Journal of Biology

BioMed Central

Minireview

Can modular analysis identify disease-associated candidate genes for therapeutics?

Jesper Tegnér

Address: Department of Medicine, Center for Molecular Medicine, Karolinska University Hospital, 171 76 Solna, Stockholm, Sweden.
Email: jesper.tegner@ki.se

Published: 28 May 2009

Journal of Biology 2009, **8**:48 (doi:10.1186/jbiol149)

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/8/5/48

Abstract

Complex diseases such as allergy change gene expression in several cell types and tissues. Benson and colleagues have now shown, in a paper in *BMC Systems Biology*, that this complexity can be studied effectively using an integrated experimental and computational modular analysis. Their strategy revealed a core of allergy-associated genes of potential therapeutic value.

Technologies are an important driver of progress in the medical sciences. Recent advances in array-based and sequence-based instrumentation have opened up new ways to monitor the inner molecular world of the cells and tissues that might be relevant to human diseases. Yet it is far from evident how these large datasets should be analyzed and how they can be integrated with other sources of data in order to become informative. Conversely, the medical community expects nothing less than a list of predictive biomarkers reflecting the risk of disease or its progression and an understanding of the cellular mechanisms involved in disease. However, comparing microarray samples from healthy and diseased individuals using a differential gene expression protocol generates a list of thousands of genes, and it is not clear which genes are important for what.

A key idea, originating from engineering science in general and computer science in particular, is the notion of 'divide and conquer', which refers to first breaking down a problem into smaller sub-problems that are simple enough to allow an analysis and then combining the solutions to the sub-problems, which gives the solution to the original problem. Modular analysis of genomic data implements this strategy by dividing the original genomic data into smaller number of modules and then conquering the reduced complexity by using these modules for prioritization to give a shorter list of disease-associated genes. Such genes could either be causal drivers of disease or secondary reactions to disease that could potentially be useful biomarkers.

Benson and colleagues, in a recent paper in *BMC Systems Biology* [1], have used a modular approach to study allergic asthma. They managed to divide the complexity and arrive at the gene encoding the interleukin-7 receptor (IL7R) as a putative key regulator in allergic asthma. Importantly, their computational analysis is accompanied by experiments. Here, I put their analysis in the context of other modular approaches and discuss the possible use of this methodology for finding and prioritizing useful candidates for therapeutics.

Dividing complex biological data into modules of disease-associated genes

Not surprisingly, there are several different ideas on how to divide and conquer high-throughput functional genomics data. I will restrict my discussion here to gene expression data, although similar remarks could be made for sequence

data. Conceptually there are two distinct problems. One is: given a module of disease-associated genes, how can we compute and/or experimentally predict which genes are good candidates for therapeutics? Before discussing this problem I will first give an overview of different approaches to the other problem: identifying a module of genes.

A module is a group of genes that are related in some way to each other and therefore a module is effectively a measure of similarity. Grouping genes into modules depends on an exact mathematical definition of similarity. For example, if similarity is defined as the distance in a network, then a graph theoretical calculation will be used. However, if gene functional associations are used, then gene similarity will be measured in terms of gene ontology (GO) or correlation in gene expression values. Therefore, different algorithms are used for dividing the genes into modules, a fact that could be confusing for the clinical researcher.

The need to reduce the complexity of the original high-throughput gene expression data was realized early on in its analysis [2]. Applying established engineering concepts, such as principal component analysis (PCA) and singular value decomposition (SVD), reduced the dimensionality of the data. Instead of analyzing scattered points (the samples) in a high-dimensional space equaling the number of genes, the data could thereby be projected into a two- to four-dimensional space. However, it turned out to be difficult to make a biological interpretation of the resulting linear combinations of large numbers of genes. This problem forced the development of different strategies in which the available knowledge on a limited number of genes could be used to predict the functions of as-yet uncharacterized genes.

The use of hierarchical clustering in the classic compendium study on yeast data by Rosetta Inpharmatics [3] grouped genes (shown as rows) by their similarity of expression across several experimental conditions (columns). Novel gene function was then predicted by inspecting genes in the same cluster as genes with known functions. Subsequent work by Eran Segal and colleagues [4] developed more statistically sound procedures for identifying robust modules using a Bayesian formalism applied to microarray data generated from cancer samples.

It became clear, however, that a similarity measure based only on correlations was insufficient, because the clusters (modules) or Bayesian modules did not have an internal network structure that could be used for a more refined analysis. As a consequence, a large number of studies addressing this problem appeared in the literature at the beginning of 2004. The idea was that if we could identify the wiring within cellular networks, various different algorithms could be applied to find 'connected groups' in such networks. Such an analysis would then provide more biological insights into the mechanisms of disease.

