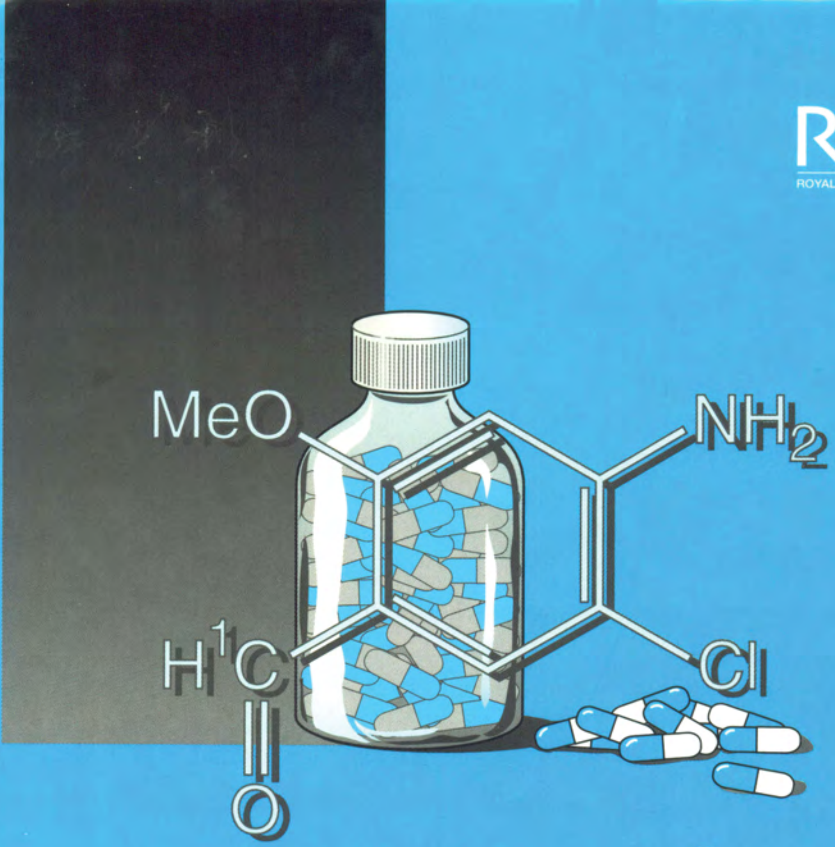
RS•C
ROYAL SOCIETY OF CHEMISTRY



drug design

CUTTING EDGE APPROACHES

edited by DARREN R. FLOWER

Drug Design

Cutting Edge Approaches

Drug Design
Cutting Edge Approaches

Edited by

Darren R. Flower
The Edward Jenner Institute for Vaccine Research, Newbury, UK

RS•C
ROYAL SOCIETY OF CHEMISTRY

The Proceedings of the meeting on Cutting Edge Approaches to Drug Design held on 13 March 2001 at the Scientific Societies Lecture Theatre, London, UK

Special Publication No. 279

ISBN 0-85404-816-2

A catalogue record for this book is available from the British Library

Published by The Royal Society of Chemistry,
Thomas Graham House, Science Park, Milton Road,
Cambridge CB4 0WF, UK

Registered Charity Number 207890

For further information see our web site at www.rsc.org

Typeset in Great Britain by Vision Typesetting, Manchester
Printed and bound by Athenaeum Press Ltd, Gateshead, Tyne & Wear

Preface

The application of computational sciences to pharmaceutical research is a discipline whose time has come. A tranche of techniques, both old and new, have recently matured into potent weapons in the war against disease. Molecular informatics – computational chemistry or molecular modelling, bioinformatics, and cheminformatics – has reached new heights of sophistication and utilitarian value within drug discovery. As an initiative to further foster and disseminate understanding of molecular informatics within the wider pre-clinical research environment, the organising committees of the Biological and Medicinal Chemistry Sector (BMCS) and the Molecular Modelling Group (MMG) of the Industrial Affairs Division of the Royal Society of Chemistry (RSC) inaugurated a series of one day meetings to address the subject. Highly technical, highly specific meetings that cover certain methodological aspects of the discipline are quite common, but we felt need for a broader and more accessible kind of conference that would serve as a gentle introduction to cutting edge approaches to drug design. This book is the proceedings of our first such meeting.

