Driving strategic improvements in biologics drug discovery with scientific informatics

Adoption of a scientific informatics system is essential to the success of every drug discovery organisation. However, making effective, informed decisions for adoption and implementation of these systems can be a complicated, multi-faceted process and will significantly impact the organisation. Good decisions can accelerate innovation and success, whereas poor decisions can hamstring a company for many years to come. Here we discuss many of the key factors to be considered when making such a decision, drawing from lessons learned from small molecule drug discovery, factoring in differences between small molecules and biologics, and highlighting the current status and trajectory of information technology trends.

Modern drug discovery, utilising the concepts of target identification, screening, design and synthesis, can be traced back to the 1940s, including the pioneering work of James Black, Akira Endo, Gertrude Elion and George Hitchings, among others. In the decades that immediately followed, the great majority of drugs developed were small organic molecules, with major exceptions being vaccines and synthetic human insulin. Whereas chemistry-based compound design could utilise Kekulé's pioneering work on carbon chain formation over a century earlier, the discovery of the basis of biological hereditary, providing the necessary knowledge for sequence analysis and recombinant biology, was not finally determined until 1953![1]. Even with the rapid progress of molecular and cellular biology that followed, the first monoclonal antibody, abciximab, was not approved in the US until December 1994[2]. Today, in terms of both numbers of on-market entities and annual sales, small molecule drug discovery still dominates over biologics[3]. However, the latter represents a significant and growing proportion of the overall drug discovery industry.

Is biology destined to forever follow chemistry? We can leave that question to the philosophers – instead, here we focus on what lessons can be learned from small molecule drug discovery that can be applied to speed up the maturity of biologics discovery, avoiding missteps and non-productive methodologies and approaches. The ultimate goal is to make biologics discovery more productive and cost-effective while developing safer, more effective drugs.

During the target identification to lead optimi-

By Dr Andrew LeBeau
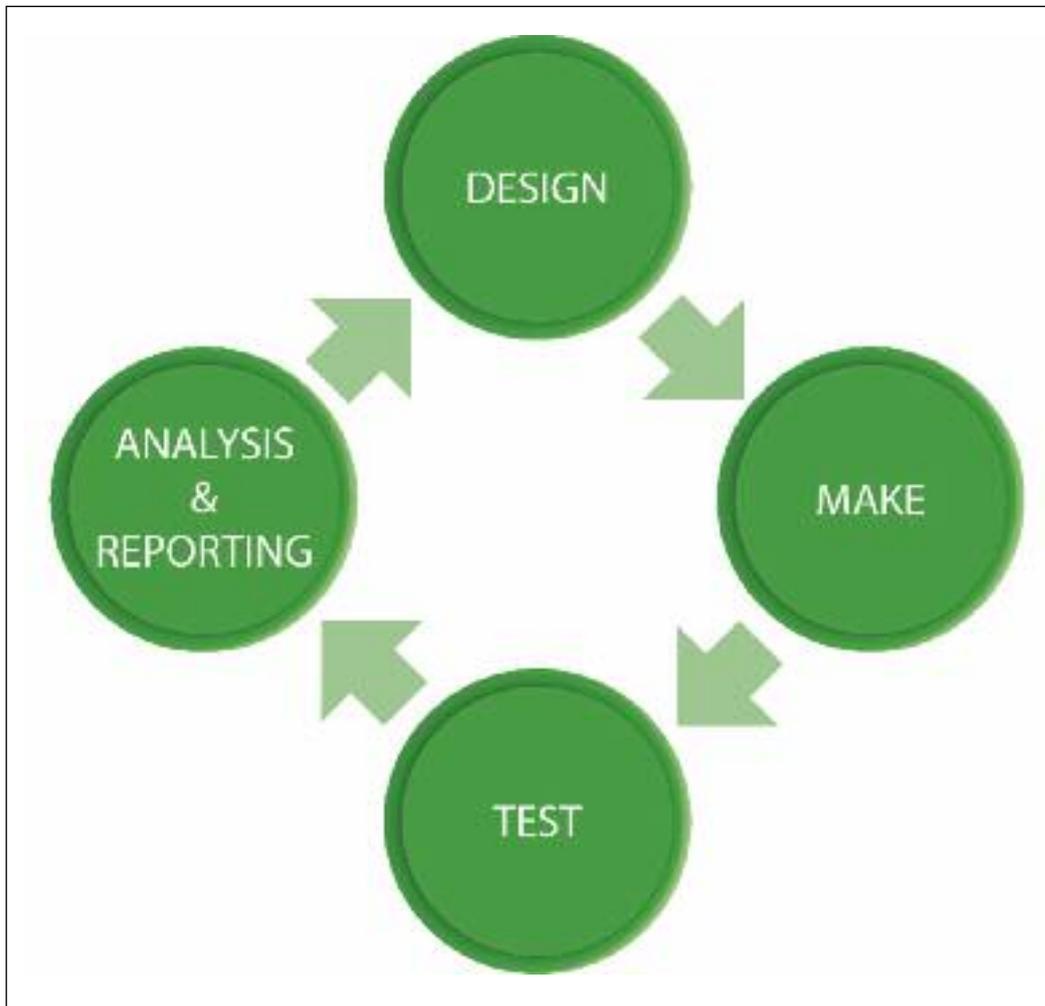
sation phase, the fundamental cycle of design > make > test > analyse and report, and back to the next design phase are shared across both small molecule and biologics discovery (**Figure 1**). An informatics system designed to help scientists efficiently gather information during the make/test phase can then be exploited to make the best decisions during the analyse/design phase. So, opportunities for adaptation of small molecule methods should exist. At the same time, the workflows used in the discovery of small molecules and biologics are different in many key respects, and so what works for the former may not work, or at least require some careful adaptation, for the latter. Finally, knowledge of the most recent and evolving trends in small molecule discovery may be useful in anticipating similar future trends in biologics. While the benefits may not be immediately realisable, awareness of these trends could be advantageous when planning and preparing for the future.

