The integration ABYSS

The long-standing disconnect between target discovery/characterisation and compound discovery is a major challenge facing the pharmaceutical industry in its quest to develop innovative therapies from the wealth of information flowing from the human genome.

By Dr Scott D. Kahn

There is enormous interest within the pharmaceutical and biotechnology industry to better utilise their data to improve the productivity of the drug discovery and development process. Notwithstanding the numerous discussions of chemistry-centric informatics (chemoinformatics[1,2]) and biological target-centric informatics (bioinformatics[3]), there remains an abyss separating these two key components in a comprehensive discovery informatics platform. This separation is a manifestation of the long-standing divide between chemistry and biology within the world's commercial drug discovery industry. The goal of this report is to examine the intersection between target- and chemistry-centric informatics, to explore how this intersection will evolve given the technological trends that will shape the pharmaceutical industry, and to start a dialog aimed at addressing the issues raised.

One of the reasons that target- and chemistry-centric informatics have yet to be actively discussed is that each of these disciplines is itself an area active with new development. A solution to the deployment challenge within chemoinformatics has recently been discussed within *DDW*[4], and the many facets of the rapidly developing informatics requirements for genomics, proteomics, and expression analysis are constantly discussed within conferences and through the literature[5].

Clearly, with all of the advances required by each of these separate informatics activities, it is easy to overlook their intersection. The consequence is that targets are selected without the benefit of previous knowledge amassed in the chemistry world, and the broad impact(s) of a chemical alteration is not easily evaluated against the wealth of knowledge amassed in the biological world. This fundamental problem can be understood through consideration of an important trend within the drug development process to account for, and exploit, patient specific variations in therapy response (pharmacogenomics[6,7]). The confluence of pharmacogenomics concepts and the drug discovery process highlights some useful starting points to explore the merging of target- and chemistry-based informatics.

In the future if drug discovery is to be performed such that variation of the protein target is addressed, a key will be a mechanistic understanding of drug action that is ultimately and intimately tied to the structure of the protein target, its co-factors, and the nature of interactions involving the protein target. This mechanistic understanding is essentially defined by the field of structural biology[8], and is a scientific discipline that is experiencing significant investment in both the academic and commercial laboratories. A major unsolved aspect within structural biology is the solution of novel

