Biological Pathways

– A pathway to explore diseases mechanism

Ming Guo

Abstract

Diseases are increasingly identified with genetic modules, such as molecular pathways. Pathway-level analysis can provide important insights for making biological inferences and hypothesis from genetic data. The computational approaches for incorporating pathway knowledge to interpret high-throughput datasets play a key role in understanding diseases mechanism from genetic studies. This paper is a review of the various methods for inferring pathway information from genetic datasets, as well as comparing pathways for different species.

1. Introduction

More and more genome-wide association studies of diseases have revealed possible connections between DNA sequence variants and various diseases. However, it is common that, only a small proportion of the disease risk is explainable by the single nucleotide polymorphisms (SNPs) identified in these studies with statistically significant difference between control and disease groups. [1]

The genes do not function alone. "For many diseases, multiple genes have been identified to collectively account for clinical phenotypes." [2] "This is most obvious in the case of genetically heterogeneous diseases such as Fanconi anemia, Bardet-Biedl or Usher syndrome, where the various genes work together in a single biological module. Such modules can be a multiprotein complex, a pathway, or a single cellular or subcellular organelle." [3] Li etc systematically studied disease-pathway relationships based on the shared genes between pathways and disease-causing genes. Disease genes are mapped biological pathways where the genes are enriched. [4] They found that, on average, over 50% of disease-causing genes are statistically mapped to pathways. "This finding reinforces the notion that disease genes are related to each other in a form of functional entity such as pathways or protein complexes." [2]

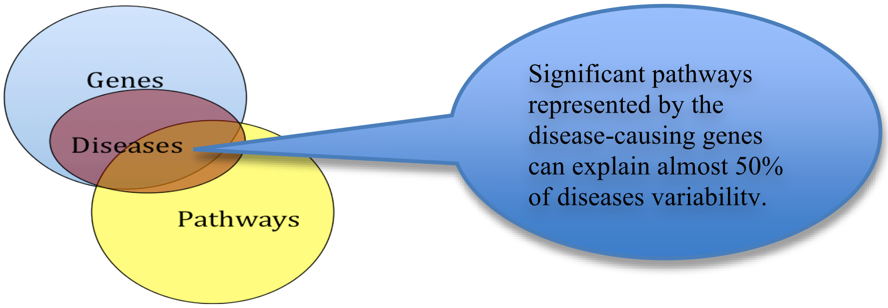
Genes
Diseases
Pathways
Significant pathways represented by the disease-causing genes can explain almost 50% of diseases variability.

Figure 1 Gene-Pathway- Disease relationship diagram

“Until recently, few computational approaches are available for incorporating pathway knowledge to interpret high-throughput datasets.” [5] In the following sections, we will review the current computational methods to map genes identified in high-throughput experiments to the relevant pathways. Such inference may provide some hints to make biologically meaningful hypothesis. We next investigate the impact of variations in gene expression data – such as tissue-specific and patient-specific gene expression – on the interpretation of significant pathways. Last but not least, we discuss the new development in the alignment of pathways from different species.

2. *What is a biological pathway?*

Biological pathways represent the biological reactions and interaction network in a cell. Each reaction is identified with its enzyme, which in turn is coded by certain gene(s). Reactome is a curated knowledgebase of biological pathways. The database includes experimentally confirmed reactions and manually inferred reactions, as well as electronically inferred reactions. The pathways are organized by the conceptual categories, and cross-referenced to a “wide range of standard biological databases, including NCBI Entrez Gene, Ensembl, UniProt databases, UCSC and HapMap Genome Browsers, the KEGG compund and ChEBI small molecule databases, PubMed and GO.” [6]

An example biology pathway is the transcription pathway. Figure 2 below shows part of the reaction diagram:

