Analysis of genetic association in Listeria and Diabetes using Hierarchical Clustering and Silhouette Index

Inti A. Pagnuco [1,2,3], Juan I. Pastore [1,2,3], Guillermo Abras [1], Marcel Brun [1,2], Virginia L. Ballarin [1]

[1]Digital Image Processing Group, School of Engineering, UNMdP
[2] Department of Mathematics, School of Engineering, UNMdP
[3] CONICET

E-mail: intipagnuco@fi.mdp.edu.ar

Abstract. It is usually assumed that co-expressed genes suggest co-regulation in the underlying regulatory network. Determining sets of co-expressed genes is an important task, where significative groups of genes are defined based on some criteria. This task is usually performed by clustering algorithms, where the whole family of genes, or a subset of them, are clustered into meaningful groups based on their expression values in a set of experiment.

In this work we used a methodology based on the Silhouette index as a measure of cluster quality for individual gene groups, and a combination of several variants of hierarchical clustering to generate the candidate groups, to obtain sets of co-expressed genes for two real data examples. We analyzed the quality of the best ranked groups, obtained by the algorithm, using an online bioinformatics tool that provides network information for the selected genes.

Moreover, to verify the performance of the algorithm, considering the fact that it doesn't find all possible subsets, we compared its results against a full search, to determine the amount of good co-regulated sets not detected.

1. Introduction

New technologies for genomic analysis measure, simultaneously, the expression of thousand of genes. An important goal in some studies is to find sets of co-expressed genes, to study co-regulation and biological functions. This task is usually performed by clustering algorithms, where the whole family of genes, or a subset of them, are clustered into meaningful groups based on their expression values in a set of experiment. These techniques provide insight on the possible co-regulation between genes, under the hypothesis that co-expression may indicate evidence for co-regulation, and these hypotheses of co-regulation must eventually be corroborated or be rejected by further experiments. Existing clustering algorithms are based, usually, on a global approach [1, 2, 3, 4, 5, 6], or on just pairwise analysis [7]. While some of them provide methods for the selection of the best number of clusters [2], they result in very large number of clusters, depending on the number of groups required or determined by the algorithm. The large size of these sets makes improbable their use as hypothesis generator.

A more comprehensive approach consist on the analysis of co-expression of all possible sub-sets of genes. While simple, this approach fails because of the need to analyze an exponential number

of candidate sets. This search of meaningful set would be impractical if the number of available genes for the analysis is large, which it is usual in this field.

In this work we present a new algorithm, that combine the particularities of clustering hierarchical algorithm with individual cluster validation based on the Silhoutte index, to generate a ranked list of gene group, avoiding the exhaustive search, but providing high quality results. As in [2] we use the Silhouette index as a quality/homogeneity measure, but instead of using it to generate the partitions, we use it to select the best sets, from several partitions, obtained by different variants of hierarchical clustering.

This approach permits the use of many different clustering algorithms, combining the best results of each one of them, avoiding the use of a full search approach, while still providing, as shown in the results, a good approach to the optimal results. In the next section we present a) an introduction to pattern recognition tools, b) the proposed algorithm, c) the result in genomic data. The final conclusions show that this algorithms could result in a useful tool to the researchers as preliminar technique to data analysis.

2. Approach

2.1. Pattern Recognition

Pattern Recognition techniques have been widely used to identify object according to different type of features [8]. There are two basic types of pattern classification, supervised and unsupervised classification. Supervised classification uses samples previously classified to design the classifier, to be applied to new patterns. Unsupervised classification groups unclassified patterns based on similarity. Clustering algorithms is a collection of techniques for unsupervised classification [9]. The clusters are formed according a similarity measure, which is usually defined as the proximity of the points according to a distance function. One of the distance function often used is the Euclidean distance:

$$d(x,z) = \| x - z \| = \sqrt{(x-z)^t(x-z)} \tag{1}$$

where x and z are two vectors. In this case, when minor is the distance between two elements, major is their *similarity*.

In this work, the objects to be classified are genes. The data consists of a set of m samples $S_1, S_2, .., S_m$ and n genes $g_1, g_2, ..g_n$ that are normally represent by a matrix of two dimension M where $M(i,j)$ represent the expression of gen g_i for the sample S_j. The expression of each gen g_i, across all samples, correspond to a row of matrix M, and it is represented by a feature vector $X_i(x_{i_1}, x_{i_2}, .., x_{i_m})$, where each value x_{i_j} represent the expression of geng_i for the sample S_j. The feature vector is then composed by m samples, so that the regions of the partition H are in space R_m.

