In silico research in drug discovery

Georg C. Terstappen and Angelo Reggiani

Target and lead discovery constitute the main components of today's early pharmaceutical research. The aim of target discovery is the identification and validation of suitable drug targets for therapeutic intervention, whereas lead discovery identifies novel chemical molecules that act on those targets. With the near completion of the human genome sequencing, bioinformatics has established itself as an essential tool in target discovery and the *in silico* analysis of gene expression and gene function are now an integral part of it, facilitating the selection of the most relevant targets for a disease under study. In lead discovery, advances in chemoinformatics have led to the design of compound libraries *in silico* that can be screened virtually. Moreover, computational methods are being developed to predict the drug-likeness of compounds. Thus, drug discovery is already on the road towards electronic R&D.

Historically, the discovery of novel drugs has been led by chemistry and pharmacology. With the advent of genomic sciences, however, biology has established itself as the main driver. Drug-discovery programs typically start with the identification of suitable drug targets (Fig. 1). Such targets are biomolecules that, in most cases, are proteins such as receptors, enzymes and ion channels. During the stepwise process of target validation, a sufficient level of 'confidence' has to be established that the target is of relevance to the disease under study and modulation of the target will lead to effective disease treatment. The initial steps of target validation are usually obtained *in vitro* and in animal models but the ultimate validation can be achieved only in clinical experiments in humans. After initial target validation has been obtained, modulators of the target have to be identified. Such modulators can be agonists or antagonists in the case of receptors, activators or inhibitors of enzymes, and openers or blockers of ion channels. This phase of so-called lead identification starts with the design and development of a suitable assay to monitor the target under study. Subsequently, high-throughput screening (HTS) exposes the target to a large number of chemical compounds (typically in the order of 10^5) that increasingly come from high-speed parallel and combinatorial synthesis[1]. Active compounds that demonstrate dose-dependent target modulation are called lead compounds when a certain degree of selectivity for the target under study can be shown and the first positive results in animal models are obtained. Such lead compounds are optimized in terms of potency and selectivity as well as physicochemical properties, and their pharmacokinetic and safety features are assessed before they can become candidates for drug development. Although most of the process of early pharmaceutical research relies predominantly on experimental work in the laboratory, the computer has become increasingly important. This review briefly describes the areas where *in silico* approaches are already operating in early pharmaceutical research and contribute significantly to drug discovery.

Georg C. Terstappen*
Angelo Reggiani
Biology Dept, GlaxoWellcome Medicines Research Centre, Via A. Fleming 4, 37135 Verona, Italy.
*e-mail: gct66554@glaxowellcome.co.uk

Target discovery, target identification and validation

One of the most important areas that *in silico* approaches are operating in is target discovery. Here, the cross-disciplinary science of bioinformatics has become essential. In 1996, a survey showed that there were 483 molecular targets of current therapies (45% receptors, 28% enzymes, 5% ion channels and 2% nuclear receptors)[2], but the sequencing of the human genome will soon be complete, revealing all potential targets for therapeutic intervention. A 'working draft' of the complete human genome sequence has already been announced (on 26 June 2000, by the *Human Genome Project* and *Celera Genomics*) and estimations for the total number of genes range from ~34 000 to 140 000 (Ref. 3). On the basis of bioinformatics analysis, successful target classes alone, such as receptors, enzymes and ion channels, can be predicted to amount to ~6500 (Fig. 2), which indicates the huge potential for target discovery. Not all of these biomolecules will become drug targets and the big challenge is to select the most relevant targets for a given disease.

In silico *gene-expression analysis*

Because only ~3% of the 3×10^9 bases of the human genome sequence actually encode proteins and *in silico* gene identification is still a difficult task[4], public and private expressed sequence tag (EST) databases represent an important source for target discovery. Such databases contain short sequence information from expressed genes, which allows their identification and which is taken as indicative of the encoded proteins[5]. The value of such databases has continuously been increased by adding sequences from different human tissues and states of development and disease: the public EST databases alone contain more than 1.6 million human sequences[6]. One important use of these databases in target discovery is to infer relative gene expression levels, simply by counting how often a given EST sequence appears in a given cell or tissue. Gene expression levels are important because the phenotype is determined by the small portion of genes that are expressed at any given time in a cell or tissue type, and changes in gene expression can be associated with disease. Thus, by comparing levels of gene expression in normal and disease states, novel drug targets can be identified *in silico*. Because these

http://tips.trends.com 0165-6147/01/$ – see front matter PII: S0165-6147(00)01584-4

