Use of archived final transcript

Members of the Duke University community, students, faculty and staff at other institutions, or members of the general public may access the digital archives. Typical research uses of interview materials include scholarly or other publications, presentations, exhibits, class projects, or websites. However there may be other uses made as well, since the materials will be available to the general public. Investigative reporters and lawyers engaged in or contemplating litigation have, for example, used the Human Genome Archive.

Your permission to post the edited, written transcript of your interview, and any related documents, to a digital archive is completely voluntary. Unless you consent to their wider use, all materials from your interview will be available only to members of the research team affiliated with this project.

The form below provides you with different options for how, when, and with whom your interview materials will be shared.

(A) ___X___ I place **no restrictions** on my interview materials.

OR

(B) ___My interview materials may be reviewed, used, and quoted by students and researchers affiliated with Duke University; *and in addition* (check all that apply):

___Researchers unaffiliated with the Center for Public Genomics may **read** the interview transcript and any related documents only after obtaining my permission.

___Researchers unaffiliated with the Center for Public Genomics may **quote** from the interview only after obtaining my permission.

___Researchers unaffiliated with the Center for Public Genomics **DO NOT HAVE** my permission to **read or quote** from the interview.

Posting interview materials to public digital archives: In spite of any restrictions listed above, I give permission for my interview materials to be made publicly available on the Internet by deposit in an institutionally affiliated archive:

___ 1 year from the date of this form

___ 5 years from the date of this form

___10 years from the date of this form

___25 years from the date of this form

___After my death

___Other: _____ (please specify a date or condition)

Signature: Michael Waterman Date: October 18, 2012

Interview of Michael Waterman

Waterman: If the connection gets weak we can cut the pictures off but it's always nice to see the other face.

Ogez: Ya. Sorry about that confusion there for a little bit. I haven't used Skype that much so I thought you just called the number.

Waterman: Well, ya you can. So it's Skype to Skype is free, Skype to a number is, you know, 2 cents a minute or something.

Ogez: Oh okay. Okay so before we started things off I just wanted to say thank you very much for agreeing to do this. It means a lot to me and I'm sure you're a busy man so thanks for spending your time doing this.

Waterman: Sure.

Ogez: And also, you are aware that you're being recorded. I have to say that.

Waterman: Yes.

Ogez: Okee doke. And so, a little bit about me. My name is Michael Ogez. I'm a freshman here at Duke and I'm a prospective math and computer science major and as part of my class "A Social and Political History of Genomics" we're trying to reconstruct a history of the human genome project and just of genomics in general through interviews with people involved so I chose to interview you because you've done so much in the field that I want to study. Okay. Did you have time to look at the questions?

Waterman: I did. Yeah. Sure.

Ogez: Can we just start with the first one, so, a little bit of background.

Waterman: Sure. I grew up on a livestock ranch on the Oregon coast. And so the professions I could see as a kid were…we didn't have electricity most of the time so the professions I could see as a kid were dairy farmer, livestock rancher, logger. And that was sort of it. I knew there were little towns nearby. I couldn't imagine being a store—you know owning a store or something. So going to Oregon State was a way to get out of that life. And I initially started in engineering because that sort of sounded like something I could get a job in. I switched out off engineering mostly because in some sense it was practical in that it was about doing things not understanding things and I was really fascinated by the facts that human beings actually had an understanding of things like chemistry and physics and so on.

Ogez: I sort of feel that same was about engineering. That's sort of why I like math and computer science more.

Waterman: I think if computer science had existed I would've almost surely gone into computer science but it didn't exist then. There was a little electrical engineering but somehow I switched into math where the hour requirements were fairly moderate so I took a lot of earth science and I also took a lot of literature and philosophy and stuff. Which is a great interest of mine.

Ogez: So later on after you had your background in math and computer science what made you choose computational biology as the area in which you wanted to apply your skills?

Waterman: Well I didn't initially—it was through Los Alamos. I went there a few times in the summer at Los Alamos National Laboratories. And one of the famous mathematicians from the Manhattan project, Stan Ulam, was there. And he had some instinct that there was something to do in this new area of biology, and he in fact brought Temple Smith to visit, and that's where I met Temple.

Ogez: And Ulam was one of the founding fathers of computational biology, right?

Waterman: Well I'm not sure that's quite right. But if you just look him up if there had been a Nobel Prize in mathematics he would've had it. He started out as a very theoretical mathematician in Poland and then found that he could do certain things in theoretical physics. There's a beautiful story about he and Keller contesting over the ideas of the hydrogen bomb. And Ulam had a key idea for that as well as for the initial project. So he was interested in this area. And his instinct was there was something to do here. Again that's how I met Temple.

