Michael S. Waterman: Breathing Mathematics into Genes >>>

Michael Waterman

Interview of Michael S. Waterman by Y.K. Leong

Michael Waterman is world acclaimed for pioneering and fundamental work in probability and algorithms that has tremendous impact on molecular biology, genomics and bioinformatics. He was a founding member of the Santa Cruz group that launched the Human Genome Project in 1990, and his work was instrumental in bringing the public and private efforts of mapping the human genome to their completion in 2003, two years ahead of schedule.

After his PhD in statistics and probability from Michigan State University, he taught at Idaho State University and visited Los Alamos National Laboratory for a short period before going to University of Southern California (USC) in 1982 to pursue a long and distinguished career in molecular biology, mathematics and computer science. The well-known "Smith-Waterman algorithm", which he developed with Temple F. Smith in 1981 for determining the degree of similarity (homology) of amino acid sequences from DNA, RNA and proteins, is catalytic in igniting the bioinformatics revolution. The formulae, which he and Eric Lander derived in 1988, are crucial for the so-called shotgun strategy for assembling genome sequences by cutting up the genome into short fragments that are easier and faster to sequence and then fitting them correctly together. In 1995, he published the first textbook Introduction to Computational Biology: Maps, sequences and genomes that laid the foundations of the new field of computational biology, of which he is considered to be the founding father. When he first went to USC, he started one of the world's first cross-disciplinary programs connecting genetics, mathematics and the information and computer sciences. With the setting up of the Center for Computational and Experimental Genomics in 2001, Waterman and his collaborators and students continue to provide a road map for the solution of post-genomic computational problems.

For his scientific contributions he was elected fellow or member of prestigious learned bodies like the American Academy of Arts and Sciences, National Academy of Sciences, American Association for the Advancement of Science, Institute of Mathematical Statistics, Celera Genomics and French Acadèmie des Sciences. He was awarded a Gairdner Foundation International Award and the Senior Scientist Accomplishment Award of the International Society of Computational Biology. He currently holds an Endowed Chair at USC and has held numerous visiting positions in major universities.

In addition to research, he is actively involved in the academic and social activities of students as faculty master of USC's International Residential College at Parkside.

Waterman has served as advisor to NUS on genomic research and was a member of the organizational committee of the Institute's thematic program Post-Genome Knowledge Discovery (Jan – June 2002). On one of his advisory visits to NUS, *Imprints* took the opportunity to interview him on 7 February 2007. The following is an edited and enhanced version of the interview in which he describes the excitement of participating in one of the greatest modern scientific adventures and of unlocking the mystery behind the building blocks of life.

Imprints: Your *PhD* was in probability and statistics. How did you get into biology?

Michael Waterman: My PhD thesis was in probability and I did my initial work in probabilistic modeling and iteration of deterministic functions. I got into biology in connection with Los Alamos. Stan Ulam, who was a mathematician, was interested in what mathematics you might need in the new biology. He brought Temple Smith to Los Alamos for a number of visits. Another scientist at Los Alamos named Bill Beyer had an NSF project for one summer and I came to work with Bill and Temple. That was how I met Temple Smith and what really started me in this area.

I: Ulam was not really a biologist.

W: Not at all. He started as a completely pure mathematician in the Polish school of mathematics famous for its problems begun in a café [Scottish Café]. He came to the US – I forgot who really brought him to Los Alamos – and worked on the Manhattan Project. He actually flowered there and

Continued on page 17

Continued from page 16

contributed to all of the amazing crowd there. He was one of those few great men who was still around many years later.

I: It seems that he had some foresight and could see that biology would be a thing of the future.

W: Yes, but I don't think he knew exactly what it was. He saw that it was intriguing and different, and it was clear to him there was something there.

I: I remember he wrote a book on "What is Life?"

W: That book was by Schrödinger. It's not really a very accurate book about the subject, but it inspired many people to take on the mathematical and other aspects of biology.

I: After Los Alamos, where did you go to?

W: When I started this work, I was a faculty member of Idaho State University. I was just visiting Los Alamos in the summer. Then in 1975 I went there permanently until 1982. In 1982, for most of the year, I had a visiting appointment in the department of biochemistry and biophysics at the University of California in San Francisco – a very biological place. Then I went to USC [University of Southern California].

I: All this while, you were still doing mathematical work?

W: I do mathematical work. At Los Alamos, I was in a statistics group analyzing energy data. Beginning with Temple Smith I also worked on stratigraphics in mathematical geology. I worked on a number of different applied problems, but because of the connections with Bill Beyer and Temple Smith, I was doing some of this work in biology, mostly algorithmic, in biosequence metrics as a hobby until finally, about 1981, 1982, I decided that that was what I would really like to concentrate on.

