Bicliques in Graphs with Correlated Edges: From Artificial to Biological Networks

Aaron Kershenbaum[1], Alicia Cutillo[1], Christian Darabos[1,3], Keitha Murray[2], Robert Schiaffino[2], and Jason H. Moore[1(✉)]

[1] Institute for Biomedical Informatics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
jhmoore@upenn.edu
[2] Department of Computer Science, Iona College, New Rochelle, NY 10801, USA
[3] Research Computing, Dartmouth College, Hanover, NH 03755, USA

Abstract. Networks representing complex biological interactions are often very intricate and rely on algorithmic tools for thorough quantitative analysis. In bi-layered graphs, identifying subgraphs of potential biological meaning relies on identifying bicliques between two sets of associated nodes, or variables – for example, diseases and genetic variants. Researchers have developed multiple approaches for forming bicliques and it is important to understand the features of these models and their applicability to real-life problems. We introduce a novel algorithm specifically designed for finding maximal bicliques in large datasets. In this study, we applied this algorithm to a variety of networks, including artificially generated networks as well as biological networks based on phenotype-genotype and phenotype-pathway interactions. We analyzed performance with respect to network features including density, node degree distribution, and correlation between nodes, with density being the major contributor to computational complexity. We also examined sample bicliques and postulate that these bicliques could be useful in elucidating the genetic and biological underpinnings of shared disease etiologies and in guiding hypothesis generation. Moving forward, we propose additional features, such as weighted edges between nodes, that could enhance our study of biological networks.

1 Introduction

Relationships of many types can be modeled as graphs, where the related entities are the nodes of the graph and the relationships between them are the edges. In order to make the model more complete and solutions more realistic, properties relevant to the specific problem can be associated with the nodes and edges, transforming the graph into a network [20]. One type of problem which has received a great deal of interest is to identify sets of nodes which are closely related, e.g., diseases with common etiologies. In the context of graph theory, this involves a search for collections of nodes with many connections between them. In the ideal case, edges between all pairs of nodes in the subset would exist. Such graphs are called cliques.

G. Squillero and P. Burelli (Eds.): EvoApplications 2016, Part I, LNCS 9597, pp. 138–155, 2016.
DOI: 10.1007/978-3-319-31204-0_10

It is often true that the nodes are of two distinct types, say A and B, and that the edges can be drawn between nodes of one type and the other [19]. For example, we might be interested in relationships between diseases and genes. The problem now becomes one of finding a collection of nodes in A that are all connected to all nodes in B. These nodes form bipartite graphs, or simply bigraphs, and are known as bicliques [11]. They may represent significant relationships between the two sets of nodes. A maximal biclique is one that is not contained in any other one; i.e., at least one of its component sets is not properly contained in the corresponding set in any other biclique.

We present an algorithm for solving the problem of finding all maximum bicliques of a bigraph. We discuss the algorithm's runtime and memory requirements as a function not only of the number of nodes and edges in the network, but also of the density of the graph (fraction of edges present), the distribution of node degrees, and the correlation among the edges.

We apply our algorithm to the analysis of two Human Phenotype Networks (HPN) [7], bigraphs models of human phenotype interactions based on shared biology. The first uses biological pathways, and the second genes, to identify common etiology in disease phenotypes. Through the use of Genome-Wide Association Study (GWAS), we have amassed a wealth of knowledge linking genetic variants to human disease phenotypes. However, our understanding remains limited by the complexity of human disease and by the epistatic, polygenetic and pleiotropic effects of genotype-phenotype interactions. Bicliques and other phenotype networking tools allow for a systematic examination of shared biology between diseases. These tools can be used to (1) identify diseases previously believed to be unrelated, (2) pinpoint genetic or biochemical pathway associations that underlie common disease etiologies and (3) identify therapeutic targets that may be applicable to multiple diseases.

2 Background

2.1 Bicliques in Research

Due to the large number of applications they model, bigraphs, bicliques, quasi-bicliques, and maximal bicliques have been studied extensively in the context of human genotype-to-phenotype relationships. Cheng and Church [4] introduced the concepts of "biclustering" both genes and conditions and illustrated their technique by identifying co-regulation patterns in yeast and human gene expression data. Tanay *et al.* [26], Wang *et al.* [27], and Liu *et al.* [15] address the same problem using bipartite graphs to model relationships between genes and conditions and identified bicliques in the bipartite graphs. Sanderson *et al.* [24] use bipartite graphs and bicliques in phylogenetics to improve the accuracy of tree construction. Zhang *et al.* [28] use maximal bicliques to identify groups of genes that are associated with related biological functions.

With the generalizability of these tools, many other applications exist both inside and outside of biology. Bicliques can be applied to the study of other factors influencing human disease, such as environmental exposures, or to the

study various other biological networks. For instance, Maulik *et al.* [18] use quasi-bicliques to study viral-host protein-protein interaction networks. Liu *et al.* [16] use quasi-bicliques to find interacting protein group pairs in protein-protein interactions. Beyond biology, Sim *et al.* [25] use quasi-bicliques to cluster groups of stocks and financial ratios to identify investment opportunities.

2.2 Some Technical Considerations

This range of applications has resulted in significant research to find efficient algorithms to identify maximal bicliques. Prisner [22] demonstrated that the number of maximal bicliques in a bipartite graph with n vertices varies according to the formula $2^{n/2}$. Zhang *et al.* [28] noted that algorithms for finding maximal bicliques follow one of three approaches: exhaustive search, reduction to the clique enumeration problem in a general graph, and reduction to the frequent itemset mining problem in a transaction database. Alexe *et al.* [1] enumerate all maximal bicliques using a consensus algorithm similar to the consensus method for finding prime implicants of a Boolean function. Makino and Uno [17] use the second approach to convert a bigraph into a general graph by adding all edges to connect all vertices within the same partition. In this case, algorithms for enumerating maximal cliques in a general graph can be applied. However, Zhang *et al.* [28] noted that this approach is neither practical or scalable. Using the third approach, the adjacency matrix of a graph can be viewed as a transaction database, and a biclique corresponds to a pair of frequent closed itemsets in the transaction database.

