Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics

Jürgen Bajorath

Over the past few years, bio- and chemo-informatics have rapidly evolved as related yet distinct disciplines. In drug discovery, it is increasingly recognized that combining and integrating these approaches is crucial for their successful application. In addition, the use of complementary experimental and informatics techniques increases the chances of success in many stages of the discovery process, from the identification of novel targets and elucidation of their functions to the discovery and development of lead compounds with desired properties. This review highlights recent trends that emphasize the role of integrated bio- and chemo-informatics research in drug discovery and discusses representative concepts and methodologies.

Jürgen Bajorath
Albany Molecular Research
Bothell Research Center
18804 North Creek Parkway
Bothell, WA 98011, USA
and Department of
Biological Structure
University of Washington
Seattle, WA 98195, USA
tel: +1 425 424 7297
fax: +1 425 424 7299
e-mail:
jbajorath@nce-mail.com;
jurgen.bajorath@
albmolecular.com

▼ The introduction of bio- and chemo-informatics as 'new' research disciplines was catalyzed by somewhat different needs[1,2]. In biology, informatics techniques were originally adapted to facilitate the processing and analysis of large amounts of genomic sequence data. By contrast, in chemistry the advent of combinatorial approaches made it necessary to increasingly employ informatics tools to guide the synthesis of novel types of libraries and, of equal importance, to manage rapidly growing compound collections. Another, earlier origin of chemo-informatics in pharmaceutical settings has been quantitative structure activity relationship (QSAR) analysis[3]. In both biology and chemistry, a previously unknown explosion of primary data and information has been a major determinant for the type of approaches that have been, and continue to be, developed. Because of the rapid progress of large-scale genome sequencing projects, culminating in a 'first draft' of the human genome[4], R&D in bio-informatics has moved on from the genome to the proteome level, formally defined as all proteins that are potentially expressed by a genome (a rather dynamic array, dependent on cellular conditions).

The focus has shifted somewhat from the evaluation and annotation of genomic sequence data to the analysis of actual gene products, consistent with rapidly increasing interest in proteomics[5]. For example, most recent estimates are that the human genome could code for ~40,000 proteins[4,6], and the expression patterns of these proteins vary greatly in different cell types. At a given time, only a fraction of potential gene products is expressed and, therefore, identification and characterization of these proteins becomes a crucial task. In general, protein-focused informatics efforts aim to better understand the cellular expression, posttranslational modifications (e.g. specific glycosylation or phosphorylation), family relationships, structures, and functions of proteins, as well as to evaluate their potential as drug targets.

Similar to these recent trends in bioinformatics, research and development in chemo-informatics have gone far beyond QSAR-like analyses, design of combinatorial libraries and management of compound databases. Whereas the evaluation and design of 'chemical diversity' dominated this field in its early days[7], new concepts are being developed to study SARs of ligands or to focus chemical libraries on specific drug targets[8]. Furthermore, the development of methods for

1359-6446/01/$ – see front matter PII: S1359-6446(01)01961-4

computational or virtual screening, both at the small molecular and macromolecular level, has become a major focal point in drug discovery settings[8], in addition to topics such as prediction and design of chemical reactions[9] or physicochemical properties of molecules[10].

What have not changed at all are the basic challenges for informatics disciplines: scientific knowledge must be derived from rapidly growing amounts of primary data of increasing complexity, be they biological or chemical in nature. Consequently, development and refinement of data processing, analysis, and visualization tools continue to have a major role in this area, just as they had from the beginning. In addition, other challenges persist. In bio-informatics, source databases primarily consist of sequence data, in chemo-informatics they consist of synthetic molecules. However, the majority of protein sequences identified in multicellular organisms, ~75%, does not yet have specific functions assigned to them[11]. Similarly, the vast majority of available compounds do not have known specific biological activities. Thus, together with database mining, annotation of biological function and identification of biological activity continue to play a key role in the informatics arena.

A global view of information-rich technologies in drug discovery

A common characteristic of contemporary drug discovery projects is their increasing complexity (reflected by the currently popular 'from gene to lead' paradigm), compared with the past where discovery efforts were largely dominated by chemistry and pharmacology. For example, much emphasis is being placed on the identification and validation of novel drug targets that could provide a direct link to specific disease states[12]. In modern drug discovery, several areas or stages are well complemented by bio- and/or chemo-informatics efforts. However, before discussing representative approaches, recent trends in drug discovery should be put into perspective.

Where are the problems, where are the opportunities?