The pharmaceutical industry is a hugely profitable global business: the total annual worldwide sales for all human therapeutic drugs is about \$350 billion, while the farm livestock health market is worth about \$18 billion and the annual sales for the companion animal health market is approximately \$3 billion. To put these huge numbers into context: \$350 billion is comparable to the yearly gross national product of Taiwan, the Netherlands, or Los Angeles County. Drug sales are increasing at about 5% a year, while the vaccine market, currently worth a modest \$5 billion a year, is increasing at about 12% per annum. The annual global investment in R&D is around \$30 billion, up from \$2 billion in 1980. As a proportion of sales, average R&D expenditure has risen from 11.4% in 1970 to 18.5% in 2001. The merged GlaxoSmithKline is, at least in terms of market capitalisation, now amongst the top few largest companies in the world, yet controls less than 10% of the pharmaceutical market.

The structure of the global pharmaceutical market is highly biased. Over 50% of all marketed drugs target G-Protein coupled receptors. This includes a quarter of the 100 top-selling drugs, which generate sales of over $16 billion per year. Many of the top one hundred GPCR targeted drugs are so-called blockbusters each earning over $1 billion dollars a year. The biggest sellers have, however, been anti-ulcer drugs that have dominated the market place for most of the last 25 years. SmithKline Beecham's Tagamet, launched in 1977, was followed by Glaxo's Zantac (launched 1983), followed by Astra's proton pump inhibitor Losec, whose global sales peaked at $6.2 billion. Putting aside these blockbusters the 'average' drug struggles to recoup its development costs. Indeed, two out of three marketed drugs fail to yield a positive return on investment.

After a long relaxed period of sustained profitability, the industry now faces a drive towards increased efficiency. The emphasis is now firmly on shortening time-to-market, yet tightening by regulatory bodies has increased the time it takes to approve each new NCE: 19 months in 2001 up from only 13.5 months in 1998. Estimates of the attrition rate within pharmaceutical R&D varies widely: figures quoted lie somewhere between 0.25 and 0.001 depending how one does the calculation. Only about 1 in 12 compounds in development reaches the market. Competition has also increased dramatically and so has the concomitant rate of mergers and acquisitions. Should current trends continue, within 5 to 10 years five companies will control about 80% of the pharmaceutical market.

Yet 40% of human disease remains incurable and many existing therapies are far from ideal. The nature of illness has, at least in the West, changed out of all recognition over the last century, and can be expected to do so again during the next hundred years. Thus the challenge to modern medicine, of which the pharmaceutical industry is a key component, has never been greater, yet neither has the technology available to address it. The post-genomic revolution – genomics, transcriptomics, and proteomics – compounded by High Throughput Screening, and the coming revolution of lab-on-a-chip super-synthesis, will deliver an unprecedented information explosion. It is only through informatic strategies that we will be able to manage and fully exploit this data overload.

The first Cutting Edge Approaches to drug design was held on March 12 2001. The meeting opened with a barn-storming performance by one of the big beasts of structure based drug design: Professor Sir Tom Blundell. This was followed by a talk by Dr Jon Terrett of Oxford Glycosystems, deputising for Dr Andy Lyall, OGS's Head of Informatics. Dr Darren Green, of GlaxoSmithKline, spoke next on the subject of Virtual Screening, followed by Dr Dave Brown of Pfizer, who described the role of X-ray crystallography in drug design. In the afternoon, we had a talk by Dr Iain McLay from GlaxoSmithKline on lead optimisation methods followed by Dr Andy Davis talking about the resurgent role of Physical Organic Chemistry in drug discovery. The day was finished off by three talks detailing applications of informatic strategies: Dr Pascal Furet (Novartis) discussed Kinase inhibitors, and Dr Peter Hunt (Merck, Sharp & Dohme) & Dr Frank Blaney (GlaxoSmithKline) discussed drug design problems in G-protein coupled receptor research.

Before we came to put these proceedings together, Dr Furet declined to contribute. Later, it became apparent that, that for various reasons, Dr Terrett, Dr Brown and Dr Blaney would also be unable to contribute to the writing of this book. In order to compensate for this, I prevailed on Professor Teresa K Attwood, incipient grand dame of British bioinformatics, to help me describe the importance of integrated bioinformatics within G-protein coupled receptor research target discovery. My own group contributed a review of an exciting development in drug discovery research: the application of computational methods to the design of vaccines. I have also included a introductory chapter, which, apart from plumbing the depths of my own ignorance, attempts to put the other chapters into some kind of context, while trying to introduce some of the concepts that will be explained later in more detail. In writing these proceedings, we have tried to stay close to the ideal of the original meeting by attempting to balance technical accuracy with accessibility and readability for the non-specialist. Readers can judge for themselves if we succeeded.

