For two decades Drs Henrik Nielsen and Søren Brunak have been developing sequence-based prediction servers at the Center for Biological Sequence Analysis. Here, they outline the Center’s considerable contributions to the field of bioinformatics and what the future holds.

To begin, could you highlight the key activities conducted at the Center for Biological Sequence Analysis (CBS)?

SB: CBS conducts basic research in the general fields of bioinformatics and systems biology. We are highly active across a spectrum of molecular level disease systems biology, chemical biology, regulatory genomics, evolutionary analysis, immunological bioinformatics and proteomics, revealing protein interaction patterns and post-translational modifications.

The Center is also interfacing to the general area of medical informatics and analysing data from the healthcare sector, including text mining of electronic patient records and biobank questionnaires.

CBS represents one of the largest, multidisciplinary basic research groups within bioinformatics and systems biology in academia in Europe. How has the Center achieved this status?

SB: We were formally established in 1993, which was relatively early in the field’s development. We were fortunate that the Danish National Research Foundation was created in 1992 – this meant novel resources could be allocated to emerging fields such as bioinformatics. Through the Foundation we were able to bring bioinformatics from the single investigator level to an effort that could take on larger tasks.

The Center has achieved a great deal since its inception 20 years ago. What have been your greatest achievements in recent years?

SB: The Center has been one of the key early players in machine learning-based bioinformatics and data-integration driven systems biology. In the latter area, a highlight has been a temporal analysis of protein complex formation during the eukaryotic cell cycle. Another has been co-abundance analysis of metagenomic data across hundreds of patients where we can establish how smaller, optional gene groups in plasmids, genomic islands and phages are interdependent on larger gene groups comprising entire bacterial genomes. Yet another success has been our involvement in the merger between molecular level systems biology and phenotypic data from the medical informatics area, such as electronic patient records.

CBS develops sequence-based prediction servers, such as SignalP. How did you develop this server?

HN: The largest task in developing a well-performing sequence-based prediction server is getting the dataset right. SignalP is based on data with experimental annotations derived from the international UniProt database; we
The Center for Biological Sequence Analysis has made great progress studying cell biology, especially in the area of protein science. Fundamental to its achievements is their development of a sequence-based prediction server. The study of biology has evolved in recent decades, one factor has been the increasingly important role of bioinformatics in its development; this has led to the interdisciplinary field of systems biology, a growing area of study that aims to characterise living systems in their entirety in order to better understand the complex relationships taking place at a sub-cellular level. This holistic approach means that different sources can be integrated in order to find novel perspectives and, ideally, solutions to biological problems. Doing so could lead to more effective ways of combating diseases such as cancer by providing a basis for more personalised targeted therapies, but it also means vast quantities of data must be managed.

Organising and analysing information on such a massive scale is key to the existence of systems biology as an academic field, and the development of more effective computational tools to do this is therefore a constant requirement. One of the leading lights in this area is the Center for Biological Sequence Analysis at the Technical University of Denmark. With 12 specialist research groups currently conducting investigations at the interface between computational biology and wet lab experimentation, the Center’s activities span the interaction and co-existence of all kingdoms of life within biological systems.

### Integrating the Data

Along with Dr Søren Brunak, the Center’s Director since its foundation in 1993, Dr Henrik Nielsen has been working towards the development of data integration tools and prediction servers, with recent work at CBS seeing the MetaRanker added to the long list of prediction servers already available for use. This latest example integrates different types of data for analysis and results in phenotype-specific gene prioritisation of the human genome. Data types the MetaRanker uses include genome-wide association studies (GWAS) data, protein interaction data, gene expression data and linkage region data. Using one type as a scaffold on which the others are integrated has presented difficulties: “Instead, the information is integrated on an equal footing basis where all data types play a role in concert,” Brunak explains.

After developing such tools, CBS then focuses upon making them as accessible as possible. Academic users wishing to use MetaRanker or any other of the Center’s existing computational tools need only upload their data via the website after which CBS’s supercomputers perform the calculation – in this case for gene prioritisation – and produce freely downloadable results.

### Proteins Predictions

Although CBS has achieved much in the area of sequence-based prediction servers, it is with regards to protein sorting where the Center has excelled – partly due to industrial interest in protein engineering. In its broadest sense, a sequence-based prediction server receives input in the form of either nucleotide or amino acid sequences, and returns predictions of functionally or structurally important features.

