CHEMOGENOMICS
a gene family approach to parallel drug discovery

Currently available drugs only target around 500 different proteins[4]. Recent reports from efforts to sequence the human genome suggest there are tens of thousands of genes[1,2] and many more different proteins. Popular estimates of the number of 'new' drug targets that will emerge from genomic research range from 2,000 to 5,000[3]. A critical question as we enter the post-genomic world is: how can the pharmaceutical industry rapidly discover and develop medicines for these new targets to improve the human condition?

By Dr Paul R. Caron, Dr Michael D. Mullican, Dr Robert D. Mashal, Dr Keith P. Wilson, Dr Michael S. Su and Dr Mark A. Murcko

In the pharmaceutical industry to date, research and early development activities have typically been organised according to therapeutic area. In organising their drug discovery efforts in this way, companies have sought to create efficiency by building a critical mass of expertise and experience in the biology of related diseases. Over the past 20-30 years this organisational approach has proved successful for many companies. While there is no doubt that this strategy produces some synergies in early stage research, greater efficiency in late-stage clinical development and marketing is the main driver for the organisation of research and development resources along therapeutic area lines.

Pharmaceutical companies have also traditionally tackled one protein target at a time in drug discovery. Over the years, increasingly sophisticated technologies and approaches have increased the efficiency of drug discovery based on single targets at a pace sufficient to keep the pipelines of many major companies well-stocked with promising development candidates. The development and application of high-throughput chemical synthesis and *in vitro* biological screening, for example, as well as new computational methods applied to QSAR, structure-based drug design and informatics, have accelerated the drug discovery process[4].

Dramatically new and different drug discovery approaches, however, are needed to take full advantage of the massive influx of targets being elucidated through genomic research. Simply stated, a therapeutic area focus and a single target drug discovery approach do not create enough efficiency to allow companies to keep pace with the massive inflows of new target information. An ideal solution would be to accelerate drug discovery by processing multiple related targets in parallel, reusing information and know-how across targets in a way that allows chemistry to be broadly leveraged. Drug discovery approaches that focus on structurally similar protein families, and thus leverage the way which particular classes of chemical compounds will interact with targets within the same family, may be just such an ideal solution.

Just as the fields of genomics and proteomics are broadly characterised as the identification and classification of all the genes and proteins in a genome, the field of chemogenomics may be characterised as the discovery and description of all possible drugs to all possible drug targets[5].

