



Cloud Computing: A Drug Discovery Game Changer?

By Paul Davie
at InhibOx Ltd

Cloud computing has the potential to transform drug discovery in many areas – heralding the dawn of no-compromise computer-aided drug discovery and making effective virtual screening a reality.

Automation versus craftsmanship? High-throughput screening was supposed to be the answer to declining pharmaceutical research productivity, but many feel it failed to deliver the benefits it initially promised. This is so often the case with a new technology. These great ideas come along, shake our world and then we wait for the promised revolution. Sometimes it comes through – often it does not. My question is: will cloud computing disrupt the drug discovery game, or just speed it up a little?

My belief is that its effect will be transformational – here's why. Let's start with a refresher, and make the link between high-throughput screening, computer-aided drug discovery and cloud computing along the way.

MODELLING & SCREENING

Back in the 1980s, the advent of computer-aided drug discovery (CADD) promised to revolutionise pharmaceutical research. Teams of expert modellers were brought together in dedicated groups in all of the world's major pharmaceutical, agrochemical and biotech companies. Software suppliers vied for leadership in the space, and hardware suppliers made a killing from selling powerful graphics workstations and compute servers. In these heady days, rational drug design was proclaimed by some as heralding a new dawn in the industry. There were success stories (the anti-flu drug Relenza, the cyclic urea anti-AIDS treatments, and so on) to encourage things along. The craftsmen modellers remained in their teams, serving the organisation, but the vision of all medicinal chemists relying on CADD technology never came to pass. Twenty-five years on, CADD still constitutes an enormously valuable component in drug discovery research – but it has not torn up the rule book.

Then in the 1990s, along came high-throughput screening (HTS). Automation had arrived and would completely revolutionise drug discovery, overturning rational design by either medicinal chemists or modellers. It was a brilliant proposition: "Let's make/buy and screen everything we can get our hands on and see what sticks!" Robots were bought, compound inventories amassed, and off we charged. There were

teething problems, of course. How did you know the compound being tested was reliably what it claimed to be on the tin? How could you handle that data deluge to tease out useful knowledge? But the market responded, as markets do, and solutions appeared. HTS took its place as a \$2 billion market in mainstream drug discovery, but still can be said ultimately to have failed to deliver on its early promise. Let's look at some of the reported statistics that underpin that proposition:

- ◆ The estimated costs associated with HTS in an average pharmaceutical drug discovery project are around \$1 million. Pharmaceutical companies each spend, on average, \$50 million per year on HTS (1)
- ◆ A majority of HTS campaigns fail to find a lead compound, and only one in a hundred new projects delivers a marketed drug (2)

Those are hardly compelling statistics.

BRINGING IT ALL TOGETHER – VIRTUAL SCREENING

The world of CADD was quick to recognise the value of the principle of HTS and soon a computer-driven equivalent appeared – virtual screening. The principle was simple: obtain or build a 3D model of the target site and dock into it as many candidate structures as you can get hold of to estimate how well they might bind; select the most promising candidates and buy or make these for physical screening. The cost of modelling the whole process is negligible compared with physical HTS, so it must be a valuable pre-cursor and the return on investment should be large – right? Wrong – if you consider the effectiveness of the first wave of virtual screening operations.

The virtual screening principle is sound, though, so we should examine its failings to date. There are three fundamental causes:

1. Impoverished data sources
2. The need for skilled practitioners
3. Simplistic, compromised modelling tools



There are three accessible sources of compound data for virtual screening: supplier catalogues, (around 6 million entries), each company's additional internal inventory (less than one million and lacking diversity) and publicly available data sets (not necessarily subject to rigorous quality control). Considering the universe of possible candidates for modelling, this is a very meagre set of data sources.

Virtual screening should not be a totally automated process, at least not with the software tools currently available off the shelf. Not enough of the knowledge of skilled practitioners has yet been encoded in the commercially available software for the job. The range of targets and challenges facing the modellers is wide, and their skill and experience is of paramount importance in determining the outcome of the virtual screening study.