Implementation of an informatics environment

One area of fundamental importance to drug discovery is the proper use of informatics to support, and not detract from, the scientific process of discovery research and overall innovation. Today, information technology (IT) is deeply integrated and a fundamental part of the drug development process, as it is with almost any other industry. But in the history of drug development, this was not always the case, given that the fields of IT and modern drug discovery emerged and evolved over a similar time period. Introduction of computational servers, electronic databases and subsequently personal computers, along with many other IT innovations, all occurred during the history of modern drug discovery, not preceding it. Adoption of IT, intended to improve drug discovery processes, also meant significant disruption as those technologies themselves were undergoing rapid evolution. These technologies are now quite

mature, having transitioned to a current phase of ongoing change, reducing overall disruption while still delivering incremental improvements. But it was with much pain that we reached this point. Implementing informatics systems that were based on rapidly evolving IT meant mistakes and detours occurred, so it is paramount that we reduce or avoid similar issues in biologics drug discovery now that the underlying technologies are more mature. Moreover, commercial providers of scientific informatics systems were themselves at early stages of development, so many drug discovery companies opted to write their own systems since full commercial-off-the-shelf (COTS) systems did not exist. That is not the case today.

But is this really an issue? If IT is mature and COTS systems are available, won't biologics drug discovery simply adopt it efficiently, perhaps having already done so? In our experience, implementation of a professional grade, COTS, scientifically-aware informatics system in biologics drug discovery remains an issue. Data that would otherwise go into an ELN at a small molecule company is still often captured in paper lab notebooks and/or in non-scientific, general-purpose tools such as Microsoft Excel, PowerPoint and SharePoint. We know from small molecule drug discovery that there are better options. Small molecule discovery had no choice as ELNs were not available pre-21st century. But that is not the case now, so why are there still biologics companies using a paper/Excel solution? Clearly there are many factors. When a new start-up emerges, there is a lot to do, and activities and resources need to be carefully prioritised. Budgets are often tight and early-stage personnel tend to be focused on science and scientific technology, not information technology. It is easy to imagine that the implementation of an ELN, and other scientific IT infrastructure, is seen as something that can wait. The very future of a start-up company is often not assured, so putting a lot of energy into a sophisticated IT environment can seem premature. Moreover, there is a history in the drug discovery industry that chemists have generally been more willing to adopt commercial software than biologists. This may be because at the time that IT was evolving rapidly, with general access to personal computers, the dominant software element in biology was open-source bioinformatics software. That legacy remains today, but in our experience is much milder and less differentiating between younger chemists and biologists.

So while waiting to implement a scientific informatics system might appear to make sense, doing so is a critical and potentially very costly mistake. Not having such a system in place undermines the effectiveness of scientists at the very time this is most critical, as decisions are informed by incomplete data, and intellectual property protection is at risk. Additionally, this makes the eventual implementation ever more challenging, as non-standardised practices become implemented, more employees are brought on board, and an increasing body of data is created without a proper data management system in place. An informatics system implemented early can easily grow with the organisation, rather than seeming to disrupt it later, by imposing much-needed standardisation of processes and digitisation of legacy data from unstructured documents. For scientific companies whose entire value is in their intellectual property, not having systems to secure, audit and track that information can expose the entire enterprise to unnecessary risk.