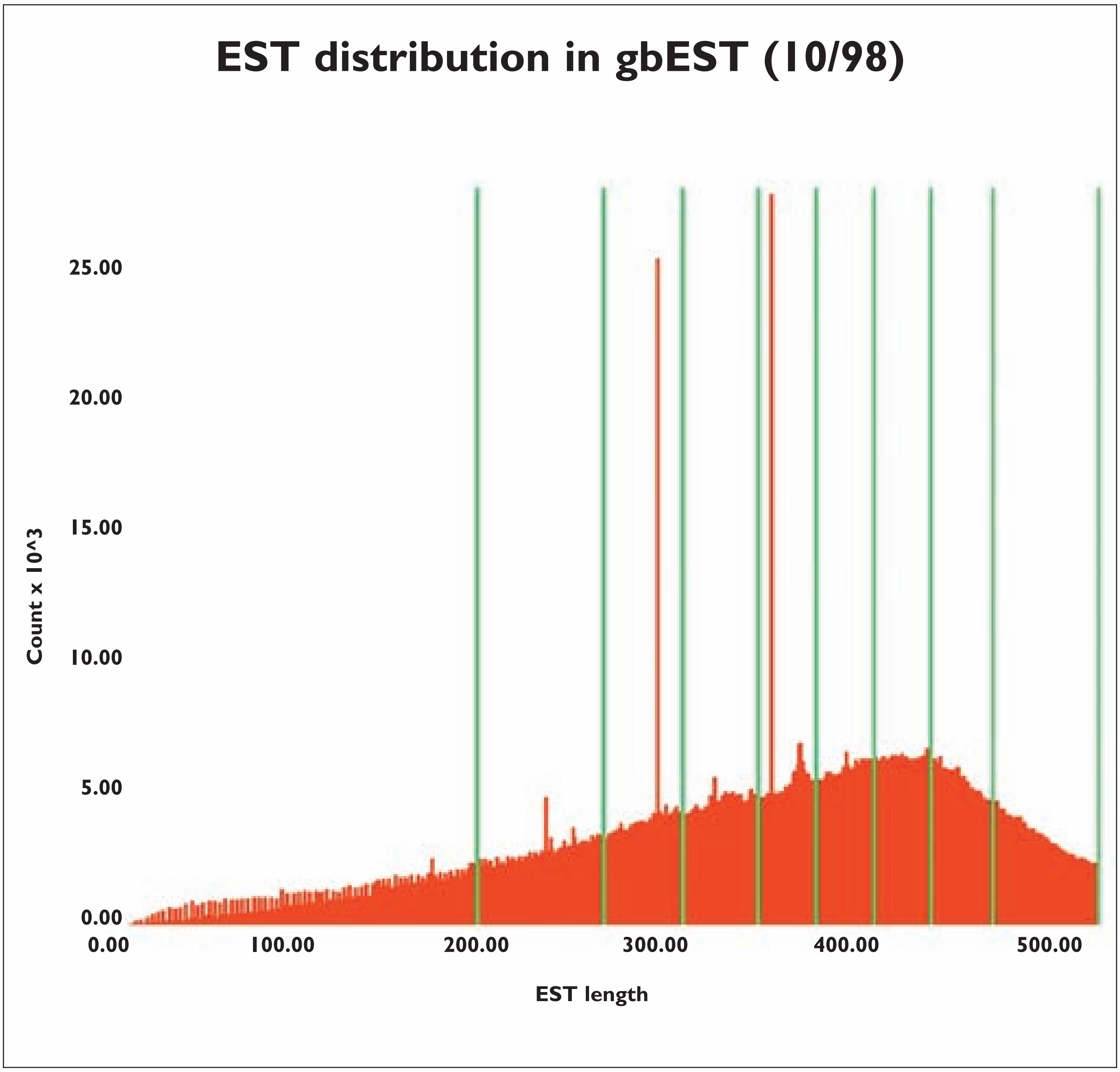

Figure 1
EST length analysis of the DNA sequence data in the public domain. Percentiles denoted by the green lines for every 10%

protein structures due to problems with obtaining crystals, or due to the environment in which the protein resides (eg, a transmembrane protein). Albeit, these challenges must be overcome if the drug discovery industry is to increase the number of innovative therapies and thereby decrease the relative number of 'me-too' drugs that are brought to market.

A leading question is whether structural biology can indeed provide a bridge between the target and the chemical compound worlds, and if so, what is the current state-of-the-art. As has been mentioned elsewhere[9,10], the ability to leverage protein structure in assigning the function of unknown protein sequences can be productive. This result follows from an observation that a majority of sequence data collected within genomics projects, typically as expressed sequence tags (ESTs) are of a length that encode the necessary structural motifs used by protein to form interaction/active sites (**Figure 1**). The real challenge, then, is to determine if available methods can parse the DNA into the expressed protein amino acid sequences, and if these amino acids sequences can in turn be converted into the appropriate three-dimensional structural motif that would provide information on the putative function of the protein (and hence the original DNA sequence). Ultimately, if successfully applied, functional information derived from 3D structure would be a good starting point to understanding the functional consequence of genotypic variations found in the human population.