Which brings us to the modelling tools themselves. The platforms in use today were developed over the last three decades and built on algorithms that optimised performance for the systems available at the time. Let's be clear about what this means. Moore's Law has accurately foreseen the doubling of computer processing power on a less than two-year cycle for the last four decades. A system designed just 10 years ago (new, in CADD terms) would have been built to deliver the most accurate results it could within a 'reasonable' time, on hardware with three per cent of the power of today's equivalents. Scientists in industry can be impatient souls, and rightly so; thus 'reasonable' means up to a day for a large task, and a minute for a small activity along the way. Go beyond these limits, and compromises have to be made in the algorithm or its parameters to make the time taken acceptable. Modelling is a very complex, multivariate job – so this meant lots of deeply-seated compromises, embedded in the systems and the way they are applied. March forward ten years and what would have taken a month could be achieved in a day with the same software – so you don't have to compromise anything like as much, now.

And then along came a cloud...

CLOUD COMPUTING

Cloud computing offers companies virtually unlimited computing resource on tap – you effectively rent as much time and power as you need at any given occasion.

Cloud computing is a general term for anything that involves delivering services over the internet; the name comes from the symbol often used to represent the internet in flowcharts and diagrams. Cloud services are sold on demand, typically by the hour; they are elastic – a user can have as much or as little of a service as they

want at any given time; and they are fully managed by the provider. The user needs nothing but a personal computer and internet access. Significant innovations in virtualisation and distributed computing, as well as improved access to high-speed internet and a weak economy, have accelerated interest in cloud computing. The flexibility of the cloud computing resource, allocated on demand, is a compelling alternative to the huge internal resources that would be needed to cope with the peak requirement typical of computational chemistry.

The business model for CADD required the purchase of the most powerful compute servers and graphics systems you could afford, along with licences for a wide range of expensive modelling software platforms, writing the cost off over a given period – usually three years. This was extremely capital intensive, so the investments were constrained. This asset was utilised as needed by the modellers, which meant it was sometimes churning jobs through at maximum capacity – while the users queued patiently to get their work started – and at other times it sat idling, waiting for the next job. At any given moment, the modellers could have used 10-times its power, or did not need it at all. This lumpy usage is necessarily inefficient.

Cloud computing acts to lift the business constraint so that there is, in effect, an immediate jump in the power available, over and above the evolutionary improvements driven by Moore's Law. So now all of those nasty software design compromises should come into question. Cloud computing ought to herald the dawn of no-compromise computer-aided drug discovery and, specifically, it can make effective virtual screening a reality. Virtual screening fits the cloud computing model perfectly because it is so inherently parallel: you pay no extra for using 1,000 CPUs for one hour, rather than waiting 100 hours for 10 CPUs to do the same job.

One only has to look at some examples to understand the potentially dramatic effect of cloud computing in pharma.

Take, for example, the treatment of the receptor site in virtual screening. The site is flexible and may well be occupied with bound water molecules. Add in the flexibility of each candidate drug, and the complexity grows alarmingly. Few virtual screening approaches even begin to attempt to account for such complexities. Yet issues such as the flexible-ligand and flexible-receptor docking have to be considered according to Dr Garrett Morris, co-author of AutoDock (3), the world's most widely used docking package. AutoDock is used in

internet-distributed biomedical grid computing projects such as IBM's World Community Grid project, FightAIDS@Home (4). He and Paul Finn, CSO at InhibOx, are pioneering cloud-enabled approaches using an internal version of AutoDock and proprietary software to bring new, necessary rigour to the art. They have built a 100 million compound database of flexible 3D candidate models, with calculated shape, stereochemistry (5), charge and physical properties to guide the virtual screening (6). It is massively computer-intensive, but it is appropriately rigorous and cloud computing makes it all tractable.

The visionaries have seen the potential for cloud computing to transform drug discovery in many areas – beyond the creation of effective virtual screening.

