Mathematical Conversations

Ron Shamir: Unraveling Genes, Understanding Diseases >>>

Ron Shamir

Interview of Ron Shamir by Y.K. Leong

Ron Shamir made significant contributions to optimization and graph algorithms and is one of the leaders in bioinformatics and computational biology whose pioneering work contributed to the historic completion of the ambitious Human Genome Project in 2003.

He went from Tel Aviv University and the Hebrew University to the University of California, Berkeley for his PhD in operations research. While he is based mainly at Tel Aviv, he has held visiting research positions at top universities and research centers in Seattle, Rehovot, Rutgers and Berkeley. A full professor in the School of Computer Science of Tel Aviv University since 2000, he holds the Raymond and Beverly Sackler Chair in Bioinformatics. He was also the head of the School of Computer Science at Tel Aviv.

He is actively involved in organizational and committee work for international scientific meetings. He has been invited to give lectures at major scientific meetings, research institutes and leading universities throughout the world. He is well-known for his tremendous energy of scholarship in reviewing activities and in serving on the editorial boards of many leading international journals in discrete mathematics, computer science, bioinformatics and computational biology – among them SIAM Journal on Discrete Mathematics, Journal of Computer and System Sciences, Journal of Computational Biology, IEEE/ACM Transactions on Computational Biology and Bioinformatics.

As the leader of an active and internationally well-known group (ACGT) on algorithms in computational genomics at Tel Aviv, he sets the direction of and contributes extensively to the research that has produced algorithms and software for gene expression analysis, genotype analysis, graph-theoretic tools for modeling biological systems and statistical software for whole genome association studies. Among other things, he continues to contribute to one of the central problems of the post-genomic era, namely the determination of the function of genes and pathways.

He was an invited speaker at the Institute's program on *Algorithmic biology: Algorithmic techniques in computational biology* held from 1 June to 31 July 2006. *Imprints* took this opportunity to interview him on 18 July 2006. The following is an edited and enhanced version of the transcript of the interview, in which he talked about the excitement of switching from a mathematical realm in theoretical computer science to uncharted virgin territory in bioinformatics and computational molecular biology. Here he also gives us an insightful glimpse of the "brave new world" of modern biological sciences and its impact on human life.

Imprints: You did your PhD at the University of California at Berkeley in optimization. Could you describe the route that took you from operations research to biology?

Ron Shamir: I did my PhD in operations research, but with a very strong computer science tendency. One of my advisors, Ilan Adler, was from the operations research department and the other, Dick Karp, was from computer science. I joined the department of computer science in Tel Aviv a couple of years after that. I worked mainly in the field of optimization problems until around 1990. At that time I was on sabbatical in New Jersey at the DIMACS [Center for Discrete Mathematics and Computer Science] institute, and I did some work on temporal reasoning – in this problem one has to place events as intervals on the time line based on given constraints on the relations between event pairs. At some point there was a workshop and I presented this work, and the late Gene Lawler was in the crowd, and he told me, "Listen, this is very appropriate for modeling physical mapping of DNA." I didn't know what DNA was or what physical mapping was. Indeed, physical mapping just deals with constraints of intervals along the line, but the intervals are not temporal events but actual DNA blocks. So I started reading biological articles and got excited about this field. It was in the early days of the Human Genome Project, and I found myself part of this new field – in the beginning only partially, but eventually this became my main research interest. The first steps of this transition occurred in Rutgers,

Continued on page 10

Continued from page 9

New Jersey, but of course, a lot of things happened in the following years.

I: Was there a group doing research in that field?

S: Practically no. There was one colleague of mine, Haim Wolfson, who was working on structural problems related to biology from the geometric point of view and who got into the field a little earlier. But even so the term "bioinformatics" was not invented yet. We have set up our first formal bioinformatics program in the year 2000 at Tel Aviv University. We came a long way since, with a very strong and active bioinformatics community across the university, including over 15 groups in computer science, life sciences, medicine, physics and engineering.

I: Did it occur to you to continue your work in the United States?

S: No. I got several offers over the years, but never seriously considered accepting them. Israel is my home country, the home of my family, of my nation, and of my mother tongue. Of course, scientifically such offers were tempting, since the research conditions in the United States are better, but science is not everything to me.