Key IT trends

Awareness of IT trends is important for understanding needs both now and in the future to support an informatics system. Even though the overall IT landscape is more mature and stable than in pre-21st century times, concerns about potential disruptions and future-proofing needs can delay informatics adoption decisions due to 'paralysis by analysis'. Personal access to technology and IT mobility has matured to a point where most researchers and other drug discovery personnel are equipped with a laptop computer and either company-provided or personal smart phone that they use as part of their work life. While some data may be stored on the laptop itself, systems-of-record data storage and much of the daily-used project data is on company-managed servers or as part of cloud-hosted software-as-a-service (SaaS) environments. This configuration has been relatively stable for a few years, and will probably remain so for the next few years, given that currently there appears to be no major disruptive technologies in the consumer space which could influence business personal computing. Mobile computing platforms such as tablet computers, including 2-in-1 laptop/tablet hybrids, have not emerged as a dominant item of personal technology over laptop computers, as was once anticipated. The business market seems to have followed suit. Between low-weight laptops with optional touchscreens and smartphones getting larger, tablets have been squeezed out. The entrenched business operating systems and software (eg Microsoft Windows and Office apps) probably played a major role as these

Figure 2
Technology adoption curve from Diffusions of Innovations by Everett Rogers. After a slow transition through the Innovators and Early Adopters, adoption of cloud computing in drug discovery organisations is firmly in the Early Majority
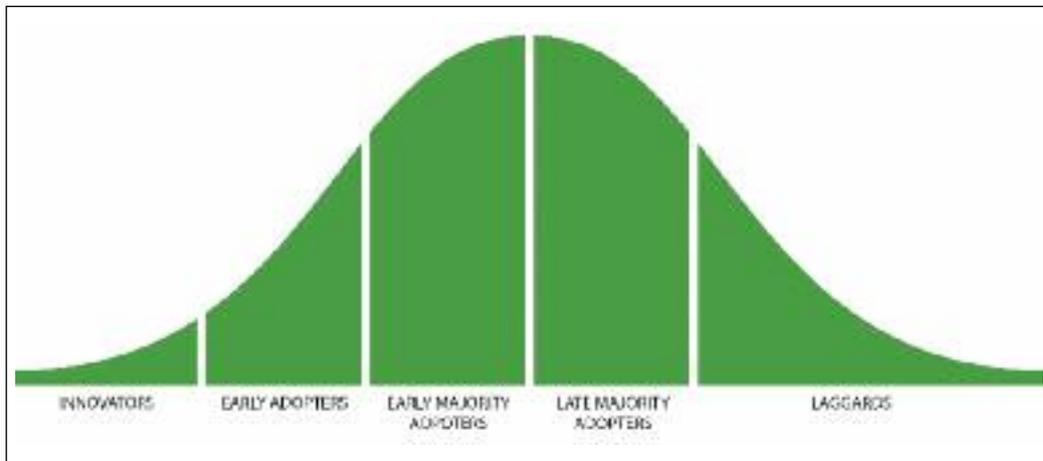
have only been available on mobile platforms relatively recently.

Cloud computing faced many years of resistance due to data protection concerns and entrenched business practices, but our observation is that the pace of adoption has accelerated significantly in the last two to three years. Utilisation of cloud has followed the typical technology adoption curve, driven initially by the pioneers and early adopters, and now we are well into the early majority (**Figure 2**). Driving adoption is that cloud computing is inherently logical for many organisations, and an additional dynamic is its utilisation to facilitate collaboration across global companies and in particular for projects involving other organisations, such as CROs and in joint discovery efforts with partner organisations. We anticipate continued adoption and full maturation of cloud computing over the next several years. As with personal computing, this technology should remain relatively stable for the next few years.

A third major IT trend whose effect and impact on businesses has been widely theorised is the social element. To what degree will (and has) the widespread consumer adoption of social computing translate to those working in the drug discovery industry? Attempts have been made to add 'social feeds' to informatics systems to facilitate collaboration and speed projects forward. There is little evidence to conclude that this technology has been broadly adopted. On the other hand, use of instant messaging (IM) apps on laptops and smartphones for more casual business interactions appears widespread. It seems that doing 'real science' requires informatics support that is more substantial and structured than social feeds allow, but IM is valuable for routine, less intellectually-demanding communication.