Each gen can be assigned to one of k possible groups, and the result of a clustering algorithm is a partition of the set of genes, as a set $W = W_1, W_2, .., W_k$, where each W_i is a group of genes with certain degree of similarity.

2.2. Hierarchical Clustering

The Hierarchical Clustering is one of most used clustering algorithms in bioinformatics [10]. In this algorithm, it is not necessary to know previously the number of classes, to divide the total set of elements according their characteristics, since it generates a partitioning tree that can be later used to generate any amount of clusters. At each iteration, the two closest clusters are joined together (based on some measure of distance) and form a new one. If the number of classes is known beforehand, the algorithm ends when you reach the K number of classes searched. If the number of classes is not previously known, the algorithm can be continue its process until there is only one large cluster with all the elements on it. The staggered manner in

which the clusters are gathering can be displayed on a tree diagram called dendrogram, where successive junctions between groups are showed with increasing distance from the root. Given a set of N items to be clustered, and an $N \times N$ distance (or similarity) matrix, the basic process of hierarchical clustering [11] is described here:

(i) Start by assigning each item to a different cluster, so that for N items, there are N clusters, each containing just one item. Let the distances (similarities) between the clusters be the same as the distances (similarities) between the items they contain.
(ii) Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now there is one cluster less.
(iii) Compute distances (similarities) between the new cluster and each of the old clusters.
(iv) Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be done in different ways, which is what distinguishes single-linkage from complete-linkage and average-linkage clustering. In single-linkage clustering (also called the connectedness or minimum method), the *distance* between two clusters is the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the *similarity* between two cluster is the largest similarity from any member of one cluster to any member of the other cluster. In complete-linkage clustering (also called the diameter or maximum method), the distance between two clusters is the largest distance from any member of one cluster to any member of the other cluster. In average-linkage clustering, the distance between two clusters is the average distance from any member of one cluster to any member of the other cluster.

2.3. Quality measures

A common problem of clustering algorithms is the validation of results. There are basically two types of validation. Internal validation, which is based on calculations performed on the resulting clusters such as the separation between them, its roundness or closely packed. This type of validation does not require additional information to that reviewed. The other is the comparison of partitions, which can be obtained compared with other partitions generated by the same algorithm with different parameters (relative validation), or with the actual partition of the original data (external validation) [12]. One of the methods of internal validation used is Silhouette index, which measures the quality of clustering as the average quality of its elements [13]. The Silhouette value for an element ranges between -1 and 1. A high Silhouette value indicates that the element is closer to its own cluster elements than to the ones that do not belong to its own cluster.

To define the Silhouette index for the points, two measures need to be defined first. Lets N be the number of clusters. Let $X \in C_k$ a point (element) that belongs to the cluster C_k, and let n_k the number of elements of C_k, then the first measure $a(X)$, the average distance of X to the points of C_k is defined by:

$$a(X) = \frac{1}{n_k - 1} \sum_{Y \in C_k, Y \neq X} d(X, Y)$$

The second measure, $b(X)$, the average distance to the nearest cluster, for a point $X \in C_k$, is defined by:

$$b(X) = \min_{h=1..N, h \neq k} \quad \frac{1}{n_h} \sum_{Y \in C_h} d(X, Y)$$

Finally, the Silhouette index $S(X)$ is defined by:

$$S(X) = \frac{b(X) - a(X)}{max(a(X), b(X))}$$

Then, the Silhouette index for a cluster is defined as the average Silhouette of its points:

$$S(C_h) = \frac{1}{n_k} \sum_{X \in C_k} S(X)$$

It can be seen that if $b(X) \gg a(X)$ (the point is closer to points in the same cluster than points in other clusters) then $S(X)$ is close to 1. If this happens for most of the points in a cluster C, then the Silhouette $S(C)$ is also close to one, indicating that it is isolated from other clusters.

In the other hand, if $b(X) \approx a(X)$, (there are points outside the cluster closer to X than points in the same cluster), then $S(X)$ is close to 0. If this happens for most of the points in a cluster C, then the Silhouette $S(C)$ is also close to zero, indicating that it is mixed with some other clusters.