I: What is the difference between bioinformatics and computational biology?

S: Actually I don't think there is a difference. It used to be thought that bioinformatics dealt more with the databases and software, and computational biology used to deal with algorithms. In the past, I used them as two distinct terms, but I do not make the distinction because people are using them interchangeably. We ended up calling our BSc and MSc programs (which should have been called "computational biology") "bioinformatics", because this is the term people are using. Semantically, there is also a technical difference. Bioinformatics is primarily informatics (computer science), and computational biology is primarily biology. But in the way people use these terms, it is the reverse.

I: Structural genomics is often considered as an investigation into the language of genes. Has linguistics or the study of human languages cast any insight into genomics?

S: I don't have much to say about this because I don't know linguistics well enough. The only aspect of it that I am aware of is natural language processing. It's not really linguistics. It deals with collecting the relevant words from large texts like the collection of millions of scientific abstracts and the like. So in that sense, the mechanics of trying to parse scientific texts has been used. I would say that, in general, such approaches were pretty successful, but of course they are not as accurate or provide "clean results" as human investigators would do on the same task. It is good as an initial filter but it lacks human intuition and broad context understanding.

I: Is it correct to say that gene determination is more of a computational and statistical problem rather than a systems problem?

S: It's a mixture. I don't think you can separate them. Gene determination using just statistical or just computational methods has been successful in a limited sense. A few years after the human genome has been completely sequenced, we still do not have the full picture of the genome because our prediction tools are not accurate enough. People have been using additional species to try to get better gene prediction. People have been using the systems approach. I think we are still a few years away from coming up with the ultimate set of genes. This can only be done with integrated methods that use what we will learn from systems biology and comparative biology approaches, and, of course, from classical experimental methods in biology.

I: Are there any general principles which help you to say that there are only 5 percent of the genes that remain to be determined?

S: Five percent is just a metaphor, not a solid number. It is a rough guesstimate based on extrapolation of what is already known. Five to ten years ago, we thought that most of the gene regulation occurs at the level of transcription, and now all of a sudden, we have this huge wealth of mRNA, siRNA, microRNA, etc. that changes the picture completely. Who knows what else there is that we are not aware of at this point? For example, there is very exciting work about ultra-conserved regions in the genome that we don't know the structure and function of and there must be a reason that they are so conserved. There is a lot of signal probably hidden at the level of the packing of the DNA and making certain regions more exposed and or accessible for transcription. There's definitely much more in the genome than what we know at this time. There are a lot of exciting revelations waiting for us. That is what makes the field so interesting.

I: In that case, it will depend a lot on the technological advances available at present.

S: Definitely, a lot of it depends on the ingenuity of experimental scientists and on technology development.

Continued on page 11

Continued from page 10

I: Has computational work on DNA nucleotides yielded general principles of genome evolution?

S: I think this is probably one of the oldest areas in computational biology because research on evolution and on phylogenetic trees has been an active research area since the 60s. It's a very old area, on the computational biology scale. I do believe we know much more about the selection forces that act on the DNA. At the time of Darwin the belief was that positive selection was probably the dominant force. The general belief now is that most of the evolution is neutral. There are exceptional cases of either positive or negative selection, but neutral evolution is now believed to be more dominant. Of course, there may be surprises in stock for us in the "junk DNA" region, which covers the majority of the genome, and of whose evolutionary role we understand very little now.

I: By neutral, do you mean that it is random?

S: Yes, most of the changes in the DNA do not affect the well-being of the phenotype; most of the mutations are neutral. Occasionally a combination of such mutations will have an effect – even a dramatic effect – on the phenotype.

I: Could it be that anything we don't understand, we say it is "random"?

S: We just don't understand it at this point. We are in the dark but it's definitely not random. Take the occurrence of ultra-conserved regions. It's definitely not random, it's statistically very significant, but we don't understand the role of these regions.

I: If I understand it correctly, much of the DNA of the human genome is considered to be "junk DNA" in the sense that they do not contain recognized functional elements. How can we be sure that indeed they do not contain "recognized functional elements"? Is it possible that they may contain such elements which we are ignorant of?