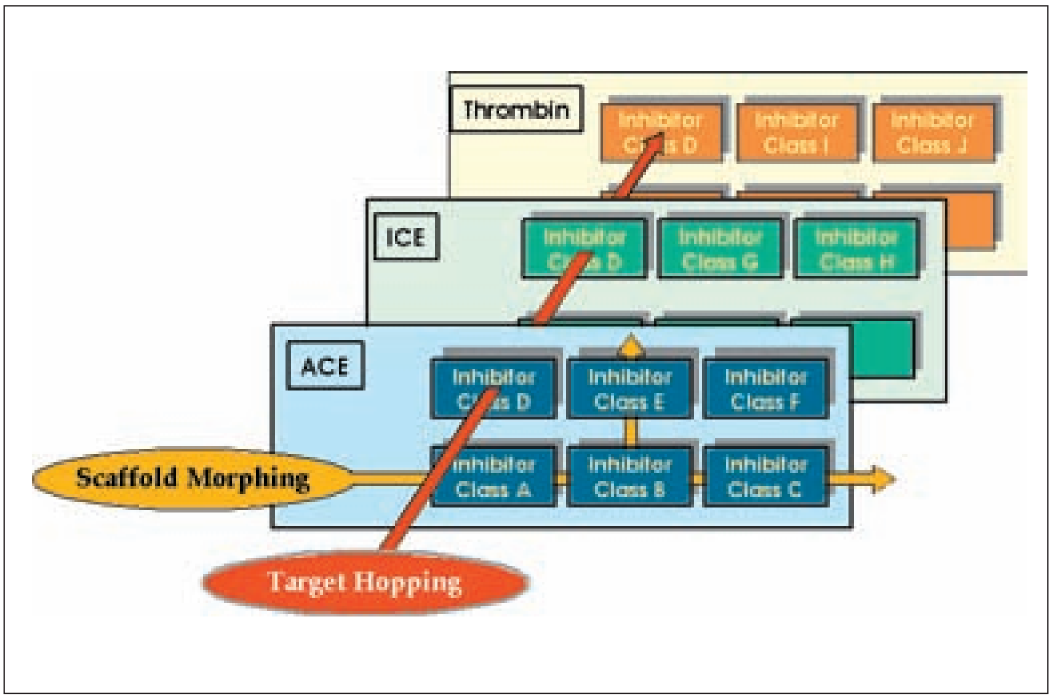


Figure 1
Scaffold morphing and target hopping are two key concepts in chemogenomics. *Scaffold morphing* is the generation of multiple, chemically distinct lead classes ('scaffolds') against any particular target. *Target hopping* is the ability of compounds from the same lead classes to interact with multiple targets – in effect, to be 'reused'. Importantly, while the scaffold *class* is reusable, *different specific compounds* from the same scaffold class will be optimal for different targets in the family

Analogous to genomics and proteomics, success in chemogenomics will require not only highly integrated technology and computational advances, but will necessitate fundamental changes in the pharmaceutical drug discovery process. Any significant progress towards this goal could generate a formidable package of patentable drug molecules.

Organising research by gene family

Industrialising parts of the drug discovery process-by incorporating parallel processing, miniaturisation and robotics-has helped to increase the efficiency of drug discovery in important ways. The requirement of practical expertise in many parts of of the drug discovery process, however, suggest that there is a limit to the efficiency that will be created by automation.

Chemogenomics is distinct from chemical genetics. Chemical genetics (sometimes called 'reverse chemical genetics') entails the use of defined chemical probes to help understand biological targets and pathways. The fundamental premise is that chemical probes, if of sufficient potency and selectivity in cellular or animal models, can be used to help understand and to prioritise those targets of the greatest therapeutic relevance. Thus chemical genetics as currently described in the literature is essentially a 'target validation' technology. Chemogenomics, on the other hand, is principally a 'chemical' technology which aims to produce new chemical entities (NCEs) – clinical development candidates – as efficiently as possible. The molecules which derive from the chemogenomics approach can of course be used as chemical probes in a 'target validation' sense as well.

A major potential source of efficiency in any process lies in the reuse of information and know-how. A gene family approach to drug discovery seeks to exploit this efficiency to its maximum. Targets within a gene family – defined by homology at the protein sequence level – will often have very similar *in vitro* assays and properties, providing some leverage of biology resources. In addition, a significant percentage of compounds designed and synthesised against one family member will be active against other family members, which can allow medicinal chemistry on multiple targets to have a common starting point. In addition to creating efficiency, reuse of chemical and biological information may produce intellectual property that is transferable among related targets. Some of these concepts have been touched upon by other groups[7-11].

Based on our experience with employing structural biology and modelling approaches together with combinatorial and medicinal chemistry, we have found that it is possible to design multiple classes of compounds to inhibit each target within a gene family. We refer to this as scaffold morphing.