The application of structural biology methods to genomic analysis requires a high level of automation; such automation was pioneered by Professor Andrej Sali in his creation of MODPIPE and has subsequently been refined in its implementation within GeneAtlas^{TM}[11]. Once automated, these techniques can then be distributed across large numbers of available computers to perform an analysis on a full genome in a matter of days for a small bacterial genome, to weeks for genomes of larger organisms. The value of this analysis is significant, yielding an incremental amount of assignment[12] for the Deinococcus Radiodurans and Vibrio Choleræ genomes of 16% and 22%, respectively, when compared to the assignments made available from The Institute for Genomic Research (TIGR)[13,14]. Importantly, much of this incremental assignment is associated with a full protein structural model that can be used to understand active site regions, and hence how potential drug molecules might interact to promote or inhibit the role of function of the protein species.

The bridge between target- and chemistry-centric informatics will result as a common framework is developed that permits both perspectives to be expressed as a view on a common database schema. While such a schema is not available today, two interesting approaches are being taken that can evolve into a merged solution as experience is gathered. In the first of these, annotations of target sequences are collected and indexed that express the consequence of the interaction of chemical compounds with the putative protein product(s). Termed chemogenomics, these annotations are derived from the chemical concepts of pharmacophores and protein-ligand binding that have been developed and validated during the last decade. As shown in **Figure 2**, the chemogenomic approach of reverse screening uses the ability of a protein structure to interact with a specific chemical compound to draw relationship between itself and other protein targets that exhibit the same response. As many such probe compounds are used to assess a protein target's interaction profile, a response 'fingerprint' can be created from which comparisons can be made. Rather than similarities based upon the composition of the protein (ie, the existence and ordering of its amino acid residues), these similarities represent a similar response profile which might serve as a basis for understanding whether and how selective compounds/drugs might be prepared.

There are several technological hurdles to overcome in making a practical chemogenomics solution. The first of these hurdles is accessing 3D protein structures that correctly reflect the variability in shape that the protein can exhibit when binding a ligand or interacting with another protein or co-factor. While related to the conformational analysis of small organic compounds, the problem with protein flexibility is that there are many more degrees of freedom, and a large preponderance of forces in subtle balance that must be accurately modelled. A second hurdle is the reliable and automatic identification of interaction and active sites. Here several approaches have been adopted that seek out invaginations lined with hydrogen bond donor/acceptor interactions; the success of such methods is closely related to the aforementioned challenge with representing the 3D structure of proteins appropriately. A final hurdle involves the accounting of the effects of post-translational modifications (of the protein target) on the interaction and/or active sites available. It is reasonable to assume that with the current focus on increasing functional knowledge of the proteome that rapid progress in both of these areas will be reported.

All of this work in creating chemical annotations would not contribute to the evolution of a comprehensive schema in the absence of a database. Fortunately, processing of genomic quantities of data can only be done efficiently within a database context, and indeed most efforts to annotate genomic sequence data tends to be captured in relation databases. While the handling of structural biological information (3D protein structure, chemogenomics, etc) is a step in the right direction, considerable progress is needed before a schema has been developed that truly encompasses a convergence of biological targets derived from sequence data and a robust handling of compound related information.

References

1 Brown, FK. Cheminformatics: What is It and How Does It Impact Drug Discovery. Annual Reports Med.Chem., 33, 1998, 375-384.

2 Hann, M, Green, R. Chemoinformatics – A New Name for an Old Problem. Curr. Opinions Chem. Bio., 3, 1999, 379-383.

3 Bishop, MJ, Rawlings, CJ. DNA and Protein Sequence Analysis. IRL Press at Oxford University Press, Oxford, 1997.

4 Ertl, P, Miltz, W, Rhode, B, Selzer, P. Web-Based Cheminformatics for Bench Chemists. Drug Discovery World, 1, 2000, 45-50.

5 See, for example, each of the issues of Drug Discovery World in 2000 for multiple articles.

6 Murphy, M. Pharmacogenomics: A New Paradigm for Drug Development. Drug Discovery World, 1, 2000, 23-32.

7 Brockmöller, J, Kirchheiner, J, Meisel, C, Roots, I. Pharmacogenetic Diagnostics of Cytochrome P450 Polymorphisms in Clinical Drug Development and in Drug Treatment. Pharmacogenomics, 1, 2000, 125-151.