For our analysis, we use a slightly modified measure of cluster silhouette. In this case, for each set C of points, to compute its Silhouette we assume that all the points that do not belong to C belong to a second cluster C^c, and the *modified* silhouette index is computed based on this 2-clusters situation:

$$b(X) = \frac{1}{n^c} \sum_{Y \in C^c} d(X, Y)$$

3. Methods

3.1. Proposed algorithm

The objective of this study is to select subsets of genes highly correlated. In the expression profile space of genes (expression through all samples), provide by a distance function, where each genetic profile is represented with a point of R_n, this corresponds to find compact sets that are separated to other points sets. A naive way to find these compact and separate groups would consist in *measuring* the compactness of all possible subset of the N profiles. A suitable measure to measure compactness may is the Silhouette Index [13], successfully used by the authors for other purposes [14, 15].

One problem of this approach is that the amount of subsets grows up in a combinatorial form with the number of genes. For example for a set of 2000 genes, the number of subsets to evaluate is 2^{2000}. For this reason it is required some technique to find the best subsets (or many of them) avoiding the exhaustive search. To solve this problem, in this paper we propose to limit the family of subsets where the search, using hierarchical clustering to generate a family of candidate subsets, and then evaluate the *modified* Silhouette index only on those subsets. An application of the hierarchical clustering algorithm generates a total of $2 * N$ subsets to process. Because there are so many variants of hierarchical clustering, an important problem is to determine the best variants of hierarchical clustering for the task of detecting co-expressed genes on genomic data.

In a previous work [16] we studied the ability of variants of hierarchical clustering (complete-linkage, average-linkage, single-linkage) to detect the best subsets using Silhouette index as quality measure. To this purposes we used four different sets of simulated data with known result. Keeping the size of the artificial sets small, we were able to analyze all possible groups, sorting them by Silhouette Index, and use that information as gold standard to evaluate the performance of variants of hierarchical clustering to detect the best subset.

As an additional approach, we combine all the variants, forming a large sets of candidate groups, still smaller than the maximum 2^N sets, and select the best groups from this set. This approach extends the search space, relative to the use of only one algorithm, maintaining the computational requirements still low. Another advantage of this approach is that new clustering

algorithms may be added to the pool, generating more candidate groups, therefore improving the overall quality of the result.

The main steps of the proposed algorithm are:

- Apply to the data some variants of hierarchical clustering.
- For each algorithm, register all the possible branches in the dendrogram three (one-gene groups are not processed).
- Compute the Silhouette index for each group (against the rest of the genes)
- Select those groups with Silhouette index above a threshold

The important step in this method, that sets it apart from previous methods, is that it does not use just one clustering (partition) of the genes, based on an algorithm, but a larger set of groups, defined by the dendrogram tree from the hierarchical clustering algorithms. Allowing the evaluation (via the modified Silhouette index) of intermediate groups (ones not showing on a *optimal* partition) avoids issues were a large group is not part of that partition, but would be a good candidate because of its compactness.

In the previous work, the analysis was aimed at finding which variant of hierarchical method detected a major proportion of the groups with highest Silhouette (determined by a full search in simulated contexts). In these analysis, and additional ones realized for this work, it was visible that no variant has better performance, except when using a combination of all the variants.

In the next sections we apply this approach, based on the combination of variants to generate the candidate clusters, and the modified Silhouette index to rank, to search for correlated genes in two real data situations.

4. Application Examples

In this section we describe two application examples, where the algorithm is applied to two different sets of data: microarray based expression, for diabetes, and QTLs, for listeria. In both cases we applied the algorithm to detect significant sets of genes/QTLs, and analyzed their significance, based on existing knowledge about them.

4.1. Listeria

In the first application example, we used data of mouse susceptibility to monocytogenes listeria [17, 18]. This study consists in the analysis of the relation between QTLs (*Quantitative trait Loci*, stretches of DNA containing or linked to the genes that underlie a quantitative trait) and the *survival* time of 120 mouses that have been infected with listeria, where *survival* is defined by a survival time of more than 240 hours. The dataset consists in 35 surviving mice and 85 un-surviving, analysing 133 QTLs for each mouse. After filtering those QTLs with missing data, only 28 QTLs were retained.