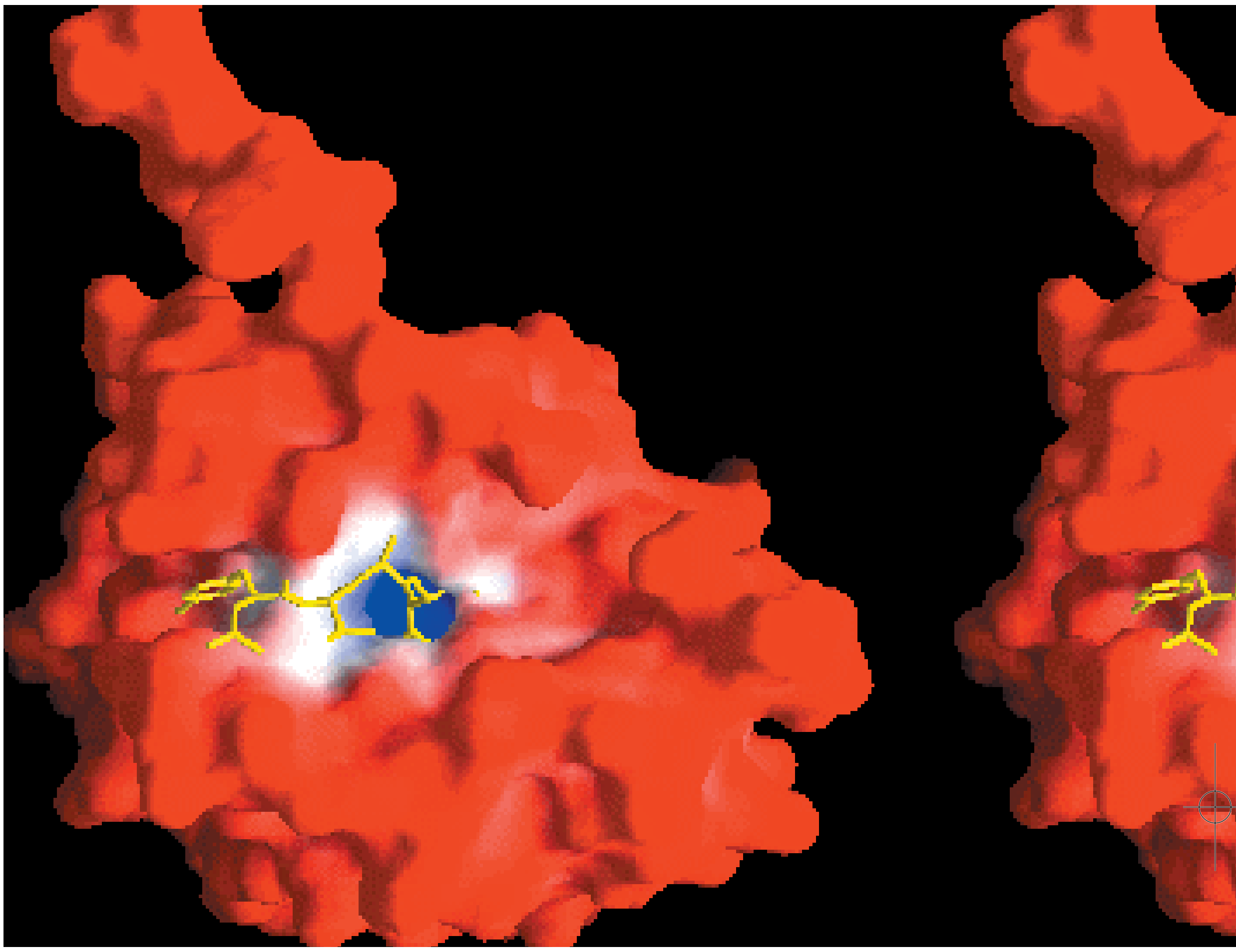


Figure 2
Sequence homology is often a good predictor of three-dimensional structural homology. On the left panel is the crystal structure of caspase-1 (ICE) colour coded by the sequence homology of a set of 10 different caspases. Blue = highly conserved sequences, white = intermediate homology, and red = low homology. On the left panel is the crystal structure of caspase-1 colour coded by the three-dimensional structural variation in the C-α, positions taken from a superposition of five different caspases. Blue = highly conserved C-α positions, white = intermediate and red = low structural conservation

Once created, the molecular libraries may be screened against multiple targets within the family, and the breadth of activity of each active chemical scaffold may be rapidly explored. This 'compound reuse' strategy is sometimes called target hopping. The combination of 'morphing and hopping' are essential for the rapid generation of multiple development candidates against multiple targets within a family (**Figure 1**).

