BIOINFORMATICS

Vol. 22 no. 14 2006, pages e117–e123
doi:10.1093/bioinformatics/btl260

Dense subgraph computation via stochastic search: application to detect transcriptional modules

Logan Everett[1], Li-San Wang[2] and Sridhar Hannenhalli*

[1]Penn Center for Bioinformatics and [2]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA 19104

ABSTRACT

Motivation: In a tri-partite biological network of transcription factors, their putative target genes, and the tissues in which the target genes are differentially expressed, a tightly inter-connected (dense) subgraph may reveal knowledge about tissue specific transcription regulation mediated by a specific set of transcription factors—a tissue-specific *transcriptional module*. This is just one context in which an efficient computation of dense subgraphs in a multi-partite graph is needed.

Result: Here we report a generic stochastic search based method to compute dense subgraphs in a graph with an arbitrary number of partitions and an arbitrary connectivity among the partitions. We then use the tool to explore tissue-specific transcriptional regulation in the human genome. We validate our findings in Skeletal muscle based on literature. We could accurately deduce biological processes for transcription factors *via* the tri-partite clusters of transcription factors, genes, and the functional annotation of genes. Additionally, we propose a few previously unknown TF-pathway associations and tissue-specific roles for certain pathways. Finally, our combined analysis of Cardiac, Skeletal, and Smooth muscle data recapitulates the evolutionary relationship among the three tissues.

Contact: sridharh@pcbi.upenn.edu

1 INTRODUCTION

Eukaryotic protein coding genes are transcribed by RNA *Polymerase-II*. To accomplish this, Pol-II is critically aided by several other transcription factors (*TF*) (Kadonaga, 2004). These TFs bind to specific DNA elements in the relative vicinity of the gene, and through cooperative interaction guide Pol-II to the transcription start site (TSS). An important long-term goal is the knowledge of groups of functionally interacting factors—*transcriptional module* (Bolouri *et al*., 2002; Thompson *et al*., 2004). Transcription modules provide an efficient mechanism to co-regulate a group of functionally related genes, for instance, specific to a tissue (Wasserman *et al*., 1998) or involved in immunity (Senger *et al*., 2004).

A combinatorial approach to transcriptional module detection uses a graph-theoretical abstraction: in a bi-partite graph of TFs and genes, where a TF is connected to its target genes, a large bi-partite clique represents a potential transcriptional module (Hannenhalli *et al*., 2003). This is precisely the problem of *clique enumeration in bi-partite graphs* (Alexe *et al*., 2000). One can attach weights to the TF-gene pairs indicating the likelihood that the TF regulates the gene. In this case a more desirable optimization is to detect *heavy* sub-graphs (Tanay *et al*., 2004). These combinatorial, enumerative approaches although effective in several biological problems (Hannenhalli *et al*., 2003), are inherently inefficient, thus limiting their application. Also, a practical extension of this abstraction should include additional types of nodes in the graph, for instance functional classes or tissues. A maximal clique in a tri-partite graph with Tissue as the additional partition would reveal tissue specific transcriptional modules. One can imagine the utility of having additional partitions representing other kinds of functional information.

Efficient algebraic approaches based on spectral graph theory have been proposed to co-cluster the two dimensional gene-expression (Ernst *et al*., 2002), and word-document (Dhillon, 2001) datasets; dense blocks in the permuted matrix represent co-clusters. The main limitation with this approach is that the co-clusters are non-overlapping and it is difficult to assess their significance. Dense sub-graph computation in general graphs has been studied in the context of identifying web communities (Flake *et al*., 2000) using network flow techniques. However, these methods focus on detecting a single most dense subgraph and are not adaptable to our specific problem domain, as will become clear later. There are approaches to detect overlapping clusters, although only in 2-dimensional, gene-expression data (Ihmels *et al*., 2002).

Another desirable feature that is lacking in current approaches is that they do not distinguish a ubiquitously connected vertex from a vertex that is highly connected to a specific subset of vertices. In our application, we would like to avoid such ubiquitously connected vertices without having to filter them out in a pre-processing step. For instance a TF like Sp1 is not interesting, unless it is more tightly connected to our genes of interest than to other genes.

Here we propose a stochastic search based approach to detect dense subgraphs while addressing the concerns discussed above. We assess the significance of our solutions based on graph randomization. Our current implementation exploits the tri-partite graph structure with an arbitrary connectivity between partitions. We have applied the tool on human whole genome TF-Gene graphs for tissue specific genes to discover tissue specific transcriptional modules. We have validated the clusters detected in Skeletal muscle based

*To whom correspondence should be addressed.