The most popular of these is Nielsen’s own SignalP server – currently version 4.1 – which predicts secretory signal peptides. A secretory signal peptide acts like the postcode of a protein, signifying that the protein it belongs to should be exported from the cytoplasm of the cell. During transport across the membrane, the signal peptide is cleaved off. When SignalP receives an amino acid sequence of a protein, it offers a prediction of whether there is a signal peptide and, if yes, exactly where the signal peptide will be cleaved: “This is incredibly useful,” Nielsen states, “because the ability to predict which proteins are exported from the cell is important for basic research, biotechnology, and drug discovery”. This is an observation borne out by the fact that the first SignalP paper on protein engineering is not only the most cited publication in CBS’s history but also the seventh most cited co-authored Danish paper ever.
INTELLIGENCE

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

OBJECTIVES

The Center for Biological Sequence Analysis at the Technical University of Denmark conducts basic research in the field of bioinformatics and systems biology. The group of 120+ employees, working in 12 specialist research groups, has a highly multidisciplinary profile. CBS represents one of the large bioinformatics groups in academia in Europe.

CONTACT

Dr Henrik Nielsen
Associate Professor
Technical University of Denmark
Building 208
2800 Lyngby
Denmark
T +45 45256124
E hnielsen@cbs.dtu.dk

Dr Søren Brunak
Director
Technical University of Denmark
Building 208
2800 Lyngby
Denmark
T +45 45252477
E brunak@cbs.dtu.dk

www.cbs.dtu.dk

SØREN BRUNAK is Professor of Bioinformatics at the Technical University of Denmark and Professor of Disease Systems Biology at the University of Copenhagen. He is the founding Director of the Center for Biological Sequence Analysis and works across traditional bioinformatics, systems biology and medical informatics.

HENRIK NIELSEN is Associate Professor at the Technical University of Denmark. He earned his MSc degree in Biology from University of Copenhagen and his PhD in Theoretical Biochemistry from Stockholm University, but has been affiliated with the Center for Biological Sequence Analysis since its start, working mainly with prediction of protein sorting.

TARGETED THERAPIES

The future of CBS, however, is not inextricably tied to the prediction of secretory signal peptides. Progress is currently being made on a meta-prediction server named EPipe which, though incomplete, is running in beta form on the Center’s website. It is a tool that takes a set of closely related amino acid sequences – mutants or splice variants of the same gene, for instance – and searches for differences in predicted features among more than 50 sequence-based prediction methods. In this way, it will be able to highlight predicted functional consequences of mutations or alternative splicing events. If this is accomplished then Nielsen is confident that EPipe will prove itself to be an integral tool in the search for functional, sequence-driven differences between human individuals, the results of which would have significant implications for the future of patient treatment.

Disease systems biology as a general field is heading increasingly in the direction of precision medicine. With targeted therapies and more personalised treatment being the next big step in patient care, CBS looks set to play an important role in driving this development: “We are in a very good position as we have access to fine-grained phenotypic information from the Danish healthcare sector and also molecular data in vast amounts,” reveals Brunak. With this information at their disposal, the chance to contribute to the aims of specifically targeted medicines could also lead to reduced toxicity in patients and a lowering in the cost of healthcare.

AN INROAD TO INFRASTRUCTURE

In its two decades as a renowned institute of biological sequencing, CBS has been awarded and continues to attract substantial funding from both in and outside of Europe, reflecting not only the relevance of the research taking place but also the collective, collaborative nature of bioinformatics itself.

Evidence of this fundamental characteristic can be found in the Center’s current engagement in the pan-European ELIXIR project. Its aim is to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine, the environment, the bio-industries and society. Along with 15 countries, the European Bioinformatics Institute is working to establish this infrastructure formally with CBS focusing on the registry and interoperability of tools. With servers like Nielsen’s own heavily used SignalP, the Center has considerable interest in making such tools interoperable so that they are easier to connect in workflows.

As a relatively young discipline, the need for a formalised infrastructure to deal with the vast amounts of data involved is pertinent. Equally important to its growth, however, is the role of centres like these as educators. CBS teaches over 20 courses, ranging from introductory modules for bachelor students to advanced courses for PhD students and secondary school teachers. Recently, CBS has also been able to contribute to the Coursera collection of online courses.

It is this attitude to teaching and career development that has made it possible for the Center to attract grants from national and international sources, as well as a substantial core contribution from the Technical University of Denmark. With an emphasis on collaboration and its eye firmly set on the future of disease systems biology, CBS is confident it will continue to play a vital role in the continued growth of bioinformatics.

The first SignalP paper on protein engineering is the most cited publication in CBS’s history and the seventh most cited Danish paper ever.