My Voyage into Mathematical Genomics

David Sankoff

In the late 1950s, science students graduating from my high school were led to believe, whether because of Sputnik or simply the charisma of the physics teacher, that the only respectable university degree at McGill University was honors mathematics and physics. An easily influenced 16-year-old, I went along and actually felt vaguely guilty about having been more interested in my 10th grade biology course. Or at least the botany parts, being less attracted to bugs, reptiles and rodents than to flowers and trees. Fortunately, while struggling through my undergraduate years, being kicked out of the honors program along the way and twice failing my Statics and Dynamics course, I spent the summers acquiring and applying tissue culture and virology skills and learning about DNA, the ongoing deciphering of the genetic code, and the discovery of mRNA, as a nepotic summer student in the lab of my uncle Lou Siminovitch at the Ontario Cancer Institute.

I made it through my mathematics Ph.D. thanks to tolerant comprehensive examiners and the encouragement of my supervisor, Don Dawson. He not only taught an advanced course on Brownian motion that I actually understood, and turned one of my weird ideas into a paper on the so-called Dawson-Sankoff inequality, but also told me to write a thesis on whatever interested me, which at the time was the phylogeny of language families, and the stochastic processes generating language divergence. I also benefitted from the large contingent of statisticians in the department, particularly Michael Stephens, who was a generous mentor to many students and who got me involved in his research on the distribution of goodness-of-fit statistics.

My first academic appointment in 1969 was at the Centre de recherches mathématiques at the Université de Montréal, and while I was continuing my work on phylogeny, my friend the late Robert J. Cedergren, a biochemist, posed a series of questions having to do with the comparison of RNA sequences. Over a period of five years, we published a series of algorithms and analyses for nucleic acid sequence comparison, multiple alignment and secondary structure prediction, along with mathematicians such as Peter H. Sellers and Vaclav Chvátal. I have written in detail about this period in the journal Bioinformatics (2000). A small number of other mathematicians, including Michael Waterman, were doing this kind of work at about the same time, but we did not fit in with any established community. I had the impression some of my colleagues thought we were crackpots and many biologists thought we were turning easy procedures (on short sequences) into unnecessary, complicated, algorithms. I did participate in mathematical biology meetings, but felt out of place with researchers in differential equations and statistics. I finally edited a collection of articles, including five of my own, with the late Joseph Kruskal in 1983. For lack of any alternatives, this volume eventually became somewhat of a classic.

At about the same time a growing number of mathematicians and computer scientists, motivated by the increasing number and length of DNA and protein sequences, joined us in this problem area, many molecular biologists became knowledgeable about algorithms, and we began to organize small meetings. By the end of the 1980s the newly named field of bioinformatics was quite busy, and we crackpots suddenly became pioneers. At about the same

time, the term "genomics" came into usage, and not long after "computational biology" was used to refer to the more mathematical aspects of bioinformatics.

In 1987, the Canadian Institute for Advanced Research set up a network of scholars across Canada in the field of Evolutionary Biology, and Cedergren and I were among the nine or ten Fellows appointed. Later on, Joseph Felsenstein and Michael Waterman became Associates of the program. At the first annual Fellows' meeting, I was listening to a talk by Monique Turmel, a biologist from Laval University comparing the order of markers on the chloroplast genomes of two algae. One of her diagrams, with lines connecting the marker positions on the two genomes, immediately suggested an interesting combinatorial statistics problem to me, which was easily solved with the help of my colleague Martin Goldstein, and indeed was well-known, as the late Sam Karlin later pointed out to me. Nevertheless this led to my 25-year preoccupation with chromosomal rearrangement and other gene order problems.

I soon expanded my statistical approach to combinatorial algorithmics. Fortunately, during the 1990s a good number of (but not all) the brilliant students and postdocs working with me found this problem area as attractive as I did. Whatever I know about algorithms, they taught me, while we formulated and solved, or almost solved, a number of comparative genomics problems, inversion distance for signed and unsigned genomes, more general genomic distances, the median problem and rearrangement phylogeny, genome halving, and the exemplar problem. These people are now all prominent figures in the field: John Kececioglu at Arizona, Mathieu Blanchette and Guillaume Bourque at McGill, Nadia El-Mabrouk at the Université de Montréal, Vincent Ferretti at the University of Toronto and David Bryant at Otago in New Zealand. Many other scholars joined us in this research endeavor, notably Pavel Pevzner and his students in California, who achieved a number of striking results.

At the same time I continued my gene-order modeling work, collaborating with Joseph Nadeau, who was briefly at McGill and Dannie Durand at Carnegie-Mellon.

When I left Montreal for the University of Ottawa in 2002, it did not take long for a number of other students and postdocs to get involved in my mathematical genomics projects. In a series of separate projects, Wei Xu and Chunfang Zheng made great strides in the median problem, including work with Eric Tannier in Lyon. Zheng also introduced guided genome halving, and developed a suite of techniques for ancestral gene order reconstruction. We became increasingly interested in flowering plants, because most of them descend from processes of genome duplication or triplication followed by massive loss of many of the extra genes. We developed consolidation algorithms, with Katherine Jahn of Bielefeld, and practical halving methods and applied them to detailed study of the grape, poplar, cereal, tomato, coffee, turnip and lotus genomes. Our biologist colleagues appreciate our ability to objectively deconstruct the processes of evolution, and I find myself finally able to indulge my high-school fascination with the botanical world!

Recent Publications:

- Phase change for the accuracy of the median value in estimating divergence time (A. Jamshidpey & D. Sankoff) BMC Bioinformatics 14, S15:S7 (2013)
- Ancient eudicot hexaploidy meets ancestral eurosid gene order (C. Zheng, E. Chen, V. A. Albert, E. Lyons & D. Sankoff) BMC Genomics 14, S7:S3 (2013)
- A consolidation algorithm for genomes fractionated after higher order polyploidization (K. Jahn, C. Zheng, Jakub Kovac & D. Sankoff) BMC Bioinformatics 13, S19:S8 (2012)
- Medians seek the corners, and other conjectures (M. Haghighi & D. Sankoff) BMC Bioinformatics 13, S19:S5 (2012)
- Listing all sorting reversals in quadratic time (K.M. Swenson, G. Badr & D. Sankoff) Algorithms for Molecular Biology 6, Doi: 10.1186/1748-7188-6-11 (2011)
- Multichromosomal median and halving problems under different genomic distances (E. Tannier, E., C. Zheng & D. Sankoff) BMC Bioinformatics 10, 120 (2009)

Books:

- Models and algorithms for genome evolution (C. Chauve, N El-Mabrouk & E. Tannier eds.), Springer, 2013
- Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families (D. Sankoff & J.H. Nadeau, eds), Kluwer, 2000

Related Links:

- My home page: http://albuquerque.bioinformatics.uottawa.ca