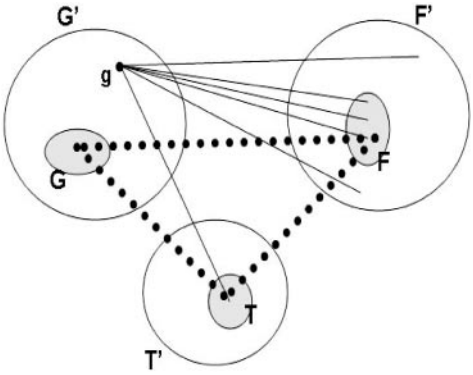

Fig. 1. Illustration of an iteration of the stochastic search.

on the literature evidence. When applied to TF-Gene-GO graph without the TF-GO edges, our approach can successfully deduce TF-GO relations, i.e. functional assignment of TFs. Similar application to TF-Gene-Pathway reveals novel TF-Pathway relations. Application to Tissue-Gene-Pathway graph using the combined datasets for Cardiac, Skeletal, and Smooth muscle recapitulates the evolutionary relationship among the three tissues and reveals novel Tissue-Pathway relations. Thus, our work presents a novel efficient approach for dense subgraphs and its application to a variety of genome wide tri-partite graphs.

2 RESULTS

2.1 Computing dense sub-graphs by Random Search—method overview

The goal of our approach is to find 'all' distinct dense sub-graphs. Because our input involves thousands of nodes and edges, our method has to be time-efficient. We adopt a *stochastic hill-climbing* approach that attempts to strike a balance between speed and premature stopping at local optima. In summary, consider a Markov chain where each state represents a potential solution (represented by an indicator variable for each node where a value of 1 indicates that the vertex belongs to the cluster). We connect each state to another state if they differ in exactly one vertex, and define the transition probability to capture the fitness of the solution. Starting from some starting state we stochastically traverse the neighborhood of this state in the state space until an optimal state is reached. We repeat this process starting from a large set of seed states to obtain several good solutions.

Although our approach is applicable to a general graph, in order to highlight the specific application to transcriptional module detection, here we illustrate the method using a tri-partite graph (Fig. 1). Let the three parts be G′ (genes), T′ (Tissues) and F′ (Transcription factors). We will also refer to these parts as G_G, G_T, and G_F respectively. Figure 1 shows the input graphs G′, T′, and F′ and a potential solution G, T, F. Intuitively, we want a solution such that nodes in G are connected to a large fraction of nodes in F and a relatively smaller fraction of nodes in F′ (same holds for all pairs of subsets). This can be captured using a log-likelihood score.

For a node g and a subset of nodes X in another partition, $N(g,X)$ is the number of nodes in X connected to g and $D(g,X) = N(g,X)/|X|$, i.e. the fraction of nodes in X that g is connected to.

The 'score' of a solution G, T, F is

$$S(G,T,F) = \sum_{u \in G \cup T \cup F} \left[\sum_{X=\{G,T,F\},\, u \notin X} N(u,X) \log \frac{D(u,X)}{D(u,X')} \right]$$

In other words, for every node, we compute its log-likelihood score with respect to each of the other partitions. In a given iteration of our stochastic search (state transition in the Markov chain), the solution can grow or shrink. Every node, both, inside and outside the current solution, is scored. The score of a node outside the current module, i.e. $g \notin G$ AND $g \in G'$, is $S(G \cup \{g\}, F, T) - S(G, F, T)$, i.e. the relative increase in the cluster score if g is *added* to the module. A node inside the module, i.e. $g \in G$ can be scored analogously as the relative increase in the module score if g is *removed*. The scores for all nodes from all partitions are normalized to a sum of 1 (after initializing the negative scores to 0). A candidate node is chosen according to this probability distribution. Note that adding or removing a single node corresponds to a state transition in our Markov chain. The procedure stops when no significant gains are achieved for several consecutive iterations.

To seed our stochastic search, we enumerate all maximal completely connected clusters with a user specified minimum number of nodes from each partition. For instance a typical value we have used is 3 genes and 3 transcription factors. We then iterate until we exhaust all seeds or reach the specified number of clusters; we choose the largest of the unused seeds and run the stochastic search algorithm to obtain a dense subgraph X; we then prune all seeds that highly overlap X to avoid finding similar subgraphs in subsequent runs. We stop after a pre-specified number of clusters (100) are identified.

Data preparation. From among the 546 vertebrate TF positional weight matrices (PWM) in TRANSFAC v8.4 (Wingender *et al*., 1996), we have extracted 221 representative PWMs (methods). This was done to minimize the bias in our clusters caused by highly similar PWMs connected to the same set of genes. For these 221 PWMs, we obtained the TF-Gene edges for all human genes using our binding site prediction method based on *Phylogenetic Footprinting* (Levy *et al*., 2002) (methods). We defined Gene-Tissue edges using an entropy-based measure of tissue-specificity (Schug *et al*., 2005) and the Novartis tissue survey data (Su *et al*., 2004). Finally Gene-GO and Gene-Pathway edges were defined using GO (Harris *et al*., 2004) and KEGG pathway resources (Kanehisa *et al*., 2002).

2.2 Tissue-specific transcriptional modules in Human—Skeletal Muscle as a case study

We identified 477 genes specifically expressed in Skeletal Muscle based on our threshold for tissue-specificity. We then applied our tool to the bi-partite graph consisting of 477 genes and 221 representative transcription factors. Figure 2 ('o') shows the cluster score distributions for this graph.