Ogez: So that sort of leads me to my next question. So the Rockefeller, the meeting there in 1979, Temple smith also attended—

Waterman: So we started I think in 1974 working together and by 75 we had a couple of manuscripts that took a while to publish.

Ogez: Which manuscripts were those?

Waterman: So they were the '76 publication "Advances in Applied Mathematics" which did the multiple alignment and the variable length insertion/deletion. And then there was maybe a '78 paper which was on evolutionary trees.

Ogez: With the multiple…

Waterman: Ya there were several things in the '76 paper. The one it's known for is the multiple gaps.

Ogez: So the 1985 Santa Cruz meeting organized by Robert Sinsheimer. Can you talk about your role in the meeting and how important was your math and computer science background?

Waterman: Ya so let me finish to answer a couple of your other questions. So we had this collaboration Smith and I. He knew much more about biology than I did at that time. And I was just interested in what were the problems, you know. And so we spent an enormous amount of time formulating problems. Because he was a physicist he had a different view of what a problem was than I did. He knew something of the biology. So it was a lot of fun. And our work substantially changed I think. A huge thing happened when he was at a Gordon conference—I think in '78—and when he came back and he told me about introns. And if u were in a world that knew about bacterial genes and then we all would've assumed that all organisms were coded like that, the fact that there were intronic pieces was an incredible shock even to my limited knowledge of biology. I thought, "How could this be?!" We immediately knew that these alignment methods we were devising were not appropriate for this new setting. And we didn't know what the problem was and we fussed around with this problem for some time—a couple of years. And Peter Sellers had produced a very elaborate but not very useful method approaching the same thing before we finally had this simple algorithm and could state what the problem was.

Ogez: That's interesting because nowadays we think of introns as just something that is so simple and we think, "Oh, of course it is that way." So that must have been quite a shock.

Waterman: Yeah, so imagine that's all we knew were bacterial genes. Who could've dreamed up that in so-called higher organisms there were these snippets of nucleotides in amongst the coding that could be cut up and thrown away. I mean, you know, this would just be ridiculous, right? And yet there it was. Why was it? Which came first? What was it doing there? I mean incredible!

Ogez: Since you were talking about developing algorithms we can skip to the 5th question about how—what was the best part in the discovery? Is it when you first think of the idea or is it when you can finally look at your work and say, "Look what we've done?"

Waterman: For me there are a couple of parts. One is sensing there is something to work on. There's something really interesting here. I don't know exactly what it is, but "Ah!". And trying to formulate a problem. And that's half of it anyway. If u can figure out what it is you're trying to do in the morass of imprecision that biology gives us. So that's, you know, finding a good solution is maybe the other half. But certainly half of it is and maybe more than half is finding the area and articulating the problem. And I really like doing that. I like the sense that there's something there and I don't know what it is yet. So then it wasn't long, you know, there was this

meeting at Rockefeller and that was—I don't know how influential it was or not. I think it was a step towards getting a national database.

Ogez: What did you exactly do at that conference?

Waterman: I just listened to people—tried to figure out what the hell was going on. The Santa Cruz meeting that you brought up was really a wonderful experience. I probably was invited because of Harry Noller who was at Santa Cruz and I knew him quite well.

Ogez: He was one of the organizers?

Waterman: He was one of the organizers and a genius biologist incidentally. And I had never met most of these people before. David Botstein who was very opinionated and kind of was a shock to me and George church who was a very young guy. Wally Gilbert who again was another very opinionated guy. David Shwarz who had just done pulsed field electrophoresis. And we've had a friendship since that time. But I was the computer science—as you say—I was the math and computer science person.

Ogez: Were you the only one?

Waterman: I was the only one. So the question for me was, "Is it possible computationally to do this job?"

Ogez: To handle all the data?

Waterman: Yeah. And you know the cost—you guys probably know this better than I do at this time—but the cost of reading a base was something like 12 dollars. And if you had to do a depth of 5, multiply 12 times 5 times 3 billion and you come up with a lot of money, right? So when we tried to figure out the cost, "Oh, suppose it could be a billion dollars." A dollar a finished base. Which turned out to be a pretty good estimate. So there was a lot of optimism there. And applying some optimism to the storage and computing time of a computer it seemed to me that it was possible to do the computing also. In the end probably about as optimistic as the 3 billion dollar price tag but in the end it worked.

Ogez: Did everyone else there sort of share your optimism? Or were there some people that didn't think it was possible?