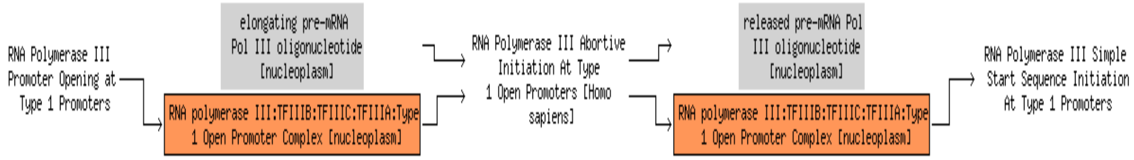
RNA Polymerase III Promoter Opening at Type 1 Promoters
elongating pre-mRNA Pol III oligonucleotide [nucleoplasm]
RNA polymerase III:TFIIIB:TFIIIC:TFIIIA:Type 1 Open Promoter Complex [nucleoplasm]
RNA Polymerase III Abortive Initiation At Type 1 Open Promoters [Homo sapiens]
released pre-mRNA Pol III oligonucleotide [nucleoplasm]
RNA polymerase III:TFIIIB:TFIIIC:TFIIIA:Type 1 Open Promoter Complex [nucleoplasm]
RNA Polymerase III Simple Start Sequence Initiation At Type 1 Promoters

Figure 2 RNA Polymerase III Abortive Initiation At Type 1 Open Promoters [Homo sapiens] [7]

The transcriptional regulators of a pathway can be identified with genome-wide binding analysis (also known as genome-wide location analysis). Investigators can identify the set of target genes bound in vivo by each of the transcriptional regulators that are encoded in a cell’s genome. [8] Lee etc. identified a gene regulatory network from yeast genome using the genome-wide location analysis. “Just as maps of metabolic networks describe the potential pathways that may be used by a cell to accomplish metabolic processes, this network of regulator-gene interactions describes potential pathways yeast cells can use to regulate global gene expression programs.” [8]

It is now becoming increasingly evident that the connection between pathways and diseases is fashioned at multiple interconnected levels and is controlled by an interplay between cell signaling, gene expression, the establishment of multifaceted transcriptional motifs and the temporal and spatial organization of chromatin in loops and domains.

3. Mapping Genes to pathways

The gene-pathway-disease model for understanding diseases mechanism requires first the identification of significant pathways from a set of genes. There are many algorithms that address this issue with different levels of sophistication. They can be grouped into four major types. One is based on over-representation of genes in certain pathways; another type is based on functional scoring of genes, which allows adjustments based on the correlation between phenotype and gene expression data; the third type also incorporates topological information of the pathways; the forth type – the gene set approach – focuses on using functional units comprised of a set of correlated genes to make pathway predictions.

3.1 Over-representation approach

This approach starts with a set of differently expressed genes and uses statistical analysis to identify the Gene Ontology terms that are over-represented in the list of genes.
This approach is limited in their accuracy because they do not incorporate "known interdependencies among genes in a pathway that can increase the detection signal for pathway relevance." [5] In addition, they treat all genes identified from the gene expression data equally, which is not necessarily true. Thus, this approach can "produce many false positives when only a single gene is highly altered in a small pathway." [5]

3.2 Functional score of genes

This type of approaches incorporates functional indicators of the genes identified in a microarray experiment. An example of this is the Gene Set Enrichment Analysis (GSEA). It "ranks all genes based on the correlation between their expression and the given phenotypes, and calculates a score that reflects the degree to which a given pathway is represented at the extremes of the entire ranked list." [9]

Many experiments use different cell lines or different conditions for a certain disease phenotype. Genes may vary in stability depending on the cell type or disease being studied. [10] Thus, one of the drawbacks of the functional score approach is that the correlation with phenotypes could be compromised due to the variance of genes expression in different tissues or cell types. More details on this will be covered in a later section.

3.3 Pathway topology

Pathway topology depicts the genetic interaction, indicating which genes are upstream of other genes, hence are prerequisites for the activation of the downstream genes. Changes in the expression level of the downstream genes may not affect the pathway as much as the upstream genes. Hence, such information is important to be incorporated in identifying the significantly enriched pathways.

Draghici et al. developed an algorithm called Impact Factor, which would identify pathways with "both a statistically significant number of differentially expressed genes and biological meaningful changes" on the pathway. [9] Part of the impact factor calculation is the perturbation factor (PF), which sums up all the PFs of the genes directly upstream of the target gene, normalized by the number of downstream genes, and weighted by the type of interaction (induction or repression). This approach was successfully applied to several data sets - such as genes differentially expressed in lung adenocarcinoma, and was shown to outperform the over-representation approach by providing biologically meaningful results.