Significance To estimate the significance of the cluster scores, we randomized the input graph and computed the cluster scores, as shown in Figure 2 ('+'). A majority of the identified clusters have a score greater than the maximum score in the randomized graph. To obtain a more stringent background, we randomized the graph 100 times and for each randomized graph we retained only the maximum cluster score after running our tool until exhaustion

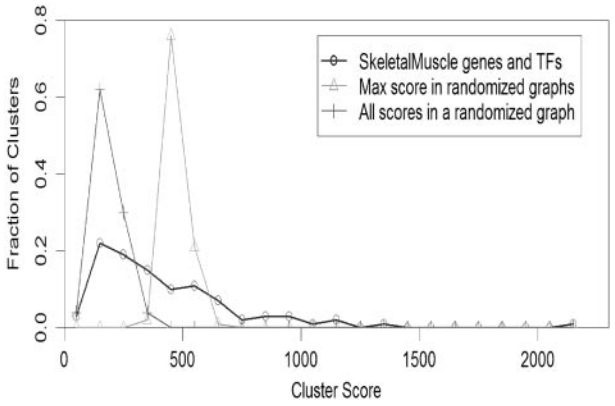

Fig. 2. The cluster score distributions for (i) clusters in Skeletal muscle specific genes and the corresponding TFs, (ii) maximum cluster score, one per randomized graph for 100 randomizations, (iii) all cluster scores for one randomized graph.

(using the same parameters as that for the original graph). As shown in the figure ('Δ'), even though the peak of real scores is to the left of the peak of max scores in the randomized graphs (this is expected since we are using max scores for the randomized graphs), there are several clusters in the input graph which score better than the maximum for any randomized graph (∼700). These 24 clusters therefore represent highly significant clusters.

Sensitivity Wasserman and Fickett have analyzed six transcription factors believed to confer muscle specific regulation (Wasserman *et al*., 1998), namely, Sp1, AP-1, Myf, SRF, MEF-2, and TEF-1. Sp1 is included in several of the top 24 clusters, including the top scoring one. In our initial grouping of positional weight matrices AP-1 belonged to a group with CREB as the representative and CREB was included among the top 24 clusters. Myf is an E-box protein and was grouped along with several other E-box proteins, with E47 as the representative, and E47 was included among the top 24 clusters. SRF was not included among the top 24 clusters but was in a cluster ranked 28, whose score is still above 90% of the background scores. MEF-2 was included in a very low scoring cluster (ranked 57). TEF-1 has a short 6 base pair binding site with very little information content, as reported in TRANSFAC and hence was not part of our input graph. However, TEF-1 is very similar to Tax/CREB in terms of binding site similarity which is in the same group as AP-1 mentioned above. Hence most of the factors analyzed in (Wasserman *et al*., 1998) are included in the high scoring clusters that we have identified.

Specificity To evaluate other factors identified by virtue of belonging to high scoring clusters, we extracted the 13 transcription factors that were included in greater than 10 of the 20 top-scoring clusters. These factors are: Sp1, MAZ, MAZR, Muscle_initiator, ETF, Churchill, EGR-1, AP2, VDR, MTF-1, Zic1, ZF5 and Spz1. MAZ (consensus: GGGGAGGG), MAZR (consensus: GGGGGGGGGGCCA), Churchill (consensus: CGGGGG) and ETF (consensus: GCGGCGG) are very similar to Sp1. ETF is a close homolog of TEF-1 (mentioned above), whereas MAZ sites are experimentally known to bind Sp1 (Parks *et al*., 1996), and MAZ is expressed in Skeletal Muscle (Song *et al*., 1998), MAZR binding site was found to be significantly enriched in 400 bp upstream of muscle genes in an independent computational analysis (Aerts *et al*., 2003). Muscle_initiator was derived by analyzing the promoters of specific Myc targets *in vivo* (Grandori *et al*., 1997). EGR-1 with SRF and Sp1 regulates muscle contraction (Irrcher *et al*., 2004). AP-2 with Sp1 regulates the muscle gene Utrophin (Perkins *et al*., 2001). VDR is involved in muscle development (Endo *et al*., 2003). MTF-1 is involved in oxidative stress response (Wimmer *et al*., 2005), an essential process in muscle. Zic1 is involved in skeletal development (Aruga *et al*., 1999). ZF5 is known to repress c-Myc (a gene involved in myogenesis) and one of the ZF5 isoforms is specifically expressed in skeletal muscle (Numoto *et al*., 1997). Thus, apart from Spz1, there is varying degree of support that all other transcription factors frequently found in high scoring clusters are involved in Skeletal muscle processes.

Although we have discussed the results only for Skeletal Muscle, we have in fact applied the tool to all tissues in the Novartis set. The score distributions follow a similar pattern relative to randomized graphs but specific analysis of the results in these tissues was not done.