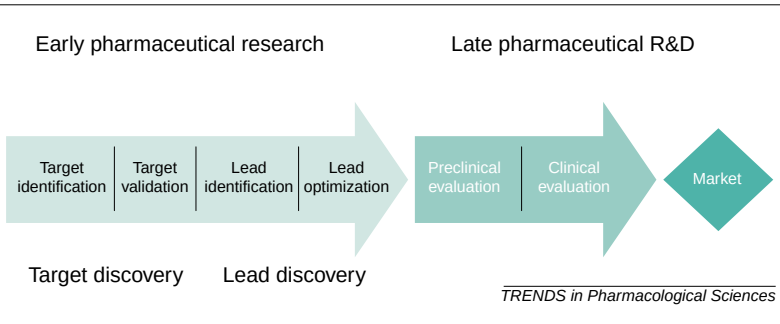

Fig. 1. The process of pharmaceutical R&D, often referred to as discovery process, can roughly be divided into an early and late phase. The early phase is mainly represented by target and lead discovery, whereas the later phase deals mainly with clinical evaluation and development.

databases contain only a limited number of sequences for every tissue or cell type (typically 5000–10 000), only moderate-to-highly expressed genes can be obtained. Several new *in vitro* technologies for high-throughput gene-expression analysis have been developed including serial analysis of gene expression (SAGE), variants of differential display and DNA microarrays (gene chips), which can overcome the limits of EST-based approaches[7]. However, most important is having a suitable database that can be queried effectively. *In silico* database searches can then identify genes that are up- or downregulated in a disease state, expressed in a particular tissue and associated with biological and biochemical pathways. The level of confidence in a target increases greatly if a disease-specific expression can be demonstrated and if the target is expressed in a tissue or cell type that is important for that disease. Routinely, data obtained *in silico* have to be confirmed in the laboratory.

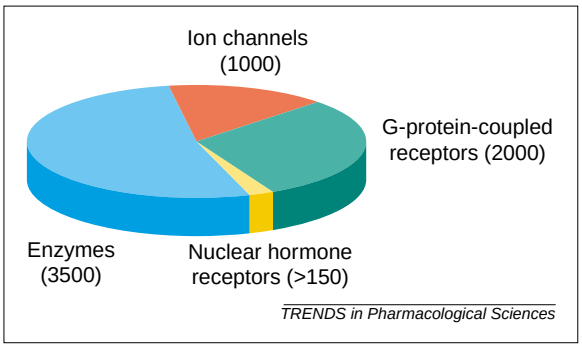

Fig. 2. Predicted numbers of potential drug targets belonging to different biochemical classes. The databases OMIM (Online Mendelian Inheritance in Man; http://www3.ncbi.nlm.nih.gov/Omim/), SWISS-PROT (http://www.expasy.ch/sprot/sprot-top.html), Incyte LifeSeq Gold Human Gene Database (http://www.incyte.com/products/lifeseq/lifeseqgold.html), GeneCards (human genes, proteins and disease; http://bioinformatics.weizmann.ac.il/cards/), TIGR Human Gene Index (http://www.tigr.org/tdb/hgi/hgi.html), and GDB (Genome Database; http://gdbwww.gdb.org/) were mined for the different target classes using a keyword search in each database. Multiple entries of the same gene within a database were counted only once to avoid internal redundancies. Results obtained in the different databases were compared with each other to avoid redundancies between different databases. To minimize annotation errors, entries annotated as 'similar to' were disregarded and the private Incyte and TIGR databases (which are supposed to contain more errors) were used only to confirm results obtained from the public databases. An average number of 10^5 human genes was used for calculations of the numbers for the different target classes. The indicated numbers of our analysis should be considered as a rough estimation.