8 Fersht, A. Enzyme Structure and Mechanism, 2nd Edition, W.H. Freeman and Company, New York, 1985.

9 Powell, K. Target Discovery and Drug Design: Extracting the Value from Genomics. Drug Discovery World, 1, 2000, 25-30.

10 Zhu, ZY, Yan, L, Edwards, DJ, Badretdinov, A, Olszewski, K. GeneAtlas: A High-Throughput Pipeline for Automated Model Building and Functional Annotation of the Genome. http://www.msi.com/life/struct_bio/webzine/00/q4/appnotes/Gene%20Atlas1.html

11 Molecular Simulations Inc, 9685 Scranton Road, San Diego, California 92121 USA.

12 Yan, L, Edwards, DJ, Szalma, S. Private communication.

Continued on page 54

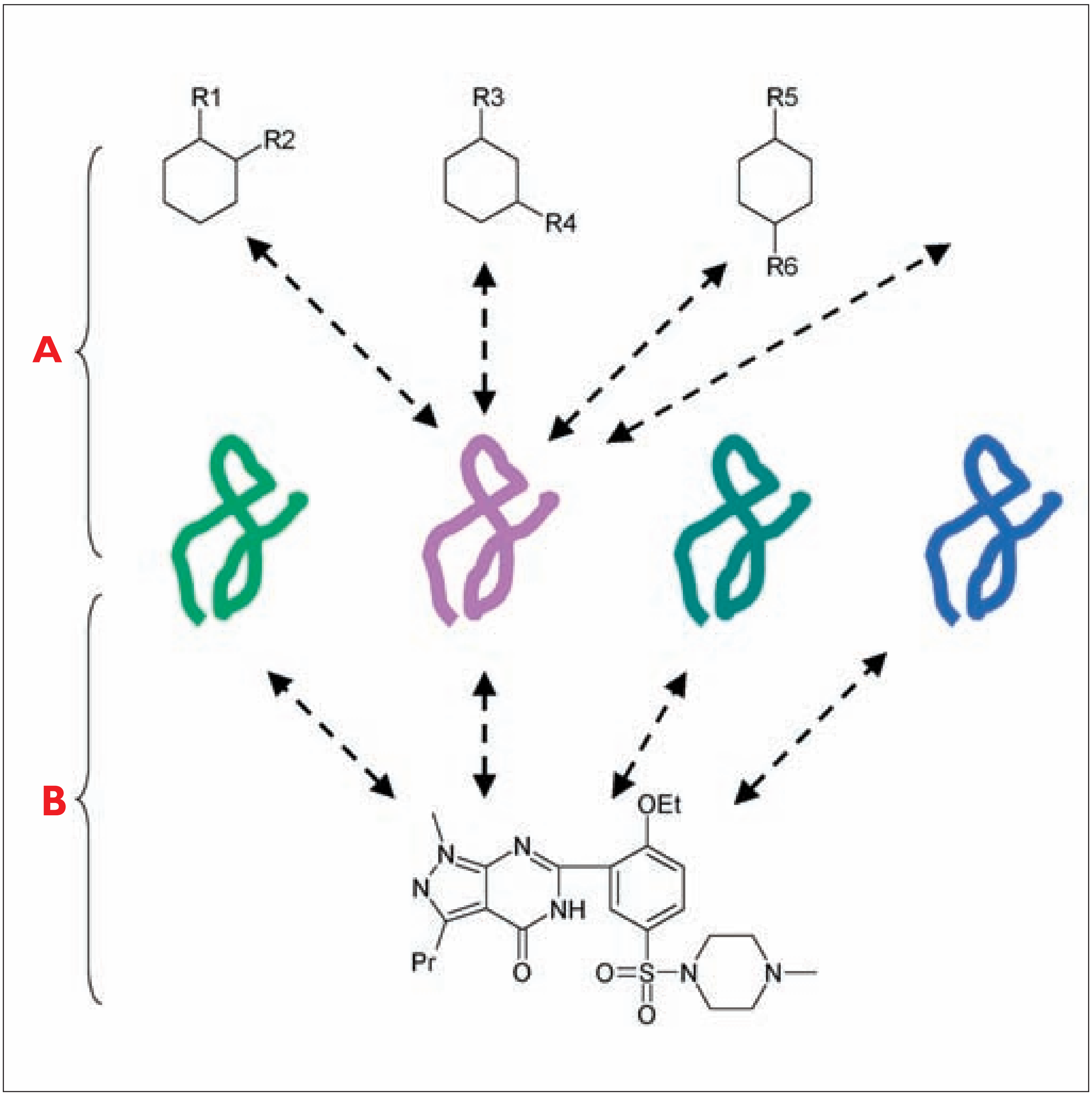

Figure 2
A Evaluation of a single protein target in its ability to interact with a variety of chemical compounds (virtual screening); the data is stored with the compound
B Evaluation of many protein targets in their ability to interact with a single chemical compound (reverse virtual screening); the data is stored with the protein target

The convergence on a unified platform for target and chemistry data is also occurring from the chemistry world. Here there are many efforts to provide more comprehensive access to all biological data known for a compound, and perhaps more importantly, such biological data is being tracked with much greater consistency (ie, assay system to assay system) and with more detailed information on the protocol and target system used. Progress here is being driven by the need of chemists to compare and relate information on a compound in ways that detailed relationships between a compound's structure and the resulting activity/selectivity can be formulated and used in a prospective manner.

A goal for the creation of a unified informatics platform that spans targets and compounds can ultimately be found in the types of problems that could be solved. As examples, consider the ability to evaluate target classes not only on their involvement in a specific biochemical pathway, but also based upon the uniqueness of their interaction with compounds that might be developed as a drug. Use of such chemistry information in target selection could predispose targets for success as compounds are tested in vivo. Conversely, an up front knowledge of a compound's profile against proteins involved in metabolism and/or toxicological pathways could avoid one of the more costly points of failure for a late-stage development candidate. Both examples suggest an improvement in overall efficiency to the drug discovery and development process.