Thanks are, of course, due to all the speakers, and their co-authors, for their astounding and outstanding efforts. I should also like to extend my thanks to the other organisers of the meeting: Dr Nicola Aston (GlaxoSmithKline, Chair), Dr Terry Hart (Novartis), both representing the BMCS, and Dr Chris Snell (Novartis). Of course, the meeting itself could not have happened without the organisational brilliance of Elaine Wellingham, to whom ineluctable thanks are due. I should also like to thank Alan Cubitt, Janet Freshwater, and the rest of the staff of RSC books, without whose help this excellent tome would never have seen the light of day.

As we have said, Cutting Edge Approaches to Drug Design (CEAtoDD) was the first of an on-going series of one-day lectures. We have already held CEAtoDD II and are planning CEAtoDD III, which will be held on March 2003. For up to date information, please visit the web-site for these meetings (currently at URL: http://www.jenner.ac.uk/CEAtoDD/CEAtDD.htm). Alternatively, visit the Molecular Modelling Group web page (URL: http://www.rsc.org/lap/rsccom/dab/ind006.htm).

Dr Darren R Flower
The Edward Jenner Institute for Vaccine Research

Contents

Molecular Informatics: Sharpening Drug Design's Cutting Edge

Darren R. Flower

EDWARD JENNER INSTITUTE FOR VACCINE RESEARCH,
COMPTON, BERKSHIRE RG20 7NN, UK

1 Introduction

The word 'drug', which derives from the Middle English word '*drogge*', first appears in the English language during the 14th century and it has, at least during the last century, become, arguably, one of the most used, and misused, of words, becoming tainted by connotations of misuse and abuse. The dictionary definition of a drug is: 'a substance used medicinally or in the preparation of a medicine. A substance described by an official formulary or pharmacopoeia. A substance used in the diagnosis, treatment, mitigation, cure, or other prevention of disease. A non-food substance used to affect bodily function or structure.' Even within the pharmaceutical industry, possessed, as it is, by a great concentration of intellectual focus, the word has come, in a discipline-dependent way, to mean different things to different people. To a chemist a drug is a substance with a defined molecular structure and attributed activity in a biological screen or set of screens. To a pharmacologist a drug is primarily an agent of action, within a biological system, but typically without a structural identity. To a patent lawyer it is an object of litigious disputation. To a marketing manager it is foremost a way to make money. To a patient – the pharmaceutical industry's ultimate end-user – a drug is possibly the difference between life and death.

Unmet medical need is, then, a constant stimulus to the discovery of new medicines, be they small molecule drugs, therapeutic antibodies, or vaccines. This unmet need has many diverse sources, including both life-threatening conditions – such as arise from infectious, genetic, or autoimmune disease – and other conditions that impinge deleteriously upon quality of life. The division between the causes of disease is seldom clear cut. Genetic diseases, for example, can be roughly divided between those resulting from Mendelian and multifactorial inheritance. In a Mendelian condition, changes in the observed phenotype arise from mutations in a single dominant copy of a gene or in both recessive copies. Multifactorial inheritance arises from mutations in many different genes, often with a significant environmental contribution. The search for genes causing Mendelian disorders has often been spectacularly successful. Multifactorial diseases, on the other hand, have rarely yielded identifiable susceptibility genes. The identification of NOD2 as causative component for Crohn's disease[1] has been hailed as a major technical breakthrough, leading, or so it is hoped, to a flood of susceptibility genes for multifactorial diseases. Unfortunately, the mode of inheritance in many multifactorial diseases is probably so complex that the subtle interplay of genes, modifier genes, and causative multiple mutations, which may be required for an altered phenotype to be observed, will, for some time yet, defy straightforward deduction.

