*Bioinformatics*, 32(11), 2016, 1758–1759
doi: 10.1093/bioinformatics/btw291
Message from the ISCB

# Message from the ISCB: 2016 ISCB Accomplishment by a Senior Scientist Award Given to Søren Brunak

**Christiana N. Fogg**[1] **and Diane K. Kovats**[2,*]

[1]Freelance Science Writer, Kensington, MD, USA and [2]ISCB, Bethesda, MD, USA

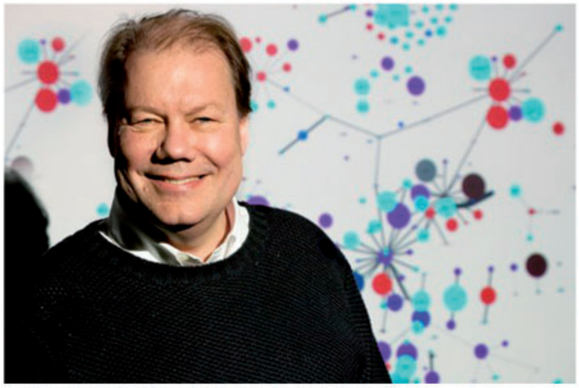
*To whom correspondence should be addressed.
**Contact:** dkovats@iscb.org

The International Society for Computational Biology (ISCB) recognizes an established scientist each year with the Accomplishment by a Senior Scientist Award for the significant contributions he or she has made to the field. This award is bestowed to scientists who have contributed to the advancement of computational biology and bioinformatics through their research, service and education work. Professor Søren Brunak of the Novo Nordisk Foundation Center for Protein Research in Copenhagen, Denmark has been selected as the winner of the 2016 Accomplishment by a Senior Scientist Award.

The ISCB awards committee, chaired by Dr Bonnie Berger of the Massachusetts Institute of Technology in the USA, selected Brunak as the 2016 winner. Brunak will receive this award and deliver a keynote address at the 2016 Intelligent Systems for Molecular Biology meeting (ISMB 2016) being held in Orlando, Florida on July 8–12, 2016. ISMB is ISCB's world class annual meeting that brings together computational biologists and interdisciplinary scientists from around the globe.

Brunak's early interest in physics began with a childhood friendship with Jakob Bohr, grandson of Nobel Laureate physicist Niels Bohr. He considers this early informal exposure to physics instrumental in developing his interest in the field but acknowledges that his physics teacher in primary school also nurtured his interest. Brunak said, 'I was primed by the fact that one of my childhood friends was Jakob Bohr. I grew up close to this family. Maybe I was therefore listening a little more to what the physics teacher would come up with. He was good at turning deep questions into something that could be understood by kids our age'.

Brunak went on to study physics formally as a graduate student but first took a detour in astronomy. He recalled, 'First I went into astronomy, but I found it increasingly difficult to explain at dinner parties the importance of astronomy'. He then completed his Master of Science in physics in 1987 at the Niels Bohr Institute, University of Copenhagen. 'I had been fascinated by computers. My masters thesis was titled *The Physics of Computation* (in Danish 'Computerens Fysik'), and I studied what happens in the computer when it computes. I was inspired by the work of Rolf Landauer and Charles Bennett at IBM. They worked on determining if you could compute without dissipating heat in reversible physical processes where no information would be discarded'. It was Brunak's interest in the work of Bennett that stimulated his interest in biology. 'Bennett used DNA transcription as an example of how a computation (a copy operation) can be done without dissipating a lot of energy. My thesis was also about computation processes in the brain, which are related to machine learning. It's also about throwing information away so what you are after is distilled out of the data. In the big data context, there is a huge information reduction need so my experience with the physics of computation has inspired me when designing machine learning algorithms that use a lot of information and end up with a yes or no, for example answering the question of whether a protein structure is helical or not at a given position in the amino acid sequence. A lot of bioinformatics is about throwing information away in a smart way so what you are really after is retained'.



Brunak completed his Ph.D. in computational biology in 1991 in the Department of Structural Properties of Materials at the Technical University of Denmark. He then went on in 1993 to become founder and director of the Center for Biological Sequence Analysis at the Technical University of Denmark, a large center that still exists. His early work in bioinformatics focused on protein structure. He recalled, 'I worked with protein structure with machine learning approaches. Meetings were small, data sets were small. We tried to get a lot out of little. We were raised in the data-poor era. The machine learning approach is not only good for boiling down but also for extracting'. Even during this era of limited data, Brunak considered computer power an important priority. ‘During my early studies in the late 1970s I started with punch cards and huge magnetic tapes. During my PhD I obtained a grant for a fast four processor Apollo 10000 machine, and I later always spent a lot of money on supercomputers so computer speed was not a problem. Now it is a real problem because we have millions of instances of a genome. We are in a situation where computer science matters in a new way. I have been around computers so long so I’ve seen a lot of special purpose hardware developed. But people always go back again and again to the general purpose computer that can take any algorithm, or do things like align sequences with any setting’.