It is important to note that here we are using QTLs instead of genes as target for our search of correlation. Because of the nature of the model used here is not restricted to expression information, the same analysis can be applied to discrete QTL studies.

To verify the performance of the algorithm in this context, considering the small amount of markers under study, we were able to perform a full search of QTLs groups (up to 7 elements), computing the Silhouette index of all of these subsets. The performance of the search algorithm (based on clustering) is measured based on the number of best subsets detected.

Table 1 shows the comparison of results between full search and combination of variants of hierarchical clustering. We can see in the table that the 8 most significant groups were properly detected by the algorithm. Still more, the top groups of 2, 3, 4, 5 and 6 elements, were successfully detected by the algorithm.

Table 1. Comparison of results between full search and combination of variants of hierarchical clustering. The best 8 groups were successfully detected by combination algorithm.

Id_All	Sil	Id_combination	Num_elem	Groups Member
1	0.90661	1	2	8,9
2	0.86923	2	2	3,4
3	0.86788	3	2	1,2
4	0.814008	4	2	25,26
5	0.736231	5	3	8, 9, 10
6	0.701954	6	3	24, 25, 26
7	0.672453	7	2	11,12
8	0.66783	8	4	1,2,3,4
18	0.592628	9	2	14,15
21	0.563286	10	2	17,18
24	0.554229	11	5	8,9,10,11,12
37	0.49119	12	3	16,17,18
44	0.452103	13	6	8,9,10,11,12,13
69	0.393421	14	5	14,15,16,17,18
71	0.387993	15	4	23,24,25,26
327	0.293433	17	6	14,15,16,17,18,20
346	0.290286	18	6	6,14,15,16,17,18

4.2. Diabetes

In this second example of biological application we run the algorithm on microarray data. This data was obtained from a previous study that analyzed the expression profiles of obese and thin subject [19]. This study presents the expression profile to 18 subjects, 13 with obesity and 5 without, using a U133A chip of Affymetrix. From the original dataset, with 22283 genes, only the most variable 1000 genes were preselected for the analysis. In this case, it is not possible to compare the resulting groups with a full search, since there are 2^{1000} possible subsets. Therefore, the resulting groups/sets were studied based on actual biological knowledge about them.

We restricted the analysis to the top eight ranked sets, described in Table 2. We verified if the genes found on these groups, have assigned similar biological functions.

It should be noted that in group 1, even if there are two different Affymetrix probes 204550_x_at and 215333_x_at, they make reference to the same gene. The same situation is repeated for group 6.

Table 2. Sets detected with higher Silhouette using a combination of variants of hierarchical algorithms.

Group	Silhouette	Size	Probes (Affy Ids)	Analysis
1	0.9548	2	204550_x_at , 215333_x_at	Unique gene
2	0.9429	3	204418_x_at , 204550_x_at , 215333_x_at	2 Genes from the same family.
3	0.82689	2	207831_x_at , 207907_at	No relationship found.
4	0.80904	2	201639_s_at , 201904_s_at	No relationship found.
5	0.79401	2	205175_s_at , 213670_x_at	Indirect metabolic Relationship
6	0.79394	2	200966_x_at , 214687_x_at	Unique gene
7	0.79345	2	200991_s_at , 202676_x_at	Common function:Protein binding
8	0.79343	3	201379_s_at , 207831_x_at , 207907_at	Common function:Protein binding

The most interesting case identified is Group 2. In this group there are 3 probes, 2 of them reference to the same gene called GSTM1 (Glutathione S-transferase mu 1). The relation

between these 2 probes and the third one, which references to gene GSTM2 (Glutathione S-transferase mu 2) is that all of them are members of the same family, and they are involved in metabolic process.

We analized these genes with the GeneMania bioinformatic tool, using the Co-expression option. GeneMania searches large biological datasets to find related genes, including protein-protein, protein-DNA and genetic interactions, pathways, reactions, gene and protein expression data, protein domains and phenotypic screening profiles [20] . Figure 1 show the result of the analysis of co-expression for groups 2, 5 and 66. The figure 1 shows a large amount of linkage between the genes of the selected groups. Most of the associations are obtained from the Gene Expression Omnibus database (GEO), but GeneMania tool only collect data associated with publications. To group 2 some functions associated are glutathione transferase activity, glutathione derivative metabolic process, peptide metabolic process, peptide binding, modified amino acid binding, etc.