Heart disease, diabetes, and asthma are all good examples of multifactorial disroders. Asthma, in particular, is, arguably, one of the best exemplars of the complex influence of environmental factors on personal wellbeing. It is a major health care problem affecting all ages, although it is not clear if the disease is a single clinical entity or a grouping of separate clinical syndromes. Asthma is a type I, or atopic, allergic disease, as contrasted with type II (cytotoxic), type III (complex immune), or type IV (delayed type). The word 'asthma', like the word 'drug', first appears in English during the 14th century. It derives from the Middle English word *asma*: a Medieval borrowing from Latin and Greek originals, although the incidence of allergic disease has been known since ancient times.[2,3] It is a condition marked by paroxysmal or laboured breathing accompanied by wheezing, by constriction of the chest, and attacks of gasping or coughing. It is generally agreed, that, over the past half-century, the prevalence of asthma, and type I allergies in general, particularly in western countries, has increased significantly. The reasons for this are complex, and not yet fully understood. Clearly, improvements in detection will have made a significant contribution to the increased apparent incidence of asthma, and other allergies, as is seen in many other kinds of condition, although this will only make a partial contribution to the overall increase. Other causative factors include genetic susceptibility; increased allergen exposure and environmental pollution; underlying disease; decreased stimulation of the immune system (the so-called hygiene or jungle hypothesis); and complex psycho-social influences. This final class includes a rich and interesting mix of diverse suggested causes, including the increasing age of first time parents, decreased family size, increased psychological stress, the increase in smoking amongst young women, decreases in the activity of the young, and changes in house design. The last of these, which includes increased use of secondary or double glazing, central heating, and fitted carpets has led to a concomitant increase in the population of house dust mites such as *Dermatophagoides farinae* and *Dermatophagoides pteronyssinus*, which are believed to be key sources of indoor inhaled aero-allergens.

Amongst the rich, developed countries of the first world – the pharmaceutical industry's principal target population – some of the most pressing medical needs are, or would seem to be, a consequential by-product of our increasingly technologized, increasingly urbanized personal lifestyles. These include diseases of addiction or over-consumption, those that characterize the West's ageing population, and those contingent upon subtle changes in our physical environment.

Certain diseases have increased in prevalence, while the major killers of preceding centuries – infectious diseases – have greatly diminished in the face of antibiotics, mass vaccination strategies, and improvements in hygiene and public health. In 1900, the primary causes of human mortality were influenza, enteritis, diarrhoea, and pneumonia, accounting between them for over 30% of deaths. Together, cancer and heart disease were responsible for only 12% of deaths. Today, the picture is radically different, with infectious disease accounting for a nugatory fraction of total mortality. Chronic diseases – the so-called 'civilization diseases' – account, by contrast, for over 60% of all deaths.

Many of these diseases, and indeed many other diseases *per se*, are preventable, and the development of long-term prophylactics, which may be taken over decades by otherwise healthy individuals, is a major avenue for future pharmaceutical exploration. Hand in hand with the newly emergent discipline of pharmacogenetics, the development of prophylactics offers many exciting opportunities for the active prevention of future disease. As Benjamin Franklin inscribed in *Poor Richard* in 1735: 'An ounce of prevention is worth a pound of cure'. However, for drugs of this type, problems common in extant drugs will be greatly magnified. 'Show me a drug without side effects and you are showing me a placebo,' a former chair of the UK's committee on drug safety once commented. As pharmaceutical products, of which Viagra is the clearest example, are treated more and more as part of a patient's lifestyle, the importance of side effects is likely to grow. A recent study concluded that over 2 million Americans become seriously ill every year, and over 100,000 actually die, because of adverse reactions to prescribed medications. A serious side effect in an ill patient is one thing, but one in a healthy person is potentially catastrophic in an increasingly competitive market place. If the industry is able to convince large sections of the population that it has products capable of preventing or significantly delaying the onset of disease, then financially, at least, the potential market is huge. Whether such persuasion is possible, and who would bear the cost of this endeavour, only time will tell.