3.4 Gene Set Approach

The methods abovementioned aim at inferring pathway representation directly from the gene expression data of each gene. All these approaches implicitly assume each gene as target for enrichment. The gene set approach, on the other hand, treats the known functionally related genes together as a group in calculating statistical significance. The null hypothesis is that genes of the same pathway are not co-regulated more strongly than a randomly selected group of genes without any functional relationship. This is more powerful than the previous approaches because the joint score of gene sets that are known to be in a functional relationship will increase the potential to detect subtle signals in gene expression data. [11] A number of methods have been proposed that work on the level of gene sets.

Rahnenf¨uhrer etc introduced ScorePAGE algorithm, which takes all the genes in a predefined pathway as a gene set. The algorithm analyzes the change of activity of a pathway for different samples or across different time points with respect to some baseline condition. It scores each pathway based on the sum of similarity measurements of all gene pairs within the pathway. The significance level of a pathway is calculated using a nonparametric permutation test by randomly permuting gene label assignments. The optimal scoring measure is adaptively determined based on statistical significance. [11]

The similar approach can be used to extend and refine partial knowledge about a pathway using the available expression data. Ihmel etc developed an algorithm based on co-regulation of genes. This algorithm is referred as the 'signature algorithm'. It receives a gene set that partially overlaps a transcription module – comprised of the co-regulated genes and the experimental conditions that trigger this co-regulation – and then provides the complete module as output. [12] Tanay took the idea one step further by identifying modules, namely, groups of genes with statistically significant correlated behavior across diverse data sources. [13]

This view of using a module of co-regulated gene set to predict functionality is supported by the modular nature of genetic diseases. [3] Spirin and Mirny analyzed the network of molecular interactions formed by proteins, nucleic acids and small molecules in a cell. These modules include protein complexes and dynamic functional units, such as signaling cascades, cell-cycle regulation, etc. They found that "molecular modules are densely connected within themselves, but sparsely connected with the rest of the network.” [14]

4. Variation in Gene Expression

Current genetic studies identify genes that belong to a certain pathway based on the assumption that genes related by a pathway show significant correlation in genetic expression. However, “variation in gene expression is extensive among tissues, individuals, strains, populations and species.” [15] Whitehead etc. conducted a comparison study on a set of 192 metabolic genes in brain, heart and liver. They found that “half of the genes (48%) were differentially expressed among individuals within a population-tissue group and 76% were differentially expressed among tissues.” [15]

4.1 Tissue-specific gene expression

It is common for diseases to be associated with genes expressed only in specific tissue types. For example, “of the oxidative phosphorylation genes differentially expressed between tissues, 92% were more highly expressed in heart or brain than in liver.” [16] This makes sense because “the primary purpose of the heart is to act as a pump, and contraction is highly dependent on oxidative metabolism.” [16] Therefore, the identified significant pathways need to be interpreted in the context of certain tissue type information.

TiGER (**Ti**ssue-specific **G**ene **E**xpression and **R**egulation) is a database for generating comprehensive information about human tissue-specific gene regulation, including both expression and regulatory data. [17]

GeneGo - a leading provider of data mining & analysis solutions in systems biology - combines all tissue-specific information into their pathway maps database. [18] Researchers can build their own custom pathway in a particular tissue and also add “disease effect” interactions to specify how interactions may be perturbed in the specific tissue. In a more general sense, one is able to specify the tissue of interest, then generate a map indicating whether genes in the map or network are known to be expressed in the specified tissue. Figure 3 shows an example pathway map where a “T” appears next to those objects on the map that are known to be expressed in the tissue specified.

4.2 Patient-specific Pathway

Vaske etc. developed an approach called PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models) to infer patient-specific genetic activities incorporating curated pathway interaction among genes. [5] They combined pathway information with multiple genome-scale measurements on a single patient sample to infer the activities of genes, products and abstract process inputs and outputs for a single pathway – the Integrated Pathway Activities (IPAs). A gene is modeled by a factor graph which depict the various types of interactions across genes, including transcription factors to targets, subunits aggregating in a complex, post-translational modification and sets of genes in a family performing redundant functions. The parameters of the model were estimated using an iterative approach based on expectation maximization (EM) algorithm. The EM algorithm converged quickly on the true dataset, which suggests that "the pathway structures and inference are able to successfully identify patterns of activity in the integrated patient data." [5]