Li *et al.* [14] suggest using frequent itemset mining techniques to mine maximal bicliques. They show that the problem of enumerating all maximal complete bipartite subgraphs is equivalent to mining frequent closed itemsets in the adjacency matrix. That is, a closed itemset and the set of transactions containing the closed itemset form a biclique. Liu *et al.* [25] point out that the large maximal biclique mining problem has size constraints on both vertex sets, while the frequent itemset mining problem puts size constraints on only one side of the transaction set. Their approach uses a divide-and-conquer technique to prune the search space.

The number of bicliques may be limited based on specific properties of the nodes and edges. For example, if the maximum node degree can be limited, the number of bicliques can become a polynomial in the largest node degree. Thus, the problem of finding all maximal bicliques is intractable for graphs allowing nodes of very high degree. It is therefore important to better understand the features that separate tractable instances of the problem from intractable ones and to better understand which of these factors are intrinsically associated with real world problems. This knowledge can help determine which types of problems are approachable by analysis using bicliques and which are not.

2.3 The Human Phenotype Network

Human Phenotype Networks (HPNs) [7,9] are mathematical graph models where the nodes represent human phenotypic traits. Edges represent genetic connections between traits, the granularity of which can be modulated from shared genetic variants (Single Nucleotide Polymorphisms, or SNPs), genes, or pathways, to name only a few. Because HPNs rely on Genome-Wide Association Study (GWAS) data, they incorporate diseases, physical and behavioral traits. HPNs are bigraphs by nature, before being projected onto the space of phenotype nodes only. In their projected form (single set of phenotype nodes), they have been successfully used to study diverse aspects of human traits and disease interactions, from pleiotropy and epistasis [8], to Type 2 Diabetes in East Asian populations [23], to environmental effects [7]. In this work we utilize two different HPNs in their bigraph (pre-projection) form. The first one is based on shared pathways between diseases and the second relies on shared genes between the traits [8].

Biological networks are generally expected to have heterogeneous connectivity placing them in the scale-free family. This means that the degree distribution follows a power-law, or exponential decay. Within the network, this translates into the presence of a minority of highly connected nodes (i.e. hubs). When the degree distribution of a scale-free network is plotted on a logarithmic scale, the resulting curve is approximately linear across the top [20].

3 Methods

In the present work, we utilize both pathway and gene-centric HPNs, using data from the May 2014 version of the NHGRI GWAS catalog [6]. To maximize coverage, HPNs included genotype-phenotype associations found in the database of Genotypes and Phenotypes (dbGaP). The pathway HPN incorporates genetic pathways information for Kegg and Reactome to build association between phenotypes/traits. The GWAS catalog and dbGaP report 1,252 traits combined, annotated with 37,681 SNPs in 16,411 loci. Bipartite HPNs rely on almost 1,000 phenotypic traits, and over 10,000 genes and almost 1,500 pathways, respectively. Throughout the analysis HPNs remain bipartite, where phenotypes and genes/pathways each represent a set of nodes. The resulting bipartite network can be projected in the space of phenotype vertices to obtain the HPN found in literature.

3.1 Definition of Terms

A **graph** $G = (N, E)$ [12] is defined by a set of nodes NN and a set of edges NE, where

$$N = \{n_i | i = 1, 2, ...NN\} \quad , \quad E = \{e_k = (n_i, n_j) | k = 1, 2, ...NE\}$$

and n_i, and n_j are contained in N. Graphs are used to model relationships. The nodes are the objects being related and the edges model the link between them. In the present work, we will only consider undirected graphs.

A **clique**, $C = (NC, EC)$ is a subgraph of a graph where NC and EC are subsets of N and E.

A **bigraph** (or bipartite network) $G = (U, V, E)$ is a graph whose node set is partitioned into disjoint sets of nodes, U and V, and edges E connect nodes in U to nodes in V; i.e., there are no edges between nodes in U and also no edges between nodes in V.

$$E = \{e_k | k = 1, 2, ...NE\}$$

where $e_k = (u_i, v_j)$ for u_i, contained in U and v_j contained in V.

A **biclique** $CB = (UC, VC, EC)$ is a subgraph of a bigraph $B = (U, V, E)$ where UC, VC, and EC are subsets of U, V and E, respectively and for all u_i, in UC and v_j in VC there is an edge $e_k = (u_i, v_j)$ for pairs u_i, in UC and v_j in VC. Bicliques model relationships between nodes in U and nodes in V. For example, U may be a set of diseases, V may be a list of genes and E may be a set of relationships between the diseases and the genes.

A **maximal biclique** is one containing nodes that are not proper subsets of the nodes in any other biclique. Maximal bicliques represent strongly related groups of nodes in G and represent groupings of the nodes into categories. Note that maximal bicliques need not be disjoint; nodes may overlap.

Networks are graphs where properties are associated with the nodes and edges. We therefore use these terms interchangeably. The search for bicliques is motivated by the desire to find groups of nodes that share one or more properties. Indeed, the problems we solve are often aimed at identifying relevant properties relating nodes to one another. In this paper, we restrict the discussion to unweighted (or all edges have the same weight), undirected edges but the algorithm we present extends naturally to the weighted, directed problems.

3.2 Biclique Detection Algorithm

We present and analyze an algorithm for finding maximal bicliques. A bigraph is defined as a triple, $G = (U, V, E)$. U and V are disjoint sets, not necessarily of equal cardinality. We will refer to a typical member of U as u_i and to a typical member of V as v_j. The members of E, $e = (u_i, v_j)$ are undirected edges so that U and V are interchangeable.