Have recently introduced experimental or theoretical tools already revolutionized the drug discovery arena? Simply put, has it become easier to discover and develop drugs? In general, the answer is no, although the magnitude of R&D efforts and their costs continue to increase[13]. This problem is not new: in the early 1980s, rational drug design (mostly referring to structure-based design approaches) was anticipated to change the drug discovery landscape dramatically and circumvent a major bottleneck, the generation of high-quality clinical leads and, thus, similar expectations were articulated in the 1990s with the advent of combinatorial and high-throughput technologies. Then, increasingly larger numbers of screens and molecules were thought to solve major discovery problems (that X-ray crystallography and computer simulations alone were unable to do a decade earlier). Thus, a reductionist approach, that is, attempting to minimize the number of experiments and shortcut discovery pathways, was replaced by the 'numbers game'.

As is often the case with major R&D trends, initial expectations were generally too high and neither paradigm quite lived up to its expectations, despite large-magnitude efforts and resource commitments. Of course, at present, the 10+ years it usually takes for novel drugs to reach the market makes it difficult to assess the real impact that, for example, novel combinatorial techniques in chemistry and biology might have later on. However, the current view is that it has not become significantly easier to produce high-quality drug candidates[13]. Thus, a paradigm shift is observed yet again, with increasing emphasis placed on the integration of bio- and chemo-informatics concepts to complement experimental discovery programs (the emphasis being to complement, not replace). An attractive feature of these new technologies is that they are capable of supporting drug discovery programs at different levels, from data management and database mining to the introduction of novel design or discovery tools. The practical and conceptual diversity of informatics approaches offers many opportunities for integration into discovery programs. For example, structure-based drug design[14] has become a component of much broader and more integrated 'structural bioinformatics' or 'structural genomics' concepts[15,16]. One of the consequences of these developments is the 're-rationalization' of drug discovery research, as predictive methods and computational models are again more frequently used, being part of the broad spectrum of informatics methodologies.

Informatics components of drug discovery R&D

Major stages of preclinical drug discovery (focused here on the identification and validation of small-molecule leads, rather than biologicals or protein drugs) are summarized in Box 1, which gives examples of computational/informatics components that have begun to significantly impact these efforts. As indicated, bio- and chemo-informatics approaches are important contributors to many stages in discovery. Several of these, for example, microarray analysis or computational prediction of ADME characteristics[17], can be applied at different stages during the discovery process.

Box 1. Major stages in preclinical drug discovery and informatics components

(1) Identification of novel drug targets:
Mining of DNA and protein sequence databases and analysis of similarities
(2) Characterization of biological function(s) of target proteins:
Analysis of sequence similarities, signature motifs, and 3D structures; DNA or protein array analysis
(3) Evaluation of the therapeutic relevance of target protein functions:
DNA or protein array analysis
(4) Identification of ligands or cellular activity of drug targets:
Searching of protein–ligand interaction databases; protein array analysis
(5) Development of *in vitro* assays
(6) Screening for active compounds (hits):
Design and evaluation of diverse compound-libraries; virtual screening and/or docking of compound collections
(7) Transformation of hits into leads:
Design and analysis of target-focused or analog libraries; similarity searching for compounds with greater potency
(8) Optimization of leads:
QSAR and ADME/Tox analysis; target 3D structure- and/or lead-based drug design
(9) Assessment of optimized leads in *in vivo* models
(10) Selection of preclinical candidates

Box 2. Crucial steps in target validation

- Novel target identified?
- Cellular expression detected?
- Gene cloned?
- Biological function determined?
- Therapeutic relevance demonstrated?
- Ligands or substrates identified?
- Expressed in recombinant and active form?
- Purified in sufficient quantities?
- Cell-based or *in vitro* assay established?
- Assay formatted for screening?

Bio- and chemo-informatics research areas with high significance in drug discovery

From genomics to proteomics: sequences, motifs, structures and beyond

The classical way to identify potential new drug targets is to search for homologs of known proteins in genome or expressed sequence tag (EST) databases[18]. For example, in addition to more global sequence similarities, therapeutically relevant receptor or enzyme families usually share functionally important signature motifs (e.g. catalytic residue positions or cofactor binding regions) that can be used as database queries in pattern-matching searches, in addition to other consensus residues or motifs derived from family-specific or phylogenetic sequence profiles. Because most functional annotations are performed at the protein (rather than the DNA) level, the availability of reliable protein sequence information is crucial. Thus, genomic sequence data must be processed and structured; open reading frames must be identified and translated into protein sequences. Consequently, current genome databases and servers (e.g. Ensembl; http://www.ensembl.org/) either already contain proteome information or, at least, establish close 'relations' to proteome databases (such as GeneQuiz[19]). However, the assignment of functions to newly identified proteins goes beyond the level of sequences and relational databases. Making the transition from sequence to three-dimensional (3D) structure, either by experimental determination or prediction/modeling, can provide substantial clues about protein function[20], more so than can be deduced from sequences alone. Major reasons for this are that protein structure is much more conserved than sequence, protein folds are recurrent and similar structural motifs often confer similar functions.

