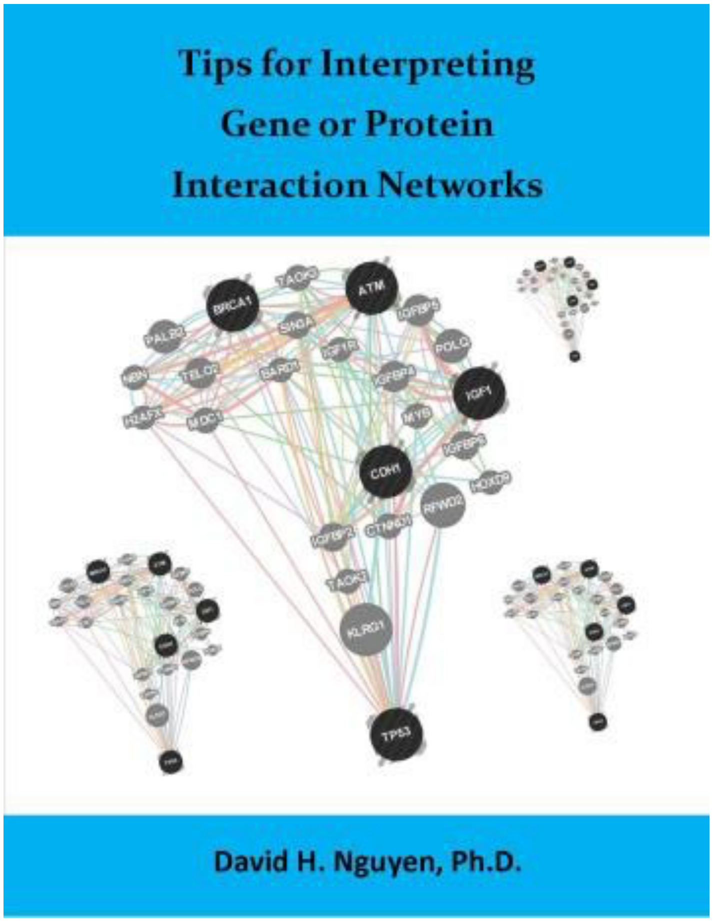
Tips for Interpreting Gene or Protein Interaction Networks

By David H. Nguyen, Ph.D.

Principle Investigator
Tissue Spatial Geometrics Lab
www.TSG-Lab.org

Email: dave@tsg-lab.org



Tips for Interpreting Gene or Protein Interaction Networks

David H. Nguyen, Ph.D.

About The Author

David H. Nguyen, PhD, is Principal Investigator at the Tissue Spatial Geometrics Lab (www.TSG-Lab.org). He creates pattern recognition algorithms to measure shape features that are hidden to the human eye. He suffers from a fictitious disease called “randomnesia,” which is characterized by the inability to see randomness, only varying degrees of disorder -- a disorder that can be precisely quantified assuming we can measure the order from which the said disorder deviates. By reading this paragraph, you might have just gotten infected with randomnesia.

Dedication

My first technical science book is indebted to…

Mr. Duane Nichols, M.S., my high school biology teacher, who taught me the joy of learning and doing science. Thank you for waking up early everyday so that we could have “zero period” at 7:00 AM. If there was a science teacher version of *Mr. Holland’s Opus*, he would be it.

And to others along the way:

Dr. Suraiya Rasheed, Ph.D.

Dr. Sue A. Ingles, Ph.D.

Dr. Gary L. Firestone, Ph.D.

Dr. Hanh Nguyen, Ph.D.

Dr. G. Steven Martin, Ph.D.

Dr. Elena Rodriguez, Ph.D.

Dr. Mary Helen Barcellos-Hoff, Ph.D.

Dr. Sylvain Costes, Ph.D.

Dr. Jian-Hua Mao, Ph.D.

Preface

The ability to interpret gene or protein interaction networks is becoming a valuable skill in biomedical research. It is useful not just to those who produce or analyze high throughput array data, but also to those who produce data from single-read-out biochemical assays such as RT-PCR, western blotting, and protein-protein interaction assays. This is because network analysis programs can be used to predict interaction partners, based on the results from multiple single-read-out assays.

This book is meant to be a resource for those who seek to understand what a network of genes or proteins means in terms of the biological processes that are at work in their high throughput data. It will certainly be of help to those who are just starting out, but may also be insightful to veterans of this trade.

