Folding, Funding and Phyre - a tool-building quest to solve one of the biggest problems in science.

Lawrence Kelley[1]

[1]Backstories

April 28, 2020

History

I decided to be a scientist when I was 7 years old after watching Carl Sagan on TV talk about neutron stars. By 16 I discovered the protein folding problem: a disarmingly simply stated problem whose solution lay at the heart of everything in biology. How hard could it be?

So a degree in Biochemistry later, I did a PhD in the lab of a computational molecular modeller (Mike Sutcliffe) and got started on the problem of clustering protein structures from NMR. I wrote my first program in C and built my first web site with my characteristically stupid pun in the name - OLDERADO: On-Line Database of Ensemble Representatives and Domains [1] - to look at the results. Each morning I would go on the internet as it then was to check the dozen or so new sites that had sprung up the night before and which had been manually added to a list known as the 'internet directory' by someone 'out there' on the new web. Google appeared about a month later. Websites were going to be a big deal. Being an early adopter of web-based science seemed like a good bet.

Then a move to London to be with Mike Sternberg, a world expert on protein structure prediction. I inherited an early protein fold recognition program (Foldfit [2]) written by Rob Russell, Mansoor Saqi, Paul Bates and Roger Sayle (of Rasmol fame) and had to improve it. The result was 3D-PSSM [3] and I made a 'fancy' website (read: 1990's black background with gold, animated logo) for other people to use the program and for me to figure out where the bugs were. It worked great - we won the automated section of the international CASP competition in structure prediction, and hundreds of people started using the site. It was one of the first of its kind, which made it easier to stand out. Throughout I was mainly making the site to help me understand how my software worked in various situations. With so many different types of data to look at, I needed a way to sensibly display this info just to understand what was going on internally. This remains a critical focus.

New techniques arose in the field and I sat on my laurels with 3D-PSSM. I neglected the site and realised only later how valuable a large internet audience is and how I had squandered it at its peak. So I vowed to try again and rebuild it from the ground up, this time properly. Using newfound skills in web development I created Phyre [4]. This did well in both scientific recognition and number of users so I was very pleased. But soon funding looked iffy and was about to run out.

Funding

I wondered if any users could help. So I asked on the Phyre page if anyone could write a letter of support for us that we could include with our application for funding. We received over 1,000 letters of support which was amazing and had a huge effect on our application. But I think it placed the funding bodies in a difficult position. Making web sites that help scientists use state-of-the-art tools is undoubtedly a good thing. These site are continuously working, 24 hrs/day reliably doing *some* work to help science. They seem like safe and sensible places to direct funding - IF they are widely used. IF they are widely used, the funding body would look deliberately negligent if they failed to fund it. But at the same time they are justified in thinking this is outside their remit. They decide on science funding, not tools. So new arms of the funding bodies started to form, dedicated to tools and resources. This is where we are today. But deciding how to handle funding for tools is a new challenge we are all still grappling with.

The quest for funding is ever-present. It is critical that tools are maintained. How do you justify the maintenance cost? It is easy to fall through the funding cracks - it's neither proper science, nor pure infrastructure. I get very paranoid about funding living in this limbo. But if user numbers and citations continue to increase, we are safe. These are the only ways to prove your value in this game. But this hopefully reflects real-world utility of the tools you create. So what's the best way to maximise the utility of these tools to researchers?

Build it, and they will come

Great tools based on state-of-the-art algorithms often go unused and unnoticed - great science gets 'hidden' in the hands of a few people competing intensely over small absolute changes in accuracy using programs designed for no one but the developers. The gains to society and science by enabling more people to use these tools are far greater. But then they need to be far easier to use if more people are to use them. Hence my focus on user interfaces.

The best way to learn what works for user interfaces is by using lots of them. And thinking about the types of question a user may have. Simply BE a user and every time you want to do something and CAN'T, write a program to do it. Accumulate these programs in a web site. This empowers me and anyone else who cares, with these new tools - a virtuous circle.

'BE the user' sounds great. But it gets harder as you spend more time with computer development and less time working on a biological problem with real proteins. I don't mean wet lab. Just computational analysis problems with specific proteins faced by real scientists. That's the hardest part to remain connected to. For now I just imagine scenarios. I imagine all the combinations we can make by connecting our tools in different ways, and build some that seem most promising. I meet users at workshops to find out what they want. But this is where the next development needs to happen. How can I communicate effectively with 50,000 users about what they want and what I can deliver? This is the biggest challenge for me right now.