However, the identification of novel proteins, and even their functional characteristics, is in itself not sufficient to render them valuable therapeutic targets. In fact, target validation (as summarized in Box 2) involves much more. Without localizing, cloning and expressing the putative target in its active form, and without identifying cellular activities and substrates or ligands, it is essentially impossible to develop reliable assays that can then be used for screening. Moreover, even if the mechanics of target validation can be handled, it usually remains to be determined whether the selected protein and its activity or interactions are therapeutically relevant. Therefore, efficient target validation strategies continue to be a major bottleneck in informatics-driven drug discovery research[13].

Although initial bio-informatics efforts only provide the basis for target validation, new methodologies are beginning to address later-stage questions. For example, for target identification, and also the evaluation of functions and their therapeutic relevance, microarray technology has an increasingly important role. DNA arrays[21] are typically used to analyze differential expression of genes as a function of specific conditions, such as genetic knockouts,

gene mutations and/or drug treatments[21–23]. In this context, the detection of disease is a major focal point of profiling methods[22,23]. Computational analysis of expression data produced by arrays is a major task for bioinformatics[23,24]. For functional assignments and analysis of individual drug sensitivity, the study of genetic polymorphisms, in particular single nucleotide polymorphisms (SNPs), has become a major topic for both genomic sequencing and array analysis[25], as part of pharmacogenomics research[26] and systematic analysis of disease genes and markers[12]. In addition, protein arrays or chips can be used to analyze protein–protein interactions or to conduct a search for ligands[27,28]. This type of analysis complements informatics efforts to assign binding specificities and functions to novel proteins. It is supported by the use of protein–protein interaction maps or databases that have been derived from exhaustive yeast two-hybrid screens[29] or theoretical genome analysis[30,31]. It is evident that these studies contribute significantly to target validation.

Virtual screening, lead identification, drug-like properties and ADME analysis

Once targets have been selected and validated, discovery projects typically move to the hit and lead identification phase, which is heavily impacted by informatics efforts. As discussed later, HTS and virtual screening are highly complementary disciplines and not mutually exclusive. Virtual screening can be performed at the macro-molecule and/or small-molecule level, depending on the information available. If a 3D structure or molecular model of the therapeutic target is available and the ligand binding site is known, a variety of docking algorithms and simulations can be employed to screen 3D databases of compounds[32] and prioritize computational hits for testing. This prioritization is done on the basis of scoring functions that account for surface complementarity, energetic terms and/or solvation models[32,33]. However, a structural template is often unavailable and, in these cases, virtual screening depends on the availability of initial hits, usually produced by HTS. These hits then serve as template molecules for 2D or 3D database or similarity searching[34] for more active compounds. For virtual screening, a variety of computational tools have been developed including diverse molecular fingerprints[34], pharmacophore queries[35], or multi-dimensional QSAR models[36]. Following lead identification, subsequent optimization efforts are driven by medicinal chemistry and, typically, are further supported by QSAR studies.

Currently, a major focal point of chemo-informatics research is the development of methods that enable the prediction of downstream characteristics of selected compounds, most importantly ADME parameters[17,37] and drug-like characteristics[37,38]. The reason behind these intense efforts is the large fall-out rate of compounds in clinical trials because of lack of efficacy, side effects or toxicity, which continues to be a significant and costly problem in drug development[13,39]. For example, the average costs involved in a Phase I clinical trial are US$1–1.5 million, whereas Phase II studies require US$2–10 million per therapeutic indication[39]. Thus, every compound that fails in a Phase II clinical trial, for lack of efficacy or safety, results in a definite loss of several million dollars. Consequently, computational analyses attempt to derive as much information as possible about desired and, perhaps more importantly, undesired molecular properties and to make use of this knowledge early in the discovery process, during chemical library design, lead identification and optimization[38]. Focal points of ADME analysis include molecular transport properties, such as passive intestinal absorption and blood–brain barrier penetration models[10,40]. Such models are often derived by QSAR-like analysis from experimental learning sets and used as filters when profiling chemical libraries or lead compounds. Also under development, but so far less common, are computational models to predict drug metabolism on the basis of interactions with drug-metabolizing enzymes, in particular cytochrome P450 isoforms. Methods to evaluate the drug-like characteristics[38] of compounds or libraries range from the estimation of simple molecular property distributions[40,41] or derivation of structural rules[42] to complex neural network simulations[43]. Similar to ADME approaches, these methods rely on the analysis and comparison of databases consisting of drugs and non-drugs (which are often randomly collected synthetic compounds). Despite the availability of molecular property or structure rules and the success of neural network methods to predict drug-like character, principles of drug-likeness are just beginning to be understood. This is not surprising considering the complexity of mechanism of action, efficacy, molecular transport or metabolic stability parameters that determine whether or not a molecule makes it through the discovery ‘pipeline’.