This book does not cover the myriad methods and nuances of computing high throughput gene or protein expression data, but is primarily concerned with interpreting the network of genes or proteins that are generated from the computation. Computing the data and interpreting the data harbor their own challenges. My favorite part is interpreting the networks, which I hope you will learn to like after reading this book. As you will see, successful interpretation of gene and protein networks relies more on your knowledge of biology than it does your knowledge of bioinformatics.

This book contains two chapters. Preceding the first chapter is a page that lists the tips for interpreting gene or protein networks. This was meant to be a cheat sheet to help you quickly reference the tips as you do your analysis. Chapter 1 provides what I consider to be the minimal background information to help someone who is just starting out with network interpretations to best understand the tips. This book was written for someone with a graduate level of biology knowledge. Chapter 2 contains 18 tips, each with a short explanation that will help you interpret interaction networks. I have divided the tips into categories that will help you understand their general purpose. The appendix contains a non-exhaustive list of free online software programs that I have found useful in my own research.

Table of Contents

Preface……………………………………………………………………………………… Pg. 3

The 18 Tips Summary Sheet…………….. Pg. 5

Chapter 1 – How to Use This Book……………………………………………… Pg. 6

Chapter 2 – The 18 Tips Explained…………………………………………….. Pg. 8

Appendix – Network Analysis Programs…………………………………….. Pg. 15

The 18 Tips Summary Sheet

Things for Which to Look

Tip #1. Things may not be what they seem.

Tip #2. Interpret the enriched biological functions in light of other experimental data.

Tip #3. Be familiar with fundamental cellular functions.

Tip #4. Know the master regulator genes of different tissue types.

Tip #5. Look for tissue or cell type specific genes.

Tip #6. Direction of regulation matters.

Tip #7. Don't be overly swayed by fold-changes, big or small.

Tip #8. Look for recurring themes.

Things to Know Ahead of Time

Tip #9. Do your homework ahead of time.

Tip #10. Keep in mind that a gene can have multiple gene symbols.

Tip #11. The database curator doesn’t know as much as you do about your research.

Tip #12. Maintain your own database of biological categories.

Tip #13. Normal tissue, mutant, or tumor?

Tip #14. Cell culture, organ culture, organism, or colony?

Miscellaneous Tips

Tip #15. Use network analysis to predict targets for, or guide, biochemistry assays.

Tip #16. Practice, Practice, Practice

Tip #17. Use multiple network analysis programs.

Tip #18. Ask for a second opinion.

Chapter 1

How to Use This Book

Instructions for Using This Book

The tips are not arranged in a sequential order of operations, though certain tips do go hand-in-hand with other ones. In those cases, I have noted which tips are connected and in what way. To a veteran, these tips may be second nature, but to someone who is just starting out I have categorized the tips into sections that will help make more sense of them. Most of the tips talk about interpreting mRNA transcripts when the word "gene(s)" is used. However, the same concepts can be applied to proteomics data, micro-RNA data, non-coding RNA data, and sequencing data. Broadly speaking, if a molecule interacts with other molecules in some specific or systematic way, then they constitute a network of molecules.

Reading this book will not make you better at interpreting networks unless you put what you learn into practice. The more successful analyses that you do, the more intuitive interpreting networks become. Understanding networks is like playing the game of Jenga. Jenga is a vertical tower consisting of rows of small wooden rectangles. Each person takes turns removing a block of wood, except from the top row, with the goal of leaving the tower standing. The person who causes the tower to fall loses. Certain blocks do not affect the stability of the tower and can easily be removed. Other blocks will pull on their neighbors when removed, distorting the shape and stability of the tower. Some blocks will immediately collapse the tower if removed. Genes and proteins that govern biological processes are like Jenga blocks. Some genes carry more weight than others.

What is a curated database?

Network analysis programs have a database of biological categories that contain gene or protein names within them. A biological category is usually a process, such as "cell division," a signaling pathway, such as "MAPK signaling," an organ, such as "brain," or a gene list that the user (you) generated. Biological categories can have distinct genes or genes that overlap with other categories.

A curated database is a database in which someone purposefully assigns genes to biological categories. Some network analysis programs have curators who read published papers and updated the categories. Others programs define categories based on other online databases that contain information about what protein domains are common among gene families, which proteins are predicted to physically interact with each other, genes induced by a certain drug treatment, etc. Regardless of which curating method is applied, genes are rationally assigned to categories.

How do network analysis programs generally work?

