Is the cloud enabling bioinformatics to become the biology solutions domain?

For many years we have used the term bioinformatics to describe, well, anything outside cheminformatics in the R&D informatics domain. It has been a 'catch-all' term to label informatics tools that handle biologically relevant information. But the tools and software that have been labelled as such have not been expansive – albeit they are biology centric.

By Dr Paul Denny-Gouldson

These tools have been focused on very specific areas such as peptide and oligonucleotide analysis – predicting things based on a sequence, alignment of sequences, etc. However, in the past few years we have seen the combination of cloud technology, new biology techniques and, importantly, the rapid growth of biological entity therapeutics – macromolecules, eg MABs, FABS, RNA, large protein, Antibody Drug Conjugates (ADCs) – drive innovation in the provision of tools to support this diverse science and move far beyond the typical sequence-based concepts that have existed so far.

In this article we look at the history and the challenges that exist in the biology field and examine how the cloud, 'appification', SaaS, micro services, web tools and other new IT architectures and endeavours are helping to change informatics support. I have used broad areas of interest to highlight the impacts and new developments that we are seeing in the market. The examples and topics are by no means exhaustive and should not be viewed as such – they are included as 'jump off points' for readers who would like to know more.

Cell and molecular biology – cutting and pasting at a molecular level

Us molecular and cell biologists are no longer plagued by having to use numerous MSoft Excel and Access databases to store and track our primers, plasmids and cell lines. We are no longer forced to design our constructs in isolation in standalone tools and, with the advent of new methods for editing genomes, biologist scientists interested in CRISPR are able to work far more effectively. Benchling (www.benchling.com), for example, has been matched with the development of a plethora of cloud-based tools to help the scientists do their work more effectively and accurately by linking to other services.

Some of the more innovative tools, such as those

provided by Deskgen (www.deskgen.com), have been developed with a supply chain in mind. The tools are connected to suppliers of reagents and molecular biology constructs so, once you have designed your new 'construct', you can order all the things you need to make by clicking the 'shopping cart' or even get them made for you. Some of the larger and more advanced providers, such as Horizon Discovery (www.horizondiscovery.com), can provide your new cell line to you within a few weeks or months.

This is more than just software as a service – it is 'science as a service' (SaaS) – bringing the best software together with a business service. Now, this is not new for the Chemistry sector – there are many long-standing and established contract synthesis houses out there – but the biology domain is jumping ahead of these by working more collaboratively with many different vendors. This type of business-to-business collaboration to deliver an aggregated service has only become economically viable for SMEs (small medium enterprises) since the advent of cloud and SaaS support. Without automation capabilities, such as those provided through the cloud, these collaborations become very tedious, manual and technically-challenging to implement. The power comes from the full and relatively easy automation of the specific scientific value chain in question.

HELM – a possible MOL format equivalent for large molecules

Cheminformatics has been supported for many years by the 'MOL' format and its derivatives https://en.wikipedia.org/wiki/Chemical_table_file. This format has allowed structural information about small molecules to be shared between applications and scientists very easily. It has had a number of iterations over the years, but essentially it has become a standard. Over the years, other formats have been developed to support chemical drawing packages such as ChemDraw (CDX) and Marvin Sketch (MRV). These formats experienced some problems around format interchange and 'information loss', so organisations typically use one of the formats as their corporate standard to avoid these issues. Many suppliers have produced chemical fingerprinting capabilities to support fast database querying of chemical structures by structural motifs – but the exchange format and display interchange format typically defaults to MOL.

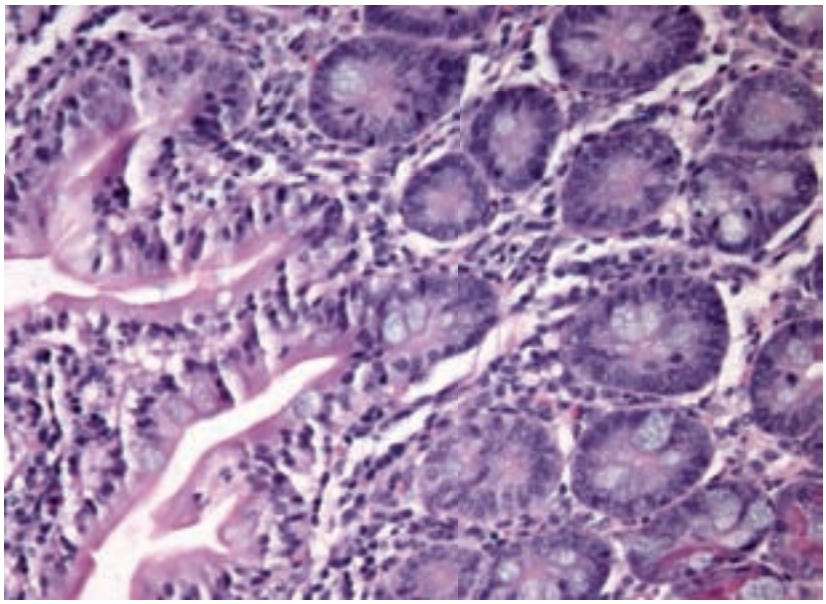
So, what happens in the biology domain when display and structural information needs to be shared between scientists and organisations? Well, until very recently it has been the dominion of well-established open source formats such as gcg, swissprot, etc. For a history of this endeavour, visit http://www.sib.swiss/sp30/the-history-of-swiss-prot. These academic programs shaped the bioinformatics landscape we see today and have democratised the manipulation of basic protein and nucleotide sequences.

