

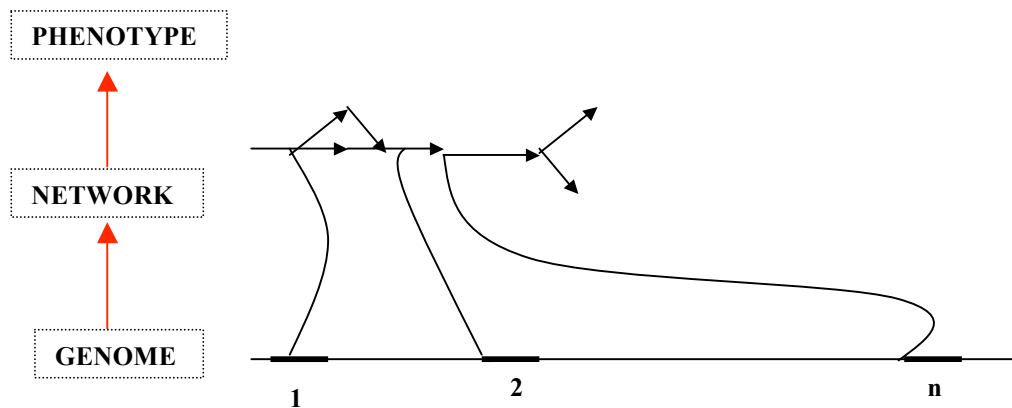
Networks and Genetic Mapping

1.9.08

Genetic Mapping attempts to home in on the positions on the genome that are causative for a phenotype of interest such as disease susceptibility. The more that could be learnt about the genotype→phenotype function, the more valuable this will be. The widespread redundancy in the genetic architecture presents a problem in this context, as many phenotypes will only be detectable by multiple events in different parts of the genomes. This could be solved by including more positions in the assumed genotype→phenotype function. However, this immediately leads to a combinatorial explosions and a serious loss of statistical power. Exemplifying with the human genome and focusing on standard genes, single gene genotype→phenotype function would need to consider ~20.000 positions, considering only protein coding genes and that these are atomic. Two-gene functions would have to consider ~2*10⁸ position pairs and even large numbers for characters depending on more genes. Clearly, a qualified reduction or weighting of which gene sets to consider, would be of value and could lead to a major boost in statistical power. Such a reduction could be achieved by the incorporation of biological knowledge and one kind of knowledge that is obtained on a large scale presently is the architecture of networks. The present project describes an idealized setup that could investigate potential benefits and problems of such approach.

How could networks help with mapping? The object is to use network properties to map genes to characters and also make statements about which sets of genes will not work as a group and should not be considered as a combination in a mapping study.

An ideal situation would be that the underlying network was known. This is almost true in a series of cases, where a universal metabolism is known and when the genome of the organism in question has been determined, then a sub-metabolism is defined. For any metabolism a set of characters that it defines should be listed. Clearly this is an ideal situation and in real applications the network for an organism is known with error and that should be included in the description.

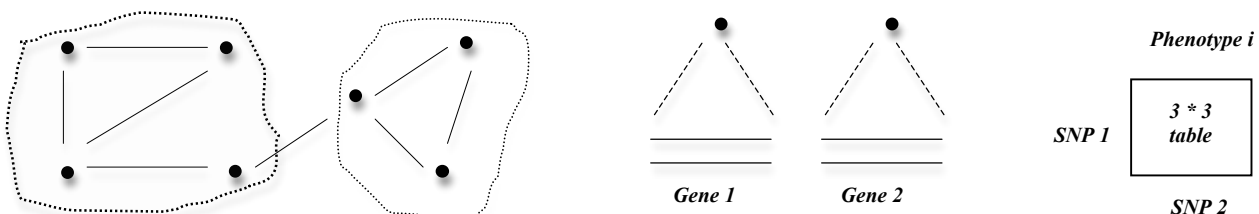


A genome has been sequenced and its genes easily found and it can then be mapped onto a Universal Metabolism. The genes found in the genome basically activate an edge (reaction).

Network characters could be defined as flux across coefficients to what is called extreme metabolisms. Extreme metabolisms are the analogues of a basis vector for vector space in a positive coefficient linear programming problem.

In most cases the underlying network is not known or partially known. The value of network based genotype->phenotype functions should be a major reduction in dimensionality and for instance characters involving k genes could be defined in terms of less than k terms. So either this must be achieved by assuming something about the general structure of the network, such as sparseness for instance number of edges per nodes is constant. An alternative could to investigate networks that are neighbours to a partially known network.

Work related to this has been done by for instance Omholt and colleagues (a series of papers), Rzhetsky and colleagues (a series of papers) and Emily et al. (2008). Omholt has investigated genotype→phenotype mappings based on dynamic models of networks. Rzhetsky has used literature mining extracted networks and high throughput networks to define clusters of genes. A disease could then be caused by any mutation in any gene in a cluster. Emily used protein interaction networks to restrict which interactions were considered in searching for epistasis. Both the above methods have been used to analyzed data. Rzhetsky and colleagues defined “candidate clusters” and Emily and colleagues found several significant SNP-SNP interactions. To illustrate the Rzhetsky and Emily approaches:



At the left is shown 7 genes with connections. Two clusters of sizes 4 and 3 could be defined according some criteria and suitable search. A disease would occur if a there was a mutation in the disease relevant cluster. In the middle the contents of two connected genes are expanded into genes with two sets of SNPs varying. To the right a table for one phenotype state is made for a pairs of SNPs. It is 3 * 3, because, there are 2 possible homozygotes and 1 heterozygote. If these tables are different from genotype to genotype, it will be a sign of epistasis.