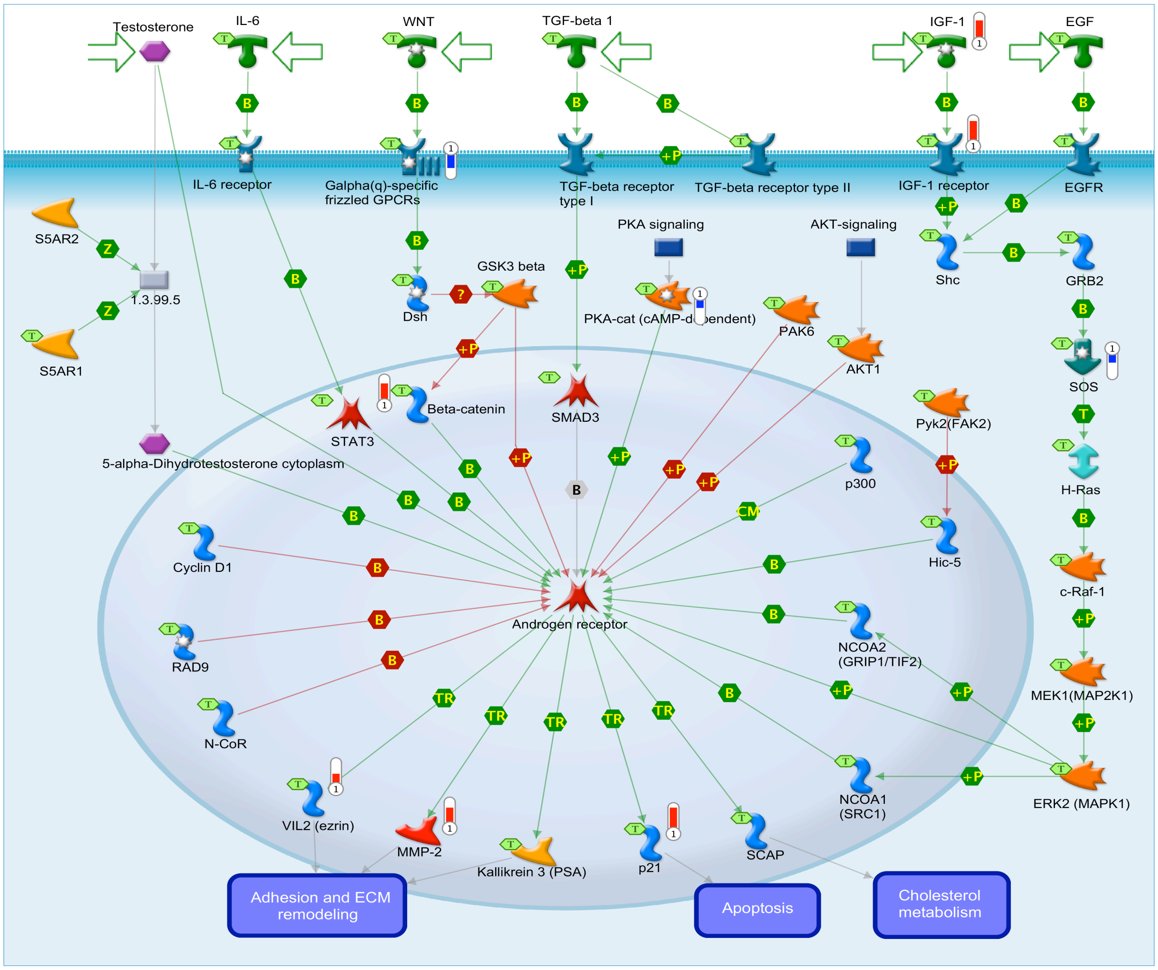

Figure 3 GeneGo Pathway map
with tissue information indicating gene expressed in the selected tissue

The output of the model is the likelihood ratio that a pathway's activities are altered in the patient. Comparing with a competing pathway inference approach called SPIA, PARADIGM was able to identify altered activities in cancer-related pathways with fewer false-positives. [5]

"The power of pathway-based approach is that it may provide clues about the possible mechanisms" underlying the difference in observed phenotypes. [5] However, the robustness of such algorithm is still to be validated as more multidimensional datasets become available in the future.