2.3 Functional annotation of TFs via tri-partite cluster detection

Here we illustrate the utility of extending the above approach to multi-partite graphs. The largest cluster in the TF-Gene graphs for Skeletal muscle specific genes includes 36 TFs and 89 genes. We constructed a tri-partite graph by including the GO biological process (GOBP) for the genes as the third partition and connecting this new partition to the 'Gene' partition only. We computed dense clusters in this graph, with minimum edge density threshold of 0.75. This resulted in 14 sub-clusters. As in section 2.2, the scores of these 14 clusters are higher than the maximum scores for 100 randomized graphs (Wilcoxon rank sum test base p-value = 3.8E-04). Although, the GO annotations in these sub-clusters are largely overlapping, the genes and TFs in the sub-clusters are not so. Nevertheless it is difficult to interpret such subtly distinct sub-clusters based on the current literature. Instead, we assessed whether we can accurately assign functions to TFs via their sub-cluster membership. Recall that we did not use any known TF-GO relationships in identifying the clusters. The 14 sub-clusters involved 21 TFs and 12 GOBPs. Thus a total of 252 TF-GOBP relations are possible. In this universe of 252 relations, 59 are directly supported by the GO annotation for the TF protein, and thus represent the positives. To predict TF-GOBP relationships, we assigned each TF in a sub-cluster to each BP in that sub-cluster, resulting in 93 predicted TF-GOBP relations. Of the total of 252 relations, the overlap between predicted 93 and known 59 relations is 33 (Hypergeometric p-value = 5.4E-04). In other words 35% of our predictions include 56% of the known relations. To evaluate the validity of the 60 predicted relations with no supporting GO annotation, we took an indirect approach. For TF x and GOBP p, we estimated the *support* for a TF-GOBP relation '$x \leftrightarrow p$', as the number of $x \leftrightarrow g \leftrightarrow p$ triplets where the $x \leftrightarrow g$ indicates a binding site for x in g's promoter, and $g \leftrightarrow p$ indicates a GOBP annotation of g as p. We expect the 60 predicted TF-GOBP relations to have a greater support than the background. For the background we used the 133 of the 252 relations which were neither predicted, nor known. Also to

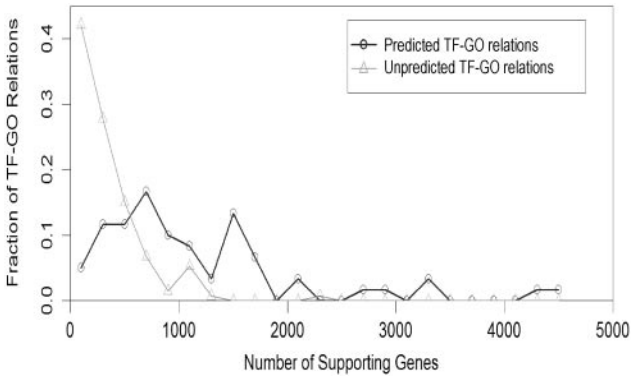

Fig. 3. Amount of indirect support for predicted TF-GOBP relations and for the background.

avoid circularity, we only used the support from genes which were *not* specific to Skeletal muscle and hence were not part of the input graph. Figure 3 shows that the predicted relations have a significantly greater support than the background (Wilcoxon rank sum test based p-value = 5.7E-15).

2.4 Co-regulation of genes involved in specific pathway

To detect specific pathways within the skeletal muscle specific modules, similar to the previous section, we constructed a tri-partite graph by including the known pathways for the genes as the third partition (instead of GO) and connecting this new pathway partition to the gene partition only. We computed 4 tightly connected clusters in this graph. One of them was intriguing in that the transcription factor ETF uniquely belonged to this sub-cluster. Recall that ETF was detected as a frequent member of high scoring clusters and is a family member of TEF-1 known to be involved in muscle processes but ETF does not have any direct evidence for involvement in muscle processes. The other TFs in this sub-cluster are p300, Sp1, AP-2 and EGR. And the genes in this sub-cluster are *Keratin 17*, *Vitronectin, Integrin-α7, Integrin-β1A,* and *cytosolic*, *malic enzyme 1*. Furthermore, the pathway 'ECM (Extra Cellular Matrix) receptor' belongs uniquely to this sub-cluster. Indeed *Vitronectin*, *Integrin-α7,* and *Integrin-β1A* belong to this pathway. ETF binding site occurs within 85 bps of a Sp1 site in the 1 kb promoter region of 4 of the 5 genes and in *Vitronectin* and *Integrin-β1A*, there are 2 distinct binding sites for ETF. Even though there is no direct experimental evidence supporting the role of ETF in the ECM-receptor pathway, we believe that the strong circumstantial evidence makes it a promising candidate to pursue for direct functional validation. Discoveries like this one can be made readily by an approach like ours that takes into account multiple types of information in an unbiased way.