There are many network analysis programs. Each calculates the degree of overlap between your uploaded gene list (with or without gene expression values) and the categories in its database differently. This book will not discuss the methods by which the strength of overlaps are calculated, but will discuss the usefulness – or not – of accuracy statistics (i.e. p-values and q-values). Depending on the goal with which you are using a network analysis program, the p-value and q-value statistics may vary in their importance. Certain tips in Chapter 2 will discuss situations in which this is the case. In general, the stronger the overlap between your gene list and a biological category, the more likely that biological category describes what is going on in your research model – though you will see in Chapter 2 that things may not be what they seem, even if there is a strong degree of overlap. Throughout this book, the term enrichment describes the occurrence of one or more biological categories in the analysis results that are produced by a network analysis program.

Chapter 2

The Tips Explained

Things for Which to Look

Tip #1. Things may not be what they seem.

The most important piece of advice that I can give about interpreting gene networks is that things may not be what they seem; meaning the biological processes that are enriched in your gene list of interest may not be what is actually happening in your research model. If you are studying gene expression of a tissue that does not have skeletal muscle, but receive “muscle development” as an enriched category, it wasn’t a mistake. Epithelial cells can undergo epithelial-to-mesenchymal transition (EMT) during which they up-regulate genes such as smooth muscle actin. Thus, muscle development may actually be EMT.

Here are other hypothetical examples of hidden meanings.

“EMT” may be a stem cell program. EMT has been linked to a process of dedifferentiation into a more stem-like state. Therefore, a enrichment for EMT may be a sign that dedifferentiation is occurring, or that stem cells or progenitor cells have become abundant.

“Organismal development” may be metastasis. The process of organismal development involves the movement of cells to different compartments and then differentiation into mature cell types. This differentiation process can involve the expression of genes that form cell-cell junctions and cell-matrix connections. These same processes are down-regulated during metastasis.

“Lysosome biogenesis” may be autophagy. Autophagy is the process in which a cell digests its own organelles. Autophagy can be a part of senescence, which is cellular aging and dormancy. Lysosomes are vesicles that engulf and digest other organelles. During autophagy, lysosomal activity increases. Thus, enrichment for “lysosome biogenesis” may actually indicate autophagy and/or senescence.

“Metabolism” may be enrichment for auxotrophic cells. Auxotrophs are cells that are dependent on a nutrient that is produced by another cell type. The ducts in a mammary gland consist of two main cell layers. The luminal cells line the inside of the duct, while the myoepithelial cells line the outside of the duct. Myoepithelial cells cannot produce the amino acid glutamine, and rely on luminal cells to secrete glutamine, which the auxotrophic myoepithelial cells take up. Luminal and myoepithelial cells have distinct metabolic profiles, as do the subtypes of breast cancer that reflect a luminal-like state compared to those that exhibit a myoepithelial state.

Tip #2. Interpret the enriched biological functions in light of other experimental data.

In light of Tip #1, enrichment for a biological category is no guarantee that that category really describes what's happening in your research model. So how might you be more certain that "organismal development" (described in Tip #1) is actually something like metastasis? Well, if the tumors for which you extracted the RNA for microarray analysis also have a high rate of metastasis, then a metastasis program is more likely to be what's really going on. Conversely, if "organismal development" appears in your results, and the genes in this category are cell-cell junction genes, you might want to investigate your tumor model to determine if metastasis has happened or if there has been an increase of circulating tumor cells that may not have metastasized to a distant organ. This tip is the second most important piece of advice.

Tip #3. Be familiar with fundamental cellular functions.

It is helpful to be familiar with fundamental biological processes. Changes in cell morphology are accompanied by changes in actin polymerization, cell-cell junctions, and cell matrix interactions.

Changes in metabolism are accompanied by changes in glycolysis and the citric acid cycle. Changes in growth rates are accompanied by activated growth factor signaling pathways. It is helpful to know both the canonical and non-canonical players in your favorite intracellular signaling pathway. Review articles are useful for gaining a general knowledge of the key components of these biological processes.

Tip #4. Know the master regulator genes of different tissue types.

Some would say that transcriptional complexity defined by the number of transcription factors, transcriptional co-regulators, and transcription factor binding elements makes complex organ systems possible. If you are studying a tissue type or organ, it is helpful to understand the master regulator transcription factors that govern normal organ development. This is especially the case if you are studying a tumor or a mutated organism. Germline tumors, such as teratocarcinomas, can differentiate into hair and teeth. Tumors originating from somatic cells do not have as much pluripotent potential, but can exhibit diverse biological activities. Thus, being familiar with the master regulators of organ development will help you understand perturbations a complex research model.

Tip #5. Look for tissue or cell type specific genes.

