Terry Speed: Good Gene Hunting

Terry Speed

An interview of Terry Speed by Y.K. Leong

Terry Speed is world-renowned for his important and numerous contributions to the applications of statistics to genetics and molecular biology, and in particular, to biomolecular sequence analysis, the mapping of genes in experimental crosses and human pedigrees, and the analysis of gene expression data. A member of the NIH Genome Study Section from 1995 to 1998, he investigated fundamental problems arising from the Human Genome Project.

He has received numerous honors from the world's leading scientific bodies and has been invited to give lectures on his research; in particular, he was a Wald Lecturer at the US-based Institute of Mathematical Statistics. He has been on the editorial boards of international statistical journals, and currently of the Journal of Computational Biology. He is also the President of the Institute of Mathematical Statistics.

He holds joint positions in the Department of Statistics at the University of California at Berkeley and in the Division of Genetics and Bioinformatics at the Walter and Eliza Hall Institute (WEHI) of Medical Research in Melbourne. Each year, he divides his time equally between the two organizations.

The following is the result of an interview of Terry Speed conducted by the Editor of *Imprints* in three stages: an "electronic interview" shortly before he came to the Institute in January 2004 as invited workshop lecturer of the Institute's program on "Statistical Methods in Microarray Analysis", a face-to-face meeting at the Institute and a final "electronic interview" after he returned to Berkeley. The result is a frank and insightful revelation of an intellectual journey from a statistical beginning shrouded in abstract algebra and mundane experimental designs to one of the world's principal centers of activity responsible for the unfolding of one of the most dramatic scientific dramas of the 20th Century.

Continued on page 17

Continued from page 16

Imprints: What did you do for your PhD?

Terry Speed: My thesis was in algebra, being entitled something like "Topics in distributive lattices". I'd started thinking I might be able to do something interesting on (elementary) probability theory within a non-classical logic known as intuitionism, and ended up studying aspects of the algebraic structure underlying that logic. I hadn't done much algebra in my undergraduate degree, so I was to some extent just catching up. It was fun. I found that I liked algebra, which was news to me, and I have continued to enjoy the algebraic aspects of what I do more than the analytical ones.

I: When and how did you get interested in applying mathematics to biology?

S: I was always interested in mathematics and in biology. My aim on leaving high school was to go do combined science and medical degrees, and to go to work in medical research. This was in 1960, the year Macfarlane Burnet, then director of my current institute, shared the Nobel Prize for medicine. On arriving at university I found that I was ok at mathematics, but less so at practical science such as lab work in biology (e.g. dissection of rats and mice, looking at cells down a microscope, etc). So I switched entirely to mathematics and statistics, but included some genetics, which didn't have labs. That was given by a former student of R.A. Fisher, Peter Parsons. I wrote an undergraduate statistics thesis on a topic of R.A. Fisher's: the survival of mutant genes in populations, elementary theoretical population genetics. Throughout my PhD I was surrounded by outstanding people in that field, but I resisted the temptation to join in. At the time I was hooked on pure mathematics… so much to learn, so little time.

I: When you first started to apply mathematics to biological problems, was there any beacon at that time in showing the way, or did you have to hack a path through virgin jungle, so to speak?

S: From my PhD until the time I went to Berkeley I did very little on quantitative biology apart from what cropped up in statistical consulting and in collaborations. It was mostly very classical, with a couple of exceptions (e.g. some baby pedigree analysis): I doubt that DNA ever got mentioned. When I was in the maths and stats part of CSIRO I was conscious that Australia's leading "genetic engineers" were also in CSIRO, doing fancy DNA-related stuff near me in Canberra. Our division did statistical consulting for them, but there was nothing involving DNA. When I asked them could we statisticians get involved in this "gene-splicing" and perhaps help them with the quantitative aspects, I was told fairly firmly: no, there's no statistics of any kind in that research, go and design some more agricultural experiments for those old-fashioned folk over in the other building, and leave us to our high-tech stuff. So we did. Of course they were wrong … elsewhere in the world bioinformatics was being created around that time (early 1980s). Naturally I wonder how things might have been different if they had been more receptive … (It's always good to get into a new area early on, so I tell my students: when the basic problems are still unsolved!)

All this changed when I went to Berkeley in 1987, for there the "routine" statistical consulting that came in the door involved DNA: molecular evolution, intragenic recombination, and other topics, still of interest today. Then I realized I had to catch up with about 30 years of molecular biology, and fast, if I wanted to have a chance of answering the questions they brought to our consulting service. Incidentally, this is one of the many good things about doing statistical consulting: you never know what might walk in the door, and it really can give you new interests, and change your research directions. Of course, it is scary too, because you are on unfamiliar or only vaguely familiar territory much of the time.

I: When you first went to Berkeley in 1987, what was the state of computational biology like? Did you have any hunch that something momentous was in the brewing?