2.5 Delineating Tissue-specific transcription factors and pathways via tri-partite clustering

Next we evaluate whether our approach can reveal subtle differences between tissues related at a gross level. We combined the TF-Gene data for genes specific to Heart, Skeletal muscle and Smooth muscle, resulting in a tri-partite graph consisting of 221 TFs, 1519 genes, and 3 tissues. Among the three tissues, Heart (574 genes), Skeletal muscle (477 genes), and Smooth muscle (666) genes, there are 117 genes in common between Heart and Skeletal muscle, 65 genes between Heart and Smooth muscle and 32 between Smooth and Skeletal muscle. This is consistent with phylogeny based results in (OOta *et al*., 1999). Even though cardiac muscle is evolutionarily closer to skeletal muscle, it is functionally closer to smooth muscle in that both cardiac and smooth muscle are involuntary. We investigated whether this evolutionary relationship is also reflected in the transcriptional modules. First, among the top 10 clusters, only 1 involved a single tissue and the other 9 involved exactly 2 tissues. Of these 9 cases, 6 involved Heart and Skeletal muscle, where 3 involved Heart and Smooth muscle. Thus there are twice as many clusters associating heart with skeletal muscle relative to smooth muscle. Second, among the top 10 tri-clusters, we recorded whether a TF belonged in a cluster with a tissue. For the three tissues in the specified order (Heart, Skeletal, Smooth), we assigned 3 binary numbers to each TF. For instance (1,0,1) means that the TF is associated with Heart and Smooth muscle but never with Skeletal muscle. The number of TFs belonging to the 7 possible binary vectors are—001(5), 010(0), 100(0), 110(29), 101(17), 011(0), 111(44). Thus most TFs are associated with all three tissues. Additionally there are more TFs uniquely associated with Heart and Skeletal muscle (29) than there are uniquely associated with Heart and Smooth muscle (17). Thus the transcriptional modules reflect the greater similarity between Heart and the Skeletal muscle. However, the statistical significance of this is not clear given the greater similarity between Heart and Skeletal muscle in terms of common genes.

Next we computed tri-clusters in the Tissue-Gene-Pathways graph in order to detect associations between tissues and pathways. A total of 15 clusters were detected, each with 2 tissues (this is because we required the seeds to have at least 2 tissues). Heart and Skeletal muscle co-associate in 8 cases, Heart and Smooth muscle co-associate in 5 cases and Smooth and Skeletal muscle co-associate in 2 cases. Furthermore, there are pathways that uniquely associate with one of the three tissues in our dataset. For instance there are 9 pathways uniquely associating with Heart and several of these have to do with immune system, e.g. *B cell receptor signaling pathway*, *Natural killer cell mediated cytotoxicity*, and *T cell receptor signaling pathway*. *Carbon fixation* pathway is uniquely associated with Skeletal muscle, and there are several pathways that uniquely associate with Smooth muscle, an overwhelming majority of which are involved in amino acid metabolism and degradation. We could not however assess the significance of these findings based on the current literature.

3 METHODS

Binding site annotation

We extracted the 1 kb regions upstream of the annotated transcripts in the hg16 release of the human genome from UCSC database (genome.ucsc.edu). We also extracted the Human-Mouse alignments for these regions. We searched the 1 kb regions using 546 binding profiles (Positional Weight Matrix or PWM) for vertebrate transcription factors from TRANSFAC v8.4 (Wingender *et al*., 1996). The search was done using the tool PWMSCAN (Levy *et al*., 2002). The initial hits were based on a p-value cutoff of 0.0002, corresponding to an average frequency of 1 hit every 5 kb scanned in the human genomic background. We filtered these initial

hits further using Human-Mouse alignments. For each hit we computed the fraction c of binding site bases that were identical between human and mouse. We retained the hits such that either p-value $\leq$ 0.00002 (1 in 50 kb) or $c \geq 0.8$. This procedure is similar to the one reported previously (Levy *et al*., 2002).

Clustering transcription factor PWMs

Pair-wise similarity computation Each PWM X is a 4 by k matrix for k-length binding site, where X_{ui} is the proportion of base u at position i, such that $\sum_u X_{ui} = 1$ (Stormo, 2000). We compute the dissimilarity or distance between position i of PWM X and position j of PWM Y using relative entropy $RE_{ij} = \sum_u X_{ui} \ln (X_{ui}/Y_{uj})$ (Durbin *et al*., 1998). For two identical positions this value is 0 and the more dissimilar the positions, the higher the RE value. However, as defined, this is an asymmetric measure and in practice we take the average of R_{ij} and R_{ji} as the distance between the two positions. Notice that according to this measure, for two positions at which the base pairs are distributed according to the background probability (say, equi-probable), their RE value will be 0, even though individually these positions are not informative. Let R_{ir} be the RE-value between column i and background probability distribution of bases. R_{jr} is defined similarly. We define the similarity between column i and column j, $S_{ij} = R_{ir} + R_{jr} - ((R_{ij} + R_{ji})/2)$. We first compute the S_{ij} for every pair of columns for all PWMs in the TRANSFAC database. These values are normally distributed with mean μ and standard deviation σ. The sum of k such S-values is also normally distributed with mean $\mu_k = \mu k$, and standard deviation $\sigma_k = \sigma\sqrt{k}$. To compute the similarity between k consecutive columns of two PWMs, we sum up the k S-values for aligned column pairs and transform this value to a z-score $= (S - \mu_k)/\sigma_k$, which makes the scores for different values of k comparable. Next, for every PWM-pair and for every alignment offset with a minimum of 6 base overlap between the PWMs (i.e., $k \geq 6$), we compute the similarity z-score ('z-value'). Using the empirical distribution of z-values for all alignments of all PWM pairs, we convert each individual z-value into a p-value, i.e., the probability of observing the z-value or higher in the background distribution; we call this the pz-value. Finally, to compute the similarity between two PWMs X and Y while allowing for the possibility that two related PWMs may be slightly shifted in positions, we slide the PWMs relative to each other such that at least 6 positions are aligned. For each such offset we compute the *pz*-value. Let *mpz* be the minimum *pz*-value over all offsets. Notice that the longer PWM pairs have a greater number of possible offsets and thus tend to achieve a low *mpz*-value. To correct for this effect, we compute the significance of the observed *mpz*-value as the random expectation of observing the mpz-value for K trials where K is the number of offsets. That is,