Certain genes are only or mainly expressed in certain tissue types or cell types. For example, immunoglobulin genes are expressed by B lymphocytes. Granzymes are expressed by T lymphocytes. Collagen X is expressed by chrondrocytes. Knowing tissue and cell specific genes will help you quickly spot things like inflammatory infiltration or differentiation.

Keep in mind that genetically engineered cell lines or organisms that have genetic knock-in, knock-down, or knock-out properties can completely change the cell and tissue specificity gene expression. Transcription factors and transcriptional co-regulators are all connected, so even small perturbations can cause the global gene expression program to deviate from the wild type condition. Thus, large perturbations from genetic engineering can nullify this tip.

Tip #6. Direction of regulation matters.

Some network analysis programs allow you to upload the actual gene scores or arbitrarily assigned gene scores (i.e. +2 for up two-fold, -2 for down two-fold) for each gene in your list of genes. Others only allow you to upload a list of genes, without information about relative expression levels or direction of expression – up-regulated or down-regulated. It will be helpful to make turn your gene list into three lists: up-regulated genes, down-regulated genes, and both up- and down-regulated genes. Running all three lists will make interpreting the enrichment categories much easier. For example, if your down-regulated gene list enriches for “macrophage activation,” then there may be fewer macrophages or depolarized macrophages in the experimental condition represented by your gene list.

Tip #7. Don't be overly swayed by fold-changes, big or small.

Have you ever seen a swarm of birds or school of fish move in unison? They change directions as if they were one organism. Gene expression works in similar ways. Depending on your research goal, fold-change cut-offs for your expression levels of each gene may or may not matter. If you are seeking to find key genes in your list for which to do knock-out experiments to test a hypothesized mechanism, then analyzing data that has high, and statistically rigorous, fold-changes is important. However, not all researchers want to approach their expression data as reductionists. Systems biology – among its various definitions – is interested in multi-parameter or multi-component changes, even if the fold-changes are small. The debate about the importance of small versus large fold-changes is beyond the scope of this book. Suffice it to say, small fold-changes may not satisfy statistical rigor to a degree of comfort for some, but what is important is whether the enriched biological categories due to many small fold-changes can be validated. The point of rigorous statistics is to make bioinformatic data trustworthy as stand-alone data. However, the experimental validation of a mechanism interpreted from genes that were modulated by small fold-changes will trump statistics. At least, in my book it does – pun intended.

Tip #8. Look for recurring themes.

After inputting your gene expression data or gene list into a network analysis program, you will receive a list of biological mechanisms that are enriched for within your gene list. Examples of biological mechanisms can be categories such as angiogenesis, cell division, cell adhesion, senescence, unfolded protein response, lymphocyte activation, etc. The more you see similar categories appear in the list of enriched categories, the more likely that the biology in common amongst those categories is actually present. However, if the network analysis program has a statistical metric, such as a p-value, that quantifies the strength of the overlap, even one category can be real in your research model.

Things to Know Ahead of Time

Tip #9. Do your homework ahead of time.

Tip #3 and #4 discuss the knowledge of biological processes that will help your interpretation. You've been doing this all along as you have been studying biology. Take the time to learn these things before or while you practice interpreting gene networks. The more information you have, the easier it will be to find insightful mechanisms.

Tip #10. Keep in mind that a gene can have multiple gene symbols.

Different model organisms have different gene symbols. It will help if you familiarize yourself with the multiple gene symbols of your favorite genes. For example, the transcription factor p53 is encoded by the human TP53 gene and the mouse Trp53. The "Gene" search option on

PubMed is a database that lists the official gene symbol and alternate gene symbols for many species.

Tip #11. The database curator doesn’t know as much as you do about your research.

The curator is the person who defines which genes belong to which biological category. The curator is not an expert in your field or your research model. The curator doesn't know the context of your research. Therefore, the curator's decision to assign genes to biological categories may be based on research papers that study a completely different research model than yours. Therefore, don't always accept the curator's assignments as definitive.

Tip #12. Maintain your own database of biological categories.

It will be helpful if you maintain your own database of biological categories. If you have gene expression data from prior experiments that were derived from your research model, those gene lists will help you understand your future array data. You can become your own curator. Certain network analysis programs allow you to store your own gene lists and then compare your new gene list to them.

Tip #13. Normal tissue, mutant, or tumor?

This question is important because it limits the interpretations that you can make. Normal tissues are generally better at maintaining tissue specific gene expression patterns. Tumors or genetically engineered research models that have a high potential for differentiation or dedifferentiation require more creativity to interpret. See Tip #5 for more information on this topic.

