Tech News

Unraveling cancer through network models

In many ways, cancer is simply a devastating natural mutagenesis experiment. Alterations to genes and their products, as well as additional downstream modifications, lead to dangerous and deadly consequences. From recent studies, we know there are a few key cancer drivers, genes such as *p53* and *Ras* that have central roles within the genetic pathways causing these devastating effects. But interestingly, these mutations don't show up in all cancers; in fact, they represent a small portion of the information that researchers and clinicians require to understand tumor biology and diagnose and treat disease.

The result: cancer researchers are now focusing on the "long tail", collecting and cataloging rare mutations occurring in 1% or fewer of cancer patients. These rarer mutations may underlie the critical functional changes within cells that characterize and define this collection of diseases. But there is a big challenge here, a double-edge sword for researchers: because of their rarity, it is actually much harder to distinguish these rare mutations from random mutations that don't affect disease.

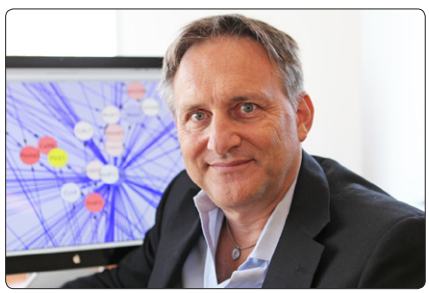
Building network context

So, how then do researchers go about locating these important rare variants? The functional consequences of mutations in the genome can often be seen in the molecules, such as proteins, that they encode. Over time, bioinformatics researchers have learned how these biomolecules interact with each other in the cell, curating protein and metabolite connections into wiring diagrams with nodes for proteins or other molecules and edges that indicate an interaction with another molecule. This has led to the development of a landscape of bioinformatics methods for understanding the misfires that cause cancer and control the disease process according to Trey Ideker a bioinformatician, at the University of California at San Diego. Researchers gather systematic information on genetic interactions from genome sequencing that can then be combined with public data on protein-protein interactions, producing more comprehensive databases featuring millions of associations between molecules. The question then, Ideker says, is how researchers can take this grab bag of interactions and introduce context to build pathway models that researchers can take advantage of to understand and diagnose disease.

Even with millions of associations catalogued, these databases are far from complete. In many ways, our current understanding of interaction networks is like navigating through a major city with a general map explains Andrea Califano of Columbia University. "It's like having a map of a city with Main St and Broadway and not actually knowing whether the city is New York or Boston."

Although understanding how proteins function together is critical for basic research, the real benefit could come from the diagnosis and treatment of disease. Here, cancer is the "killer app"—an important problem that motivates the need for networks.

"It's not just a good idea for solving these diseases, but required for solving these diseases," says Ideker. So, as large data sets emerge from cancer genome sequencing projects such as the NIH/NHGRI's Cancer Genome Atlas project (TCGA), bioinformatic analysis that integrates that information into the context of protein networks is essential for helping researchers to make sense of the deluge of cancer data.



Andrea Califano says obtaining detailed reference maps for cells remains a challenge. Image courtesy of Chris Williams.

Using networks for analysis

A primary motivation for the TCGA project was to understand the similarities and differences among various types and subtypes of human cancers. Researchers involved in the TCGA are currently looking at hundreds of samples from each of more than 20 tumor types to identify rare mutations involved in cancer.

One bioinformatics approach to understanding the possible effects of specific rare mutations is to create a "heat map"—a graphical representation of mutations in context with nearby neighbors in a protein network. HotNet, an algorithm developed by Ben Raphael and his colleagues at Brown University, is one such algorithm. The idea is straightforward: mutations to a single gene confer a certain amount of "heat" to a pathway. If no nearby genes are mutated, only that single mutation is interesting to researchers. However, if 4 or 5 genes that are only mutated occasionally but are closely linked in the network, those mutations propagate heat among them creating a "hot zone", and implicating that area of the network. "The idea behind all these approaches is to implicate your neighborhood," explains Ideker.

When Raphael and his colleagues used HotNet for their TGCA analysis of the ovarian cancer genome (1), they observed well-known signaling pathways such as p53 and Ras. But they also pulled out the Notch signaling pathway based on a combination of individual, infrequently mutated genes. While other experimental evidence had suggested Notch might be involved in various cancers, Raphael emphasizes, it's a nice example of how computational tools can help point researchers toward a biologically relevant hypothesis.

Other findings have implicated genes that were not so well known to researchers. TCGA analysis of ovarian and kidney cancer samples (2) identified hotspot genes that don't line up with any current experimental hypotheses for cancer. "Are they real? Are they not? It requires some additional experimental work," Raphael says. Even though the algorithm generates results that are consistent with experimental data, which lends credibility, "ultimately what we're doing is generating hypotheses."

These algorithms based on molecular networks could allow researchers classify tumors into subtypes based on overlapping hot zones. According to Ideker, patients might not have mutations in the same genes, but if they have mutations in genes that are closely connected within a network, that information might help researchers understand subtypes of the disease even if patients lack similar mutation profiles.

A core issue for this type of analysis is how best to simplify available biochemical data into a form that takes into account biological function, allowing researchers to model the effects. The sheer amount of information researchers have collected in curated networks (e.g., REACTOME, BioCarta, WikiPathways, KEGG, and NCI-PID) can be overwhelming, with data on gene expression, copy number, epigenetic state, neighbors in a pathway, transcription factors, and more notes Josh Stuart of the University of California Santa Cruz. Collaborating with David Haussler, also at UC Santa Cruz, Stuart has

developed an algorithm called PARADIGM which takes all that available data on a gene and transforms it into a single number to indicate whether the gene is active in the cell or not (3).