In the algorithm, we represent a biclique as a triple (L, R, A) where L is a subset of U, R is a subset of V, and A is the set of nodes, u_m in U that are adjacent to u_i, a given node in L (think of L and R as simply being left and right sets of nodes, with no particular significance to left and right.) For a biclique, $c_i = (L_i, R_i, A_i)$, we refer to L_i, R_i, and A_i as the biclique's L-set, R-set and A-set, respectively.

We say two nodes, u_i and u_m, in L are adjacent if they share at least one neighbor, v_k, in R; i.e., if there exist one or more pairs of edges (u_i, v_k) and

(u_m, v_k) that are members of E in G. We define a **singleton biclique**, s_i, as a biclique where L contains a single node.

As an example, consider Fig. 1. Let the nodes in the top row be U and the nodes in the bottom row be V. As shown in red, nodes 8, 9, 3 and 4 form a biclique (which is not a singleton biclique), C, because both 8 and 9 are connected to both 3 and 4. Note that for nodes to form a biclique all the left nodes have to be directly connected to (not just adjacent to them by the definition above) all the right nodes. Nodes 3 and 4 are both directly connected to nodes 0, 5, 8 and 9.

We thus have:

$$L_c = \{8, 9\} \quad ; \quad R_c = \{3, 4\} \quad ; \quad A_c = \{0, 5\}$$

Note that we do not include 8 and 9 in AC because they are already in L_c.

A singleton biclique, s_i, is a biclique where L contains a single node. In green in Fig. 1, we show an example of a singleton biclique: [{14} ; {7, 10, 12} ; {5, 9, 11, 13, 15, 17, 18}] where the 3 sets inside the square brackets are respectively L_c, R_c and A_c.

The algorithm maintains a collection of maximal bicliques found so far and a collection of candidate bicliques to be expanded into larger bicliques. The expansion of biclique is a set of bicliques formed by adding each member of its A-set to its L-set.

The collection of maximal bicliques is managed as a map where each key is the R-set of a biclique and the associated value is the biclique itself. The candidate collection is managed as a priority queue. In order for this to work, sets of nodes have to be comparable. Since they are sets, we can test for inclusion and equality. In addition to this, the algorithm also has to know when two sets are not comparable. Two sets are said to be not comparable if they are not equal and neither includes the other. We will refer to the map of bicliques as MB and the queue of candidate bicliques as QC.

MB contains all bicliques that are currently maximal. As the algorithm proceeds, new bicliques are added and to MB and bicliques in MB may be replaced by bicliques that dominate them. A biclique $c_i = (L_i, R_i, A_i)$ dominates biclique

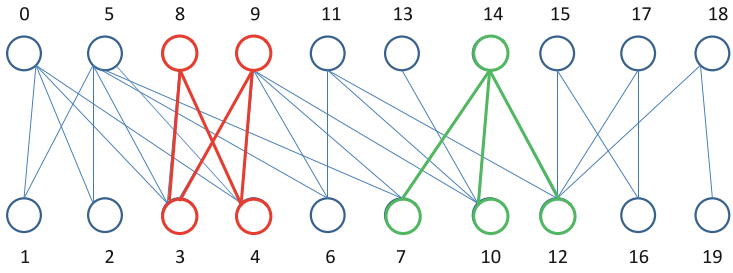

Fig. 1. Biclique schematic. L nodes 8 and 9 form a biclique with R nodes 3 and 4, with adjacent nodes 0 and 5. A singleton biclique is formed between L node 14 and R nodes 7, 10, and 12 (Color figure online).

$c_j = (L_j, R_j, A_j)$ if L_j and R_j, are contained in L_i ,and R_i. If $L_i = L_j$, and $R_i = R_j$, then $c_i = c_j$. Note that since the A-sets are derivable from the R-sets, if R_j is contained in (or equal to) R_i then A_j is contained in (or equal to) A_i.

The algorithm is implemented from the following basic operations.

Merging Two Bicliques: The merger, M_{ij}, of two bicliques, c_i and c_j is defined as

$$M_{ij} = c_k = (L_k, R_k, A_k)$$

where

$$L_k = L_i \cup L_j \quad ; \quad R_k = R_i \cap R_j \quad ; \quad A_k = A_i \cap A_j$$

Expansion of a Biclique: The expansion, EXP_i of a biclique c_i is the set of bicliques M_{ij} formed by merging $c_i s_j$ with each singleton biclique, s_i, in the A-set of c_i. Thus,

$$EXP_i = \{M_{ik} | k \in A_i\}$$

In its simplest form, the algorithm begins by putting all the singleton bicliques into QC and also into MB. It then pops c_i from the front of QC and expands it. Each M_{ik} in EXP_i is tested. Specifically, the biclique, c_m, (if any) whose key equals the R-set of M_{ik} is retrieved from MB. If is null, M_{ik} is put into MB. If is not *null*, it is merged with M_{ik}, forming M_{im} , and M_{im} replaces c_m in MB.

As the algorithm proceeds, new biclique are added to MB and previously found bicliques are replaced by ones that dominated them. As a biclique is expanded, it is removed from QC. When newly found bicliques (either new ones or merged ones that dominate previously found ones) are placed in MB they are also added to QC. The process terminates when QC is empty.

Given the original graph $G = (U, V, E)$, we form a set, $T = U$; i.e. $t_i = u_i$. We then order the t_i based on their degrees in G, smallest first. Other ordering are possible; e.g. by degree, largest first. For each t_i we will form a graph $G_i = (U_i, V_i, E_i)$ which we will use to find all the maximal bicliques in G_i.

We form U_i first by removing from U all nodes, u_j not adjacent to t_i in G, i.e., where adjacency is defined as above by the existence of a path through some v_j in V_i. We then remove from U_i all t_j for $j < i$. V_i contains all v_j in V for which an edge (u_i, v_j) exists in G. E_i is the subset of E whose endpoints are in U_i and V_j.