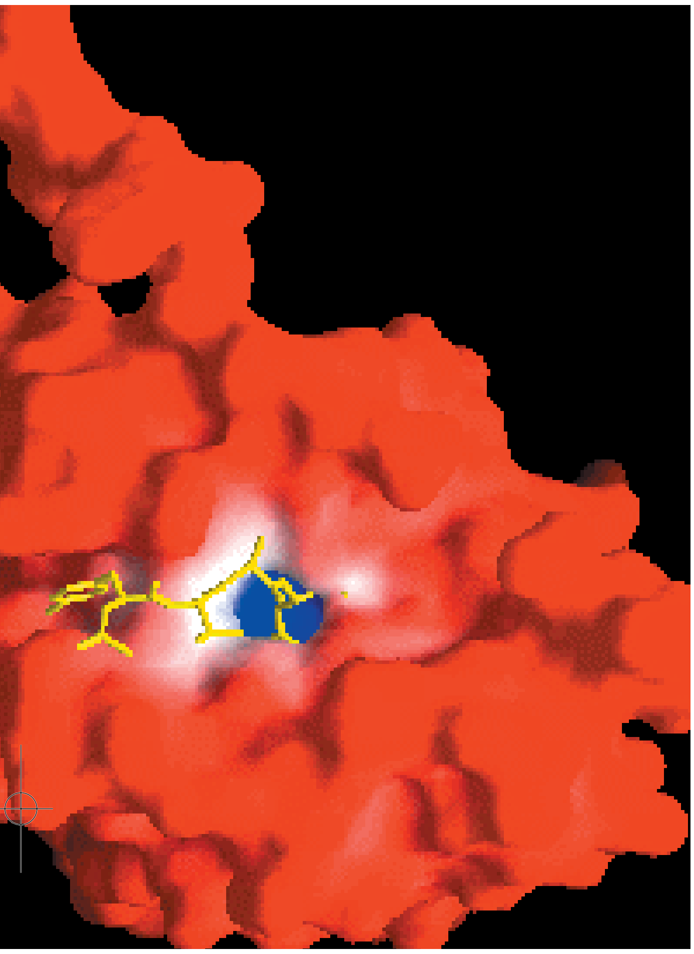
The three-dimensional structures of approximately 15,000 proteins are now publicly available and there is a recent surge of interest in the public and private sector to actively obtain representative structures for novel proteins. The combination of this raw data and refined homology modelling tools is now enabling the structures of a large number of pharmaceutically relevant protein targets to be predicted as well as the shapes and physical properties of potential ligand binding sites. The ability to map genomic data on to protein structures provides the framework linking three billion As, Cs, Gs and Ts to drug design chemistry.

The completed genome sequence enables the identification and classification of all members of a gene family into subfamilies based on a number of criteria: overall sequence homology, domain structure, and/or transcriptional regulation. This genome-wide perspective distinguishes chemogenomics from traditional gene family research.

The need for therapeutic area knowledge

A central tenet of the chemogenomic approach is that efficiencies will result from reuse of information and compounds, driven by the overlap between chemical space and the active sites of the protein family members. Maximal increases in efficiency and productivity can only occur when this knowledge is concentrated in a single discovery unit which is organised along areas of 'chemogenomic space'. We believe it is possible to have this organisational structure from project initiation through first in man studies.

A distinct advantage of this discovery structure



occurs for targets with multiple potential therapeutic indications. Often, the optimal compound characteristics (intravenous versus oral dosing; formulation; specificity) will differ across potential therapeutic indications. During lead optimisation, differing compound characteristics across multiple therapeutic can be explored fully, whereas in a traditional pharmaceutical organisation discovery efforts are usually confined towards only a single therapeutic indication. While the therapeutic area organisational structure is sub-optimal for a chemogenomic approach to discovery, there remains a clear rationale for organising late stage development (phases II-IV) and commercial operations along therapeutic lines.

How genomic data can drive drug discovery

Having complete knowledge of all the members of a given gene family provides a new perspective on the drug discovery process. The availability of complete gene sequences information for an entire target family, together with a representative subset of protein structures, allows one to build three-dimensional models for the entire protein family and to map the interactions of a given substrate or inhibitor and specific residues in the target even when detailed structural data is unavailable. This provides a firm understanding of the subset of residues which provide key inhibitor interactions, and enables the prediction of inhibitor specificity. For example, inhibitors which make strong interactions with unique or 'rare' residues are likely to demonstrate more target specificity. We and others have demonstrated that single amino acid changes are sufficient to generate specificity in protein kinases[12-15].

It must be appreciated that high-resolution structural data is not required in the chemogenomics strategy. Naturally, the ability to map protein sequence on to inhibitor-target co-complex structures provides a fundamental link between the genomic sequence information and the medicinal chemistry required for drug design. However, at the amino acid level, it is possible to utilise various protein folding prediction methods and mutagenesis data to build respectable models of most proteins. These models would be sufficient to provide guidance for chemistry both with respect to potency and specificity. The recent publication of the structure of a mammalian GPCR and the increasing number of publications on membrane-bound proteins[18-20] suggest that sufficient structural information to construct such models will be available.