Prediction of gene function

Elucidation of gene function *in silico* is another important field for bioinformatics in target discovery. Typically, a new DNA sequence [which could come from the above gene-expression analysis (e.g. a gene upregulated in a diseased tissue)] is first subjected to similarity searches [e.g. using basic local alignment search tool (blast)][8] in sequence databases such as the public GenBank (nucleic acid level) and SWISS–PROT (protein level), and often its function can be derived from similarities and homologies to sequences of known function (Box 1). In about 30–35% of cases, where no clear functional prediction is possible, several recently developed computational methods in comparative genomics can help to deduce specific functions[9,10]. Usually a target is much more attractive if its function can be demonstrated. Function can point to a biochemical and/or pathophysiological pathway that the target is involved in, thus shedding further light on its relevance for the disease under study. In addition, knowledge of the function is usually necessary for subsequent development of an assay to monitor the modulation of the target. If the target is found to belong to a highly 'tractable' structural class (such as receptors, enzymes or ion channels), its value increases even more because these target classes have demonstrated therapeutic utility. Databases can be mined directly for novel genes belonging to such 'tractable' structural classes. In this case there is no *a priori* information regarding disease relevance associated with the target. Further disease relevance of a novel target can be derived from its chromosomal localization using *in silico* polymerase chain reaction (PCR)[11]. If the target is mapped to or close to a locus that has been associated with the disease under study, the target itself might be associated with that disease.

Lead discovery, lead identification and optimization

Once sufficient 'confidence' in a target has been obtained, lead discovery usually starts with the development of a suitable assay to monitor the target and identify modulators (lead identification). Enormous advances in HTS have been made and ultraHTS (uHTS) allows the screening of 100 000 compounds on a target per day[12]; however, if less compounds could be tested without compromising the probability of success, the cost and time would be greatly reduced.

In silico *library design and virtual screening*

In silico (or virtual) compound library design operates to reduce the number of compounds to be tested, and two basic applications can be distinguished: diversity and structure-based design. Diversity design aims to select a smaller sub-library from a larger compound library in such a way that the full range of chemical diversity is best represented[13]. When no structural information about the target and/or target ligands is available, diversity design is the method of choice.

Box 1. Prediction of gene function exemplified by the identification of the human frizzled-3 (*FZD3*) seven-transmembrane-domain receptor[a]

The example below outlines the different steps of *in silico* analysis and the software tools that might be used. It should be kept in mind that a relatively easy case has been chosen to illustrate the procedure because the complete mouse sequence was already known.

A mouse sequence was used as the starting point for *in silico* analysis. The mouse frizzled-3 protein sequence (SWISS-PROT accession number Q61086) was employed as a query to search GenBank for human expressed sequence tags (ESThum subdivision) and human genomic sequences (HTGShum subdivision) using the TBLASTN tool (http://www.ncbi.nlm.nih.gov/blast). Homologous EST sequences were retrieved from the ESThum covering the full coding sequence for the predicted human frizzled-3 protein (FZD3) and two genomic sequences comprising the entire coding exon complement of the candidate *FZD3* locus were similarly retrieved. The pairwise alignments between the mouse protein and the human genomic sequences (dynamically translated in all six reading frames) were used to deduce the intron and exon positions on the genomic sequences. Splice site consensus sequence prediction was carried out using the Neural Network Splice Site Prediction Tool (NNSPLICE0.9; http://www.fruitfly.org/seq_tools/splice.html). In this way, the genomic structure of the human *FZD3* gene locus, which comprises six exons and five introns, was defined. The multiple alignment program CLUSTAL W (Ref. b) (http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html) was used for subsequent comparisons of the mouse and human deduced amino acid sequences. The human protein showed 98% identity with the mouse protein, thus identifying the novel sequence as the human homologue of the frizzled-3 receptor.

References

a Sala, C.F. *et al.* (2000) Identification, gene structure, and expression of human frizzled-3 (*FZD3*). *Biochem. Biophys. Res. Commun.* 273, 27–34

b Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680

The different computational methods for compound selection are mainly based on compound similarity clustering, grid-like partitioning of chemical space or the application of genetic algorithms[14]. The results of such *in silico* diversity selections (*in silico* screening) are smaller sub-libraries of manageable size with a high degree of chemical diversity that are then subjected to HTS *in vitro*.