S: Definitely, yes. There are probably a lot of functional elements that we are not aware of either because we don't have the technologies to identify them or we simply didn't ask the right questions. What happened during the last 5 to 10 years has shown us that our knowledge is very limited. For example, what happens now with the genome-wide chips is that we used to think that expressions happen only in the regions coding for genes, and now all of a sudden we have evidence showing that a lot of expressions is going on in non-coding regions, contrary to what we thought before. The same thing happens with the binding sites of transcription factors. We used to look at them only in promoter regions. Now when you look at the binding in a genome-wide fashion, using ad-hoc chips and some of the techniques developed here in Singapore; you see that there is binding all over the genome, very far from known promoters. So definitely there is a lot of function out there that we are not aware of.

I: It may seem daunting for an outsider to go from biology into computational biology or from computer science into biology. From your experience, what is the least painful, if not the best, way to make such a transition?

S: Luckily for the young generation of students, there are already integrated programs. For example, in Tel Aviv University, we have, back in 2000, set up an undergraduate program where students get a full double major in computer science and biology plus a specialization in bioinformatics. So they can really speak both languages. We now also continue a similar program into the graduate level. For the young generation, it is simple. The transition for someone already educated in one of the three fields is indeed not easy. The different disciplines use different languages, both in terms of terminology and in terms of culture. I came from computer science and mathematics. To me a proof is something where you can write "QED" at the end. Once you've done it, the problem is solved. In biology, the notion of a proof is very different. A proof can be re-proved and un-proved. The notion of a definition that we cherish in mathematics does not exist in biology. The terms keep changing their meaning because of new light that is shed on them. A lot of the difficulties in the beginning were created since each area had its own culture and its own way of thinking. There are some cultural barriers in between. Many biologists of the previous generation are not that eager to try to speak the mathematical language. Many mathematicians are not eager to speak biology.

The transition that I went through – going from computer science and mathematics to computational biology – is easier than the transition required for a biologist if he or she does not have any basic training in computation, because first of all, the language of computation is very formal and very hard to pick up in an informal way. You really have to go to classes or digest the textbooks. Also, mathematics is very structured. You cannot learn "B" before you learned "A". Almost everything is very dependent on prior theory, in contrast to biology which is much more "flat". Another advantage for someone making the transition from the computational side is that biologists have wonderful textbooks – they are very clear and full of illustrations. The textbooks in computer science or mathematics are much less

Continued on page 12

Continued from page 11

friendly; so trying to learn from the literature in mathematics or computer science, if you don't have any prior training, is much harder than trying to learn biology from the books. Even so, it is not easy. It's a difficult process. If you ask, "What is the easiest path?" the easiest path is to be 18 now. Then you can learn it in an integrated way. There are very good programs both in Europe and in the United States. I don't know if you already have such programs here for undergraduates.

I: We have started to some extent, but we are at quite an elementary stage and still developing.

S: I would suggest – I don't know whether you have considered it or not – even if you don't train the next generation of biologists in computational biology, train them in computation. Have them learn one or two courses on basic principles of computation so that they will have basic knowledge in this "foreign language". It is worth the effort to include it into the biology curriculum. Also, give mathematics and computer science students one or two courses in life sciences, just the principles, so that they will be closer in language. Then, in graduate school, if they decide to go into the other area, they will have an easier start.

I: What about asking computer scientists to write better books?

S: That would be great, but you need people who will be willing to put in the effort. I think it's mainly a marketing issue. There are very good books for software manuals, simply because there are many thousands of people who will buy it. Biology is taught to millions of students. Computer science students are ten-fold or twenty-fold fewer; so there is not enough demand. There is not sufficient reward to simplify these texts; it's really hard work to turn something complicated into friendly and simple concepts. And in the end, there is only so much that you can simplify – mathematics is a formal language and a formal discipline.

I: How much benefit has the complete mapping of the human genome yielded to the medical and clinical sciences?