Nevertheless, the science of delivering macromolecules as therapeutics has put many more demands on the drawing, capture, display and searching of biomolecules. To properly represent these entities, support for non-natural bases and peptides, conjugated polymers and small molecules, domains, structurally modified and cross-linked sequences is absolutely critical. Therapeutically active large molecules can be represented at a very high level in the historical formats – but critical information is now lost. Non-natural peptides or oligonucleotides are typically represented as an 'X' or one of the letters not already used. But what the X is or means is unknown to the user and also the software applications that come across them. Cross-linking information can be added as annotations, but there is no real standard for this and it is not easy to visualise. 3D structures can be shared using the PDB (Protein Databank) format – but all of this is very 'lossy' and with the loss comes the critical problem – you lose information critical to the understanding of the molecule and its behaviour.

The massive increase in biologics focus by pharma and biotech has driven the requirement to better support the drawing, underlying representation, interchange and computational description of macromolecules but without the loss of information described above. HELM (hierarchical editing language for macromolecules) has been one standard that has quickly become a foundation for many new applications that support the design and cataloguing of macromolecules in a way that enables exchange and registration.

What does HELM do and why is it so important?

HELM (http://www.pistoiaalliance.org/projects/hierarchical-editing-language-for-macromolecules-helm/) has been developed to deal with all the subtleties of biologics, such as post-translational modifications, cross linking of domains, linkers and ambiguity. It allows molecules to be defined, stored, computationally compared and, importantly, shared in a standard manner between scientists and organisations – in the same way chemical structures are shared using MOL. Some ask: why not use MOL for large molecules? When they are small it does work, but as soon as they grow above



20-30 peptides or nucleotides, they become unmanageable and cumbersome to work with.

At the core of any standard is the evolution of tools to support the drawing of the biomolecules. The acceptance of the HELM standard has seen some commercial products emerge that provide good user experience and functional capabilities http://www.quattro-research.com/projects/helm/ https://www.chemaxon.com/products/bioeddie/. These tools are being developed with the searching of structures and bioregistration in mind. Drawing a molecule is the first step in providing true, chemistry-like, structural searching capabilities and true structurally-based registration processes, such as duplicate checking, application of business rules etc. These subsequent structural searching-based steps require representation and fingerprinting concepts to enable searching, both of which the above vendors are supplying. This enables the typical biomolecule questions such as:

"Find me all the molecules with a cross link between domains 1 and 3 at residue positions 4 or 8 where I have a substructure of a conjugated molecule similar to X and modified non-natural residues containing Y and no post translational modification regions." Not as simple as it first seems – this needs structural representations as well as sequences.

Others, such as Biovia (www.biovia.com), Core (www.coreinformatics.com), Genedata (www.genedata.com), IDBS (www.idbs.com) and Labkey (www.labkey.com), have taken a more inventory-like approach to the registration of biological entities, where the asset is tracked and information about it stored. These do not provide the structural search capabilities, but do provide valuable traceablity and genealogical information required in the biologics R&D field.

The advent of genomic sequencing – driving improvement in compute power and adoption of cloud

The historical sequence alignment tools like BLAST and FASTA (https://en.wikipedia.org/wiki/BLAST) that were developed more than 30 years ago have formed the foundation of many things – one of which is the support of genomic sequencing. There are two tricky things when considering this science: the actual sequencing – the determination of the nucleotides and their order – and the piecing of the data back together to generate a full sequence. Many forget that the majority of the genomic sequencing methods produce billions and billions of DNA fragments that are sequenced (https://en.wikipedia.org/wiki/DNA_sequencing).