Structure-based library design is biased by structural requirements for activity on a particular target and needs prior information of the target structure (e.g. X-ray or nuclear magnetic resonance). The goal is to select from existing compound libraries or to design compounds with three-dimensional complementarity (i.e. shape, size and physicochemical properties) to the target-binding site. In the latter case, new approaches can directly guide the design of virtual combinatorial libraries, which are first screened *in silico* for target complementarity, thus reducing the number of compounds that will have to be synthesized and tested *in vitro*. The combination of structure-based design and combinatorial chemistry, which is also called 'combinatorial docking'[15], and the different computational tools and methods have been reviewed elsewhere[16]. It can be expected that the hit-rate (rate of compounds found to be active on the target under study in a dose-dependent manner) of such focused libraries will be higher than that of diversity screening. Indeed, a recent example using the protease cathepsin D showed the hit-rate of the focused approach to be more than twice as high as that of diversity screening[17]. Unfortunately, most X-ray structures available are for enzymes because membrane proteins such as receptors and ion channels are exceedingly difficult to crystallize. For such targets, a focused approach can still be employed if ligands (natural or synthetic) are known for the actual target under study or for the respective target class. For example, if a common pharmacophore model can be derived from this information, virtual screening of compound libraries is performed to define a subset of biased compounds that is then subjected to HTS *in vitro*. Alternatively, *in silico* screening of the pharmacophore against virtual libraries is carried out and interesting compounds are synthesized for HTS using combinatorial chemistry. Both diversity and structure-based screening can be performed in an iterative manner. In this case, the results of *in vitro* HTS are analysed *in silico* [e.g. using programs such as SCAM (statistical classification of the activities of molecules)][18] to derive rules that can be used for the rational selection of further molecules to be tested *in vitro*.

Prediction of drug-likeness

When lead molecules have been identified, they have to be optimized in terms of potency, selectivity, pharmacokinetics (i.e. absorption, distribution, metabolism and excretion (ADME)] and toxicology before they can become candidates for drug development. Because the high overall attrition rate in drug discovery is caused mostly by the non-'drug-likeness' of the compounds identified[19], the early analysis in this respect is becoming common practice. *In silico* approaches to predict pharmacokinetic parameters (ADME) were pioneered by Lipinski *et al.*[20] By studying the physicochemical properties of >2000 drugs from the WDI (World Drug Index, Derwent Information, London), which can be assumed to have entered Phase II human clinical trials (and therefore must possess drug-like properties), the so-called 'rule-of-five' was derived to predict oral bioavailability (intestinal absorption) of a compound that can be considered as the major goal of drug development. If the hydrogen bond donors are <5, hydrogen acceptors <10, relative molecular weight <500 and lipophilicity (logP)<5, the compound will probably be orally bioavailable. Additional methods that need further validation have recently been reviewed[21]. For compounds targeted to the CNS, another important aspect is blood–brain barrier (BBB) penetration. On the basis of predictive models described by Abraham *et al.*[22], a simple two-variable equation has been devised that allows rapid automated *in silico* screening of (virtual) libraries for compounds with a potential to cross the BBB (Ref. 23). The equation is based on the polar surface

Box 2. Future directions

In target discovery, genetics and polymorphism analysis will have a major impact. A dense map and database of single nucleotide polymorphisms (SNPs) is being developed by public and private efforts that should allow easier disease-association studies[a,b]. Bioinformatics will be crucial for data analysis, which is expected to lead to the identification of susceptibility genes that have potential as drug targets, and to contribute to pharmacogenomics. The latter attempts to identify the genetic determinants of different drug responses across a population, to allow the development of 'individualized therapy'[a]. In addition to the benefits that can be expected for patients (e.g. fewer or no side-effects and a better response to the drug treatment), such drug therapy based on the individual patient's genetic profile might lead to a drastic reduction of cost and time in clinical trials. As genome sequences are being completed and annotated in the 'post-genomic' era, the field of proteomics is becoming increasingly important. *In silico* analysis will be crucial to extract value, especially in the light of recent advances in protein microarrays[c].

In lead discovery, major efforts are being devoted to further refine the prediction of drug-like molecules and to predict pharmacokinetic and toxicological properties of compounds. The latter will also be based on data obtained from large-scale gene-expression analysis using DNA microarrays (toxicogenomics)[d]. For example, a particular expression pattern induced by a novel compound can be compared with patterns obtained with known toxic compounds, which might lead to specific toxicity predictions. Such predictions should lead to improved selections of combinatorial and high-throughput screening (HTS) libraries, thus increasing the probability of success in drug discovery.