To third and forth group have not documented relationships. From this analysis, they could be good candidates for further analysis of co-regulation, or other biological relationship.

The fifth group shows interesting relationships. The first Affymetrix Identifier (205175_s_at)reference to a locus NSUN5P1. This locus represents a transcribed pseudo gene of a nearby locus on chromosome 7, which encodes a putative methyltransferase. Diseases associated with NSUN5P1 include Williams-Beurem Syndrome. The other group member references to the KHK gene (213670_x_at). The KHK gene encodes ketohexokinase that catalyzes conversion of fructose to fructose-1-phosphate. The product of this gene is the first enzyme with a specialized pathway that catabolizes dietary fructose (GeneCard information [21]). Due to the fact that NSUN5P1 is a pseudo gene, we can not do an analysis of relation with the KHK gene (probe 213670_x_at). For this reason we search genes associated to Williams-beuren disease with pathways related to the KHK gene.

In this search we found the MLXIPL gene that encodes a basic helix-loop-helix leucine zipper transcription factor. This protein forms a herodimeric complex and binds and activates, in a glucose-dependent manner, carbohydrate response element (ChoRE) motifs in the promoters of triglyceride synthesis genes. This gene is deleted in Williams-Beurem Syndrome. Both genes are involved in metabolism process. Figure 1 show the result of a co-expression analysis between MLXIPL and KHK gene, using the GeneMania bioinformatic tool.

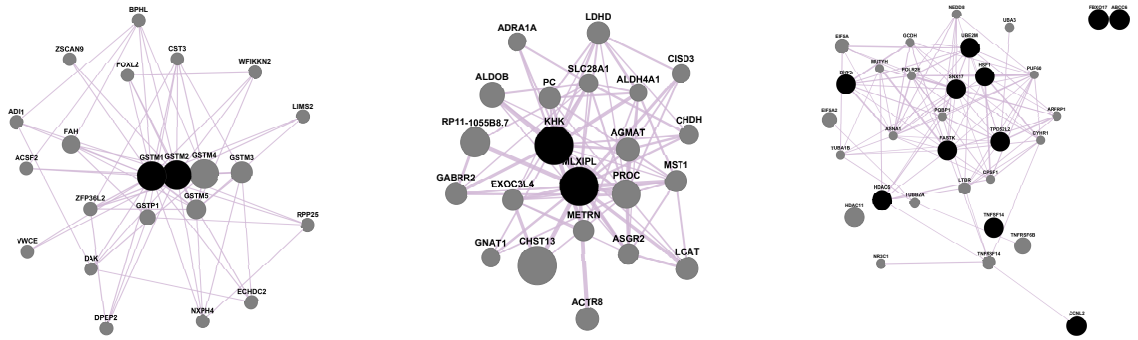
For groups 7 and 8 we used GeneCards for analysis. GeneCards is a searchable, integrated database of human genes that provides information on all known and predicted human genes [21]. For both groups their members have a common function, that is protein binding, according to GeneCards.

Other case analyzed was a big group with eleven elements, this group had been ranked in the position 66, and its Silhouette index is 0.7044, large value based on our past experience with the Silhouette index. We searched the elements and ran a co-expression analysis with GeneMania. In figure 1 show the relation between some elements. To verify the quality of the result we select randomly 11 genes and ran the same analysis, the result obtained was a graph more dispersed (Graphic not included here).

5. Discussion

In this work we apply a simple but powerful method to identify groups of genes/markers that are compact and well separated from other genes/markers, using the Silhouette index to rank the sets, and a combination of several variants of hierarchical clustering to search the best sets. In one case, due to the small size, the resulting groups were compared to the results from full search. In the second case, the resulting groups were analyzed using standard bioinformatics tools, verifying strong relationship between the genes in the top groups.

The algorithm provides a balance between search time and detection rate. It avoids the full



(a) (b) (c)

Figure 1. Results from GeneMania tool: a) Graph for group 2, where GSTM1 and GSTM2 are closely related; b) Graph for group 5, where KHK and MLXIPL gene are closely related; c) Graph for group 66, where the 11 elements are related

search, which can be impractical for large number of genes, with the cost of missing some good sets, but it is able to detect most of the top ranked sets, which is not usually possible by using only one clustering algorithm.