Chemogenomics: the caspase example

An example of an interesting gene family is the caspases, cysteine proteases with specificity for cleavage after aspartyl residues. Interleukin-1ß converting enzyme (ICE) has been shown to be essential for cytokine processing and is currently being pursued as a drug target[25]. There are also roughly a dozen caspases with sequence homology to ICE, and while the exact function of all of these is not known, it is clear that some of these have an important role in regulation of apoptosis[26]. Structural insights through x-ray crystallography facilitated the rapid identification of selective inhibitors of these other potential drug targets. Experience with ICE has also been applied to the other caspase family members: expression, purification, assay development, crystallisation, and structure determination of these homologs[27].

The x-ray crystal structures of the caspases also nicely highlight the way in which sequence conservation can be a good predictor of three-dimensional structural conservation (**Figure 2**). This is important

References

1 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

2 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. Science 2001, 291:1304-1351.

3 Drews J. In Human Disease – from Genetic Causes to Biochemical Effects. Edited by Drews J, Ryser S. Blackwell; 1997:5-9.

4 Drews J, Ryser S. The role of innovation in drug development. Nat Biotechnol 1997, 15:1318-1319.

5 Caron PR, Mullican MD, Mashal RD, Wilson KP, Su MS, Murcko MA. Chemogenomic approaches to drug discovery. Curr Opin Chem Biol 2001, 5:464-470.

6 Schreiber SL. Chemical genetics resulting from a passion for synthetic organic chemistry. Bioorg Med Chem 1998, 6:1127-1152.

7 Ohlstein EH, Ruffolo RR, Elliott JD. Drug discovery in the next millennium. Annu Rev Pharmacol Toxicol 2000, 40:177-191.

8 Lehman J, Baxter A, Brown D, Connolly P, Geysin M, Hayes M, Howard R, Knowles J, Lee M, Lyall A, et al. Systematization of Research. Nature 1996, 384 Supp 7:5.

9 Frye SV. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chem Biol 1999, 6:R3-7.

10 Thorpe DS. Forecasting roles of combinatorial chemistry in the age of genomically derived drug discovery targets. Comb Chem High Throughput Screen 2000, 3:421-436.

11 Debouck C, Metcalf B. The impact of genomics on drug discovery. Annu Rev Pharmacol Toxicol 2000, 40:193-207.

12 Wilson KP, McCaffrey PG, Hsiao K, Pazhanisamy S, Galullo V, Bemis GW, Fitzgibbon MJ, Caron PR, Murcko MA, Su MS. The structural basis for the specificity of pyridinylimidazole inhibitors of p38 MAP kinase. Chem Biol 1997, 4:423-431.

Continued on page 62

Continued from page 61

13 Lisnock J, Tebben A, Frantz B, O'Neill EA, Croft G, O'Keefe SJ, Li B, Hacker C, de Laszlo S, Smith A, et al. Molecular basis for p38 protein kinase inhibitor specificity. Biochemistry 1998, 37:16573-16581.

14 Fox T, Coll JT, Xie X, Ford PJ, Germann UA, Porter MD, Pazhanisamy S, Fleming MA, Galullo V, Su MS, et al. A single amino acid substitution makes ERK2 susceptible to pyridinyl imidazole inhibitors of p38 MAP kinase. Protein Sci 1998, 7:2249-2255.

15 Gum RJ, McLaughlin MM, Kumar S, Wang Z, Bower MJ, Lee JC, Adams JL, Livi GP, Goldsmith EJ, Young PR. Acquisition of sensitivity of stress-activated protein kinases to the p38 inhibitor, SB 203580, by alteration of one or more amino acids within the ATP binding pocket. J Biol Chem 1998, 273:15605-15610.

16 Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science 1998, 280:69-77.

17 Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, et al. Crystal structure of rhodopsin: A G protein-coupled receptor. Science 2000, 289:739-745.

18 Sukharev S, Betanzos M, Chiang CS, Guy HR. The gating mechanism of the large mechanosensitive channel MscL. Nature 2001, 409:720-724.

19 Dinarello CA. Interleukin-1 beta, interleukin-18, and the interleukin-1 beta converting enzyme. Ann N Y Acad Sci 1998, 856:1-11.

