John Cottrell, David Creasy & Darryl Pappin

Pioneers in proteomics data analysis reflect on what to focus on in computational biology.

In 1993, five papers—authored independently by William Henzel, Peter James, Matthias Mann, Darryl Pappin and John Yates—showed that peptide masses measured by a mass spectrometer could be used as a 'fingerprint' to identify proteins. This insight spurred global analyses of the proteome and the need for new computational tools. At the 2012 international meeting of the Human Proteome Organization (HUPO), John Cottrell and David Creasy received the HUPO Award for Science and Technology for developing an algorithm created by Pappin into the widely used Mascot software package for peptide identification and protein inference. *Nature Biotechnology* spoke with Cottrell, Creasy and Pappin to understand Mascot's route to successful commercialization and glean insights into current challenges in computational biology.

How was Mascot commercialized?

David Creasy and John Cottrell: In the early 1990s, Darryl Pappin at the Imperial Cancer Research Fund (now Cancer Research UK) was one of the first people to recognize the possibility of protein identification by digesting a protein with an enzyme, such as trypsin, measuring the molecular masses of the peptides and matching these measured masses against masses calculated from the entries in a protein sequence database—the technique we now call peptide mass fingerprinting. These ideas were implemented in a program called MOWSE, for Molecular Weight Search, and published in 1993. We licensed MOWSE and founded the company Matrix Science [London] in 1998 and a year later launched the Mascot product.

Did you take outside funding?

D.C. & J.C.: Using our savings and not paying ourselves for the first year or two, it was possible to get going without any outside investment. This may not be an option if you want to start a hardware company or sell products or services to large numbers of consumers. For scientific software, it is very possible, and we consider starting without outside investors the smartest decision we ever made.

What lessons might your experiences hold for other fields of biology, such as genomics?

D.C. & J.C.: Both genomics and proteomics have been driven by advances in measurement technology, which generated new types of data at greater scales. For instance, when we first started developing Mascot, we thought that 300 spectra in a single search would be a reasonable limit. How wrong we were! By 2003, it was possible to analyze 1 million spectra in a single search, and now there is effectively no limit. When there have been big advances in experimental technology, as there have been in proteomics and genomics, new ideas in computational biology have been needed.

Are there big opportunities today for commercialization of academic 'omics software?

D.C., J.C. & Darryl Pappin: In August 2001, Genome Technology magazine profiled 'proteinformatics'—not a term that ever caught on—in an article that featured search engines from seven companies. Only three are still around today. Matthias Mann was prescient in saying, "I don't think there's a big space for a lot of companies thriving on developing this software." In general, there has been a low survival rate of scientific software companies. Is genomics different? I don't know. But what hasn't changed is that existing computational bottlenecks and high-activity areas of computational methods development often hint at the need for better experimental techniques. For instance, genome assembly is an important computational problem in genomics today that is being addressed by new computational methods and by new experimental methods for assaying long DNA molecules.

How did you keep your software company viable for 14 years?

D.C. & J.C.: We've deliberately stayed very small. We started with two people and now have 11. This seems about right for the size of our market. The number of labs who are potential customers for our software is probably in the tens of thousands. The Genome Technology article speculated that proteomics software could become a multi-billion dollar per year market; the reality is more like a multi-million dollar per year market. Of course, we shouldn't underestimate the advantage of entering the field early enough to get well established.

What about cloud computing and offering software as a service?

D.C. & J.C.: People often suggest offering Mascot as a service (e.g., $1 per search). In recent years, this is usually associated with the supposed benefits of running Mascot 'on the cloud'. In a presentation made in 2004, Darryl León of Lion Biosciences [Heidelberg, Germany] noted that this model of supplying software suffers from secure data transfer issues and offers limited control in the development and implementation of applications. All of the software service companies surveyed in his presentation have either gone bust or moved away from selling software. In contrast, what has succeeded are relatively small companies offering specialized applications for which the core technology existed at the inception of the company.

David Creasy & John Cottrell

What advice would you give to those looking for important computational problems to solve?

D.C., J.C. & D.P.: Take a road less traveled. For instance, we know of at least 50 search engines for database matching of MS/MS data. There are a great many interesting and important problems in mass spectrometry–based proteomics that have not received anything like this amount of attention. For example, we really need some new ideas on protein inference. Data-independent acquisition and scoring high accuracy MS/MS data are two areas where it could be said that the software is lagging behind the hardware. Search software for other biopolymers, such as sugars and lipids, is lacking. As we mentioned, important computational problems usually arise from the need to process large quantities of interesting biological data. But perhaps a high-risk, high-reward approach would be to identify the so-called 'important' computational problems and eliminate them by developing new experimental technologies.

Interviewed by H. Craig Mak, Associate Editor, Nature Biotechnology