S: Tremendously, and it's only the beginning. For example, we know by now, as a by-product of the Human Genome Project, the causal genes for the majority of the Mendelian diseases. A tremendous amount of knowledge that we now take for granted wasn't there without the Human Genome Project. It has already made a tremendous difference and it will continue to. For example, the hapmap projects which aim to map all the single nucleotide polymorphisms, are still under way. They have already revealed millions of mutations which make the difference between all of us – different features, different life expectancy and so on. Getting such information would have been inconceivable without the reference human genome. We talk about "the human genome" as if it is a unique genome, but it's just a reference. But once you have a reference, you can start zooming in on individual mutations to see how they relate to diseases. I think we are going to reap the benefits of this visionary project for many decades.

I: This sounds like a reductionist view in biology – that everything can be reduced to the genes.

S: Well, not everything. If you can explain 50 percent of diseases just by looking at the genotype and the other 50 percent by other causes, this is a great step forward. In 5 or 10 years, you will be able to have your full genome sequence, for a reasonable cost, and the doctor can tell you, "Look, you don't have to worry about smoking because with your gene combination, it will not make any difference. On the other hand, you should be very worried about your cholesterol or whatever." I don't think the genes are everything, but they account for quite a lot. They will tell us whether someone is more likely to have a particular disease than others, and if a certain lifestyle is going to make a difference for him or her in terms of quality of life. Of course, all this is a blessing but also a curse because the genetic information must be used and not abused. To a large extent, this is not only a thing of the future but is already here today. People have been doing pre-natal tests to identify all kinds of defects, and they will be able to do much more – and more post-natal tests in the future. We can't stop this knowledge, so we'd better use it for the best of our understanding.

I: Maybe in the future we will be able to look at a person's genome and say that he or she will have a stroke at a certain age.

S: I don't think it will be able to tell you that – but it can give you probabilities. You will be able to tell someone that changing the lifestyle will make a big difference in her case. Definitely. Eventually, it's all probabilistic. There are relatively few cases of combinatorial fate. It's up to us. The more we know, the more we can control it.

I: Is there a gene that determines the lifespan of an individual?

S: Probably much more than one gene. In mice, scientists found a gene that affects longevity very significantly. We

Continued on page 13

Continued from page 12

know that this has a lot to do with the shrinking of the telomeres during the life of a creature. It's not a single gene but quite a few genes affect longevity.

I: In that case genetic engineering can lengthen lifespan.

S: That's one of the dreams. I think real genetic engineering in humans is still far ahead, but in principle, we might be able to do so.

I: It seems that in biblical times people had long life spans by modern standards.

S: I think that they just counted differently... They talk about Abraham and Sarah, and Sarah had a baby when she was 90. They probably lived, in modern way of counting, to the age of 30 to 35. Life expectancy at that time was much shorter than it is today.

I: Research in genomics and proteomics usually involves multi-disciplinary team effort whereas the tenure system in the universities (at least in the United States) centers round individual achievements. For that reason, a prominent biologist has said that progress in modern biology will come from institutes of research rather than the universities. Do you agree with this viewpoint?

S: No. First of all, the university system is not that blind to joint effort. Credit will be given to several partners in case of joint work. Even in publications from research institutions, there is the first author, the second author and the last; so credit is not spread totally equally. Institutions outside universities have other advantages over the universities because they may be more flexible, and resources can be moved more easily, but I don't think the issue of credit for work is the primary issue. In my own university we at Computer Science School have a lot of joint projects with colleagues in the Medical School and the Life Sciences Faculty. If we are successful, then everybody takes credit for the success. The question of how this credit is partitioned is there, but it is not specific to universities. I don't think that the tenure system is an obstacle to interdisciplinary research.

I: For the younger faculty, the emphasis is on single papers rather than joint papers. In that sense, a younger faculty will not want to risk going into another field.

S: But on the other hand, I have some young colleagues in my university, some of whom were my students in the past. In bioinformatics and computational biology, a lot of what they do is joint work. On one hand, it's not single-author papers, but on the other hand, they are involved in more projects, so they have more papers. It balances out. How many research projects can you carry out just by yourself? If you work with other people, you can be involved in more projects because you only do part of each project. I may be naïve about it, but I don't know of cases where this is the main obstacle.

I: You are heavily involved in many projects at the same time. How do you manage to do it?