5. Pathways Update and Alignment

The knowledge of pathways was traditionally scattered throughout the literature and hard to access systematically. In recent years, more and more catalogized pathway knowledge databases have become publicly available. Some of the databases that include pathway topology are Kyoto Encyclopedia of Genes and Genomes (KEGG) [19] MetaCore [20], HumanCyc [21] and the National Cancer Institute (NCI) Pathway Interaction Database [22]. Other websites, such as PathwayCommon [23] is a hub for nine different pathway databases and allow users to search as well as execute other computational operations on top of these pathway databases.

Most of the pathway maps are manually curated by PhD biologists. Keeping these databases updated will require massive collaboration efforts. BioCarta [24] and WikiPathways [25] are two public platforms dedicated to the curation of biological pathways. "Updates to these databases are expected to improve our understanding of biological systems by explicitly encoding how genes regulate and communicate with one another." [5]

Since different pathway databases often use different data structure and terminology, it is often not easy to discover conserved pathways between different species. Gene Ontology (GO) provides a hierarchical structure of concepts on molecular function, biological process, and cellular component. Mapping the genes in a pathway to their ontology terms, and utilizing the semantic structure of Gene Ontology to define the similarity of genes provide a basis for pathways alignment. [26] Gamalielsson, etc. developed an algorithm called GOSAP for finding semantic local alignments when comparing paths in biological pathways where the nodes are gene products. The output is the scored and ranked pathway alignment. This algorithm can be extended to the analysis of pathways where nodes are not only enzymes, but any kind of gene product.

6. Conclusion

Pathway identification and alignment are very useful for inferring biological mechanism of diseases from high-throughput genetic data. Although there is much more future work needed to validate biological impact of significant pathways, the computational tools for pathway analysis will nevertheless help us generate biologically meaningful hypotheses.

As the whole genome sequencing technique becomes more and more easily available, our ability to generate integrated datasets across different 'omics' databases will allow more comprehensive inferences made based on the pathway analysis. These could lead to new ways of understanding diseases mechanism and improvements in treatment.

References

1. Donnelly P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature* 456, 728-731.
2. Loscalzo J, Kohane I, Barabasi AL. (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol* 3: 124.
3. Oti M, Brunner HG (2007) The modular nature of genetic diseases. *Clin Genet* 71: 1–11.
4. Li Y, Agarwal P (2009) A Pathway-Based View of Human Diseases and Disease Relationships. PLoS ONE 4(2): e4346. doi:10.1371/journal.pone.0004346
5. Vaske C., et al. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26: 237-245.
6. http://www.reactome.org/
7. http://www.reactome.org/cgi-bin/eventbrowser?DB=gk_current&FOCUS_SPECIES=Homo%20sapiens&ID=112055&
8. Lee T. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298: 799-804.
9. Draghici S., et al. (2007) A systems biology approach for pathway level analysis. *Genome Res.* 17: 1537-1545
10. Lee S, Jo M, Lee J, Koh SS, Kim S (2007) Identification of novel universal housekeeping genes by statistical analysis of microarray data. *J Biochem Mol Biol*. 2007 Mar 31;40(2):226-31.
11. Rahnenf˙uhrer J., Domingues F., Maydt J., Lengauer T. (2004) Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology* Vol. 3 Issue 1 Article 16
12. Ihmel J., et al. (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics* 31:370-377.
13. Tanay A., Sharan R., Kupiec M., Shamir R. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS* 101:2981-2986.
14. Spirin V., Mirny L. (2003) Protein complexes and functional modules in molecular networks. *PNAS* 100:12123-12128.
15. Whitehead A., Crawford D. (2005) Variation in tissue-specific gene expression among natural populations. *Genome Biology* 2005, 6:R13
16. Weiss L, (Ed) (1983) *Cell and Tissue Biology: A Textbook of Histology* 6th edition. Baltimore, MD: Urban and Schwarzenberg.
17. http://bioinfo.wilmer.jhu.edu/tiger/
18. http://www.genego.com/
19. http://www.genome.jp/kegg/
20. http://www.genego.com/metacore.php
21. http://humancyc.org/

22. http://pid.nci.nih.gov/
23. http://www.pathwaycommons.org/pc/
24. http://www.biocarta.com/support/howto/path.asp
25. http://www.wikipathways.org/index.php/WikiPathways
26. Gamalielsson, J., Olsson B. (2008) GOSAP: Gene Ontology Based Semantic Alignment of Biological Pathways. Int J *Bioinform Res Appl*. 4(3):274-94.