Important amongst civilization diseases are examples that arise from addiction and over-consumption. While obesity undoubtedly has a genetic component, it also results from a social phenomenon, with a significant voluntary component, related in part to improvements in the quality and availability of food. Likewise, diseases relating to the addiction to drugs of misuse (tobacco, alcohol, and other illegal drugs, such as heroin or cocaine) give rise to both direct effects – the addiction itself – and dependent pathological impairment, such as lung cancer or heart disease. There is a need to intervene both to address and to mitigate the behaviour itself, primarily through direct drug treatment, with or without psychological counselling, and to address its resulting harmful physiological by-products. Caring for these consequent phenomena has now becoming a major burden on health services worldwide. As individuals, people find dieting difficult and giving-up strongly addictive substances, such as tobacco, even more difficult; pharmaceutical companies are now beginning to invest heavily in the development of anti-obesity drugs and nicotine patches, *inter alia*, as an aid to this endeavour. For example, the appetite supressant anti-obesity drug Reductil or Sibutramine – a serotonin, norepinephrine, a dopamine reuptake inhibitor – has recently been licensed by the National Institute for Clinical Excellence in the UK. Vaccines are also being developed to alter the behavioural effects of addictive drugs such as nicotine and cocaine.[4,5] Xenova's therapeutic vaccine TA-NIC, a treatment for nicotine addiction, has recently entered Phase I clinical trials to test the safety, tolerability and immunogenicity of the vaccine in both smokers and non-smokers. TA-NIC is thought to be the first anti-nicotine addiction vaccine to be clinically tested. Other therapies for nicotine addiction include skin patch nicotine replacement, nicotine inhalers or chewing gum, or treatment with the nicotine-free drug Bupropion. A Xenova anti-cocaine addiction vaccine, TA-CD, is currently in Phase II clinical development. We shall see, as time passes, that this type of direct pharmaceutical intervention, targeting the process of addiction rather than just treating its outcome, will doubtlessly increase in prevalence.

Box 1 *A Global Plague*

From its original introduction into Europe at the close of the 15th century, partly as a treatment for disease, the success of tobacco as a recreational drug has been astounding. Today, smoking can be justly called a global plague. It is the number one cause of respiratory disease and the single most preventable cause of death in the industrialised west. Some estimates indicate that worldwide smoking leads to more deaths per annum than AIDS, alcohol, car accidents, homicide, and suicide. Current figures would suggest that approximately 1 in 6 people in the world smoke: about 1.1 billion smokers out of a total of 6.0 billion. Of these, 50% will die prematurely from tobacco-related illness. Half will die in middle age with an average loss of life expectancy of 20–25 years. This means that in excess of 500 million, or about 10% of the existing population, will die from smoking related diseases: 27% from lung cancer, 24% from heart disease, 23% from chronic lung diseases, such as emphysema. The remaining 26% will die from other diseases including other circulatory disease (18%) and diverse other cancers (8%). Although its incidence amongst men has slowly decreased since the late 1980s, lung cancer remains the most prevalent cause of cancer deaths in the USA causing approximately 85% of bronchogenic carcinoma. It remains a deadly disease with 5 yr survival rates of only 14%. Approximately 17 million smokers in the USA alone attempt to quit each year.

In the First World, approximately one third of all people aged fifteen years and up smoke, with the percentage increasing sharply in Asia, Eastern Europe and the former Soviet States. Consumption trends indicate that smoking prevalence is reducing in developed countries (down 1.5% per annum in the United States, for example) while increasing in less developed countries (up 1.7% per annum). Based on current trends, the World Health Organisation estimates the death toll from smoking will rise to 10 million people per year by 2025. Currently two million deaths occur each year in developed countries and 1 million deaths occur each year in less developed countries. By 2025, this ratio will alter to 3 million deaths per year in developed countries and 7 million deaths per year in less developed countries. In 1950, 80% of the men and 40% of the women in Britain smoked, and tobacco deaths were increasing rapidly. There have been 6 million deaths from tobacco in Britain over the past 50 years, of which 3 million were deaths in middle age (35–69). There are still 10

Continued on p. 5

million smokers in Britain, of which about 5 million will be killed by tobacco if they don't stop. World-wide, there were about 100 million tobacco deaths in the 20th century, but if current smoking patterns continue there will be about 1 billion in the 21st century. The harmful effects of smoking have been well understood since at least the middle of the 19th century[6–10] but it was only with the solid epidemiological evidence of Richard Doll in 1950 that the link between smoking and lung cancer firmly and finally established. In the following fifty years, the links between smoking and innumerable other diseases have become clear.