Now, how can such networks be found using only a small number of experimental samples with a large number of genes? This is an impossible problem from the point of view of engineering system identification, because the number of possible networks consistent with the data is prohibitively large [5]. The key simplifying insight came from Ideker and Lauffenburger [6] and was later developed by Nicolas Luscombe and colleagues in a pioneering paper [7]. Here, the edges (or connections) in the network were simply defined by transcription factor binding experiments, and gene expression data were used to select the subsets of edges that were active under different conditions.

This idea of defining edges in a network using a static scaffold has since been reused using various data types (protein-protein interaction data, pathways from a database, text mining and DNA variants). The network of interest is then defined by combining the gene expression data with the scaffold, leaving only the active edges. By searching through such an active network using graph algorithms it is then possible to define 'more' connected parts in a well defined manner, thereby providing modules with an intrinsic network structure.

All the above approaches basically begin with a large, complex dataset, which is then simplified by dividing the data into smaller modules. Interestingly, Vidal and colleagues [8] demonstrated that this process can be reversed. They instead began with four well characterized breast cancer genes and, by using these ideas, constructed a module in which the genes were 'close' as defined by expression and proteomic data in several species.

Finding an allergy-associated module

Benson and colleagues [1] have now contributed to a disease-oriented modular analysis by combining several of the above ideas in a novel manner, as summarized in the flow chart in Figure 1. First, because allergic disease involves multiple cells in different tissues and because no prior characterization of key genes was available, they turned to several different sets of gene expression microarray data in order to find a reference disease-associated gene around which they could construct a module. Using the idea that disease-associated genes tend to interact, they could search for other disease-associated genes that were 'close'. For this purpose, the authors used a graph algorithm that identified a connected clique of 103 disease-associated genes from the microarray data.

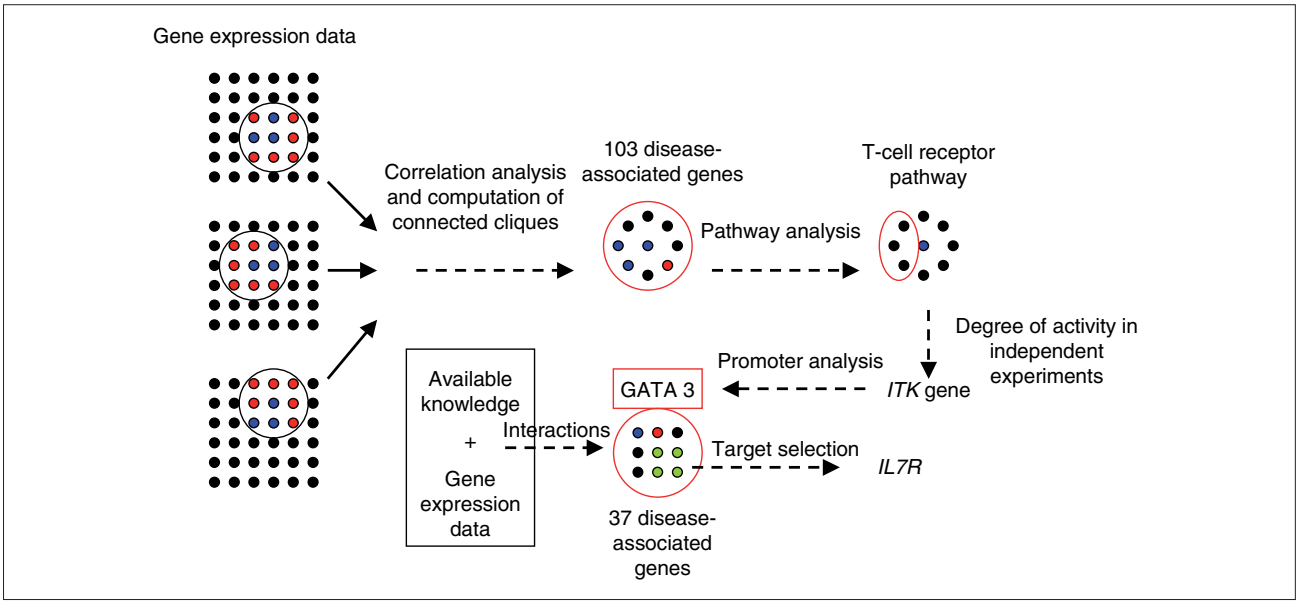

Figure 1
Flowchart of the modular analysis by Benson and colleagues [1]. Integration of several public gene expression datasets revealed a group of shared (blue) and closely connected clique (red and black) disease-associated genes. A subset of these genes were found to share the T-cell receptor signalling pathway, an observation that was then validated by independent experimentation. To identify a transcription factor (GATA3) regulating one of this subset, the *ITK* gene, a promoter analysis was performed. The final module of 37 disease-associated genes consisted of genes listed in public databases as having relevant expression patterns and interacting with GATA3.