Despite the funding complexities, we got funding for Phyre2 [5]. THIS time I can do it properly, I thought. Use the most up to date facilities in the browser to make it look nicer, easier to use, and add those new ideas I had or were suggested to me by users at workshops (PhyreAlarm, BackPhyre, One-2-one threading). I also tried to improve how a user could look both at sequence and structure at the same time in the browser to analyse a range of features. This led to the development of Phyre Investigator (one module in Phyre2) and honed my skills at javascript for the future. This most recent paper has again done well in the citation game which I hope reflects its usefulness to researchers 'out there' on the web.

Folding

It's important to remember the reason I've done any of this, and that reason is the mystery of protein folding. The person that solves the protein folding problem (at least I hope it's a person) won't do it with pen and paper. It will be someone extremely adept at using a variety of tools that probably already exist individually, and who will put them together in the right way. That is generic problem solving. Make this easier for people, and problems in general get solved faster. With folding my main fear is this being one of the first major science problems where artificial intelligence gets there before us. Google DeepMind is after it, and we've seen what they've done to the game Go.

Protein folding is the molecular biology equivalent of the Goldbach conjecture: The impossible task that no one would sensibly pay you to work on but you want to work on because of its elegant simplicity and deep importance. So you need to do something useful while you ponder the impossible. You find the most useful thing you can do for everyone else and for you. Keep it as close as you can to folding whilst maintaining an audience size that justifies your existence - i.e. homology modelling.

Problems that no one has been able to solve for decades typically don't fall over from expected directions of attack. To be a long-standing problem means most lines of attack that occur to people have been tried. So new lines must be found by looking further afield, in other areas of science where an analogous problem may have been thwarted. But to see the mapping between a problem in say physics or economics or mathematics, to the folding problem requires a decent understanding of both areas. So that means trying to learn physics, signal processing, quantum mechanics, maths, computer science, AI, etc. All in the hope of seeing a new angle from which to attack the problem. And it's fun and fascinating. But you need to make yourself useful to the world and justify your salary. My most recent direct stab at folding involved eigen decomposition of protein contact maps. As you may surmise, I have not yet been contacted by Stockholm.

Where to publish?

Well it's not really science is it? Its tools to do science. So it doesn't sit easily with most journals. You can submit an 'Application Note' which is about a page describing your web server. This doesn't really count as a proper scientific paper in many eyes. Or often authors will use their tool to do some rapidly thrown together analysis that makes what would otherwise be an application note pass for a normal scientific article. Neither of these scenarios is ideal. The way I see it, when I've made a tool I haven't answered a biological question, but I've made it easier to answer many biological questions. In the end we went for Nature Protocols which is aimed at step by step instruction for how to use previously published tools. It fits, albeit somewhat uncomfortably.

Future

So Phyre2 has done very well in attracting users and now I'm on to Phyre3 which will be out shortly (end of 2017). Again it's a full redesign, this time with two people on it, me on the front end and my colleague Stefans Mezulis on the back-end, using the browser to its best, new web tech, polished with a new engine (the PhyreEngine, naturally) and more hardware for faster processing of bigger genomes. Also an entirely new tool called PhyreRisk is being developed: a portal for analysing disease, mutations, structures and complexes. But the primary focus is still the same: utility. What do people want or need to do and what can be done computationally to help them. The better we can match what can be done computationally with what researchers want to do, the faster we will make progress in all of science. And maybe we'll beat DeepMind to the answer to protein folding.

References

1. OLDERADO: On-line database of ensemble representatives and domains. LA Kelley, MJ Sutcliffe, Protein science 6 (12), 2628-2630 (1997)
2. Recognition of analogous and homologous protein folds - assessment of prediction success and associated alignment accuracy using empirical substitution matrices. Russell RB, Saqi MA, Bates PA, Sayle RA, Sternberg MJ. Protein Eng. Jan;11(1):1-9. (1998)

3. Enhanced genome annotation using structural profiles in the program 3D-PSSM. LA Kelley, RM MacCallum, MJE Sternberg. Journal of molecular biology 299 (2), 501-522. (2000)
4. Protein structure prediction on the Web: a case study using the Phyre server. LA Kelley, MJE Sternberg. Nature protocols 4 (3), 363-371. (2009)
5. The Phyre2 web portal for protein modeling, prediction and analysis. LA Kelley, S Mezulis, CM Yates, MN Wass, MJE Sternberg. Nature protocols 10 (6), 845-858 (2015)