The algorithm proceeds by processing the G_i in succession as described above, finding all the bicliques in G_i. Because U_i does not contain any previously processed G_i removed previously processed t_j we will not find any maximal bicliques previously found when processing G_j. Thus each maximal biclique is found exactly once.

3.3 Biclique Detection Algorithm Complexity Analysis

Our algorithm, as described above, starts with singleton nodes and expands each one by adding adjacent nodes to them one at a time via a depth first search. Its runtime complexity and memory requirements can be modeled as

$O(N_{gc} \times E_{pc})$,where N_{gc} is the number of bicliques generated and E_{pc} is the effort required to generate a biclique and check it for dominance. E_{pc} can be seen to be linear in the size of the biclique, specifically the size of the L-set, R-set and A-set. These are in the worst case linear in the number of nodes in the network. N_{gc} on the other hand could grow exponentially with the number of nodes in the network and is therefore our principal concern. N_{gc} must grow at least linearly with Nc, since every maximal clique must be generated and checked. Ideally, N_{gc} would grow only linearly with N_c.

4 Results

4.1 Artificial Networks

We ran experiments on artificially generated networks using a desktop computer with 16 GB of memory and a 2.4 GHz. processor using CentOS Linux. Our principal goal was to determine the functional dependence of the complexity on the size of the network, the average nodal degree, the variation in nodal degree and the amount of correlation among the edges.

Experiment 1 examined the effect of network density on the number of maximal bicliques and the complexity of our algorithm. Figure 2 summarizes the results of this experiment for 100 and 200 node networks. Within each set of networks we varied the density between 0.05 and 0.5 for the 100 node network (Fig. 2a) and between 0.025 and 0.25 for the 200 node network (Fig. 2b). We recorded the number of maximal bicliques, N_c, in the network, and the algorithm's running time.

Both time and maximal bicliques increased exponentially with density and at comparable rate. This indicates that the algorithm is not expanding many nodes that are not producing maximal bicliques. The number of generated bicliques also converges to a rate that is roughly the same as that of the maximal bicliques

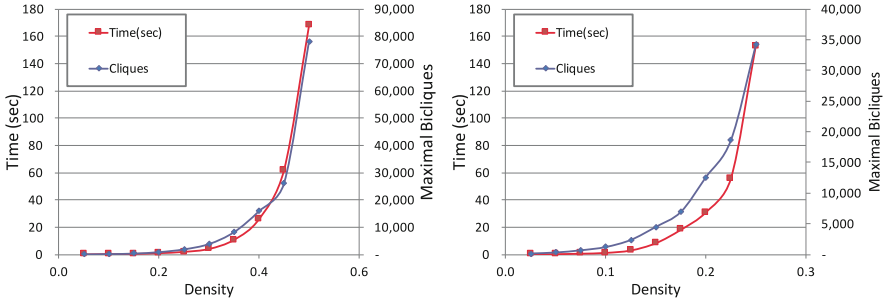

Fig. 2. Effect of network density on complexity. Time and number of maximal bicliques as a function of graph density for network sizes (a) N = 100 and (b) N = 200.

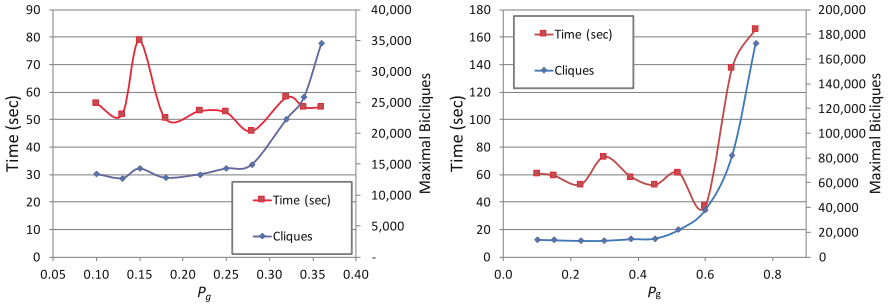

Fig. 3. Effect of correlation on complexity. Time and number of maximal bicliques as a function of P_g, the probability of an edge between nodes within a group, for network size N = 400 and groups (a) NG = 4 and (b) NG = 8.

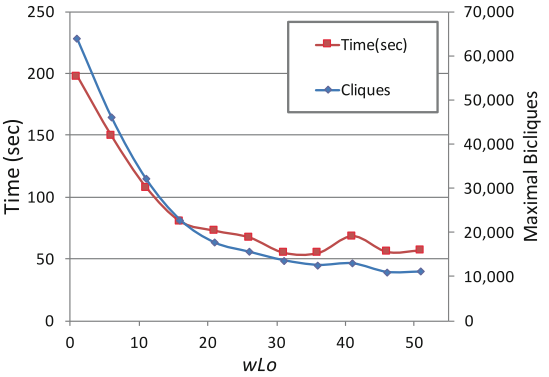

Fig. 4. Effect of node weight distribution on complexity. Time and number of maximal bicliques as a function of *wLo*, the minimum node weight in the network, for network sizes (a) N = 200 and (b) N = 2000. Note that as *wLo* increases, weight distribution narrows and degrees of nodes becomes more uniform.

N_c, but its value is approximately N times the number of cliques expanded, which is approximately twice N_c. This difference is due to the fact that in its current form, when the algorithm expands a node, n, it generates a new node corresponding to each node adjacent to n.

Experiment 2 explored the effects of correlation. We divided the nodes into groups of equal sizes and varied the probabilities of edges between nodes in the same group and nodes in different groups. We generated random networks with 200 and 400 nodes and in each case divided into groups of nodes of different sizes ($NG = 4$, $NG = 8$). P_g, the probability of an edge existing between two nodes in the same group was varied from 0.1 to 0.8, while the probability of an

edge existing between nodes in different groups was set to a value that made the overall density 0.2.

