Pathway Signatures

by Harvey J. Greenberg

How can operations research (OR), traditionally applied to management problems, help us to understand biological systems? OR emerged from WW II as not only a grab-bag of methods, but more importantly as a multi-disciplinary approach to problem-solving. Modern systems biology shares those same problem attributes, and the OR community is increasingly contributing to this frontier of medical research.

I entered computational biology when I visited Sandia National Laboratories (SNL) in 2000. Seeing the work of SNL researchers and learning about the problems showed me how operations research applies, particularly mathematical programming. This became the focus, not only of my new research, but also of my teaching and service. The following year I created the University of Colorado Center for Computational Biology and initiated a series of workshops and new courses in the departments of Mathematics, Computer Science and Engineering, and Biology. I was the main beneficiary! I continued to learn through research, teaching, and many new collaborations, particularly within the CU medical research community. In 2003 I visited Bernhard Palsson’s Systems Biology Research Institute at San Diego and later visited Leroy Hood’s Institute for Systems Biology. That is when I learned about systems biology and how much more OR can contribute. It is perhaps fortuitous that when I was a student, OR was almost synonymous with systems engineering (which has a different meaning in today’s world of computers).

One trick I learned from OR practice is to turn questions around, “Why is this true?” We do not fully understand why certain pathways are used, and research has focused on estimating parameters of biochemical interactions among molecules. Turning this around, I ask, “Why is this particular pathway used?” This gave rise to what I call pathway signatures – the natural circumstances that trigger one particular pathway but not others.

I setup a mathematical program and identified pathways that are optimal for a particular objective (defined by parameters). An example of an objective is to maximize ATP production in a metabolic network; another is to maximize reliability (probability of successful regulation) in a cell-signaling network. A simplified version is to choose the cone of linear coefficients associated with an extreme pathway for which that pathway is uniquely optimal, hence its signature. Among the infinitely many, I favored certain properties (from biological principles) and chose signature values that minimize total similarity between pairs of extreme pathways. This is key – the dissimilarity strengthens the idea of a signature, separating one pathway from the others.

Non-extreme pathways have more than one representation as combinations of extremes, and I studied approaches to see what signature offers the most insight into natural choices. With help (and data) from Palsson’s Lab, I enumerated extreme metabolic pathways (viz., red blood cell), on which I based...
my original studies. Another pursuit is to consider nonlinear objectives, such as quadratic growth, for which each non-extreme pathway is optimal. Understanding why it is better than any of the extreme pathways that define it might shed light on the underlying biology. From the few experiments I ran, redundancy was a paramount signature criterion. For flux balance analysis this means that the defining input-output matrix of the reactions had metabolite generation/consumption dependent upon others. At the very least, we can use this inference as a measure of approximation error. When the linear approximation is reasonable, we can investigate how the system is affected by the imbalance that occurs without fixed system-outputs.

Another part of the story is the Pathway Inference Problem: Infer knowledge about pathways with incomplete infor-

mation about their parts and interactions. In particular, there has been research into assembly – turning genes into pathways. I turn this into building a highway system and ask, “What junctions and connections provide the most efficient map for travel?” In this sense, pathway inference, or construction, is about optimal routing. Again, the idea of a signature is to find a biologically meaningful objective for which the network is designed (or revised) optimally. Much has been done in road and computer network design using OR with multiple objectives: minimize cost, minimize travel time, and maximize reliability, to name a few. There are many optimization principles at work in the design of life, so the natural laws of economics fit well into the OR descriptive motto: “the science of better.”

In conclusion, I see the future of systems biology broadening its network constructions to mixed-scale. The immune system, for example, needs to be represented by interactions among molecules, cells, tissues, and organs. A drug designed to block some pathways or enhance others can assume multiple targets where its behavior depends on the environmental context of the organism. We have seen such [re-]design problems in OR!

Links:
http://gcrg.ucsd.edu/
http://www.systemsbiology.org/

Please contact:
Harvey J. Greenberg, Professor Emeritus
University of Colorado Denver
E-mail: hjgreenberg@gmail.com

Oops I Did it Again.....
Sustainability in Sequence Analysis via Software Libraries
by Knut Reinert

Maybe you like Britney Spears. Maybe even her music. Maybe you are a Britney Spears fan working as part of a group on sequence analysis algorithms in computational biology. But am I right in assuming that you don’t like to hear the above song title quoted by your coworkers or programmers when they could have been spending their time doing something productive or creative? If I am right, then you might want to read on because I will tell you how, or your coworkers can avoid reinventing the wheel or writing a lot of inefficient scripts in sequence analysis.

Next generation sequencing (NGS) is a term coined to describe recent technological advances in DNA sequencing. Next generation sequencing allows us to sequence about 200,000,000 base pairs within approximately one week. That’s a two with many zeros. To explain it in different terms, whilst the human genome project spent many years and billions of dollars to sequence about 30 billion base pairs, we can now perform the equivalent amount of work within a day for a handful of dollars.

DNA as mass data has wonderful properties. While the sequencing machines churn out terabytes of data, a single byte of this data can be very important. It can decide whether you have a disease or not, whether the drugs you take do their job or not, whether you live or die. So we have to treat the data carefully. It is important to scientists working in the life sciences. At the same time, however, we need to process it fast and efficiently, after all it fills the racks of terabyte disks quickly.

In order to analyse the ever-growing volume of sequencing data it is essential that scientists in the life sciences and bioinformaticians work closely together with scientists in computer science. This can often be problematic since both communities approach the problems at hand quite differently (see Figure 1). While the the experimentalists have a holistic top-down view on what they want to achieve in a particular analysis, computer scientists usually work on a specific, small part of a larger analysis problem. On the computer science side this often results in highly efficient, but specialized algorithms that may not necessarily reflect the reality of real world data. Efforts from the life sciences side, in contrast, may result in analysis pipelines that compute a solution to the problem but are not state-of-the-art in run time or memory consumption, and hence cannot be applied to the large data volumes. The goal is obviously to use fast implementations of efficient algorithms to be able to cope with the volume of sequence data that NGS can produce. This can be achieved through algorithm libraries that collect efficient implementations of state-of-the-art algorithms, the work of algorithm designers and skilled program-