This leads to two problems: storing all the data and piecing all the pieces fragments back together to get the overall full genome sequence is a computationally intensive and algorithm optimisation problem. Both of these issues have been addressed by the academic community and sequencing technology companies. But, where to store all this data? This is an innovation that has also been supported by the cloud storage age. Many of the instrument providers also provide cloud storage for the data produced by their instruments, giving immediate access to the cloud infrastructures for alignment and annotation as the instrument creates the data.

There are regulatory implications to be considered when storing strong patients' genome information, but the overall message is that the cloud has been an enabler for this domain. Furthermore, alternative sequencing technologies that produce a single strand or small number of strands for sequencing (nanopores) do not need the alignment compute capability – but still require annotation compute power.

So we can see that the genomics age has been matched with a cloud age – whether there is causality there I am not sure – but genomics has definitely benefited and is benefiting continually from the 'cloudification' of our IT infrastructures.

The emergence of engineered biological systems – build your own organism and test it *in silico*

Another part of understanding how biological systems work is getting to grips with the systems that run cells and organisms. New informatics

approaches have become possible with the ability to run massively computationally intensive calculations routinely – commonly called simulations. 15-20 years ago, scientists would have been required to build their own compute centres – this was a fun hobby for many and your importance would be measured as a factor of how many CPU cores you had at your disposal! This world has gone following the advent of 'cloud' and burst computing, making it easier for scientists to access high-performance environments on a pay-as-you-go business model.

This low-cost simulation capability enables the scientific community to develop computational models of how organisms work at a process level, and given that, develop tools to help design a new organism to do a specific job , such as produce a specific molecule to a given specification. This is of great interest in the biologics development community as many of the complications in producing a biologic is that mammalian and bacterial cells are used to produce them. Using natural cell lines (albeit engineered to a certain level) gives rise to issues of contamination and variability of the final product. These two factors make the process and quality very complicated to manage.

Biosimulation – where scientists try to simulate a given system to see how something will behave – has been around for quite a while, typically when a drug is introduced. There are some interesting examples where this capability has been put to very good use and given rise to commercial tools – SIMCYP (www.certara.com). Cloud only makes this stuff go faster, and given scientists' requirement for everything to take a second, this is good news!

Systems biology and synthetic biology take things a little further. Systems biology is summarised here https://en.wikipedia.org/wiki/Systems_biology. Rather than trying to reduce things to a simplified, but still accurate, form, systems biology takes a more holistic approach and tries to incorporate as much of the system's variables as possible. It leads to the concept of trying to understand how things work at a far deeper level and thus looks at things such as metabolic and cell signalling pathways. As Scotty from *Star Trek* would say "to simulate these environments Captain, we need more power". Another cue for the 'cloud high performance compute' phrase.

But, when we consider the concept of 'build your own organism' this kind of knowledge, and the simulation capabilities, are critical. You do not want to spend time building something in a test tube that is not going to work. This field is emerging and the concept of Synthetic Biology is being developed for systems more complicated than the typical 'molecular biology' design problems. For a given designed protein, make me a system that will make it, ie the plasmid, cell line, etc from component parts – promoters, terminators, restriction sites, etc. This area

has attracted work and SBOL (Synthetic Biology Open Language) has been developed to help standardise communication and terminology (https://en.wikipedia.org/wiki/Synthetic_Biology_Open_Language). Some providers already produce tools that help scientists through this process such a Genedata Biologics (www.genedata.com) and Vector NTI (www.thermofisher.com) but the majority of the online web based tools, ie those amenable to cloud deployment, are delivered as open source via academic institutes.

Biologics development – optimising the production of the therapeutics

The understanding of what is involved in manufacturing a biologic is also rapidly evolving. The issues of the manufacturing process development and therapeutic optimisation should not be underestimated given the greater complexity of continuous production and bulk manufacturing of things produced by a biological entity.

Biological therapeutics are, in the main, part manufactured by cells – biological entities – so controlling the all-important properties of the molecule is extremely tricky, notwithstanding the purification of those products post-production. Many factors affect stability, purity, aggregation etc., all of which are derivatives of the cell line and the environment in which they are cultured. So the need for end-to-end bioprocess data, coupled with agile molecular and cell biology tools becomes paramount.

We have considered many aspects of the science of cell and molecular biology above and the impact of tools and cloud approaches to help optimise a given biological entity. They are all used to make a biological entity's characteristics amenable to use a therapeutic (humanised, stable, potent, immune neutral, etc). However, the cell and molecular tools are also used to make a given entity amenable to be produced in a given system (mammalian, bacterial) which is given the term 'bioprocess'.