Waterman: I don't think there were any. In that setting I don't think there was anybody who was a big skeptic. There were some serious—you know—there were some very feisty discussions about everything. One of my favorite parts of that whole meeting was at the end we were talking about the cost—you know— how would the U.S. pay for this and is this something we could do? And someone—I don't remember who it was—brought up I think the aircraft carrier cost—the price of an

aircraft carrier. And from that moment on it seemed to me we could afford the sequencing of the genome and know about our genetic heritage even if it were only for observational reasons and it had no health information at all. Just to know what we're carrying around in us and what we're passing on. it seemed to me incredibly worth it compared to what were spending on military spending.

Ogez: What was it, a couple hundred million dollars?

Waterman: You know I don't remember but ballpark.

Ogez: So about your work with Smith. You have your publication in 1981. We sort of already touched on this but can you take me through how you figured out the correct algorithm and just the process of figuring out the correct answer.

Waterman: You know we fussed around a lot. We came up with some things that didn't work and the magic of putting the zero in and you go, "Oh yeah," and then, you know, you can write as soon as you see that. Now I can make a precise partition statement of the problem and write a very simple paper on it. Had somebody said, "Guys, here's this explicitly stated problem," we would have solved it, you know, that afternoon. We hadn't had a good—we didn't have a good formulation of what we were doing. And incidentally a key part of that—us being able to make that step—was that…do you know the name Margaret Dayhoff?

Ogez: I do not.

Waterman: She was a woman who had come from physics and she had the protein database identification resource—I think it was called—at Georgetown. And she was a big figure in this area and a very tough lady. We were coming up with these metric measures between sequences and the Needleman-Wunsch measure is a similarity measure. We thought everybody should be using the metrics and she said, "Well what's the difference? I don't care unless you show me there's a difference." And I sat down to prove that these weren't equivalent notions and I instead proved that they were equivalent notions—that there is a duality, if you measure entire sequences, there is a dual version of the problem where maximum similarity and minimum distance are the same. It's a very simple calculation but that really set me back. This arrogance we had about metric spaces in mathematics—this implicit arrogance was kind of destroyed by that. So we were very open to similarity measures and that's key for our '81 algorithm. If you look at pieces with 0 distance—4 A's is zero distance from 4 A's. 15000 A's is zero distance from 15,000 A's. And yet those are very different statements. And so if you measure similarity you're counting up the number of matches or something, 15,000 and 4's are very different numbers to tell apart.

Ogez: And so you mentioned that since you didn't really know what the problem is you were sort of in the dark messing around for a little bit—

Waterman: We invented an algorithm that would later be useful for determining repeats in sequences (laughs).

Ogez: Did things usually work out like that? Where you don't really know—

Waterman: Absolutely, and you fuss around try to make formulations.

Ogez: So my next question was: most mathematical publications are single-author the creative act is solo, whereas in science it's usually an entire lab team working on the project. So how does research in computational biology compare?

Waterman: Yea although there are a lot of collaborations in computer science and more in mathematics these days. But, you know, I think, for example, in my case, in those days, I mean, I couldn't have worked in this area without him. And he critically needed my ability to write a coherent sentence as well. He wrote somewhat elaborate sentences that were hard to analyze sometimes. We would tease each other about this. Later, I'm guessing in the early '80s, I learned a lot more biology and it became a different thing, but I think it's people bringing different skill sets to bear on the same problem and the ability to communicate and work across these boundaries, which I really like.

Ogez: So it is more of a collaboration.

Waterman: It's definitely a collaboration.

Ogez: In your opinion how important of a role did the mathematicians and computer scientists play in the human genome project?

Waterman: Without that work there wouldn't be a sequenced human genome. I mean just in terms of collecting and quantitating the data, figuring out how much you believe the "a" in the fifth position of the sequence is a mathematical and statistical problem. And obviously if we didn't have computers modern biology simply wouldn't exist. Every lab refers to a database and runs various kinds of analyses.

Ogez: Ya, what would exist without computers these days?

Waterman: Ya it's hard to say. It's sort of like thinking about what if introns weren't there.

Ogez: What do you think were the signal contributions to the field of computational biology so far and what sense of what might happen in the future do you have?

Waterman: Well I certainly think sequencing the human genome was a huge contribution and there are lots of others. It's hard for me to bring one out that's of that magnitude, but sort of just open up an issue of science and you'll find one. But

you know today there's much more data. As we learn more about biology, which is so complicated, the problems become much more complicated. You know, epigenomics, methylation, its basis in the genome and analyzing that data. The computational biology follows the technology.

Ogez: It seems to me going forward all the problems are incredibly complex and there's no way to do them without computers.

Waterman: That's for sure. I'll occasionally get asked about the future of computational biology, which you started asking about. And I turn that around, saying, "I will predict it with 100 percent certainty if u tell me the future of biotechnology. " Because really this analysis and computation follows the data, follows where we can take measurements and how we can take those measurements.

Ogez: What advice would u have for someone like me considering a career in computational biology?