For all three IT trends, we seem to be in a period of relative stability, with mostly incremental changes and improvements in these and other IT capabilities, without expectations of major innovations and disruptions. Disruptive technologies are often observed in the consumer market first. Current major consumer trends include virtual reality (VR), Internet of Things (IoT) and Machine Learning (ML). For any technology to become widely adopted, it must provide a compelling advantage over existing capabilities and the drug discovery industry tends to be quite conservative, so the bar for adoption is particularly high. Considering each of these in turn, it is not immediately obvious what compelling advantage VR could provide. This remains nascent technology in the consumer space and likely to make major inroads initially in the gaming industry. For IoT, in the consumer space this means making connections between household appliances, environmental systems, etc with the internet, to allow control and information exchange. In many ways, the drug discovery industry is ahead with this technology, as laboratory instruments are connected (perhaps not always elegantly) to information management systems to allow the data generated by the instruments to be analysed, combined and interpreted. Similarly, the use of ML already has a significant history in drug discovery companies, particularly for small molecule discovery, where the development of data models to analyse drug-likeness is well established.

Overall we anticipate a period of relative stability in IT over the next several years allowing organisations looking to adopt a system to have confidence that the system will not be obsolete or need major upgrading soon after adoption.

Similarities and differences between small molecule and biologics drug discovery

Both small molecule and biologics drug discovery share the fundamental cycle of design > make > test > analyse and report, and back to the next round of design. An informatics system to support either type should include the same core software applications, comprised of data-capture applications, including an ELN and associated applications for compound and experiment design, a registration system for entity management and intellectual property (IP) protection, and laboratory instrument management and compound/consumable inventory application(s) to manage and support experimental execution. These are complemented by data management, analysis and visualisation applications such as an assay data management system for managing screening data and first-line analysis, and an analysis and visualisation application for in-depth analysis and interpretation of the data. Tying everything together should be an application to search and browse the current project status and data. Accessory software applications include tools for drawing and representing scientific entities on users' interfaces and in databases, tools for workflow management, and scientific plug-ins to non-scientific productivity applications such as Microsoft Office apps. When selecting the elements of an informatics system, it is critical to ensure that the elements can support the type of discovery being performed. In the case of entity registration systems, these have traditionally been distinct for small molecules and biologics because the nature of the entities themselves are so prominent in how the applications function and the underlying workflows differ due to the need to understand the lineage as part of determining the uniqueness of a biologic, compared with just knowing the structure for a small molecule. Conversely, assay data management systems often support both types of drug discovery, because the nature of the screening data may be less dependent on entity type and therefore a single system might work well in both cases.

While the overarching processes involved in small molecule and biologics discovery are similar, there are significant differences in operational details that impact the selection and implementation of an informatics environment. First and foremost, the nature of biologics entities are different from small molecules, being larger (up to three or more orders of magnitude) and designed and synthesised in diverse ways, which often involves live animals from which the initial creation of the can-

didate biologic is not fully controllable in the way that synthetic chemistry allows. Even when the synthesis of biologic compounds is done artificially, it is usually a much more complicated process than for small molecules. Therefore, for small molecule discovery the chemical structure of the candidate(s) and related structures are essentially the entirety of the IP, in biologics the process by which the candidate compound(s) are created also form part of the IP because they represent information about how to generate the compound(s) reproducibly. An informatics system for IP registration and protection must accommodate this distinction.

Another key consideration is recognising that comparing small molecules and biologics in the way the industry and informatics vendors support it is problematic. Small molecules represent a homogenous class of organic compounds of very similar size and composition, the synthesis of which occurs when two or more precursor molecules of the same class are combined in a reaction to create the target molecule. This means that informatics systems designed to support small molecules can be made very specific to deal with this type of entity. Biologics on the other hand represent a heterogenous class of compounds that can vary greatly in size, structure and composition. Learning from small molecule discovery, it is important to view each major class of biologics (ie antibodies, small peptides, vaccines, etc) as a class in its own right and develop informatics systems that can specifically handle whichever entity type(s) are the focus of the discovery programme. Informatics vendors must support this approach as well.