20 Marks N, Berg MJ. Recent advances on neuronal caspases in development and neurodegeneration. Neurochem Int 1999, 35:195-220.

21 Wei Y, Fox T, Chambers SP, Sintchak J, Coll JT, Golec JM, Swenson L, Wilson KP, Charifson PS. The structures of caspases-1, -3, -7 and -8 reveal the basis for substrate and inhibitor selectivity. Chem Biol 2000, 7:423-432.

for the chemogenomics approach because the combination of structural and sequence information enables rapid drug design progress within families even without having the high-resolution structures of every target of interest within those families.

Measures of success

The success of a chemogenomics approach should be quantifiable using a number of parameters – such as the number of patents, the number of compounds synthesised to get to the clinic, and ultimately increased numbers of approved drugs and sales. It is expected that patent applications generated with a chemogenomic background would be able to support claims which cover a broader range of chemical space than average and would be able to include a comprehensive list of defined molecular targets and indications. Increased efficiency in identifying potent chemical leads with 'drug-like' properties should accelerate the process of driving these leads into clinical development candidates with the desired pharmacologic and pharmacokinetic parameters without compromising quality. As the first of the candidate molecules generated using this integrated approach enter the clinic in the near future, the ability of chemogenomics to increase overall productivity in the pharmaceutical industry will start to be apparent in the number of new molecular entities approved by the FDA within five years.

Summary

The chemogenomics approach, where gene sequence information is combined with protein structure and/or models to link to chemical inhibitors, is designed to fully utilise the sequence information by considering large families of gene targets at once. This highly parallel approach depends on well-established methods such as combinatorial chemistry, high-throughput screening, computational chemistry, structural biology and bioinformatics, all of which drive the efficient re-use of information, reagents, methods and know-how as research teams move from one group of targets to the next.

It is still an open question whether a gene family focus is more efficient than a 'traditional' approach. However, early results from our research into the protein kinases and the caspases support our opinion that a gene family approach can provide a more efficient process for generating late-stage development candidates. Clinical and marketing expertise in specific therapeutic areas is also of great importance to the success of any new drug, and traditionally pharmaceutical companies have organised their R&D efforts to capture the advantages of this expertise. It is essential that gene family initiatives fully utilise such specialised knowledge in particular therapeutic areas while at the same time not limiting the pursuit of targets in other therapeutic areas. These particular challenges, while complex, are not insurmountable with foresight and skillful planning.

Acknowledgement

We would like to thank Michael Partridge for his help with revising this manuscript. DDW

Paul Caron *is the Director of Informatics and leads the kinase target selection group at Vertex. Paul holds a PhD from Johns Hopkins University in biochemistry and was a post-doctoral fellow at Harvard. He joined Vertex in 1994.*

Michael Su *is the Senior Research Fellow and Head of Biology, and co-project head of the Vertex kinase programme. Michael holds a PhD from Duke University in molecular biology and was a post-doctoral fellow at Harvard. He joined Vertex in 1990.*

Keith Wilson *holds the title of Senior Research Fellow and Head of Structural Biology. Along with Dr Su, Keith is co-project head of the Vertex kinase programme. Dr Wilson received his PhD in structural biology from the University of Oregon and joined Vertex in 1992.*

Robert Mashal *holds the title of Program Executive for Multi-Drug Resistance and is responsible for clinical oncology for kinase inhibitors. Robert received his MD degree from Johns Hopkins University and was an instructor at the Harvard Medical School. Prior to joining Vertex in 1998, Dr Mashal worked at McKinsey & Company.*

Michael Mullican *is a Principal Investigator in the Department of Medicinal Chemistry and co-ordinates all chemistry activities on the Vertex kinase programme. Mike received his PhD in organic chemistry from the University of Kansas. Prior to joining Vertex in 1992, Dr Mullican worked at Parke-Davis Pharmaceuticals in Ann Arbor, Michigan.*

Mark Murcko *holds the position of Vice-President and Chief Technology Officer, and also chairs the Vertex Scientific Advisory Board. Mark received his PhD in organic chemistry from Yale. Prior to joining Vertex in 1990, Mark worked at Merck Sharpe and Dohme in West Point, Pennsylvania.*