This simulated the situation where the network contained groups of nodes with different properties (e.g., different types of diseases) and where the edges between nodes in the same group were correlated. Figure 3 summarizes the results of Experiment 2. We see overall complexities similar to those observed in Experiment 1, and while the correlation does have some effect, we did see some increase in the rate of maximal clique generation with the smaller sample size (N = 200). In the larger sample size ($N = 400$), however, we observed in Fig. 3 an exponential increase in the number of maximal bicliques generated, with the rate of growth increasingly sharply around $P_g = 0.3$ ($NG = 4$, Fig. 3a) and $P_g = 0.5$ of 0.5 ($NG = 8$, Fig. 3b).

Experiment 3 examines the effect of the distribution of node weights varied linearly between *wLo* and *wHi*. We varied *wLo* from a value of 1 to 50 and varied *wHi* to maintain a network density of 0.2 in all cases.

We generated two sets of networks, one with 100 nodes and the other with 200. Figure 4 summarizes the results for N = 200. As the degree distributions becomes wider, the number of maximal bicliques becomes larger, increasing by more than a factor of 4. This effect was more pronounced in the larger network and thus is shown here.

In summary, we found that the density of the network is the factor most seriously affecting complexity and by working with smaller networks we were able to examine denser networks without consuming an inordinate amount of computer time and memory. It is our intent to optimize the algorithm implementation and run it on larger problems in the near future.

4.2 Bicliques and Correlation in the HPNs

We ran our algorithm on both versions of the HPN. The pathways network containing 807 phenotypes, 1444 pathways and 82,855 edges, with a density of 0.07 and the gene HPN containing 916 phenotypes, 10,011 genes and 55,223 edges, and a density of 0.006. In both cases, the degrees of the nodes vary widely. However, there are hundreds of phenotypes associated with more than 200 pathways/genes. As discussed in Sect. 4.1, this gives rise to a large number of maximal bicliques and both the runtime and memory requirements of the algorithm rise correspondingly.

We ran an experiment in which we trimmed the network by removing all nodes exceeding a maximum degree. Table 1 and Fig. 5a report results for the pathways HPN, where the maximum degree varied between 100 and 200. Results for the gene HPN follow the same trend and are only presented in Fig. 5b, for which we varied the maximum degree between 100 and 500.

It can be seen in Table 1 that as the maximum degree increases, the fraction of phenotype nodes retained increases from 73 % to 84 % of the total number of phenotypes while the number of retained edges increases from 8 % to 28 %.

Table 1. Effects of varying maximum node degree on HPN pathways network.

Maximum degree	100	125	150	175	200
# of phenotypes	589	593	633	654	676
# of edges	6,259	9,837	14,264	17,820	23,068
# of generated bicliques	96,391	498,802	2,063,365	7,075,168	26,550,050
# of expanded bicliques	5,693	12,690	28,618	63,892	180,303
# of maximum bicliques	2,001	4,944	11,705	27,728	85,481
Running time:	2.845	16.843	90.436	413.109	9,020.638

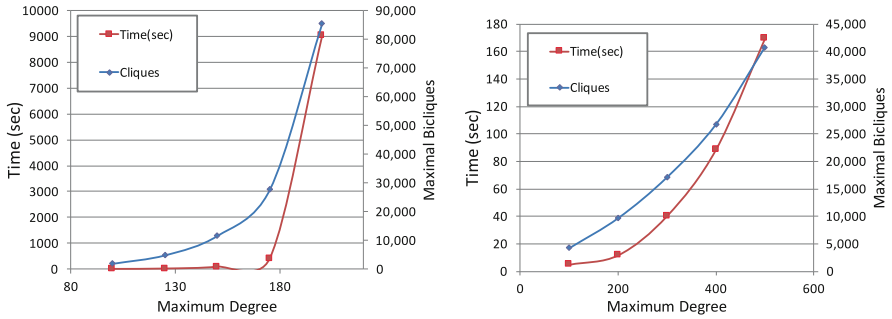

Fig. 5. Effect of maximum node degree on network complexity in pathway- and gene-centric HPN. Time and number of maximal bicliques as a function of the networks maximum allowed degree for (a) HPN pathway network and (b) HPN gene network.

Thus, most of the nodes have degrees less than 100 while most of the edges are associated with nodes with degree over 200. The number of expanded and maximal bicliques increases multiplicatively as the maximum degree increases by a constant. This is exponential behavior.

Note that the number of expanded bicliques and the number of maximal bicliques as seen in Table 1 grow at the same rate with the number of expanded bicliques, with their ratio converging towards 2. This shows that the algorithm expands only about 2 bicliques for every maximal biclique found. Thus, the complexity of the algorithm in terms of the number of bicliques expanded is linear in the number of maximal bicliques, but both are exponential in the number of nodes. The algorithm's complexity cannot decrease below linear in the number of maximal bicliques since it generates all of them. Therefore, the algorithm's complexity is minimal.

The number of generated bicliques, however, increases at a higher rate than the expanded or maximal bicliques. This is due to the implementation of the algorithm, which in expanding bicliques considers all adjacent nodes and eliminates most of the generated bicliques because they are dominated.

The running time increases even more rapidly than the number of generated bicliques. This is most probably a consequence of Java's Virtual Machine Garbage Collecting technology that automatically releases memory used by the software. As the number of objects in memory increases, in our case, the number of bicliques, so is the time necessary to free unused memory blocks. This effect is most evident for the case where the maximum degree is 200. Thus, to truly reduce the complexity and make more problems tractable, even after optimizing the implementation, we must reduce the density of the graph. This is treated in more detail in the Conclusions section below.

