NESS 2010, Harvard University

How to understand the cell by breaking it – methods and software for network analysis of gene perturbation screens

Florian Markowetz,
http://www.markowetzlab.org

[1] Title slide

[2] How to understand a complex system? Physicist Richard Feynman once said: "What I cannot create, I do not understand". This quote stresses the importance of action for understanding. A complex system is not understood solely by passive contemplation, it needs active manipulation by the researcher. In biology this fact is long known. Functional genomics has a long tradition of inferring the inner working of a cell by breaking it. "What I cannot break, I do not understand" is the credo of functional genomics research.

[3] Breaking the system. Experimental interventions into the normal activity of a cell can take many different forms. External stimulations like heat-shock change the global environment of the cell; small molecules can block or stimulate receptors in the cell membrane that activate signaling pathways in the cytoplasm; drugs can influence proteins. Even natural genetic variation and SNPs can be thought as gene perturbations. What we are focusing on in this tutorial are gene perturbations by knockouts or RNA interference.

[4-9] Examples of low- and high-dimensional phenotypes. Once a gene is perturbed (by a knock-out or by RNAi) the effects on the cell phenotype must be measured. Many screens use a single reporter (or a small number of reporters) as phenotype. Here, they will be called low-dimensional phenotypes. Examples are cell viability versus cell death (basically a binary phenotype), growth rates or the activity of reporter constructs, e.g. a luciferase downstream of a pathway of interest. Low-dimensional phenotyping screens can identify candidate screens on a genome-wide scale and are often used as a first step for follow-up analysis. Their results are hard to relate directly to specific gene functions, interactions and pathways. Low-dimensional phenotyping screens are the topic of the first half of this part of the tutorial. The second half of this part of the tutorial will be about high-dimensional phenotyping screens, which evaluate a large number of cellular features at the same time (and not only a single one or a few as low-dimensional screens do). Examples are high content screens using microscopy and cell imaging, and also global gene expression or protein measurements taken after gene perturbations. Observing system-wide changes promises key insights into cellular mechanisms and pathways that can not be supplied by low-dimensional screens. Different software packages are available for data-handling, normalization and pre-processing of phenotypic read-outs, in particular at www.bioconductor.org.
Where computation comes in. Phenotyping screens produce parts lists of which genes show which particular phenotype and what cellular function they could contribute to. The (biological and computational) challenge is to use phenotypic information to infer the internal cellular networks regulating these phenotypes. Computation contributes to this challenge on many levels: from data handling, low-level analysis, quality control and data normalization (all topics we will not talk about in this tutorial) to sophisticated methods of network reconstruction and inference (which are the particular topic of this tutorial).

Analysis of low-dimensional phenotypes

Downstream phenotypes carry only indirect information on networks. The following observation is very simple, but influences all further analysis and the interpretation of results. What we are interested in are pathways and mechanisms in the cell. But what we observe are effects either on the whole cell (viability or growth) or downstream reporters that are regulated by the pathway of interest. That means: we do not have direct information of how the genes and proteins interact to produce this phenotype. We only have indirect information from downstream effects. The goal of computational analysis is to bridge this gap and to infer gene function and to recover pathways and mechanism from observed phenotypes.

GO overrepresentation analysis is a simple way to link phenotypes to gene function. Strong phenotypes are called ‘hits’ and result in a gene list that can be mapped onto the GO hierarchy. A hyper-geometric test quantifies whether the number of hits in each GO node is unexpectedly large (or small). This is exactly the same procedure that is standard in many microarray studies to interpret lists of differential genes. Many tools exist, even online (e.g. function.princeton.edu/GOLEM).

Gene Set Enrichment Analysis. In overrepresentation analysis, a strict cutoff is applied to define hits in the screen. An alternative is to assess whether genes with a given function (e.g. in one GO node) show a trend to have strong phenotypes. This can be quantified by several statistics, most commonly used is the KS-statistic (Subramanian et al, 2005). This analysis (as well as the overrepresentation analysis) is not restricted to GO, but can applied to any collection of gene sets, as eg. MSigDB at the Broad.