I: Could you tell us how you got into collaboration with Eric Lander that led to those famous formulae in physical mapping?

W: Eric Lander was still in the Business School at Harvard when he became interested in biology. After that he was at MIT and the Whitehead. Eric had written a proposal to a private foundation and that foundation asked me to look at the proposal. I met him in that connection and so I knew him. He became more and more interested in molecular biology; initially it was going to be neurobiology. We had an acquaintance and our joint paper came at the beginning of the genomics revolution. A t that time he had an office at the Whitehead Institute [for Biomedical Research]. I was visiting and a copy of PNAS [Proceedings of National Academy of Science] was on his desk, it had to do with the first papers on physical mapping, including a paper on physical mapping of yeast by Maynard Olsen, and John Sulston had a paper on physical mapping of C. elegans. The progress reported by these papers was slower than what people expected and the C. elegans paper, which used a different method, had a simulation study. Eric and I thought we should be able to do something with it. So we started thinking about the problem and I realized that it was a coverage problem. I remembered a little book I had seen – Geometrical Probability by Herbert Solomon – on coverage. We went to the MIT library in the middle of the night. We found the book but it didn't really pertain to the kind of problems we were looking at. Actually the problem we worked on was about car parking. Of course, you have a long street on which you park cars at random locations, allowing parked cars to overlap. What is the distribution of the coverage of the long street by this process? We found it was easier to work on the problem directly and we published our first paper. And then later on, Eric proposed a problem that was a little bit different and involved using short unique DNAs to anchor the clones (cars!) for the coverage. There was a paper by Arratia, Lander, Tavare and myself. Arratia and Tavare knew that the Poisson process was the right way to look at the problem. I really learned a lot from these guys.

I: I think Eric Lander once mentioned that he left mathematics for economics first and then biology because he thought that mathematics was too "monastic" and cloistered. In contrast, do you think that you are a mathematician at heart?

W: I went to USC holding tenure at the math department with a joint appointment in biology. Now it's the opposite; actually I hold a tenure in the biology department, but in my heart I'm still a mathematician and statistician. To pick up Eric's comment, I think there are certain people in mathematics who will be motivated to work in other fields. I'm certainly one of those. I like working on problems that require different strategies. One of the appealing things about working in biology is that there is more than one person interested in what you are doing. There is a community aspect of it which is very important to me and keeps me going. There is a very wonderful aspect to working in teams. I think at my age I would find it a very hard time continuing to try to prove things in ergodic theory had I stayed in that area.

I: In contrast, pure mathematicians tend to work by themselves.

W: By themselves, with the door closed. I can do that, but I kind of like talking ideas around and even with people telling I am wrong.

Continued on page 18

Continued from page 17

I: When you wrote the first textbook, published in 1995, that laid the foundations of computational biology, did you expect those dramatic developments in the field to occur within the following 5 years?

W: At that time, I didn't think we would have the human genome sequence by the year 2000. I knew there was going to be great progress, but things happened that I didn't expect.

I: When you wrote the book, has the project already started?

W: The first discussion, which I was lucky enough to be part of, was in Santa Cruz in 1985. There were a dozen people and I was the "computer person" in the group. In fact, there was a California magazine which had an article about this. In 1985, it seemed to be feasible, the computations seemed just possible. Someone brought up the cost of a military ship, and that put the cost of the project in perspective. The project started in 1990 and it was planned to take 15 years (or more) to finish it and push it further.

I: Was the book to some extent motivated by the project?

W: In part, by the data. I had been teaching a course since 1983. While teaching it, I was writing the book and correcting the chapters. I was trying to write the book for several years.

I: Was there any book before that?

W: There was a book [edited] by David Sankoff and J.B. Kruskal [Time warps, string edits and macromolecules]. It was a book based dynamic programming with various applications to genome sequences. That was an important predecessor. There was a book, a rather naïve book, looking at information theory approaches by Lila Gatlin, which was published in 1972.

I: Which is more crucial to the theoretical techniques used in gene mapping: the "better" algorithm or the "faster" computer?

W: I think, both, especially with the difficult mapping problems people attack today, not finding a mutation in a single gene but studying a complex phenotype involving multiple genes. It's not clear how far we will get. There are other extremely important computing facilities of which you have to take advantage.

I: Like parallel computing?

W: That's everywhere today, there is no doubt. In fact, students we recruit to our program ask about what computational facilities would be available to them.

I: Is it possible to break up the problem into small parts to work on?

W: Certainly for some problems it is possible to do it that way. Most of this parallel computing activity in biology is not sophisticated, but for problems where the processors have to communicate it is much more difficult.