S: Momentous is a bit strong. It was clear that very interesting things were happening on the genetics and molecular biology front in 1987. PCR had just been invented, and was helping people generate lots of interesting data, the human genome project was starting to get talked about, mitochondrial Eve was in the air (later on the cover of *Time*), the first large-scale human genetic map was published, and so on. A big player in the genetic mapping world was Eric Lander, who I was told was a former pure mathematician. (He is now an even bigger player: a key member of the public human genome project, now forging ahead with grand plans in this post-genome era.) I missed his visit to Berkeley, but got to read his papers nevertheless. Also, I knew that Sam Karlin was very active in the field, and I quickly became aware of the many contributions of Phil Green, another former pure mathematician, and Mike Waterman, an ex-probabilist. So people from our area were already key players, and I might have thought "Why not me too?" But I just plugged away, trying to find a niche, thinking that perhaps I was already too late, that all the basic problems were solved! However, the forces that kept me involved were biologists. They were (and always are) so keen to use the latest and best methods, to be first to use a new technique in their particular corner of the subject, so if you are willing to try to help them, as consulting statisticians tend to be, you get swept along. You find yourself explaining and using Lander & Botstein's program for QTL mapping, Phil Green's CRIMAP, Sam Karlin's BLAST calculation, Mike Waterman's alignment algorithm, and with a little luck you eventually have an idea, or get a student interested, and away you go. In the decade 1987 - 1997 I learned two things: that the basic problems were not always easily solved, and that you are never too late for the next train (as Piet Hein used to say). In the mid-1990s microarrays came along, and after a while just watching, I tried my luck with some basic problems.

Continued on page 18

Continued from page 17

I: I understand that Berkeley now has a number of groups and programs that attract mathematicians, biologists and others to do interdisciplinary research. Could you tell us something about these groups and programs?

S: That is a tough question to answer briefly, and we have a web site devoted to answering it. I'd rather just refer the interested reader to it:
http://computationalbiology.berkeley.edu/

Perhaps I should add that we do have a wealth of what I abbreviate as "compbio" activity at Berkeley. So much that I could spend all my time going to seminars and sitting in lab meetings, and never find time to do my own work. That's almost my fate, but I have a number of wonderful PhD students who keep me active. I get people from other departments onto their thesis committees, I'm on the committees of students from outside statistics, and I ask my students to talk to or even collaborate with students from other departments and go to their lab meeting, and I have a small number of close collaborators of my own. If I play my cards right - and I'm still learning - I can benefit from this profusion of activity, and not get completely swamped.

I: If a mature mathematician wants to work on fundamental problems in biology or computational biology, how much biology does he or she need to master?

S: Short answer: lots. Longer answer: even more. Longer answer still: how much does a person need to know about the internal combustion engine to be a good motor mechanic? How much mathematics do you need to know to teach a course on group theory? Of course you can get by, perhaps get a few papers published in journals with little or no biology, if someone else has done the job of phrasing the problem in mathematical form for you. Then you may solve it, but that's not doing computational biology, that's doing mathematics.

No offence meant, but I sometimes say (and am doing so again now) that if you ask that question, you are already doomed (not to go far in computational biology). To put it another way, if you are not genuinely interested in biology, at least to the extent that you enjoy learning what you can about the area in which you are working, then it's probably not a great idea persisting in that area. Do what you like, I say.

I: Has statistical genetics discovered any general rules and principles about the mechanism of gene formation or combination? Do you see any parallel in the state of biology now and the state of physics one hundred years ago in the sense that there are many empirical rules and observations but a paucity of underlying theory?

S: Genetics has lots of general rules, and lots of exceptions. As for your second question, I don't accept the implicit assumption that physics is a useful model for biology. Perhaps it's just my lack of imagination, but I don't see us understanding life any century soon. We might think that physics has made great leaps towards understanding the universe at the level of particles and the universe, what with nuclear weapons, space travel, and laser scanners at supermarkets, but in my view this is easy compared to understanding cells. Wait a few centuries and you'll see what I mean.

I: Could you give us some idea of the problems on which you are working. What is your most memorable achievement?

S: That's hard for the reasons I outlined above. I'm not working on the Speed program or conjecture or hypothesis, I'm thinking (when I get the time) what I can do at all with some problems, and what I can do a little better with others. What are those problems? Well, they are always parts of bigger problems that belong to other people: what genes, if any, have their patterns of expression changed in the brains of people with bipolar disorder, in comparison with otherwise similar healthy people. What gene expression patterns change as we age? Finding ways of distinguishing real from apparent gene expression differences, in a variety of contexts, occupies a good deal of my time. At the "continuing challenge" level that is my aim: to distinguish the real from the apparent. This, of course, is a statistical problem with no single, final answer. There are others: I help people analyze their data to get better measurements of the things they want to measure. Then I sweat over questions such as: how can we tell this method of analyzing the data gives a better measurement than that method? My most memorable achievement? I'm still waiting. I hope to make a little progress on problems like the ones I just mentioned, and if I did, that would be memorable.

I: It is often said that this century will be the century of molecular biology. In your opinion, how much of this is hype and how much of it is scientifically justified?

S: Perhaps the best answer here is yes. That is, yes, it is hype, and yes it is scientifically justifiable. Beyond that I don't care to go. But don't count physics out.

I: Is there some kind of mathematical definition for a gene?