Tip #14. Cell culture, organ culture, organism, or colony?

This is an important question when it comes to interpreting biological mechanisms. For example, the elastase enzyme cleaves intracellular proteins, but is also secreted by neutrophils. In the context of a tumor, gene expression data from tumors treated with an elastase inhibitor need to be interpreted with this in mind. Is the effect of the inhibitor primarily due to its effect on epithelial elastase or due to the enrichment of tumor infiltrating neutrophils that are not altering extra cellular matrix composition? The real answer may be somewhere in between, and both may be important mechanisms of the inhibitor within the observed phenotype.

If your gene list is from epithelial cells grown in 2D culture and one of the enriched biological categories is lymphocyte activation, it could be that your cells are contaminated or that your cells are just producing cytokines that have been known to activate lymphocytes. A biological category, such as lymphocyte activation, can be enriched in your gene list even if not all of the genes in that category are on your list. Look closely at the genes.

Interpreting the mechanism represented by a biological category of secreted proteins (proteases, growth factors, lipoproteins, etc.) can be tricky – and interesting. Research models that have infiltrating cell types or infiltrating species – such as a microbial community – make the interpretation of mechanism even more complicated. Bacteria can pass genetic information to each other through horizontal gene transfer, so it is difficult to know which species is expressing which gene in which way without further experiments.

Miscellaneous Tips

Tip #15. Use network analysis to predict targets for, or guide, biochemistry assays.

Single-read-out assays, such as western blotting or qRT-PCR, usually detect only one protein or mRNA at a time. Protein-protein and protein-nucleic acid interaction assays can give readouts for multiple affected genes, but these assays are still considered the opposite of high throughput assays that measure changes in tens to tens of thousands of genes at once. Network analysis programs are mainly used for identifying biological mechanisms in high throughput data, which I call “descriptive network analysis.” However, they can also be helpful to researchers that generate data from single-read-out assays. If you have information about how several genes or proteins are behaving in your research model, then network analysis programs can help you identify potential binding partners, regulators, transcriptional targets, etc. Some network analysis programs will show you the genes that are predicted to interact with the genes on your lists. These predictions are based on published data, but also on predicted binding sites between canonical protein interaction motifs. Some network analysis programs will also show you the nature of the connections between the genes in your list and the genes in their database of biological categories: direct binding, transcriptional regulation, epigenetic regulation, etc. Network analysis programs are helpful in making educated guesses about what else is happening based on data from single-read-out assays.

I once knew a scientist who studied intracellular signaling. He taped an 8.5 x 11 inch paper above his desk that contained drawings of the signaling pathway that he studied. As he made new discoveries in his research and read new papers, he would add new genes and draw in new arrows. Before network analysis programs were invented, this was the “old school” way of making predictions about what proteins might interact with other proteins, or which signaling pathways might cross-talk with each other – it still works. Network analysis programs have made it much easier to generate an educated guess about what proteins or gene promoters your favorite protein may be affecting.

Tip #16. Practice, Practice, Practice

Interpreting gene and protein interaction networks is like solving puzzles. The more you learn and practice strategies, the more intuitive it becomes. The more you do it, the better you’ll get.

Tip #17. Use multiple network analysis programs.

Each network analysis program has its strengths and weaknesses. Running the same data in two or more programs will reveal biological categories – and thus mechanisms – that are may have been missed by one program. The fact of the matter is, not every research paper needs to present an exact mechanism that has been extracted from the interaction data. Gene and protein interaction data is very complex, so there can be multiple right answers to this question – so to speak. The stringency that applies to your research depends on the goal of the bioinformatic analyses and how they are integrated into your other data.

Tip #18. Ask for a second opinion.

Different people have stores of background knowledge. Having multiple people interpret the same network will increase the chance of finding interesting mechanisms.

Appendix

This is not an exhaustive list, just the ones that I've used.

Free Network Analysis Programs

- Gene Set Enrichment Analysis (GSEA)
  http://www.broadinstitute.org/gsea/index.jsp

- DAVID Bioinformatics
  http://david.abcc.ncifcrf.gov/

- L2L
  http://depts.washington.edu/l2l/

- GeneMania
  http://www.genemania.org/

- EGAN
  http://akt.ucsf.edu/EGAN/

- ConceptGen
  http://conceptgen.ncibi.org/

- MsigDB
  http://www.broadinstitute.org/gsea/msigdb/index.jsp

Subscription-based Network Analysis Programs

- Ingenuity Pathway Analysis
  http://www.ingenuity.com/