Dave Powers, Lilly's Associate Information Consultant for Discovery IT, was recently quoted as describing a project the firm executed using Amazon's Elastic Compute Cloud (EC2) service as follows (7): "We were recently able to launch a 64-machine cluster computer working on bioinformatics sequence information, complete the work and shut it down in 20 minutes. It cost \$6.40. To do that internally – to go from nothing to getting a 64-machine cluster installed and qualified – is a 12-week process."

There are already many examples of bioinformatics tools being deployed on EC2. These include HMMer for protein sequence analysis and BLAST for general biological sequence analysis; in fact, pre-built Amazon EC2 images are publicly available for these tools. Researchers at the Biotechnology and Bioengineering Center at the Medical College of Wisconsin have created a scalable virtual proteomics data analysis cluster (VIPDAC) that exploits cloud computing services, and they now distribute a pre-configured Amazon Machine Image (AMI) containing the OMSSA and X!Tandem search algorithms and sequence databases (8).



Paul Davie is Chief Executive Officer at InhibOx Ltd, an Oxford University Chemistry Department spin-out, pioneering the application of cloud computing to computer-aided drug discovery. He has a long and successful track record in commercial roles in computer-aided drug discovery. He held support and sales roles at Chemical Design before going on to build and manage the European sales, marketing, support and

consulting operations at Oxford Molecular. Paul went on Accelrys to establish their Consulting Division and serve as European General Manager, before becoming Chief Operating Officer at InforSense. He then founded and was CEO at Secerno, a successful database security company, before returning to research informatics with his consulting company, Davinger. Paul has an MBA and read Chemistry at Oxford University. Email: paul.davie@inhibox.com

Cloud computing is beginning to gain acceptance in pharmaceutical and life sciences companies, with GSK, Pfizer, Eli Lilly & Co, Johnson & Johnson and Genentech all quoted as using cloud computing resources.

IS CLOUD COMPUTING A GAME CHANGER?

I think it is. And it's not a question of automation versus craftsmanship – it's a combination of the two.

The scientists behind developing CADD systems have always been grasping for more computing power, to make practical their developments. Ten years ago, Professor Graham Richards at Oxford University was leading the screen-saver project (9), pulling spare CPU power from over three million PCs across the world to drive virtual screening of cancer drug candidates. Cloud computing makes that amazing vision a practical, commercial reality. At last, computing power is available and cost-effective enough to cause the design of modelling systems to be fundamentally reassessed. It should herald the development of no-compromise CADD and effective virtual screening – and that could indeed change the drug discovery game.

References

1. *Industrialization of Drug Discovery*, Handen JS (Ed), CRC Press, 2005
2. *Chemogenomics in Drug Discovery, Methods and Principles in Medicinal Chemistry*, Vol 22
3. Morris GM, Huey R, Lindstrom W *et al*, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, *J Comput Chem* Dec, 30 (16): pp2,785-2,791, 2009
4. FightAIDS@Home 2010, <http://www.worldcommunitygrid.org/research/faah/overview.do>
5. Armstrong MS, Morris GM, Finn PW *et al*, Molecular similarity including chirality, *J Mol Graph Model* 28: pp368-370, 2009
6. Ballester PJ, Finn PW and Richards WG, Ultrafast Shape Recognition: Evaluating a new ligand-based virtual screening technology, *J Mol Graph Model* 27: pp836-845, 2009
7. Mullin R, The New Computing Pioneers, *Chemical & Engineering News* 87 (21): pp10-14, 2009
8. Halligan BD, Geiger JF, Vellejos AK *et al*, Low Cost, Scalable Proteomics Data Analysis Using Amazon's Cloud Computing Services and Open Source Search Algorithms, *J Proteome Res* 8 (6), pp3,148-3,153, 2009
9. Richards WG, Virtual screening using GRID computing: the Screensaver Project, *Nature Reviews Drug Discovery* 1:pp551-555, 2002