The T-cell receptor signaling pathway turned out to be a pathway shared by these 103 genes, as detected by the Ingenuity Pathway Analysis tool, which identifies physical, transcriptional and enzymatic interactions from the literature [1]. Experimental analysis of this pathway in patient-derived cells revealed strong activation of the *ITK* gene, which is also known to be located in the genomic susceptibility region for allergy. Combining a promoter analysis of the *ITK* gene with expression data revealed that the transcription factor GATA3 regulated *ITK*.

Finally, using available databases, 47 genes were identified as interacting with GATA3. The expression data were used to filter out 10 inactive genes, thus leaving a final module of 37 disease-associated genes around the GATA3 transcription factor [1]. The construction of this module was accompanied by several experimental tests at various stages, providing confidence to the analysis.

Conquering the modules - selecting therapeutic targets within the module

The problem of selecting therapeutic targets within a module has not received much attention in studies that have used a modular approach for reducing complexity. There are various ideas from graph theory on how to compute mathematically defined properties, such as clustering and connectivity in large networks, which then could suggest which nodes are essential. However, essentiality is not necessarily equivalent to disease association. Experimental investigators have instead performed target selection using the full dataset in combination with extensive experimental testing. This is, by most measures, an inefficient and expensive procedure.

The analysis by Benson and colleagues [1] is important because it highlights the difficulty of selecting a disease-associated target from a module of 37 genes despite the elegant prior reduction of complexity. They resorted to using a connectivity criterion, selecting the *IL7R* gene because it had the largest number of connections, and they were also able to demonstrate that perturbing the *IL7R* gene affected other genes and the T-cell phenotype. There are probably several other disease-associated genes in their module that warrant further experimental investigation.

Beyond allergy - translation to the clinic

Benson and colleagues [1] have introduced a useful procedure for defining a module of disease-associated genes. As with most complex diseases, the study of allergy is complicated by the fact that the disease affects several cell types and

tissues. The process of identifying such modules therefore requires the kind of stringent experimental validation as was performed by the Benson team [1]. Despite their careful analysis, because there are other transcription factors for the *ITK* gene that are active in the expression datasets there is a significant risk that several disease-associated genes remain that were not captured in their module.

The second step of selecting a gene for therapeutics from a module is even more problematic because we are currently lacking systematic tools for this selection problem. Furthermore, it is not unlikely that an efficient therapy could require targeting of several disease-associated genes simultaneously. However, the number of combinations of three genes that can be chosen from a small ten-gene module, for example, quickly exceeds what is experimentally feasible to study.

In conclusion, Benson and colleagues [1] have devised an interesting method for finding disease-associated genes, but it needs to be evaluated on other complex diseases. Their study also makes clear that the problem of prioritizing disease-associated genes within a module for therapeutic studies in the clinic is still unsolved.

Acknowledgements

I thank the Swedish Research Council for support.

References

1. Mobini R, Andersson B, Erjefält J, Hahn-Zoric M, Langston M, Perkins A, Cardell L-O, Benson M: **A module-based analytical strategy to identify novel disease associated genes shows an inhibitory role for interleukin 7 receptor in allergic inflammation.** *BMC Systems Biol* 2009, **3:**19.
2. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data and modelling.** *Proc Natl Acad Sci USA* 2000, **97:**10101-10106.
3. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
4. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition specific regulators from gene expression data.** *Nat Genet* 2003, **34:**166-176.
5. Tegnér J, Björkegren J: **Perturbations to uncover gene networks.** *Trends Genet* 2007, **23:**34-41.
6. Ideker T, Lauffenburger D: **Building with a scaffold: emerging strategies for high- to low-level cellular modeling.** *Trends Biotechnol* 2003 **21:**252-262.
7. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes.** *Nature* 2004, **431:**308-312.
8. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual JF, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Solé X, Hernández P, Lázaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M: **Network modelling links breast cancer susceptibility and centrosome dysfunction.** *Nat Genet* 2007 **39:**1338-1349.