5 Biomedical Insights into Genotype-Phenotype Bicliques

We performed a preliminary biomedical analysis of bicliques from the gene HPN. Phenotypic data includes disease-specific and non-disease phenotypes, such as body mass index and blood glucose levels.

Though comprehensive analysis of phenotypes and genes yielded over 80,000 edges, various methods can be used to filter this input in future studies. Trimming nodes with degree 100 or more, for instance, eliminates many non-specific phenotypes such as body weight and creatinine, which have been included in various GWAS studies and which may impart little useful information in elucidating the basis for human disease. On the other hand, trimming these nodes can eliminate important diseases such as multiple sclerosis and coronary artery disease, which have been extensively studied and, due in part to their complex etiologies, have a large number of genetic and pathway associations. It may therefore be advantageous to rely upon clinical expertise to eliminate phenotypes extraneous to a study, or to focus studies on a predetermined subset of phenotypes and therefore limit the number of nodes analyzed, i.e. knowledge-driven filtering.

5.1 Assessment of the Relevance of Biclique Output

For a large set of data, output can be restricted to bicliques containing a specified number of phenotype, gene or pathway-specific nodes. For instance, we restricted output to bicliques containing at least 40 (Table 2a) or 50 (Table 2b) genes and 2 phenotypes to generate 72 and 24 bicliques, respectively. We further trimmed the output to bicliques containing at least one disease-specific phenotype – i.e. we eliminated bicliques containing only non-disease phenotypes, such as gambling and heart rate. We hypothesized that bicliques containing a large number of genes would match phenotypes with known clinical, biological, or genetic overlap and examined the bicliques to assess our hypothesis.

To summarize the findings, the phenotypes in each biclique were divided into three different categories in Table 2: redundant, meaning phenotypes are clinically equivalent or one is a subtype of another; directly correlated, meaning phenotypes are pathologically or physiologically linked and thus expected to occur in the same patient; or distinct phenotypes, meaning phenotypes may

share underlying genetic or molecular pathways but are not expected to occur in the same patient.

Note that in future studies, disease ontology can be used to filter these redundant or directly correlated phenotypes. In the present study, we chose not to filter these phenotypes in order to assess the validity of the algorithm in matching these phenotypes. As such, there are several noted redundancies seen between matched phenotypes in these analyses.

More interesting for clinical research are connections found between the distinct phenotypes in Table 2. Most of these phenotypes do in fact have a biochemical, pathological or clinical correlation. For instance, Crohn's disease and Ulcerative Colitis are clinical subtypes of inflammatory bowel disease but have overlapping symptomatology and are often misdiagnosed; genetic markers may have some diagnostic utility in distinguishing these diseases, and our algorithm could be useful in identifying markers that are shared versus distinct [5].

Beyond the ability of the algorithm to replicate known clinical and biological correlation, the algorithm provides an efficient tool for investigating candidate genes and pathways that underlie phenotypes with shared etiologies. For example, the algorithm can be used to identify candidate genes underlying etiological overlaps between multiple sclerosis (MS) and type I diabetes (TIDM), which tend to co-occur in families, though their genetic makeup (HLA patterns) are thought to be mutually exclusive, with some variants being a risk factor for one disease and protective against the other [3,13]. As shown in Table 3, our algorithm matched MS and TIDM to several genes already known to influence the risk of both diseases, including HLA DRB1 and HLA DRQ1. Our algorithm also matched the diseases with nine other HLA loci that could be used as candidates for future analysis of variants.

5.2 Assessment of a Disease-Limited Data Set and Hypothesis Generation

In addition to TIDM and MS, there is a high degree of connectivity between various types of autoimmune disease, as can be seen in Table 2. The development of autoimmune disease is genetically and environmentally complex and much is yet unknown. We can use bicliques and past research to draw hypotheses regarding candidate genes and pathways involved in these diseases.

For instance, recent research has shown that abnormal dopamine levels are observed within brain cells of mouse models and within tissues of human patients with autoimmune disease [21]. To identify candidate genes involved in this dopamine dysregulation, it may be useful to examine bicliques that contain autoimmune diseases as well as psychiatric disorders, which are largely driven by dopamine dysregulation.

Examination of several bicliques in Table 3 reveals a gene called NOTCH4 in several bicliques containing autoimmune disease as well as schizophrenia. As reflected in additional bicliques that are not shown, GWAS has implicated NOTCH4 in autoimmune diseases including multiple sclerosis, asthma, systemic sclerosis, lupus, ulcerative colitis and rheumatoid arthritis. In various studies,

Table 2. Clinical assessment of phenotypes pairs matched by bicliques. Phenotypes matched in bicliques containing two phenotypes and a large number of genes: a minimum of fifty (a) or forty (b). The phenotypes are divided into categories based on clinical characteristics. Redundancy indicates that the phenotype pair is the same clinical entity or that one phenotype is a subtype of the other. Direct correlation indicates that the phenotypes are patho-physiologically related and therefore expected to occur in the same patients. Distinct phenotypes indicates separate clinical entities that may be biologically or genetically associated and have been paired in bicliques according to the algorithm. Note: IBD stands for inflammatory bowel disease and MS stands for multiple sclerosis.