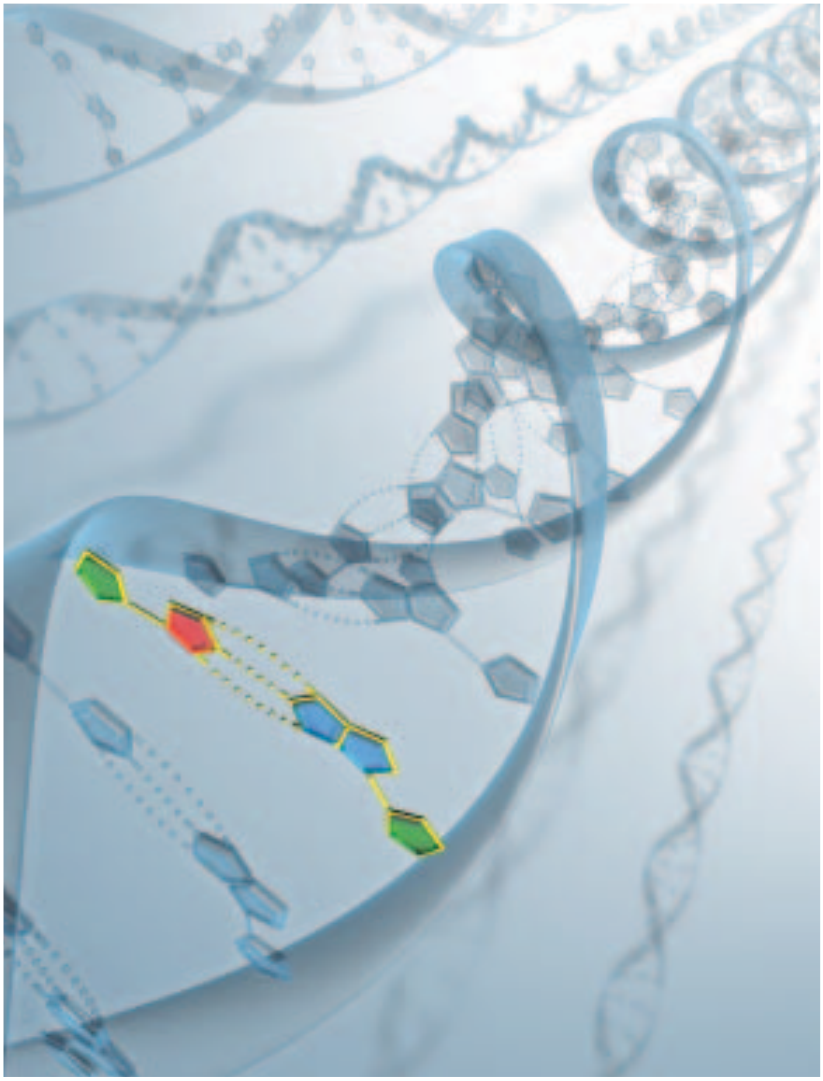
Systems and solutions exist for capturing various aspects of a bioprocess development lifecycle – from the process development and data management with IDBS' BPES (www.idbs.com), through to the analysis of ongoing development projects with dashboarding tools such as Discoverant (www.biovia.com) and Spotfire (www.perkinelmer.com).

Bioprocess also brings an opportunity to use the other element that cloud offers: easy access and support for AI (augmented intelligence https://en.wikipedia.org/wiki/Intelligence_amplification) and BDA (Big Data Analytics as defined by the three Vs – velocity, volume and variability https://en.wikipedia.org/wiki/Big_data).

Bioprocess is characterised by having all three of the Vs present. Velocity: This is particularly prevalent in the areas where real-time monitoring of fermenters is used. Given that the production of the biologic is often continuous, the system needs to be monitored in real-time for changes in any number of variables. Volume: Sub-second monitoring of large numbers of variables delivers a lot of data. Finally, Variability is guaranteed as the types and lists of variables constantly changes as understanding of the process critical parameters develops, ie you do not always know what will affect the end product or process, so you have to collect all the data.

This means that monitoring the fermentation and purification processes can have big data analytics applied to it. Keeping watch on data trends and trajectories in real-time on the large data landscape requires cloud-related capabilities – large compute and large data storage capabilities. After all, the genomics sequencing providers have solved these types of problems already and leveraged cloud to do it.

You can also see potential for AI (augmented and artificial intelligence). Augmenting the decisions made by scientists by providing an environment to identify the key variables that affect the important process outcomes and product quality is of great value. Whether artificial methods are used to analyse data to identify hidden relationships or augmented methods are used to support the scientists and enable their decisions to be made faster and with more certainty (based on statistical analysis) is a moot point – both have their place.