Biologics drug discovery occurs in two primary organisational configurations. The first is organisations that are solely focused on biologics, often a single biologic entity type. While there are some long-established biologics-only companies, many are relatively young, small and often in start-up mode. The second organisation type is where an established small molecule drug discovery company has moved into developing biologics, or plans to do so. Although the desired endpoint may be similar in both cases, the trajectory for acquiring a biologics informatics environment in these two cases typically differs significantly. As noted above, biologics-only companies often need to replace a paper-based system, or one which is a non-scientific, electronic system, such as Excel, PowerPoint and SharePoint. In this case the choices will focus on whether to implement a system iteratively, piece by piece, or take a more strategic view and create a vision for the end goal and develop a plan to get there (which can be done incrementally). In our

Informatics

References
1 Watson, JD, Crick, FH.
Molecular Structure of Nucleic
Acids: A Structure for
Deoxyribose Nucleic Acid.
Nature 1953; 171 (4356)
737-738.
2 The Epic Investigators. Use
of a monoclonal antibody
directed against the platelet
glycoprotein IIb/IIIa receptor in
high-risk coronary angioplasty.
The EPIC Investigation. The
New England Journal of
Medicine. 1994; 330 (14):
956-961.
3 Van Arnum, P. New Drug
Approvals Reached 21-Year
High in 2017.
https://www.dcatvci.org/5001-
new-drug-approvals-reached-
21-year-high-in-2017. Accessed
3/20/2018.
4 Brown, RD. Advanced
Analytics and Visualisation for
Biology Big Data. SLAS 2018
Presentation.

experience, the latter approach is likely to lead to a better ultimate outcome. In the case of a small molecule discovery organisation implementing a biologics discovery programme, a scientific informatics system is typically already in place. Existing legacy systems, and established vendor relationships, are likely to be factors in choosing the system to support biologics. Existing applications should be used where they can support both small molecules and biologics, and these can then be supplemented with software applications specifically tailored to supporting biologics discovery, either from vendors already used or others.

As scientific knowledge and supporting experimental technology evolve, opportunities arise for more complex and data-rich experimental methods. High content screening (HCS) has existed for some time, but biologics drug discovery has accelerated the development and adoption of HCS as it provides the opportunity to utilise a richer set of biological processes for the functioning of the candidate compound. Additionally, the richer biological functioning adds to the complexity of checking for undesirable side-effects. The result is richer, multi-channel data from the instruments, imposing additional requirements on analysis software to handle the larger volumes and complexity of the data.

Further challenging analysis and visualisation capabilities of informatics software is that biologic drugs have a basis in nucleic acid sequences, which in their native form can be exceptionally large. Combined with complex, multi-channel assay data, this imposes severe challenges on informatics systems in general, and analysis and visualisation applications in particular. The volume of data being generated and analysed is growing exponentially for a given investigation. For most software applications, performance will scale at best linearly with the size of the dataset being analysed (computational performance) and visualised (display performance). So if the data being analysed are 10x, 100x or even larger in scale, the performance characteristics of the software, and the associated user experience, will decline dramatically. In the worst (and not uncommon) cases, software applications will be swamped by the data volumes being generated today, and will fail due to memory limits, etc. Providers of biologics software applications must anticipate and respond to increased data volume needs with specific software engineering approaches and not rely on hardware performance improvements to hide performance deficiencies. Analysis and visualisation applications need to perform such that their performance in unaffected or minimally impacted as the scale of data increases, a sit-

uation referred to as zero-order scaling[4]. This means the applications may need to be recoded with better algorithms, potentially at the machine-code level, and to take advantage of modern hardware architectures, graphical programming units and programming languages. Drug discovery companies seeking to implement an informatics system must ensure performance metrics are fully established and verify that proposed solutions meet or exceed these.

Summary

Learning from past successes and challenges, and understanding trends and trajectories, are valuable assets in any decision-making. Initial adoption of a discovery informatics system, or making significant changes to supplement or upgrade an existing one, are major decisions, and will likely impact the organisation, positively or negatively, for many years. Such decisions clearly should not be taken lightly, but at the same time delaying or not making the decision at all is not a viable option. Doing so will almost certainly negatively impact the business, as the gaps will be filled by non-standard, ad-hoc solutions that will become harder to replace over time.

Here we have highlighted several key elements to factor into a decision, including lessons learned from small molecule discovery, current and near-future IT trends, key differences between small molecule and biologics entities and considerations to deal with the increasing volumes of data arising in biologics discovery programmes. There are other factors to consider as well, not able to be covered here, but the key takeaway should be that a scientific informatics system is essential and armed with this and additional information an organisation should approach that decision with confidence.   **DDW**

Dr Andrew LeBeau is Senior Manager of Biologics Marketing at Dotmatics. He joined Dotmatics in October 2017, bringing more than 15 years of experience working in the life sciences industry. At Dotmatics, Andrew leads efforts to highlight and promote the capabilities of Dotmatics software to support the rapidly growing and evolving field of biologics drug discovery.