In passing, we might mention that so-called diseases of over-consumption are only recalling some of the environmental disease effects prevalent in earlier ages. Heavy smoking has similar effects on the lungs to the conditions experienced by people living in countries in the cold northern climes during earlier eras. For example, dwellers in Iron Age roundhouses, Anglo Saxon and early Medieval great halls lived in large communal environments, within these domestic settings, and contended continually with large open fires, creating a high particulate atmosphere. The physiological effects of such exposure would recreate those of a heavy smoker. Yet, in other some respects their health was surprisingly good, their diet compensating, at least in part, for other factors. For example, meat – in the form of beef, mutton, and pork – was the principal component of the Anglo Saxon diet. Meat, obtained from lean, free range animals, contained, in those times, three times as much protein as saturated, and thus cholesterol bearing, fat; a ratio reversed in modern factory farmed animals. Height is often taken as an indicator of the efficacy of diet, and the Anglo Saxons were, unlike, say, the diminutive Georgians or Victorians, as tall, at least as a population, as people at the beginning of the 21st century.

The ageing population apparent in western countries is, amongst other causes, a by-product of the increased physical safety of our evermore comfortable, urbanized, post-industrial environment. Together with decades of enhanced nutrition and the effects of direct medical advancement in both medicines and treatment regimes, this has allowed many more people to exploit their individual genetic predisposition to long life. Estimates based on extant demographic changes would suggest that by 2050 the number of the super-old, *i.e.* those living in excess of 100 years, would, within the USA, be well in excess of 100,000. In terms of its implications for drug discovery, this has led to a refocusing of the attention of pharmaceutical companies onto gerantopharmacology and the diseases of old age. Examples of these include hitherto rare, or poorly understood, neurodegenerative diseases, such as Parkinson's disease, or those conditions acting *via* protein misfolding mechanisms, which proportionally affect the old more, such as Alzheimer's disease. The prevalence of stroke is also increasing: approximately 60,000 people die as the result of a stroke annually in England and Wales and approximately 100,000 suffer a non-fatal first stroke. However, the relative proportion of young people suffering a stroke has also increased. Here, 'young' refers to anyone under 65, but stroke is not unknown in people very much younger, including infants and children. Indeed, 250 children a year suffer a stroke in the United Kingdom. This disquieting phenomenon may, in the era of routine MRI scans, simply reflect the greater ease of successful detection amongst the young as well as the old.

Looking more globally – though the danger is still real in developed western countries – new or re-emergent infectious diseases, such as AIDS or tuberculosis, pose a growing threat, not least from those microbes exhibiting drug and antibiotic resistance. As the world appears to warm, with weather patterns altering and growing more unpredictable, the geographical spread of many tropical infectious diseases is also changing, expanding to include many areas previously too temperate to sustain these diseases. The threat from infectious disease, which we have seen has been largely absent for the last 50 years, is poised to return, bringing with it the need to develop powerful new approaches to the process of anti-microbial drug discovery.

From the foregoing discussion, we can identify a large array of new, or returning, causes of human disease, which combine to generate many accelerating and diversifying causes of medical need. These come from infectious disease, which have evolved, with or without help from human society, to exhibit pathogenicity, but also from diseases of our own creation, such as those resulting directly, or indirectly, from addiction or substance abuse, to other disease conditions, which have not previously been recognized, or have not been sufficiently prevalent, due to our ageing population or changing economic demographics. Patterns of disease have changed over the past hundred years and will change again in the next hundred. Some of these changes will be predictable, others not. Medical need is ever changing and is always at least one step ahead of us. Thus the challenge to medicine, and particularly the pharmaceutical industry, has never been greater, yet neither has the array of advanced technology available to confront this challenge. Part of this is experimental: genomics, proteomics, high throughput screening (HTS), *etc*., and part is based on informatics: molecular modelling, bioinformatics, cheminformatics, and knowledge management.

2 Finding the Drugs. Finding the Targets

Within the pharmaceutical industry, the discovery of novel marketable drugs is the ultimate fountainhead of sustainable profitability. The discovery of candidate drugs has typically begun with initial lead compounds and then progresses through a process of optimization familiar from many decades of medicinal chemistry. But before a new drug can be developed, one needs to find the targets of drug action, be that a cell-surface receptor, enzyme, binding protein, or other kind of protein or nucleic acid. This is the province of bioinformatics.

2.1 Bioinformatics

Bioinformatics, as a word if not as a discipline, has been around for about a decade, and as a word it tends to mean very different things in different contexts. A simple, straightforward definition for the discipline is not readily forthcoming. It seeks to develop computer databases and algorithms for the purpose of speeding up, simplifying, and generally enhancing research in molecular biology, but within this the type and nature of different bioinformatic activity varies widely. Operating at the level of protein and nucleic acid primary sequences, bioinformatics is a branch of information science handling medical, genomic and biological information for support of both clinical and more basic research. It deals with the similarity between macromolecular sequences, allowing for the identification of genes descended from a common ancestor, which share a corresponding structural and functional propinquity.