Brunak’s early bioinformatics studies looked at both structure and function and were not limited to sequence properties. Machine learning was integral to these studies, and he went on to write an authoritative text on the subject with Pierre Baldi in 1998, titled *Bioinformatics: A Machine Learning Approach*. Brunak developed several widely used algorithms rooted in machine learning including NetGene, which predicted introns and exons and splice sites, and SignalP, a signal peptide predictor. He recounted, ‘This was the time of the genome project, so we started doing exon and intron and splice site prediction using this method called NetGene. Both SignalP and NetGene were interesting in that they integrated several different predictors and exploited the same data from different angles. With NetGene, we had a splice site predictor and an exon predictor and we put them together and we got a much better algorithm out of it than staying just in just the splice site or coding/non-coding domain. In SignalP we also used the same data in two different ways’.

Brunak recalls some of the surprises of his early research. ‘My first *Nature* paper was a small paper in 1990. It was a paper where we predicted splice sites using machine learning with neural networks. We noticed a group of splice sites that the network really did not want to learn. We just kept training it and it still would not learn them. We started looking at them and it turned out that half of them were database errors, and the other half were more interesting, they were errors made by experimentalists when they interpreted their [sequence] gels. They had put the splice site in the wrong place. The would learn the rare, but true GC donor sites very late, but still learn them. It was an interesting paper that showed the power of machine learning–that it could be a little more clever than the quality of the data. *Nature* was getting tough on GenBank for removing errors, and here was a computational approach for cleaning up data sets. We used the same technique with SignalP to identify likely errors. [We thought] either it’s an error or super unusual and therefore interesting. We could see in some databases, with signal peptides, that 10–15% of the data was wrong’. Brunak saw this tedious work as an important contribution to cleaning up data sets and spent several years on this effort.

During the Human Genome Project era, Brunak recognized with many others in the field the limits of gene prediction from sequence information alone. But his research using neural networks alluded to some of our present day understanding of the complexities of genomes. Brunak said, ‘It’s not surprising now that gene prediction was not 100% successful. Now we know that there’s transcription everywhere and that what constitutes a gene is highly complex. In 1992, we had a paper in *The Journal of Molecular Biology* (*JMB*) examining the ways how a neural network looks for gene features in order to produce a prediction. It turned out when we predicted introns and exons, it looked for a specific GC-rich signal. It was not easy to get a paper accepted into *JMB*, especially when you were trying to deconvolute theoretically neural network parameters into some biological signal. The pattern it looked for was perhaps known to a referee as an early example of an enhancer. Part of the reason of the success of the machine learning approach is that we didn’t need to know upfront the features that were behind biological mechanisms’.

Brunak’s research focus has shifted direction in recent years during this era of large scale genome projects. In 2007, he was a co-founder of the Novo Nordisk Foundation Center for Protein Research at the University of Copenhagen. The Center’s main goal is to look for proteins of therapeutic value, and they are developing approaches that fit into a healthcare context. Brunak leads the translational disease systems biology group, which looks at genome, proteome and health data, where some cover the entire Danish population. Brunak explained, ‘I am interested in disease trajectories, the order in which you get diseases, comorbidities and follow-on diseases. If you get type 2 diabetes, you won’t get the same complications as your neighbor. There are certain trajectories that are more probable than others’. For the entire Danish population, almost all personal information, including education, job status and health records, are tied to a Dane’s personal identification number. As such, researchers including Brunak have an abundance of unique data to work with, and much of his work has focused on boiling down this data into meaningful observations. ‘My contribution is to put patients into progression groups and interpret proteomics data. We for example group diabetics and will see how their trajectories differ. Having the ability to work from the molecular side and having health data is presumably going to be powerful. We have data from 11 million people living and dead. We also essentially have the family tree from the entire country because it’s encoded in the personal identification number’.

Brunak’s enduring contributions to computational biology and bioinformatics have spanned his career, and given the scope of his recent work, he is certain to make a lasting and valuable contribution to the field.