Pros and Cons of overrepresentation and enrichment analysis. Both methods result in a list of p-values describing how significantly certain gene sets were represented in the observed phenotypes. Advantages: this is a standard analysis and offers a comprehensive first overview. Many people call it ‘unbiased’ and ‘hypothesis-free’. But there is also a danger to it: it is generic and not tailored to your screen. Good data analysis asks specific questions. A ‘hypothesis-free’ method can thus only be a very first starting point for a deeper exploration of the data. Also, all enrichment methods rely on known gene sets and cannot uncover new pathways or components. They treat pathways as bags of unconnected genes and proteins, even though the connections within a pathway and between pathways are most important to understand phenotypes. Additionally, the results are correlations of the form ‘this set
of genes usually shows strong phenotypes’ which can not easily be interpreted causally.

[19] **Map phenotypes to networks.** One way to overcome the limitations of overrepresentation analysis is to integrate the observed phenotypes with other sources of information like experimentally determined or hand-curated networks. This approach can be exemplified by two recent studies. Lee et al (2008) created a network in C.elegans by integrating many different data sources. Then they mapped RNAi hits onto this network and showed that they fall into connected. In this way, Krishnan et al (2008) illustrate a complex dependence of West Nile virus on host cell physiology, requiring a wide variety of molecules and cellular pathways for successful infection.

[20-21] **Finding high-scoring subnetworks.** Whereas the last two examples found the sub-networks containing many hits using by-eye-informatics, there are bio-informatical tools to automate the search and assess the significance of the results. The hits are spread over the network (any network, more on this later) and algorithms determine subnetworks which are highly connected and show a large number of hits. Efficient search algorithms exist even for large networks, see Dittrich et al (2008) and MATISSE (Ulitsky and Shamir, 2005)

[22] **Predicting phenotypes.** An additional application of mapping hits to a network is that the known phenotypes can be used to predict phenotypes of genes not included in the study. A simple idea is ‘guilt-by-association’: a gene connected to many hits should also show a strong phenotype. A gene far away from any hits will not show a phenotype. The success of this strategy depends on the quality and coverage of both the screen and the linkage in the network.

[23] **Which networks?** There are a large number of possible networks to choose from in almost all model organisms and Human. Examples are the human PPI network (Stelz et al, 2005), networks from automatically analyzing experimental literature (eg Ma’ayan et al, 2005) or a growing number of computationally inferred networks that integrate many different data sources (Lee at al, 2004; Myers at al 2005; Guan et al 2008).

[24-26] **Summary and transition.**

**Analysis of high-dimensional phenotypes**

[27] **High-content screens (HCS).** Examples of changes in morphology after gene perturbation from Boutros and Ahringer (2008): a) Drosophila neural cell line; b) human cells: raw and processed images; c-f) whole organism phenotypes of C.elegans, Drosophila, planaria. Computational issues: handle data size and number of images to process, detect and classify phenotypes. The rapid development of microscope-based technologies facilitates the acquisition of trillions of cells in thousands of diverse experimental conditions. The massive size of these datasets often precludes human analysis.
[28-29] **Clustering.** A simple first analysis step for any kind of high-dimensional phenotypic profiles is (hierarchical) clustering. This provides a first impression of structure in the data. In this sense it is related to the overrepresentation and enrichment methods we discussed for low-dimensional phenotypes.

[30-31] **Query based gene ranking.** In many applications researchers already a few key genes in mind that they are especially interested in. One of the analysis goals is to relate the phenotypic profiles of these ‘pet’ genes to the phenotypes of all other tested genes. Sometimes this task is approached by a clustering method (“let’s first cluster all the data and then see in which clusters our pet genes fall ..”) but a more direct method is query-based gene ranking as exemplified by PhenoBlast (Gunsalus et al 2004). All phenotypic profiles are ranked according to how similar they are to the query (i.e. the pet genes).

[32] **From clusters to pathways.** Clustering gives us a broad overview, but not details of pathway structure. Methods to infer pathway structure follow now.

[33] **We only see a slice of the full picture.** Most data computational biologists work with is usually gene expression data and a great deal of research went into graphical models applied to microarray data (see eg Markowetz and Spang 2007 for a review). However, gene expression phenotypes (like low-dimensional phenotypes) offer only an indirect view of pathway structure due to the high number of non-transcriptional regulatory events like protein modifications. For example, when silencing a kinase we might not be able to observe changes in the activation states of other proteins involved in the pathway; the only information we may get is that genes downstream of the pathway show expression changes. Thus, phenotypic profiles may provide only indirect information about information flow and pathway structure.

