_9$ is equally likely as generating $s_6 \rightarrow s_8 \rightarrow s_9$ in any trial of randomized BFS.

A randomized BFS generates a subset of the shortest paths between modifier and chr. 21 genes. The pseudo-code for a randomized BFS, [16], is presented in Fig. 2. The algorithm is a variant of the regular BFS that randomizes the order of successors before inserting them into a queue. Randomized BFS generates shortest traces of all the gene interactions from the initial state to nodes that represent chr. 21 genes (is_chr_21_gene). During the search, when we encounter a node, s, labeled with a chr. 21 gene (line 5) we report the path from the initial state to the node, s (line 6). We then continue searching for paths connecting the initial state to nodes labeled with other chr. 21 genes. Note that we maintain a set of visited nodes and never visit the same node twice (lines $10-12$). If there is more than one shortest path from the initial state to a node with a particular chr. 21 gene, each trial of randomized BFS generates a subset of those traces.

```
procedure Random_BFS($s_0$)
 1: $Visited := \{s_0\}$
 2: $Queue$.enqueue($s_0$)
 3: while $(Queue) \neq \emptyset$ do
 4:     $s := Queue$.dequeue
 5:     if is_chr_21_gene($s$) then
 6:         print "$Report\ Path$"
 7:     $X_{succ}$ := get_successors($s$)
 8:     $s$ := randomize_elements($X_{succ}$)
 9:     for each $s' \in X_{succ}$ do
10:         if $s' \notin Visited$ then
11:             $Visited := Visited \cup \{s'\}$
12:             $Queue$.enqueue($s'$)
```

Fig. 2. Pseudo-code for randomized breadth-first search.

V. Results

The analysis in this paper is designed to answer the following research questions: (a) How many chr. 21 genes are connected to a modifier gene? (b) What is the length of the shortest path between a chr. 21 gene and the modifier gene? and (c) What are the gene interactions between a chr. 21 gene and the modifier gene? We summarize the analysis for the SHH modifier gene and chr. 21 genes in the following paragraphs.

To answer the research questions (a) and (b), we run a single trial of randomized BFS starting from the SHH modifier gene in the Hedgehog signaling pathway. During the search, whenever we encounter a chr. 21 gene, we mark the chr. 21 gene as reachable from the SHH modifier gene and note the length of the trace. Recall that a randomized BFS returns the length of the shortest trace from the SHH gene to the particular chr. 21 gene.

Randomized BFS finds 38 chr. 21 genes that are between 3 to 7 steps away from the SHH gene in the labeled transition graph. Among the reachable chr. 21 genes, there are genes that play a significant role in Alzheimer's disease and other cancers. This result is especially interesting to researchers because virtually all individuals with DS have indicators of Alzheimer's disease by 40 years of age.

We run several trials of randomized BFS to answer research question (c). The trials result in a large number of traces

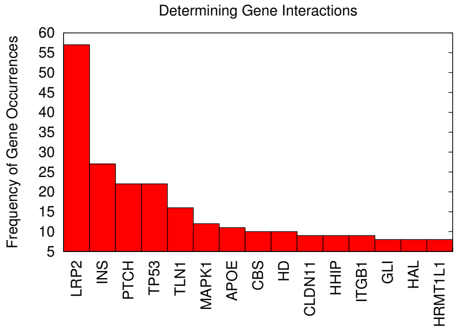


Fig. 3. Gene occurrences in SSH modifier and chr. 21 genes traces.

between the SHH gene and chr. 21 genes. The traces aid us in analyzing the different gene interactions that lead from the SHH gene to chr. 21 genes. For example, consider two arbitrary traces of gene interactions: $a \rightarrow b \rightarrow c_1$ and $a \rightarrow b \rightarrow e \rightarrow c_2$. Since gene b occurs in both traces it is more likely to affect a certain phenotype. In essence, to answer research question (c), we count the number of times a gene uniquely occurs in multiple traces.

In Fig. 3, for an arbitrary randomized BFS trial, we count the total occurrences of genes found in traces between the SHH modifier gene and chr. 21 genes. The results in Fig. 3 are exciting to biologists due to the high occurrence frequency of LRP2, INS, and MAPK1 genes. These genes have not been previously considered by biologists in the expression of DS phenotypes. The results in Fig. 3 open new avenues for empirical research by providing a list of genes closely related to SHH modifier and chr. 21 genes.

Interestingly, TP53 (involved in cancer) and APOE genes have been previously linked to DS phenotypes by researchers. The fact that our analysis relates TP53 and APOE as frequently occurring in traces between SHH modifier and chr. 21 genes provides anecdotal validation to our results. It gives hope that the other genes, such as LRP2, discovered in this analysis might affect DS phenotypes. Furthermore, genes like PTCH and GLI are part of the Hedgehog signaling pathway and are expected to have a high occurrence in the traces as seen in Fig. 3. This provides further anecdotal evidence that our analysis is biologically sound.