Box 2 *What is Bioinformatics*?

Bioinformatics is one of the great early success stories of the incipient informatics revolution sweeping through the physical sciences. Bioinformaticians find themselves highly employable: indeed many eminent computational biologists have had to re-badge themselves with this particular epithet. Their services are much in demand by biologists of most, but not yet all, flavours. But what is bioinformatics? One definition is 'Conceptualizing biology in terms of molecules (in the sense of physical chemistry) and then applying 'informatics' techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to understand and organize the information associated with these molecules, on a large scale'. A more tractable definition than this, which seems more uninterpretable than all embracing, is 'the application of informatics methods to biological molecules'. Many other areas of computational biology would like to come under the bioinformatics umbrella and thus get ready access to grant funding, but the discipline is still mostly focused on the analysis of molecular sequence and structure data.

Bioinformatics, as do most areas of science, relies on many other disciplines, both as a source of techniques and as a source of data (see Figure 1). Bioinformatics also forms synergistic links with other areas of biology, most notably genomics, as both vendor and consumer. In the high throughput post-genomic era, bioinformatics feeds upon these data rich disciplines but also provides vital services for data interpretation and management, allowing biologists to come to terms with this deluge rather than being swamped by it. It is still true that bioinformatics is, by and large, concerned with data handling: the annotation of databases of macromolecular sequences and structures, for example, or the classification of sequences or structures into coherent groups. Prediction, as well as analysis, is also important, not least in trying to address two of the key challenges of the discipline: the prediction of function from sequence and the prediction of structure from sequence (see Figure 2). Although these two are intimately linked, there is nonetheless still an important conceptual difference between them. One can discern three main areas within the traditional core of bioinformatics: one dealing with nucleic acid sequences, one with protein sequences, and one with macromolecular structures (see Figure 3).

At the very heart of bioinformatics is the Multiple Sequence Alignment (see Figure 4). With it, one can do so much: predict 3D structure, either through homology modelling or *via de novo* structure; identify functionally important residues; undertake phylogenetic analysis; and identify important motifs and thus develop discriminators for the membership of protein families. The definition of a protein family, the key step in annotating macromolecular sequences, proceeds through an iterative process of searching sequence, structure, and motif databases to generate a sequence

Continued on p. 8

corpus, which represents the whole set of sequences within the family (see Figure 5). Motif databases, of which there are many, contain distilled descriptions of protein families that can be used to classify other sequences in an automated fashion. There are many ways to characterize motifs: through human inspection of sequence patterns, by using software to extract motifs from a multiple alignment, or using a program like MEME to generate motifs directly from a set of unaligned sequences. A motif or, more likely, a set of motifs defining the family can then be deposited in one of the many primary motif databases, such as PRINTS, or secondary, or derived, motif database, such as INTERPRO (see Figure 6). This brief digression into the nature of bioinformatics has been very much a simplification, as is readily seen in Figure 7, which shows some of the greater complexity that is apparent in a less drug design-orientated view of the discipline.

Within the drug discovery arena bioinformatics equates to the discovery of novel drug targets from genomic and proteomic information. Part of this comes from gene finding: the relatively straightforward searching, at least conceptually if not always practically, of sequence databases for homologous sequences with, hopefully, similar functions and roles in disease states. Another, and increasingly important, role of bioinformatics is managing the information generated by micro-array experiments and proteomics, and drawing from it data on the gene products implicated in disease states. The key role of bioinformatics is, then, to transform large, if not vast, reservoirs of information into useful, and useable, information.