We exemplify this problem here in a cartoon example. Imagine five genes A-E one three biological levels (DNA, RNA, Protein) and as a node in the model. For each gene, DNA is transcribed into mRNA which is translated into protein. Now imagine that these genes contribute to a cascade in a signaling pathway: expression of gene A leads to protein A, which activates B, which activates C, which activates D, which is a transcription factor binding to gene E and activating it. We now measure the expression of all five genes on microarrays. Even though it could be argued that mRNA expression is a reasonable approximation of the amount of protein in the cell, gene expression measurements can never capture activity states of proteins (which we actually would like to know to understand signal flow in the cell). Thus, what we observe in gene expression data is that changes in the expression level of A correlate with changes at E. The other three genes don’t seem to be involved in the pathway, their mRNA concentrations don’t change since the pathway works on the protein level. Now what happens if we knock-out gene C? The information flow through the cascade would be interrupted and we would see a change of expression at E. Thus the intervention gives us an additional arrow in the model leading from C to E. But still our model has not much to do with the ‘correct’ answer: a cascade from A to E. The reason is that we can’t measure all changes of activity in the pathway, but only effects on down-stream genes.
[34] **Information gap between pathway and data.** More abstract, the way to think about gene perturbation screens is in a two-level model. One level contains the components of the signaling pathway which are experimentally perturbed. The second level contains the downstream effect reporters, i.e., the phenotypic profiles, which could be morphological changes or gene expression changes. In most cases we do not have direct observations of perturbation effects on other pathway components. We need to make inference from the observed effects on downstream reporters.

[35] **Consequences for network reconstruction.** Our discussion of the information gap has profound consequences for any approach to reconstruct cellular pathways. Most currently used approaches are based on analyzing dependencies between genes using different measures for correlation. This can only work if the measurements about the pathway genes are informative. However, due to the information gap, the only information is in the measurements of the downstream genes. The pathway genes themselves ‘stay flat’ when other pathway genes are perturbed. This means: we need an approach to use the information in the downstream effects to infer the wiring diagram between the pathway genes. One way to this is by looking at the subset structure in the observed phenotypes as motivated by the next slide.

[36] **Subset structure.** We will build the following observation: Perturbing some genes may have an influence on a global process, while perturbing others affects subprocesses of it. Imagine, for example, a signaling pathway activating several transcription factors. Blocking the entire pathway will affect all targets of the transcription factors, while perturbing a single downstream transcription factor will only affect its direct targets, which are a subset of the phenotype obtained by blocking the complete pathway. Different examples show how pathway features (forks, cascades, interacting TFs, or joint regulation) are reflected in subset relations between the observed phenotypic profiles.

[37-45] **Nested Effects Models** (NEMs) are a graphical model especially designed for indirect information and high-dimensional phenotypes. NEMs reconstruct features of the internal organization of the cell from the nested structure of observed perturbation effects. They take as input (1) a set of candidate pathway genes and (2) high-dimensional phenotypic profiles of downstream effects of perturbing the candidate genes. NEMs return as output a directed graph best representing the regulatory hierarchy responsible for the observed effects.

It is illustrative to compare NEMs with ‘naively’ linking causes with effects. Cause-effect graphs assume no internal structure and can result in very dense graphs if the overlap of observed effects is big. NEMs, on the other hand, result in a sparser representation by assuming an internal structure between the perturbed genes. The sparseness of the graph is equivalent to assuming a reduced number of regulatory relationships.

[46] **2-step experimental design.** The techniques we discussed in the two parts of this talk can be combined in the same study. First, a genome-wide screen identifies key genes representative for pathways and cellular mechanisms involved in the
phenotype. In a second step, this smaller set of genes is screened for high-dimensional molecular phenotypes (microarrays, protein modifications, ..). The second dataset is the starting point for inferring a pathway diagram.

[47-50] Summary, acknowledgements and all that.

### Relevant publications

http://www.connotea.org/tag/ISMB2010tutorial

- Eaton, D. & Murphy, K. Exact Bayesian structure learning from uncertain interventions, AI & Statistics, 2007


R Development Core Team. R: A Language and Environment for Statistical Computing 2007