$$P(X,Y) = 1 - (1 - mpz(X,Y))^K.$$

Clustering PWMs based on the P-values Given a p-value threshold (we use 0.005), all PWMs can be represented as a network where PWMs correspond to the nodes and two nodes are connected if their similarity p-value is below the threshold. We then compute the so-called bi-connected component in this graph. A bi-connected component is a connected component of the graph that remains connected if any of the nodes are removed. Each bi-connected component corresponds to a cluster. In other words if two PWMs belong to same cluster, they must have at least two independent lines of evidence that they are related (i.e. paths in the graph). Each cluster thus obtained represents a family of PWMs with similar DNA binding specificity. We selected the median of each cluster as the cluster representative. Out of 546 PWMs, 442 were grouped into 117 clusters, and with 104 singletons, this procedure resulted in 221 representative PWMs.

Tissue specific genes

For each gene g and each tissue t, we say *g is specific for t* if its expression level in t is considerably higher than in other tissues, using the following procedure from (Schug *et al*., 2005). We use the Novartis GeneAtlas expression dataset (Su *et al*., 2004): the dataset has 79 different types of human tissues (two replicates each). The hybridization experiments are done using the Affymetrix HG-U133A (33689 probesets) and GNF1B (11391 probesets) platforms. Let $w(g,t)$ be the average expression level of probeset g in tissue t (*not* log2-transformed) over the two replicates. For each probeset, the relative expression level for tissue t is $p(t \mid g) = w(g,t)/\sum_{\text{tissue } i} w(g,i)$. The entropy of gene g is

$$H(g) = -\sum_t p(t \mid g) \log_2 p(t \mid g).$$

The categorical specificity of gene g and tissue t is $Q(g \mid t) = H(g) - \log_2 p(t \mid g)$. A low Q score implies gene g is highly specific for tissue t: $H(g)$ is low when the expression level of g is concentrated in a few tissues, whereas $p(t \mid g)$ is high when g is highly expressed in t. We empirically chose a value of 10.5 as the cutoff for $Q(g \mid t)$, as the density of the gene-tissue specificity begins a sharp increase at a higher Q. A more stringent value of 7 was suggested in (Schug *et al*., 2005). We then remap the association from Affymetrix probeset IDs to RefSeq IDs.

KEGG and GO annotation data

We built the associations between genes (refseq ID), KEGG pathways, and GO terms as follows. We downloaded data from the KEGG server that contained the association data between KEGG pathways and NCBI GI numbers. We downloaded the association data between GO terms and NCBI GENE IDs from the NCBI server. The mappings from GI numbers and GENE IDs to RefSeq IDs are obtained from NCBI. The mapping is inclusive: for example, if KEGG pathway x is associated with GI number y, and y is mapped to RefSeq IDs a, b, and c, then x is associated with a, b, and c.

Graph randomization

To determine the significance of cluster scores we find clusters by an identical process on randomized graphs with the node degrees identical to the real graph. The graph randomization process is performed by swapping edges with non-edges under a condition that preserves the degrees of all nodes. Specifically, a quadruple of nodes (w,x,y,z) qualifies for this swapping condition if it meets the following criteria (Yeger-Lotem *et al*., 2004): (i) both w and x reside in the same partition A, and both y and z reside in another partition B; (ii) there exists an edge between w and y, denoted as $E(w, y) = 1$, and also an edge between x and z, denoted as $E(x, z) = 1$; and 3. There exists a non-edge between w and z, and a non-edge between x and y, denoted as $E(w,z) = 0$ and $E(x,y) = 0$ respectively. If the quadruple of nodes meets these criteria, we then swap the edges by setting $E(w,y) = E(x,z) = 0$ and $E(w,z) = E(x,y) = 1$.

We sufficiently randomize an edge set between two partitions by selecting a pair of nodes from each of the two partitions at random and swapping the edges between these nodes if the above criteria are satisfied. This process is repeated until the number of successful swaps is twice the total number of edges. The number of swaps required to sufficiently randomize a graph was determined by measuring the hamming distance from a representative graph after each swap operation was performed.