S: I've been very lucky to have wonderful students. In the last few years, I was also heading the School of Computer Science in addition to running my group and teaching, etc. The secret is really to have wonderful students. You don't need to guide them on every little detail. Once the group has a critical mass, there is joint effort and there is a lot of assistance by the mature students to the younger ones. Also, it's more fun to do more diverse things. I may be doing a little too much, but I have 4 or 5 different areas that I try to be active in. As long as each of these areas is exciting to me and as long as I have such wonderful students, I will continue. As my group is quite large, I seldom work on my own. I work with others, mainly students and also colleagues. Students do individual projects, they get individual credit for them and write theses. It's mostly individual work but it's done in a framework of a supportive and unified group.

I: If I may say so, mathematicians are quite notorious in working mainly on their own without getting involved with others.

S: You are right. It's a different culture in computational biology. In my early years, my papers had only one or two authors, but my papers of today may have 4 or 6 or sometimes 10 authors. Part of it is because it is the culture of a different field. Part of it is because the projects are more complex and have more aspects and require more diverse expertise. They are not as deep as pure mathematics projects but they are complex and therefore there are many people and sometimes several groups involved.

I: You hold a number of patents. Do they pertain to the algorithms or the source codes of the software?

S: Only algorithms. The codes are typically protected by copyright, which is a different type of protection. Actually all these issues are handled by the technology transfer unit at Tel Aviv University. They define what justifies patenting and also copyrighting for software. All the tools we develop are completely free for academic use, and we make an effort to make our tools useful for the academic community.

Continued on page 14

Publications >>>

Continued from page 13

Occasionally, there is interest from the pharmaceutical and biotech industry. In that case, licensing and patenting have to be addressed. But for me, the issue of patenting is of low priority. It is more important for me that our tools will be useful to others. Our software is not as robust and as convenient as commercial software. We don't have the resources to do this, and I also cannot give academic credit to students for doing work that is purely technical, like graphical interfaces. This work is more appropriate for a company. If at some point, some company would like to take the algorithms and the basic software and package them into something fancy, that would be great. We package the software to make it useful for us, and also, we hope, useful for other academic groups and occasionally to pharmaceutical companies.

I: Have you ever gone back to your original field in optimization?

S: I never really left it. I still find it interesting and I still try to find the optimization or graph-theoretic problem behind any computational problem that we address. Over the years, I realized that you have to compromise in terms of elegance what you do in order to be useful to the biologists. Perhaps 99 percent of the problems in bioinformatics are NP-hard, and only occasionally you can develop approximation algorithms for them.

I: Typically, in spending your time, do you want to dwell on the theoretical aspect of the problem or do you want to find something that works?

S: According to my training, I would, when I just started out, devote 100 percent of my time to the theoretical aspect of the problem. But if you really want to get new findings in biology, you have to compromise: you will not have time to prove everything rigorously, and you need to develop codes and not just algorithms, because the algorithms by themselves are not useful to the biologists or the medical people. But I still think of many of these problems in terms of optimization. Interestingly, in Operations Research, there is a strong emphasis on modeling. You have a real-life problem and a big challenge is to formulate it mathematically in a useful way – for example, as an optimization problem in integer programming. In recent years, I realized that in biology, a big and sometimes crucial part of the research is getting to the right problem formulation. In that respect, I am more appreciative now of the emphasis on modeling than I used to be when studying operations research. In addition to optimization, I find myself doing much more statistics than I was trained to, since the bioinformatics area requires it.

I: Is computer simulation done in bioinformatics?

S: Some people do it, particularly for modeling the dynamics of networks. There is also a lot of the use of Monte Carlo methods (I don't know whether you would call them simulation in the strict sense). When it is very difficult to theoretically analyze a particular distribution of outcomes, you can just sample it and see how the results are distributed. It's quite efficient in practice. Of course, there is also the whole field of molecular simulation where you try to study the dynamics of folding and interactions between molecules and which is a huge area that requires tremendous computational resources.

I: From a simple-minded point of view, is it possible to have a model to simulate the rules of combination of the genes by random selection from a large pool of the building blocks of genes?

S: In principle, probably yes, but we are still very, very far away from that.