(a) Bicliques,containing 50 or more genes and 2 phenotypes.		
Redundancy (4)	*Direct correlation (3)*	*Distinct phenotypes (4)*
IBD - Crohn's disease	Osteoporosis - Bone density	ADD/ADHD - Gambling
AIDS - HIV-1	Asthma - QT interval	Schizophrenia - Gambling
Coronary dx - Restenosis	Sudden death - Resting HR	Breast neoplasms - Breast size
IBD - Urcerative colitis		Crohn's - Ulcerative colitis
(b) Bicliques containing 40 or more genes and 2 phenotypes.		
Redundancy (7)	*Direct correlation (7)*	*Distinct phenotypes (11)*
IBD - Crohn's disease	Osteoporosis - Bone density	ADD/ADHD - Gambling
AIDS - HIV-1	Asthma - Respiratory function	Schizophrenia - Gambling
Coronary dx - Restenosis	Asthma - Heart rate	Breast neoplasms - Breast size
IBD - Ulcerative colitis	Asthma - QT interval	Crohn's dx - Ulcerative colitis
Cardiac hypertrophy - Cardiomegaly	Sudden death - Resting heart rate	Alzheimer's - IgG glycosylation
Pancreas cancer - Pancreas neoplasms	Dialysis mortality - Type 2 Diabetes	Alzheimer's - lipids
Biliary cirrhosis - Primary BC	Prostate cancer - Erectile dysfunction	Type 1 Diabetes - MS
		Behcet Syndrome - MS
		Crohn's disease - MS
		Lupus - MS
		IBD - MS

NOTCH4 has also been established as a candidate gene for schizophrenia and bipolar disorder [10], and is thought to influence the age of onset of disease as well as response to antipsychotic treatment [2]. Drawing on this previous knowledge and the results of our data analysis, we hypothesize that NOTCH4 may

Table 3. Example maximal bicliques containing two phenotypes and corresponding genes. Note: maximal bicliques contain variable numbers of phenotypes and genes.

Left nodes (phenotypes)	Right nodes (genes)			
Multiple sclerosis type I diabetes	HLA-C	BAT1	PPIAP9	ZBTB12
	HLA-S	BTNL2	RPL3P2	RPL32P23
	HLA-DQA2	MICA	TAP2	FCRL3
	HLA-DRB1	NOTCH4	PRM3	SLC44A4
	HLA-DRB5	TNXB	TNP2	CFB
	HLA-DRA	CLEC16A	BAT3	TYK2
	HLA-DRB9	HNRNPA1P2	CLECL1	C2
	HLA-DOB	IL2RA	KIAA0350	C16orf75
	HLA-DQB2	BACH2	ZFP36L1	IL7R
	HLA-DMB	HCG26	BAT2	MICB
Rheumatoid arthritis schizophrenia	BDP1P	GUSBP1	HLA-DRB5	HLA-DRB9
	CDH12	HLA-DQA1	LASS6	NOTCH4
	DKFZp667F0711		PRKCQ	SALL3
Multiple sclerosis lupus	BTNL2	RNF39	ZBTB12	MTCO3P1
	DDR1	TNXB	WASF5P	PVT1
	DHFRP2	HLA-DRA	CBLN2	HLA-DRB1
	HLA-B	BACH2	HLA-S	HLA-DRB9
	HLA-X	HCG18	RPS2P6	FENDRR
	IER3	HLA-DOB	FOXF1	HCG26
	MICA	HLA-L	IRF8	TCF19
	MICB	HLA-DQB2	CLEC16A	TNFAIP3
	NOTCH4	EHMT2	HLA-DQA2	TRIM26
	PSORS1C1	BAT2	HLA-DQB1	AIF1
	BAT3	CFB		

also play a role in dopaminergic pathway dysregulation underlying autoimmune etiology. Investigation of genetic and pathway overlap between distinct diseases can be used in biomedical studies to develop such hypotheses.

6 Conclusions and Future Work

We found that the complexity of our algorithm is the product of the number of nodes and the number of generated bicliques. As hypothesized, network density is the major determinant of computational complexity. Node degree distribution

and correlation among edges are secondary factors, except when resulting in a significant number of nodes of high degree.

To address complexity in large data sets, it may seem beneficial to trim the network by removing nodes of high degree. However, it is important to assess the significance of these nodes in the context of the investigation, in our case the biological and genetic underpinnings of disease. It may also be beneficial to partition the network into distinct parts; for instance, to analyze one set of diseases and their pathway and gene associations.

In phenotype-pathway analysis, we found that few of the bicliques contained a large number of phenotypes and pathways. Many of the bicliques contained a small number of phenotypes links to a large number of pathways, or vice versa. Many bicliques included redundancies between similar phenotypes and pathways, as expected, and such nodes could be eliminated or combined to reduce the size and complexity of the problem, if clinically or scientifically appropriate.

In the phenotype-gene analysis, the algorithm linked many phenotypes that were redundant, closely related, or have been associated in clinical or biological studies. The algorithm further reveals genes that may underlie these associations and also provides a tool for genotype-phenotype hypothesis generation.

As we apply the algorithm to additional investigations in biology and medicine, we will evaluate the techniques that can be guided by biological properties. The maximal biclique problem described above was defined on an unweighted graph. The problem can be extended naturally to one where an edge is given an associated weight, indicating the strength of the relationship between nodes or where a node is given a weight based on its biological significance.

Potentially, these insights can be used to enhance our understanding of the biochemical basis for disease and to identify useful targets for drug repositioning between diseases. Besides phenotype-gene interactions, the algorithm can be used to investigate single nucleotide polymorphisms (SNPs), as well as environmental and other risk factors associated with disease.

Acknowledgments. This work was supported by National Institutes of Health grants LM009012, LM010098, and EY022300.