VI. Threats to Validity

The gene regulatory networks in the databases are not currently tagged with phenotype or developmental stage information. This causes the labeled transition graph built from the gene regulatory networks to be an over-approximation of the system. The analysis, however, is aimed to provide biologists a set of genes possibly involved in DS phenotypes. A small set of genes makes it feasible to empirically determine whether the gene affects a DS phenotype. Empirical studies filter any false positives generated during the analysis.

VII. Conclusion and Future Work

This work defines a technique to build a labeled transition graph from different biological databases. On this graph we perform a reachability analysis using randomized BFS to find gene interactions between modifier genes and chr. 21 genes. The analysis for the SHH modifier gene and chr. 21 genes gives an interesting set of genes for designing empirical studies. The same analysis can be used for determining gene interactions in various modifier genes for DS and other ailments. The representation of the different gene regulatory pathways as single labeled transition graphs lends itself to a more refined analysis. As future work, biologically interesting questions can be posed in temporal logic to find traces that satisfy the property.

References

[1] R. Redon *et al.*, "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, 2006.

[2] R. J. Roper, L. L. Baxter, N. G. Saran, D. K. Klinedinst, P. A. Beachy, and R. H. Reeves, "Defective cerebellar response to mitogenic Hedgehog signaling in Down's syndrome mice," *Proc. Natl Acad Sci*, vol. 103, no. 5, 2006.

[3] J. R. Arron *et al.*, "NFAT dysregulation by increased dosage of DSCR1 and DYRK1A on chromosome 21," *Nature*, vol. 441, no. 7093, pp. 595–600, 2006.

[4] "KEGG database." [Online]. Available: http://www.genome.jp/kegg

[5] "Biocyc." [Online]. Available: http://biocyc.org

[6] "Metacyc encyclopedia of metabolic pathways." [Online]. Available: http://metacyc.org

[7] R. Bunescu, R. Ge, R. Kate, E. Marcotte, R. Mooney, A. Ramani, and Y. Wong, "Learning to extract proteins and their interactions from Medline abstracts," in *ICML-2003 Workshop on Machine Learning in Bioinformatics*, August 2003, pp. 46–53.

[8] P. Shannon *et al.*, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, 2003.

[9] "Reactome - a curated knowledgebase of biological pathways." [Online]. Available: http://www.reactome.org/cgi-bin/frontpage

[10] G. Bader, I. Donaldson, C. Wolting, B. Ouellette, T. Pawson, and C. Hogue, "Bind–the biomolecular interaction network database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 242–245, 2001.

[11] J. Heath, M. Kwiatkowska, G. Norman, D. Parker, and O. Tymchyshyn, "Probabilistic model checking of complex biological pathways," in *Proc. Computational Methods in Systems Biology (CMSB'06)*, ser. Lecture Notes in Bioinformatics, C. Priami, Ed., vol. 4210. Springer Verlag, 2006, pp. 32–47.

[12] H. Kuwahara, C. J. Myers, M. S. Samoilov, N. A. Barker, and A. P. Arkin, "Automated abstraction methodology for genetic regulatory networks," *Transactions on Computational Systems Biology*, vol. 4220, pp. 150–175, 2006.

[13] M. Kwiatkowska, G. Norman, D. Parker, O. Tymchyshyn, J. Heath, and E. Gaffney, "Simulation and verification for computational modelling of signalling pathways," in *WSC '06: Proceedings of the 38th conference on Winter simulation*. Winter Simulation Conference, 2006, pp. 1666–1674.

[14] D. L. Dill, M. A. Knapp, P. Gage, C. Talcott, P. Lincoln, and K. Laderoute, "The pathalyzer: a tool for visualization and analysis of signal transduction pathways," in *First Annual RECOMB Satellite Workshop on Systems Biology*, ser. Lecture Notes in Bioinformatics. Springer, 2005.

[15] D. Eppstein, "Finding the k shortest paths," in *IEEE Symposium on Foundations of Computer Science*, 1994, pp. 154–165. [Online]. Available: citeseer.ist.psu.edu/eppstein97finding.html

[16] M. B. Dwyer, S. Person, and S. Elbaum, "Controlling factors in evaluating path-sensitive error detection techniques," in *SIGSOFT '06/FSE-14: Proceedings of the 14th ACM SIGSOFT international symposium on Foundations of software engineering*. New York, NY, USA: ACM Press, 2006, pp. 92–104.