I: What are the prospects of quantum computers in genome mapping in the future?

W: I have no idea, not a clue.

I: Do you have some guess?

W: No, I honestly don't. I mean, it's cool, but it will be some time before the quantum computers exist. It's counter-intuitive, some of these properties of quantum computers.

I: Your algorithm and other ideas on sequence alignment have also been applied to linguistics, human language development and even consumer purchasing patterns. Are you surprised by this, and do you know whether anyone has tried applying them to imaging or pattern recognition?

W: I'm not too surprised by it. I remember years earlier in Idaho trying to use the alignment algorithm to compare two different poems which clearly had a related source. I myself was trying in this direction. One of the earlier persons in this area, David Sankoff, has always had a very serious interest in linguistics, and so this connection was there all the time. For image matching, the alignment algorithm has been generalized to multi-dimensional objects – but it's not so clear how you make this work properly. Alignment has become part of pattern recognition people. Motivated by the problems from biology, people in the area looked at approximate string matching, at the statistical distribution of random strings, and more and more elegant string matching algorithms appeared. This area motivated a lot of work.

I: So much of the DNA in the gene is "junk stuff" without any apparent functionality. It seems hard to believe that Nature is so "wasteful" in her designs. Could it be that there is something we don't understand about this "junk" DNA?

W: I'm sure there is a lot we don't understand about the junk DNA. In very recent years, there are all of these microRNA genes that are around and have important applications such as regulating gene expression which no one knew about

Continued on page 19

Continued from page 18

a few years ago. But by the numbers I've seen, maybe 6 percent of the human genome is under positive selection, and we know maybe about 2 percent of it that is critical to the organism. There is a tremendous amount we don't know.

I: So most of the selection is not in the positive direction.

W: You are asking, "What about that other 94 percent or 90 percent?" You know, it may be like a typical mathematician's office, stacked with papers that you may never have to look at again, that you might use, that you don't throw out. I think that's the difference between the Executive Office where there is hardly clutter and the working mathematician's office.

I: Has there been any progress on this junk DNA?

W: People are all the time looking for patterns in it. Some of the answers may be in how the DNA packs into the cells, the accessibility of the DNA that initiates the copies. It is not clear.

I: Do you think that it is ever possible in the distant future to use extremely powerful computers to simulate how Nature experimented successfully with the nucleotides and other building blocks of life in producing the first primeval life form?

W: It's a question we will probably never know the answer to. They're fascinating questions. Personally, I very much like this idea that the original information molecule was RNA with DNA absent, but there's not too much known for computing the origins of life.

I: Could life have started as a kind of random process?

W: Many people believe that. In the naïve calculations, we take a protein molecule, 100 long and take 20 to the 100-th power, and say that's how it happened. Or that calculation and others like it are used to argue for its impossibility. These arguments are spurious in my opinion. Just how we got the original self-reproducing molecule at the origin of life is a really fascinating question.

I: Can we do some simulation? After all, the rules of combining are known.

W: Maybe. But people also try to figure out what the environment was, of course. That is one of the key ingredients. The complexity is enormous and then there's that billion years of the early earth that you have to catch up.

I: Are there any expectations for the next great conceptual breakthrough in biology?

W: I personally don't have any predictions. I wish I did. I'm still amazed by these small RNA molecules, genes that are so important and we didn't know about them until just a few years ago. And I'm sure there will be something else like that we just haven't thought of.

I: Does it mean that the RNA is more important than the DNA.

W: Well, if first operating molecules in the cell were RNA, that would make RNA very important. One of the key pieces of evidence relates to ribosomes which are assemblies (or machines) made up of sixty some proteins and three structural RNAs. Ribosomes translate messenger RNA into proteins. Harry Noller showed that ribosomes can function without all the proteins, just with the structural RNAs. This is quite surprising and suggests to me that RNA may have been there before proteins.

I: Are there any models for the origin of life? I remember Freeman Dyson once proposed some kind of model.

W: People are always writing about that. But I don't follow it carefully.

I: What advice would you give to someone who wants to study computational biology?

W: I feel that it's important to learn as much basic chemistry, basic physics and basic statistics as the student can. The basic facts are extremely important and some depth in mathematics and biology is also required, of course.

I: Do you have any PhD students?

W: Three PhD students and one post doc at this point. We are attracting students into this area who really come prepared. They know what they want to do and they take some serious courses … 15 years ago, people often had a degree in a different area and then converted to computational biology.

I: Are there any special programs in this area?

W: We have a computational biology PhD program within the biology department. The students take courses in biology, mathematics, statistics and computer science. They work hard!