References

1. Alexe, G., Alexe, S., Crama, Y., Foldes, S., Hammer, P.L., Simeone, B.: Consensus algorithms for the generation of all maximal bicliques. Discrete Appl. Math. **145**(1), 11–21 (2004). Graph Optimization {IV}
2. Anttila, S., Illi, A., Kampman, O., Mattila, K.M., Lehtimaki, T., Leinonen, E.: Interaction between NOTCH4 and catechol-O-methyltransferase genotypes in schizophrenia patients with poor response to typical neuroleptics. Pharmacogenetics **14**(5), 303–307 (2004)
3. Atkinson, M.A., Eisenbarth, G.S.: Type 1 diabetes: new perspectives on disease pathogenesis and treatment. Lancet **358**(9277), 221–229 (2001)
4. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proceedings of the International Conference Intelligent Systems for Molecular Biology, vol. 8, pp. 93–103 (2000)

5. Cleynen, I., Boucher, G., Jostins, L., Schumm, L.P., Zeissig, S., Ahmad, T., Andersen, V., Andrews, J.M., Annese, V., Brand, S., Brant, S.R., Cho, J.H., Daly, M.J., Dubinsky, M., Duerr, R.H., Ferguson, L.R., Franke, A., Gearry, R.B., Goyette, P., Hakonarson, H., Halfvarson, J., Hov, J.R., Huang, H., Kennedy, N.A., Kupcinskas, L., Lawrance, I.C., Lee, J.C., Satsangi, J., Schreiber, S., Théâtre, E., van der Meulen-de Jong, A.E., Weersma, R.K., Wilson, D.C., Parkes, M., Vermeire, S., Rioux, J.D., Mansfield, J., Silverberg, M.S., Radford-Smith, G., McGovern, D.P.B., Barrett, J.C., Lees, C.W.: Inherited determinants of Crohn's disease, ulcerative colitis phenotypes: a genetic association study. Lancet (2015)
6. Darabos, C., Desai, K., Cowper-Sal·lari, R., Giacobini, M., Graham, B.E., Lupien, M., Moore, J.H.: Inferring human phenotype networks from genome-wide genetic associations. In: Vanneschi, L., Bush, W.S., Giacobini, M. (eds.) EvoBIO 2013. LNCS, vol. 7833, pp. 23–34. Springer, Heidelberg (2013)
7. Darabos, C., Grussing, E.D., Cricco, M.E., Clark, K.A., Moore, J.H.: A bipartite network approach to inferring interactions between environmental exposures and human diseases. In: Pacific Symposium on Biocomputing, pp. 171–182 (2015)
8. Darabos, C., Harmon, S.H., Moore, J.H.: Using the bipartite human phenotype network to reveal pleiotropy and epistasis beyond the gene. In: Pacific Symposium on Biocomputing, pp. 188–199 (2014)
9. Darabos, C., White, M.J., Graham, B.E., Leung, D.N., Williams, S.M., Moore, J.H.: The multiscale backbone of the human phenotype network based on biological pathways. BioData Min. **7**(1), 1 (2014)
10. Dieset, I., Djurovic, S., Tesli, M.: Up-regulation of NOTCH4 gene expression in bipolar disorder. Am. J. Psychiatry **169**, 1292–1300 (2012)
11. Gaspers, S., Kratsch, D., Liedloff, M.: On independent sets and bicliques in graphs. In: Broersma, H., Erlebach, T., Friedetzky, T., Paulusma, D. (eds.) WG 2008. LNCS, vol. 5344, pp. 171–182. Springer, Heidelberg (2008)
12. Gondran, M., Minoux, M., Vajda, S.: Graphs and Algorithms. Wiley, New York (1984)
13. Lernmark, A.: Multiple sclerosis and type 1 diabetes: an unlikely alliance. Lancet **359**(9316), 1450–1451 (2002)
14. Li, J., Li, H., Soh, D., Wong, L.: A correspondence between maximal complete bipartite subgraphs and closed patterns. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 146–156. Springer, Heidelberg (2005)
15. Liu, J., Wang, W.: Op-cluster: clustering by tendency in high dimensional space. In: 2003 Third IEEE International Conference on Data Mining, ICDM 2003, pp. 187–194, November 2003
16. Liu, X., Li, J., Wang, L.: Modeling protein interacting groups by quasi-bicliques: complexity, algorithm, and application. IEEE/ACM Trans. Comput. Biol. Bioinform. **7**(2), 354–364 (2010)
17. Makino, K., Uno, T.: New algorithms for enumerating all maximal cliques. In: Hagerup, T., Katajainen, J. (eds.) SWAT 2004. LNCS, vol. 3111, pp. 260–272. Springer, Heidelberg (2004)
18. Maulik, U., Mukhopadhyay, A., Bhattacharyya, M., Kaderali, L., Brors, B., Bandyopadhyay, S., Eils, R.: Mining quasi-bicliques from HIV-1-human protein interaction network: a multiobjective biclustering approach. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**(2), 423–435 (2013)
19. Milner, R.: Bigraphs and their algebra. Electron. Notes Theor. Comput. Sci. **209**, 5–19 (2008)

20. Newman, M.: Networks: An Introduction. Oxford University Press Inc., New York (2010)
21. Pacheco, R., Contreras, F., Zouali, M.: The dopaminergic system in autoimmune diseases. Front. Immunol. **5**, 1–17 (2014)
22. Prisner, E.: Bicliques in graphs I: bounds on their number. Combinatorica **20**(1), 109–117 (2000)
23. Qiu, J., Darabos, C., Moore, J.H.: Studying the genetics of complex diseases with ethnicity-specific human phenotype networks: the case of type 2 diabetes in east asian populations. In: 5th Translational Bioinformatics Conference (2014)
24. Sanderson, M.J., Driskell, A.C., Ree, R.H., Eulenstein, O., Langley, S.: Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. **20**(7), 1036–1042 (2003)
25. Sim, K., Li, J., Gopalkrishnan, V., Liu, G.: Mining maximal quasi-bicliques to co-cluster stocks and financial ratios for value investment. In 2006 Sixth International Conference on Data Mining, ICDM 2006, pp. 1059–1063, December 2006
26. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. Bioinformatics **18**(Suppl 1), S136–44 (2002)
27. Wang, H., Wang, W., Yang, J., Yu, P.S.: Clustering by pattern similarity in large data sets. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2002, pp. 394–405. ACM, New York (2002)
28. Zhang, Y., Phillips, C.A., Rogers, G.L., Baker, E.J., Chesler, E.J., Langston, M.A.: On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. BMC Bioinform. **15**, 110 (2014)