Computers can then use those single values in place of the original data to come up with predictions of how genes work within a cell. For a cancer data set, this means predictions can be made as to whether tumors with a particular genetic profile are likely to have better or worse outcomes or predict drug targets based on data from cell lines.

Stuart's algorithm is being used as part of the automated pipeline for TCGA data being funneled through Firehose, the computational pipeline used at the Broad Institute.

Expanding networks

Although network analysis is improving, it is still hampered by the many protein-protein interactions within the cell that remain unmapped. (See "Biochemical identification of protein-protein interactions") Here, Califano says, the major challenge for the field is getting detailed reference maps for cells, particularly those of different lineages where different regulatory processes occur.

For the PARADIGM algorithm, Stuart and his colleagues are slowly adding new interactions to the networks they use. A quarter of human proteins so far have been noted to regulate another gene or gene product in the curated networks,. Taking advantage of available high-throughput data, Stuart estimates that approximately 50% of proteins are included in the PARADIGM analysis networks.

Other researchers are attempting to use computational methods to uncover undocumented protein-protein interactions. Recently, Califano collaborated with Columbia biochemist Barry Honig to search for potential interactions between proteins using the tertiary structure of the proteins (4). That data alone is not context-specific,

Trey Ideker works on developing protein network maps. Image provided by Stephanie Mirkin.

Biochemical identification of protein-protein interactions

Researchers use a variety of assays to establish if two proteins interact. Two high-throughput approaches to identifying interactions are yeast-2-hybrid screens and affinity purification-mass spectrometry. Researchers can also predict an interaction between two proteins if both proteins contain domains known to interact with one another. Researchers might also look at whether genes are co-expressed in a cell and whether they're localized to the same cellular compartment. While those last two aren't definitive evidence for an interaction, Stein says, they can provide support for an interaction predicted using another method.

Each method comes with advantages and limitations. For example, though widely used, yeast 2 hybrid screens have several downsides. These screens probe interactions in an in vitro milieu of reagents and antibodies and may miss transient interactions or ones that involve membrane proteins. Not to mention that since this is an in vitro assay, the screen might pick up an interaction that wouldn't normally occur in the cell. So, researchers are most confident in defining a protein-protein interaction when multiple lines of evidence from different experimental approaches support that it occurs. -SW

Califano notes, so the researchers took advantage of additional data sets, such as gene expression analysis, to refine their findings. Using this approach, protein interactions in adenocarcinomas were analyzed. Although they've shown that it's accurate, the method only shows some overlap with known interactions. As a result, Califano says, "we know that there is a tremendous amount of work to do."

Even if researchers haven't found the connections between particular proteins or genes from a known biochemical interaction, data patterns can help point to how genes might be connected. Here, the idea of mutual exclusivity can be a huge help. If two proteins interact to form an essential complex within a cell, mutations in one of the proteins could disrupt the complex, leading to disease. But in diseased cells you'll rarely see mutations to both genes. Raphael and his colleagues have developed an algorithm called Dendrix (de novo driver exclusivity) to look for statistical relationships in genomic data that could reflect these patterns. He has already used this algorithm with TCGA data for samples of acute myeloid leukemia, finding clear patterns of mutual exclusivity (5). And they are now analyzing data from other cancer types as well.

In addition to finding distinctive features and classifying tumor types, networks can also help researchers identify new patterns of mutations occuring in multiple cancers, similar to the *BRCA1* and *BRCA2* mutations that occur in both breast and ovarian tumors. These hotspots could suggest new targets for therapies that target multiple tumor types.

Looming challenges

Tumor samples present unique challenges in network analysis. They are heterogeneous, cautions Lincoln Stein of the Ontario Institute for Cancer Research, producing yet another layer of biological complexity.

A patient sample can include normal and tumor cells, with the tumor cells possessing multiple subtypes in some instances. Understanding that heterogeneity could be incredibly important for patients, Stein adds. In some cases tumors generate different subtypes from a common ancestor, while other tumors come from multiple independent tumor modules. Heterogeneity can affect the clinical outcome. For example, if a patient has one tumor subclone with an EGFR mutation but another subclone that doesn't include that mutation, a targeted inhibitor might kill only some of the cells in the tumor. Computational methods can help researchers tease apart some of these complications, Stuart says, but biochemical techniques that produce data from single cells would provide cleaner data to start with.

For Ideker, moving network models into a healthcare setting remains a major priority as such networks would assist clinicians in making sense of the rare mutations that show up in their patients. "No one is unique. These mutations are hitting the same regions of the network." And this is one time that it is good to be like everyone else.

References

1. **The Cancer Genome Atlas Network.** 2012. Integrated genomic analyses of ovarian carcinoma. Nature, *474*: 610-615.
2. **The Cancer Genome Atlas Network.** 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature, *499*: 43-49.
3. **Ng, S. et al.** 2012. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics, *28*:640-646.
4. **Zhang, Q.C. et al.** 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. *490*:555-560.
5. **The Cancer Genome Atlas Research Network** 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. New England Journal of Medicine, *368*: 2059-2074.

Written by Sarah Webb, Ph.D.

BioTechniques 55:105-107 (September 2013)
doi: 10.2144/000114072