The effectiveness of this algorithm is related to the ability of the Silhouette index to score properly compact groups, which are at the same time separated from other groups of genes.

6. Conclusion

With this analysis we verified the results of the proposed tool, and we considered that is useful for biologists or researchers in computational biology interested in generating new hypotheses about the co-expression of genes, or genomic markers like QTLs, which are not provided in most standard analysis tools. This algorithm will generate quickly a set of good groups on base of Silhouette index, and combines the advantages of each variant of hierarchical clustering algorithm. Future work includes the analysis of other indices of group quality, besides Silhouette, and the application to new datasets, including SNPs.

Acknowledgement

This work was partially supported by CONICET and FONCYT.

7. References

[1] L. Heyer, S. Kruglyak, and Shibu Yooseph. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 1999.

[2] K. S. Pollard and M. J. van der Laan. New methods for identifying significant clusters in gene expression data. *Joint Statistical Meetings - Biometrics Section-to include ENAR & WNAR*, 2002.

[3] H. Peng, F. Long, M. B. Eisen, and E. W. Myers. Clustering gene expression patterns of fly embryos. *IEEE, ISBI*, 2006.

[4] T. Nguyen and R. Nowakowski I. Androulakis. Unsupervised selection of highly coexpressed and noncoexpressed genes using a consensus clustering approach. *OMICS*, 2009.

[5] T. Nguyen, J. Mattick, Q. Yang, M. Orman, M. Ierapetritou, F. Berthiaume, and I. Androulakis. Bioinformatics analysis of transcriptional regulation of circadian genes in rat liver. *BMC Bioinformatics*, 2014.

[6] W. De Mulder, M. Kuiper, and R. Boel. Clustering of gene expression profiles: creating initialization-independent clusterings by eliminating unstable genes. *Journal of Integrative Bioinformatics*, 2010.

[7] A. Feltus, S. Ficklin, S. Gibson, and M. Smith. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an arabidopsis case study. *BMC System Biology*, 2013.

[8] *Pattern Clasification and Scene Analysis*. Wiley-Interscience; 2 edition (November 9, 2000), 2000.
[9] *Clustering: revealing intrinsic dependencies in microarray data*. 2005 Hindawi Publishing Corporation, 2005.
[10] D. Chiang, P. Brown, and M. Eisen. Visualizing associations between genome sequences and gene expression data using genome-mean expression profile. *Bioinformatic*, 2001.
[11] Johnson S. C. Hierarchical clustering schemes. *Psychometrika*, 1967.
[12] Dalton L., Ballarin V., and Brun M. Clustering algorithms: On learning, validation,performance, and applications to genomics. *Current Genomics*, 2009.
[13] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987.
[14] J. Pearson and et al. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single nucleotide polymorphism association studies. *The American Journal of Human Genetics*, 2007.
[15] J. Hua, D. Craig, M. Brun, J. Webster, W. Tembe, K. Joshipura, M. Huentelman, E. Dougherty, and D. Stephan. Sniper-hd: improved genotyped calling accuracy by an expectation-maximization algorithm for high-density snp arrays. *Bioinformatic*, 2007.
[16] G. Abras, J. Pastore, M. Brun, and V. Ballarin. Deteccin de conjuntos significativos de genes via silohuette. *CAIS 2010, 1er Congreso Argentino de Informtica y Salud, 38 JAIIO*, 2010.
[17] V. L. Boyartchuk, K. W. Broman, R. Mosher, S. D'Orazio, M. Starnbach, and W. F. Dietrich. Multigenetic control of listeria monocytogenes susceptibility in mice. *Nature Genetics*, 2001.
[18] K. W. Broman, V. L. Boyartchuk, and W. F. Dietrich. Mapping time-to-death quantitative trait loci in a mouse cross with high survival rates. *Technical Report MS00-04, Department of Biostatistics, Johns Hopkins University*, 2000.
[19] J. Pihlajamki, T. Boes, and Eun-Young Kim et al. Thyroid hormone-related regulation of gene expression in human fatty liver. *J Clin Endocrinol Metab*, 2009.
[20] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 2008.
[21] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. Genecards: A novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 1998.