S: The history of the notion of a gene is almost the same as the history of genetics, at least in the period since the field had a name, which is essentially the 20th Century. So bear in mind that for biologists, the notion of gene is an ever-changing one. In the 15 years or so since the advent of large-scale genome sequencing, there has in essence been a mathematical definition of a gene, because people have used mathematical models to "find" genes - putative genes might be a better way to put it - in genomic DNA sequence. Of course, the definition of a gene computational biologists use is at best a crude approximation to what biologists understand by the term "gene", but if the computational definition does the job, i.e. if it finds "real" genes, no-one is going to mind too much. Nevertheless, a model that works in one context

Continued on page 19

Continued from page 18

is still just a model, and is necessarily different from the real thing, so no one should expect that the computational person's gene model embodies all important aspects of a gene.

With that preamble, let me say that a (protein-coding) gene could be defined as a DNA sequence structure consisting of a number of parts (promoter, transcription and translation start and stop signals, exons, introns, etc), each having characteristic features (base composition, dependence, specific motifs, …) and all arranged in a characteristic way. The complete specification today might be given via a generalized hidden Markov model (GHMM), with a given state space and set of parameter values, but do realize that this would hardly be recognized by a biologist. And furthermore, that before 1993, the GHMM view didn't exist. At that time a mathematician's gene might have been a neural network, and perhaps after 2010, the GHMM model will have been superceded by another, more complex mathematical object, embodying some aspects of alternative splicing, say, or post-transcriptional or post-translational regulation. This is standard in the history of science.

I: Could you share with us some of your experiences in learning biology?

O: I was always interested in evolution, and share the now conventional view of T. Dobzhansky: "Nothing makes sense in biology except in the light of evolution." Of course this doesn't mean everything that gets said in an evolutionary context is sensible or correct. Indeed I frequently think "What a total 'just-so' story" (or thoughts to that effect) about some evolutionary pronouncements, but it is undoubtedly true that intricate and wonderful biological phenomena can be made even richer, and more insights gained, by putting everything into an evolutionary context. And the comparative method (my definition: exploit historical or evolutionary considerations whenever you can) is a great thing. So that's experience number one.

The second experience I'd like to share concerns controls. Good experimental practice in biology typically involves the use of lots of controls: positive, negative, perhaps also calibration controls. In this context, controls are aspects of the experiment where the experimenter "knows" the outcome. For example, if you are obtaining aspects of a DNA fingerprint with an assay kit, you prepare a blood sample (say) in a prescribed way, and then you do the assay. The positive controls should give you the expected positive result (e.g. a bright pink spot) in a clear and unambiguous way, while the negative controls should unambiguously give the expected negative result (e.g. a blank spot). Crudely, if all went well, you should get something where there should be something, and nothing where there should be nothing, and the appropriate scale at the appropriate place. Such controls play an enormously important role in biological experimentation, and my point is this: it would be wonderful to have controls all the time, in all circumstances, and if we don't, wonderful if we could devise them. Statistics has a great need for controls, and so have many sciences that clearly don't (you can think of them).

My third experience concerns facts and interpretations. I've learned that facts and interpretations are different but more similar than we might like, and that Joe Friday's "Just the fact, Ma'am" is at best a gross oversimplification. Naturally, scientists like facts: that's why they do experiments. But they also like to draw conclusions: what do these facts suggest might be going on? Some of my most enjoyable experiences sitting in biology lab group meetings have been listening to discussions of alternative interpretations of the same set of facts, and of planning the collection of more facts, in an attempt to narrow down the range of interpretations. In such discussions you can see argument as the nature of the fact. It is even more interesting when one realizes that from time to time discoveries are made which were totally unexpected, for this reveals that no sensible interpretations of the data could have been made within the old framework, and the "fact" that an experiment delivered had to be refined before it could be interpreted. I like it a lot when dichotomies are revealed to be illusory. In physics people like to go on about relativity: how great it was when such and such an experiment involving the transit of Venus demonstrates the validity of some theory, and they have a few more examples. In biology this sort of thing happens almost daily. New, unexpected phenomena abound: restriction enzymes, retrotransposons, introns, microRNAs, … (look at the discoveries which have gained people Nobel Prizes in medicine over the last 30 years). Each can force a refinement of the "facts" (for example, was this or that controlled for? Was a certain contaminant present?) and a re-evaluation of the interpretation. That makes learning biology a great experience.

I: After all those years in Berkeley, you have now decided to spend half of each year in your home country (Australia). Is there any motivating reason for this?

S: The answer here is quite simple. My wife and I moved from Australia to Berkeley in 1987 for "a few" years. After a few more years than a few years, her pressure to return to Australia built up. Initially I was not very enthusiastic about most job possibilities back in Australia. I really wanted to stay in Berkeley. Then I found a job (my present one at WEHI) that I could get excited about, and my first thought was: can I do both? The answer so far seems to be yes, but it is an issue that gets revisited every year. One view would simply be that what evolved is a compromise, and like many compromises, there is always a tendency to want to go towards the simpler "pure state". As the guy who fell off the cliff said to someone half-way down: so far so good!