To analyze real data, four questions need to be addressed: Firstly, what are the nodes of the graph? Secondly, what are the edges of the graph and how are nodes and edges labelled? Thirdly how to define characters in terms of networks? Fourthly, how do you do genetic analysis (pedigrees and association mapping) with such characters? (None of the methods presently deal with the fact that used networks are part of a larger network.)

- The set of nodes we are after must be the possible components in dynamic or causal model relevant for the disease. Since there often is no prior model, one would have to be very encompassing. The maximal set of nodes could be all nucleotides in the genome (6×10^9). Reasonable reductions could be only consider, 1) the nucleotides with possible functions i.e. discard neutral/junk nucleotides, 2) only nucleotides for which there is variation in the sample and 3) lump all SNPs that is associated the same gene. These reductions should bring us in steps down to $1-2 \times 10^4$ nodes. Clearly, further reductions would be possible, if we were willing to assume that for instance only metabolic enzymes were relevant. Ignoring positions, where there is no variation can be unreasonable as such could well be an important component in a model for an organism or disease that is common to all individuals in question. However, for the present purposes, it seems fully adequate.
- The edges could be pairs of nodes (proteins?) mentioned in the same article or interacting in the same PIN. Both of these would lead to simple undirected unlabelled graphs. But genes can be labelled using gene ontologies and SNPs according to their genomic annotation. This would allow for the definition of sub-networks and testing specific hypotheses about which sub-networks are relevant for the phenotype.
- Both Rzhetsky and Emily use the simple graphs mentioned above. Rzhetsky uses the graph to define clusters of genes, where any mutation could affect the phenotype. Emily use the edges to define which epistatic effect will be considered. Both are computationally intensive, Rzhetsky has to search the graph for clusters and Mathieu has to investigate all edges. It is clearly that these two sets of network based characters represent the simplest ways of defining characters and more complicated characters should be investigated, including using explicit information based on hypothesis about the disease.
- Algorithms for data analysis. The Emily computations are simple, since they don't involve population studies or genealogical structures, but still demanding. The Rzhetsky analysis needs to be done in both pedigree and association mapping setting.

Case study: Autism. Autism is a multifactorial and heterogenous disease (Abrahams and Geschwind. 2008).

1. Genetic Data are available for cases and controls:

Genetic datasets: for autism – AGRE (<http://www.agre.org/>): for example, GENOTYPE DATA:

High-density SNP (Affymetrix 5.0) data on 777 families contributed by the Autism Consortium.

High-density SNP (Illumina Hap550) data on 943 families contributed by Children's Hospital of Philadelphia.

(contact researcher liaison)

-Whole Genome Scan and Fine-mapping data on 356 families.

-Genome-wide High-density 10K SNP data on 426 families from the Autism Genome Project (AGP).

-High-density SNP (2007 SNPs) data of Chr.17 on 219 families contributed by the UCLA Collaborative Autism Genetics Project.

2. A wide variety of networks are available. We generated and made publicly available two very large networks of molecular interactions: 49,493 mouse-specific and 52,518 human-specific interactions. These networks were generated through automated analysis of 368,331 full-text research articles and 8,039,972 article abstracts from the PubMed database using the GeneWays system. Our networks cover a wide spectrum of biomedical interactions, such as *bind*, *phosphorylate*, *glycosylate*, and *activate*; 207 of these relation types occur more than 1,000 times in our unfiltered multi-species data set. Because mouse and human genes are linked through an orthological relationship, human and mouse networks are amenable to straightforward joint computational analysis.

Both genetic and network data are available, but will clearly be expanding in coming years.

Work plan and organisation. The student will spend the first 6 months learning the techniques behind this methods and replicate key results of previous studies. 6 months will be used make tools allowing simulation of data on a pedigree and from a population under different network based genotype→phenotype models. The following year will be used to investigate the power of the methods on simulated data and to analyze autism data using the already published models. The last year will be used to explore more advanced network based genotype→phenotype models. The student will be placed in Oxford Centre for Gene Function in Jotun Heins' group. The student will be 2 years in Oxford and have 2 6 months stay with Prof. Andrey Rzhetsky at the University of Chicago.

References

- Abrahams, BS and Geschwind, (2008) "Advances in Autism Genetics: On the threshold of a new neurobiology" *Nature Review Genetics* 9:341-355
- Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet.* (9):699-710. Featured Review. Sep 8 (2007)
- David Deutscher, Isaac Meilijson, Martin Kupiec & Eytan Ruppin 2006. "Multiple knockout analysis of genetic robustness in the yeast metabolic network" *Nature Genetics* - 38, 993 - 998 (2006)
- Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) "Global reconstruction of the human metabolic network based on genomic and bibliomic data" *Proc Natl Acad Sci U S A.* 104(6):1777-82
- Emily, M. et al. (2008) "Using Biological Networks to Search for Interacting Loci in Genome wide Association Studies" Manuscript
- Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci U S A.* 2008 Mar 18;105(11):4323-8
- Iossifov I, Zheng T, Baron M, Gilliam TC, Rzhetsky A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res.* 2008 Jul;18(7):1150-62.
- Omholt SW, Plahte E, Oyehaug L, Xiang K. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics.* 2000 Jun;155(2):969-80.
- Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboué PA, Weng W, Wilbur WJ, Hatzivassiloglou V, Friedman C. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform.* 2004 Feb;37(1):43-53.