In summary, significant progress is being made at the key intersection between the target-centric world of biology, and the compound-centric world of chemistry. This intersection is postulated to be embodied within the field of structural biology, and moreover it is concluded that a deep knowledge of structural biological concepts will drive the creation of a comprehensive informatics platform. One outstanding problem still not addressed involves the ability of an informatics to adequately capture the variation of chemical function over time, ie, proteins being expressed during growth, up and down-regulation cycles, responses to external stimulants such as light, food and drugs, post-translational modifications. Such important biological perspectives must be captured, and fortunately some progress has been made[15]. Thus, the complex problem of spanning the abyss between target- and compound-centric informatics is really just the tip of an iceberg. DDW

__Dr Scott Kahn__ joined Molecular Simulations Inc in 1990, and during the last 10 years he has developed and marketed a variety of new technologies aimed at extracting correlative information on biologically active compounds. He was a founding contributor to the SPARTAN and CATALYST^{TM} programmes that are widely used throughout the pharmaceutical and biotechnology industries. Dr Kahn is currently the Senior Vice-President of Life Sciences at MSI. He obtained his bachelor of science degree in chemistry from Rider College in New Jersey, his PhD in theoretical chemistry from the University of California, Irvine, and he did post-doctoral research at Cambridge University in the UK. Before leaving to work in the software industry, Dr Kahn was an assistant professor in organic chemistry at the University of Illinois in Urbana-Champaign.

Continued from page 51

13 White et al. Science 286, 1999, 1571-1577.

14 Heidelberg et al. Nature 406, 2000, 477-483.

15 KEGG: http://star.scl.genome.ad.jp/kegg/