Waterman: Obviously you can't learn everything, unfortunately. I think that systems of differential equations are increasingly going to play an important part as systems biology moves ahead. But take some good physics courses, take some good chemistry courses, take a genetics course or two. Your next step after your undergraduate school—where to go to graduate school and how to aim your career—that is going to be an important one.

Ogez: You mentioned differential equations. How does that apply to—?

Waterman: So if you look at neurobiology, the data recorded in neurobiology is down to microseconds. They can do some amazing modeling because they have amazing measurements. Imagine with all these components in the cell and all these things that are going on and we have better and better measurements of these things. How these reactions take place is a matter of chemical reactions, right? The diffusions and chemical reactions. So if you just look up—what am I trying to think—there's this area where people construct little genetic circuits. There's a word for this I'm missing. So they have systems in the test tube they've engineered so they can watch the feedback loops work and so on.

Ogez: Synthetic Biology?

Waterman: Synthetic biology! Thank you! (laughs)

Ogez: (laughs) I should know that I'm taking a class called synthetic genomics.

Waterman: Fantastic. Well synthetic genomics is much more complicated than those guys do, but when they have these reduced systems they can really use differential equation modeling.

Ogez: Because the presence of the different chemicals—the rates depend on the presence of the other ones because of the inhibitors—

Waterman: Exactly. And often times biologists draw these cartoons in genetic systems and often what actually happens is more complicated and less intuitive. Which is more future employment for you, right?

Ogez: (laughs) Yea. So I guess the last question is what research are you doing right now?

Waterman: Ah. The last few years I've been doing work on something that sounds really simple. You take a couple of sequences and count, say, the four letter words. Four is a variable, but a small number. The four letter words in the sequences and you look at the statistics to compare sequences with. So it doesn't depend on…it's independent of the position of the words. The idea being that if the sequences are identical the word counts will be more similar. There was a statistic that people had used where you took the inner product—you took the number of quadruple "a's" in one sequence, multiplied by the number of quadruple "a's" in the other sequence. You do that for all the four to the fourth four letter words. Add that number up and that ended up being a very bad statistic for reasons that were intellectually interesting. So we started working in this area and now with all the genomics and so on going on we're still pursuing that. Another answer is that—I didn't know if you looked at how our algorithms work, these dynamic programming algorithms where you make these tables.

Ogez: I read the 1981 publication. I can sort of follow that one. I tried to read the one you did with Lander and I couldn't follow that one at all.

Waterman: Physical mapping. That was a genome project one too. So these dynamic programming algorithms you fill out this table.

Ogez: You make the matrix H and you have—is it an upper triangular matrix? I'm forgetting.

Waterman: So for just regular sequence alignments you take out the zero and that was the algorithm before and the structure and so those algorithms are quadratic. They take time proportional to the product of the length of the sequences. Because you're filling out the table. And various people have looked at the structure of those matrices and it's always seemed like there was more going on and I recently saw somebody give a talk where they had really looked at the algebraic structures of these things. I'm really looking forward to going back to that part of my past and really trying to understand what was going on with those matrices all those years. Very much a mathematician's area but I work with a guy in Denmark. We're planning to go back and shovel this ground once again.

Ogez: So what would you say you like better—math or computer science? Or is it the same because they're so intertwined? Do you like math or computer science more?

Waterman: I can't separate them. I used to give lectures to math and statistics departments and I would emphasize the computer science and I would give lectures to computer science departments and I would emphasize statistics. They didn't used to understand, but today computer scientists have figured out statistics. I don't really see that it's different.

Ogez: I can't think of anything else.

Waterman: Okay.

Ogez: Hmm… I'm trying to think. I've got you on the line. I've got to make the most out of this but I'm just blanking right now. Thank you so much for your time and for agreeing to do this.

Waterman: Good and I know your professor fairly well.

Ogez: You do? You know BCD? How do u know him?

Waterman: O just from…ask him. I don't remember how we met. We've known each other for some time.

Ogez: Yea he mentioned that you would be a good person to interview because you were really nice and that he was friends with you and then he gave me some correspondence you guys had about international databases from a couple years back.

Waterman: Couple of years. Sure. (laughs)

Ogez: You guy have met up at various conferences…?

Waterman: Ya. Well good luck. This is a wonderful area to start out in. While it looks like it's quite developed, I think it's just in its early days.

Ogez: And sorry for being a little bit awkward with the interview. It's my first interview so I'm just nervous and I'm sort of scrambling around for questions and stuff.

Waterman: Well I suspect my answers are at least as awkward as your questions so there we are.

Ogez: Well thank you very much and I'll let you go.

Waterman: Okay. Good luck.

Ogez: Thank you.

Waterman: Alright. Bye.