References

a McCarthy, J.J. and Hilfiker, R. (2000) The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat. Biotechnol.* 18, 505–508

b Marshall, E. (1999) Drug firms to create public database of genetic mutations. *Science* 284, 406–407

c Walter, G. *et al.* (2000) Protein arrays for gene expression and molecular interaction screening. *Curr. Opin. Microbiol.* 3, 298–302

d Nuwaysir, E.F. *et al.* (1999) Microarrays and toxicology: the advent of toxicogenomics. *Mol. Carcinog.* 24, 153–159

area [PSA (a measure of the solvation potential of the polar groups)] and the lipophilicity of a compound. Predicting the toxicity of compounds, another important aspect, has been reviewed elsewhere[24]. Sufficient high-quality and reliable data are not yet available; thus the predictive ability of the underlying models is limited and needs further development. In addition to their use in the lead optimization phase, the computational techniques described can be used early on to select a subset of compounds for screening or to guide combinatorial library design.

Although the virtual library design and virtual screening approaches described above are integral to today's lead discovery, in many cases the whole corporate compound collection (which might contain millions of chemical compounds) is used in HTS on a given target. Whether *in silico* approaches improve the identification of high-quality lead compounds and ultimately the delivery of new drugs to the market has yet to be evaluated.

Concluding remarks

In silico approaches contribute significantly to early pharmaceutical research and are especially important in target and lead discovery. It can be anticipated that this contribution will increase substantially in the near future (Box 2). The need for timely adaptation and application of *in silico* approaches in pharmaceutical research has clearly been recognized and is expected to improve further the overall efficiency of drug discovery.

Acknowledgements
We thank A.M. Capelli (Chemistry Dept) and F. Caldara (Dept Biology) for helpful discussions and comments on the manuscript.

References

1 Balkenhohl, F. *et al.* (1996) Combinatorial synthesis of small organic molecules. *Angew. Chem. Int. Ed. Engl.* 35, 2288–2337

2 Drews, J. (2000) Drug discovery: a historical perspective. *Science* 287, 1960–1964

3 Aparicio, S.A.J.R. (2000) How to count human genes. *Nat. Genet.* 25, 129–130

4 Lewis, S. *et al.* (2000) Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.* 10, 349–354

5 Marra, M.A. *et al.* (1998) Expressed sequence tags – establishing bridges between genomes. *Trends Genet.* 14, 4–7

6 Schultz, J. *et al.* (2000) More than 1,000 putative new human signalling proteins revealed by EST data mining. *Nat. Genet.* 25, 201–204

7 Lennon, G.G. (2000) High-throughput gene expression analysis for drug discovery. *Drug Discov. Today* 5, 59–66

8 Altschul, S.F. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

9 Galperin, M.Y. and Koonin, E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotech.*18, 609–613

10 Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* 10, 359–365

11 Schuler, G.D. (1998) Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.* 16, 456–459

12 Roberts, B.R. (2000) Screening informatics: adding value with meta-data structures and visualization tools. *Drug Discov. Today* (HTS Suppl. 1), 10–14

13 Gorse, D. and Lahana, R. (2000) Functional diversity of compound libraries. *Curr. Opin. Chem. Biol.* 4, 287–294

14 van Drie, J.H. and Lajiness, M.S. (1998) Approaches to virtual library design. *Drug Discov. Today* 3, 274–283

15 Boehm, H.J. and Stahl, M. (2000) Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* 4, 283–286

16 Li, J. *et al.* (1998) Targeted molecular diversity in drug discovery: integration of structure-based design and combinatorial chemistry. *Drug Discov. Today* 3, 105–112

17 Kick, E.K. *et al.* (1997) Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol.* 4, 297–307

18 Rusinko, A., III (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* 39, 1017–1026

19 Darvas, F. *et al.* (2000) Diversity measures for enhancing ADME admissibility of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 40, 314–322

20 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

21 Walters, W.P. *et al.* (1999) Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* 3, 384–387

22 Abraham, M.H. *et al.* (1994) Hydrogen bonding. 33 factors that influence the distribution of solutes between blood and brain. *J. Pharm. Sci.* 83, 1257–1268

23 Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of drug-likeness. *Drug Discov. Today* 5, 49–58

24 Cronin, M.T.D. (1998) Computer-aided prediction of drug toxicity in high throughput screening. *Pharm. Pharmacol. Commun.* 4, 157–163