4 DISCUSSION

The problem of efficient computation of tightly connected clusters in a network has been studied in several biological as well as non-biological contexts. As we have argued, however, the current approaches are either (i) computationally inefficient, (ii) detect *one optimal* cluster, (iii) find a few *disjoint* bi-clusters, or (iv) do not discriminate against ubiquitously connected nodes. The trivial approach to mask the best solution and repeat the process to find other solutions leaves us with the problem of finding the best way to mask current clusters and is not at all obvious. All previous applications in biology are limited to two partitions, typically

genes and expression conditions and there remains a need to extend this to multiple partitions. Our emphasis has been on developing an adaptable and general approach to finding meaningful clusters in a collection of interrelated heterogeneous datasets.

The problem of identifying dense subgraphs in a general graph (not necessarily a multipartite graph) has been studied in other contexts using combinatorial approaches. These approaches aim at finding the optimum (densest) subgraph. One can model this problem in a way that is amenable to a *Monte Carlo Markov Chain* (*MCMC*) technique, like *Gibbs sampling*. Briefly, we can model the edges in the graph as being generated by two distinct probability distributions depending on whether the edge belongs to the (unknown) dense subgraph or not. The unknown parameters including the edge probabilities and the cluster membership can be iteratively estimated. In fact one can also design an *Expectation Maximization* (*EM*) using the above setup. Although we have modeled the problem as a Markov chain, we have decided to search for a locally optimal cluster using a stochastic hill-climbing approach. The main reason for this is the adaptability/generality of the approach to a graph with arbitrary number of partitions and arbitrary connectivity. Any given problem domain entails different types of entities (partitions) with a different level of connectivity between partitions. It was thus important to design the method in a configurable fashion and our particular approach allows that. We have not discussed, due to lack of space, the various configuration parameters that our current implementation allows. For instance, in principle, we can have a specific schedule for selecting edges from different partitions to influence the detected clusters if we had an *a priori* knowledge. Our cluster score can be easily extended to weight edges or weight partitions and this kind of adaptability is difficult to achieve with a more standard approach like EM or Gibbs sampling. Our current implementation is 'work in progress' and this work illustrates the utility of such a tool. A fully configurable tool for finding dense subgraphs will be published in future work.

Efficient generation of seeds presents the computational bottleneck. We have followed a simple enumerative approach, given the seed size relative to different partitions. For a seed size of k in one of the partitions, we enumerate all k-vertex sets in that partition and look for neighboring vertices in other partitions in search of a seed above a specified size. This can become prohibitive for a partition with several hundred vertices and $k > 4$. By carefully choosing the partition to enumerate over, we have tried to counter this problem to some extent.

There are very few examples of experimentally determined transcriptional modules, thus making a large-scale evaluation of computational methods difficult. However, we have shown using a variety of validation approaches, that (i) the cluster scores are highly significant, (ii) we can detect almost all of the established TFs involved in Skeletal muscle specific expression, (iii) almost all of the highly frequent TFs have literature evidence for involvement in Skeletal muscle gene regulation, (iv) using a TF-Gene-GO graph, we can successfully assign function to TFs, (v) in a combined set of 3 tissues, the detected transcriptional modules support evolutionary relationship between Cardiac, Skeletal and Smooth muscle, and (vi) novel hypotheses regarding TF-Pathway and Tissue-Pathway can be generated using our approach.

Besides applying our tool to additional datasets, our future plan includes (i) Extensive simulation studies and incorporation of other score functions that account for edge weights, and (ii) extending the current implementation to a graph with arbitrary number of partitions.

ACKNOWLEDGEMENTS

L.W. was supported by an NIH postdoctoral training grant in Computational Biology Authors wish to thank Larry Singh for his comments on the manuscript. S.H. was supported by the University of Pennsylvania startup funds.

REFERENCES

Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. *et al.* (2003) 'Toucan: deciphering the cis-regulatory logic of coregulated genes'. *Nucleic Acids Res.*, **31** (6), 1753–64.

Alexe,G., Alexe,S., Foldes,S. and Hammer,P. (2000) Consensus algorithm for generation of all maximal bi-cliques. DIMACS tech report: 1–20.

Aruga,J., Mizugishi,K., Koseki,H., Imai,K., Balling,R. *et al.* (1999) 'Zic1 regulates the patterning of vertebral arches in cooperation with Gli3'. *Mech. Dev.*, **89** (1–2), 141–50.

Bolouri,H. and Davidson,E. H. (2002) 'Modeling DNA sequence-based cis-regulatory gene networks'. *Dev. Biol.*, **246** (1), 2–13.

Dhillon,I. S. (2001) *Co-clustering documents and words using bipartite spectral graph partitioning*. Knowledge Discovery and Data Mining.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Endo,I., Inoue,D., Mitsui,T., Umaki,Y., Akaike,M. *et al.* (2003) 'Deletion of vitamin D receptor gene in mice results in abnormal skeletal muscle development with deregulated expression of myoregulatory transcription factors'. *Endocrinology*, **144** (12), 5138–44.

Ernst,J., Heun,V. and Voll,U. (2002) *Generalized Clustering of Gene Expression Profiles—A Spectral Approach*. International Conference of Bioinformatics, INCOB'02, Bangkok, Thailand.