What do we do to piece all these component parts together to solve proper business problems?

The science of biology is constantly moving as we understand more and more about how things work. And at each point, new algorithms and tools are developed to help deal with the nuances and new science. However, taking each of these and making them useful for a scientist often means daisy-chaining various of the point algorithms and tools together to deliver applications, to reduce the amount of 'programming' and copy paste operations required to get a result of meaning. Many biologists have developed bioinformatics skills to cope with this but that does not scale, and scientists who are using the tools to get to an end decision point do not want the hassle.

Historically, we pipelined the tools and algorithms together using tools such as Knime (www.knime.org) and Pipeline Pilot (www.biovia.com). These tools are great at 'enabling a few to support the many' but they rely on the availability of underlying tools. Some of the established tools are not really cloud deliverable – in that respect I mean cloud deployable, web browser deliverable – but this is being delivered by some providers such as Knime, DNAnexus (www.dnanexus.com) at a level that means they can be used in a regulated environment – something critical for biologics development. The further you go down the R&D process, the importance of reproducibility and validated systems goes up. So here the tools are not cobbled together by individual scientists and just rolled out. The model is one of solution delivery by a provider, internal or external, where they manage the solution and ensure that it works, is tested and maintained.

Doing this in the pipelining world is tricky but it is exactly what the new kids on the block are doing. They package up various tools, algorithms

and services – some may be on the cloud already – and deliver those to their customers as a supported, validated service, meaning that they can be trusted and leveraged in GxP areas. We are seeing a rapid acceptance of cloud in validated environments, and it is likely to make people's lives easier, not harder as is always first imagined.

Why? Because every service used can be kept alive and online in the cloud at very little cost, unlike running things on-premise. The trick is keeping track of every version of every service used to deliver a given solution or pipeline. With cloud tools this is very easy, meaning you can have full traceablity of what has been used to get a given result and run it and check it easily, even when a service has been updated to a new version. This is guaranteed to revolutionise our thinking and capabilities, not only of biology informatics solutions in GxP areas, but of all solutions in GxP and regulated areas.

So what does all this mean?

From the descriptions and observations above, it is clear that the cloud and SaaS have had a dramatic influence on biology informatics. But it is more than that. The demand to produce large molecule therapeutics by the biotech sector has also put commercial pressure on the community to step up and deliver better fundamentals around how molecules are drawn, represented and shared. This gauntlet laid down by the industry has been met by precompetitive collaborations and endeavours that have really started to deliver value to the community as a whole.

The availability of massive compute power and storage capacity at affordable rates has also meant that everyone can have access to the latest sequencing and manipulation tools – not just those who can afford to build their own big compute infrastructures. Furthermore, the delivery of the new algorithms and tools in a cloud 'micro service' manner means that brand new business models are emerging that five years ago would have simply not have been tenable.

The drive of the pharma industry to deliver the predicted 80% of the world's therapeutics as large-molecule 'biologics' over the next 10 years has resulted in a commercial imperative that better tools be developed and made available quickly. The science and process of producing biologics is complicated and the commercial benefits of streamlining the process means we look for new ways to interpret the bioprocess data – and here we step into the realms of big data analytics and AI – both of which are a derivative of the high-level cloud concepts of coping with massive amounts of data and calculations in real-time.

The micro service concept, which is intrinsic to cloud application delivery and modern architectures, provides the ability for applications and solutions that combine tools from any provider to be delivered, validated and maintained far more easily, albeit with the same caveats as current architectures around traceability and backwards compatibility. This fundamental shift in capability (on-demand provision of compute power, almost limitless storage capacity, better application and solution architectures, easier validation, etc) which can be applied to the biology solutions domain, coupled with the tremendous, well-documented, commercial business demands requires a single question be asked: "How fast can we adapt to the change and leverage it?" DDW

Dr Paul Denny-Gouldson is VP, Strategic Solutions at IDBS. He heads the overall strategic planning for the various market verticals and scientific domains that IDBS works in. By working both internally across IDBS and externally with customers and partners, Paul is charged with the exploration and creation of new products, solutions and service offerings. He joined IDBS in 2005 as part of the acquisition of his ELN company and has spearheaded the drive to make E-WorkBook Suite the market leader. Prior to this, Paul founded a number of companies focused on combining science, technology and business. He started his career as a Post Doc and subsequently Senior Scientist at Sanofi-Synthelabo Toulouse (now Sanofi) for just under five years, where he managed a multidisciplinary molecular and cell biology department. Paul obtained his PhD in Computational Biology from Essex University in 1996, and has authored more than 25 scientific papers and book chapters.