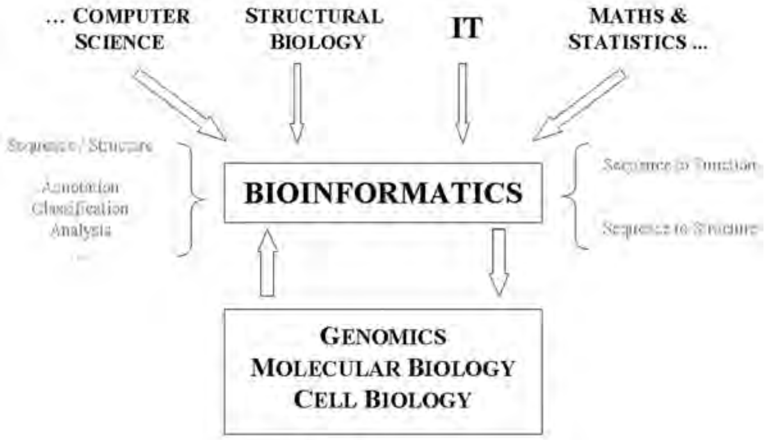


Figure 1 *Bioinformatics in its Place. Core bioinformatics makes a series of synergistic interactions with both a set of client disciplines (computer science, structural chemistry, etc.) and with customer disciplines, such as genomics, molecular biology, and cell biology. Bioinformatics is concerned with activities such as the annotation of biological data (genome sequences for example), classification of sequences and structures into meaningful groups, etc. and seeks to solve two main challenges: the prediction of function from sequence and the prediction of structure from sequence*

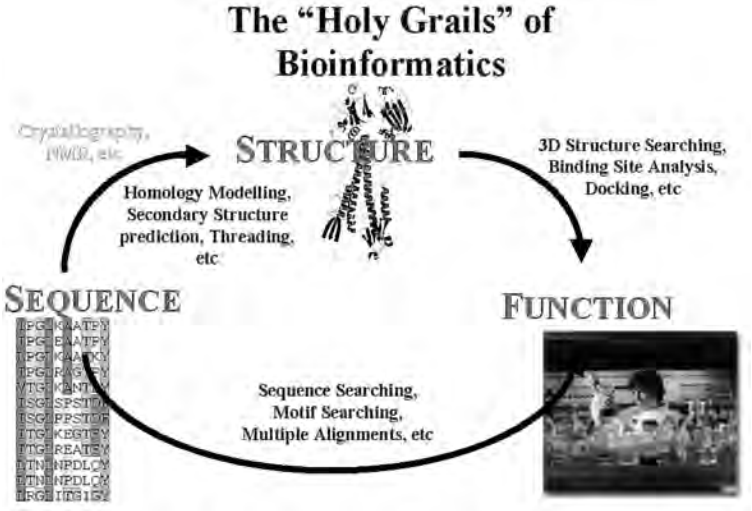

The "Holy Grails" of Bioinformatics
Crystallography, NMR, etc
STRUCTURE
3D Structure Searching, Binding Site Analysis, Docking, etc
Homology Modelling, Secondary Structure prediction, Threading, etc
SEQUENCE
FUNCTION
Sequence Searching, Motif Searching, Multiple Alignments, etc

Figure 2 *The 'Holy Grails' of Bioinformatics. Core bioinformatics seeks to solve two main challenges: the Holy Grails of the discipline. They are the prediction of Structure from Sequence, which may be attempted using secondary structure prediction, threading, or comparative modelling, and the prediction of Function from Sequence, which can be performed using global homology searches, motif databases searches, and the formation of multiple sequence alignments. It is also assumed that knowing a sequence's structure enables prediction of function. In reality, all methods for prediction of function rely on the identification of some form of similarity between sequences or between structures. When this is very high then some useful data is forthcoming, but as this similarity drops a conclusion one might draw becomes increasingly uncertain and even misleading*

Academic bioinformaticians sometimes seem to lose sight of their place as an intermediate taking, interpreting, and ultimately returning data from one experimental scientist to another. There is a need for bioinformatics to keep in close touch with wet lab biologists, servicing and supporting their needs, either directly or indirectly, rather than becoming obsessed with their own recondite or self referential concerns. Moreover, it is important to realize, and reflect upon, our own shortcomings. Central to the quest to achieve automated gene elucidation and characterization are pivotal concepts regarding the manifestation of protein function and the nature of sequence–structure and sequence–function relations. The use of computers to model these concepts is limited by our currently limited understanding, in a physico-chemical rather than phenomenological sense, of even simple biological processes. Understanding and accepting what cannot be done informs our appreciation of what can be done. In the absence of such an understanding, it is easy to be misled, as specious arguments are used to promulgate over-enthusiastic notions of what particular methods can achieve. The road ahead must be paved with caution and pragmatism, tempered, as ever, by the rigour for which the discipline is justly famous.

One of the most important recent trends has been the identification of so-called '*druggable*' receptors. As databases of nucleic acid and protein sequences