Flake,G. W., Lawrence,S., Giles,C. L. and Coetzee,F. (2000) 'Self-Organization of the Web and Identification of Communities'. *IEEE Computer*, **35** (3), 66–71.

Grandori,C. and Eisenman,R. N. (1997) 'Myc target genes'. *Trends. Biochem. Sci.*, **22** (5), 177–81.

Hannenhalli,S. and Levy,S. (2003) 'Transcriptional regulation of protein complexes and biological pathways'. *Mamm Genome*, **14** (9), 611–9.

Harris,M. A., Clark,J., Ireland,A., Lomax,J., Ashburner,M. *et al.* (2004) 'The Gene Ontology (GO) database and informatics resource'. *Nucleic Acids Res.*, **32** (Database issue), D258–61.

Ihmels,J., Friedlander,G., Bergmann,S., Sarig,O., Ziv,Y. *et al.* (2002) 'Revealing modular organization in the yeast transcriptional network'. *Nat. Genet.*, **31** (4), 370–7.

Irrcher,I. and Hood,D. A. (2004) 'Regulation of Egr-1, SRF, and Sp1 mRNA expression in contracting skeletal muscle cells'. *J. Appl. Physiol.*, **97** (6), 2207–13.

Kadonaga,J. T. (2004) 'Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors'. *Cell*, **116** (2), 247–57.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) 'The KEGG databases at GenomeNet'. *Nucleic Acids Res.*, **30** (1), 42–6.

Levy,S. and Hannenhalli,S. (2002) 'Identification of transcription factor binding sites in the human genome sequence'. *Mamm. Genome*, **13** (9), 510–4.

Numoto,M., Yokoro,K., Yasuda,S., Yanagihara,K. and Niwa,O. (1997) 'Detection of mouse skeletal muscle-specific product, which includes ZF5 zinc fingers and a VP16 acidic domain, by reverse transcriptase PCR'. *Biochem. Biophys. Res. Commun.*, **236** (1), 20–5.

OOta,S. and Saitou,N. (1999) 'Phylogenetic relationship of muscle tissues deduced from superimposition of gene trees'. *Mol. Biol. Evol.*, **16** (6), 856–67.

Parks,C. L. and Shenk,T. (1996) 'The serotonin 1a receptor gene contains a TATA-less promoter that responds to MAZ and Sp1'. *J. Biol. Chem.*, **271** (8), 4417–30.

Perkins,K. J., Burton,E. A. and Davies,K. E. (2001) 'The role of basal and myogenic factors in the transcriptional activation of utrophin promoter A: implications for therapeutic up-regulation in Duchenne muscular dystrophy'. *Nucleic Acids Res.*, **29** (23), 4843–50.

Schug,J., Schuller,W. P., Kappen,C., Salbaum,J. M., Bucan,M. *et al.* (2005) 'Promoter features related to tissue specificity as measured by Shannon entropy'. *Genome Biol.*, **6** (4), R33.

Senger,K., Armstrong,G. W., Rowell,W. J., Kwan,J. M., Markstein,M. *et al.* (2004) 'Immunity regulatory DNAs share common organizational features in Drosophila'. *Mol. Cell*, **13** (1), 19–32.

Song,J., Murakami,H., Tsutsui,H., Tang,X., Matsumura,M. *et al.* (1998) 'Genomic organization and expression of a human gene for Myc-associated zinc finger protein (MAZ)'. *J. Biol. Chem.*, **273** (32), 20603–14.

Stormo,G. D. (2000) 'DNA binding sites: representation and discovery'. *Bioinformatics*, **16** (1), 16–23.

Su,A. I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K. A. *et al.* (2004) 'A gene atlas of the mouse and human protein-encoding transcriptomes'. *Proc. Natl Acad Sci. USA*, **101** (16), 6062–7.

Tanay,A., Sharan,R., Kupiec,M. and Shamir,R. (2004) 'Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data'. *Proc. Natl Acad Sci. USA*, **101** (9), 2981–6.

Thompson,W., Palumbo,M. J., Wasserman,W. W., Liu,J. S. and Lawrence,C. E. (2004) 'Decoding human regulatory circuits'. *Genome Res.*, **14** (10), 1967–74.

Wasserman,W. W. and Fickett,J. W. (1998) 'Identification of regulatory regions which confer muscle-specific gene expression'. *J. Mol. Biol.*, **278** (1), 167–81.

Wimmer,U., Wang,Y., Georgiev,O. and Schaffner,W. (2005) 'Two major branches of anti-cadmium defense in the mouse: MTF-1/metallothioneins and glutathione'. *Nucleic Acids Res.*, **33** (18), 5715–27.

Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) 'TRANSFAC: a database on transcription factors and their DNA binding sites'. *Nucleic Acids Res.*, **24** (1), 238–41.

Yeger-Lotem,E., Sattath,S., Kashtan,N., Itzkovitz,S., Milo,R. *et al.* (2004) 'Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction'. *Proc. Natl Acad Sci. USA*, **101** (16), 5934–9.