The interface between bio- and chemo-informatics and experimental programs

Many bio- and chemo-informatics methodologies are complementary to each other, or even conceptually similar, which is well illustrated by some of the informatics components of the drug discovery process. Several of these informatics approaches are closely linked to experimental discovery efforts.

Analysis and exploitation of HTS data

HTS continues to be the source of vast amounts of assay data. These results not only present a considerable challenge for computational data analysis and management but also provide a significant knowledge base for drug discovery. For example, subsets of large screening libraries or compound decks that are found to bind preferentially to certain classes of therapeutic targets (e.g. G-protein-coupled receptors or serine proteases) can be identified and further explored. Alternatively, general cellular toxicity of chemical libraries can be estimated from multiple cell-based screens. Informatics approaches are intimately involved in screening programs and have a dual role (and in this case, distinguishing between bio- and chemo-informatics more or less becomes a semantic, rather than scientific, issue). First, an informatics infrastructure is required to efficiently record, analyze and archive the data produced by many screens on diverse targets[44]. In addition, statistical methods have been developed, most notably recursive partitioning[45] and binary QSAR[46], that correlate the distribution of active and non-active compounds in screening data sets with molecular structure and properties. On this basis, predictive models of biological activity can be calculated[45–47] and, in turn, used for virtual screening of other compound sets, thus illustrating the complementary nature of HTS and computational screening efforts.

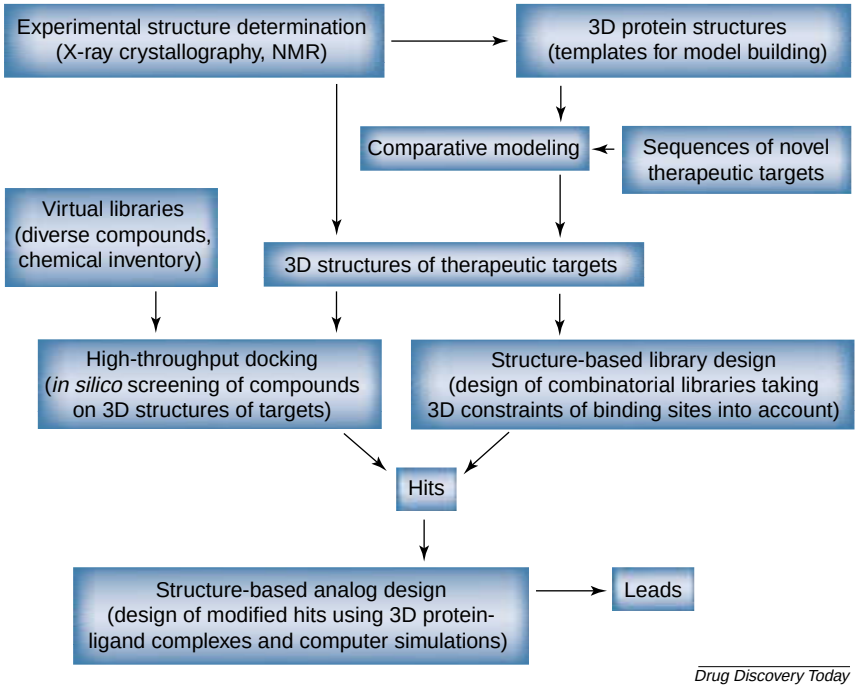

Figure 1. Integration of structural and informatics methods. The diagram illustrates how different computational approaches complement each other in the context of structure-based drug design. Experimentally determined three-dimensional (3D) structures of protein targets aid in structure-based design efforts and/or serve as templates for modeling of related targets (for which only sequences are available). Provided the ligand binding site of a target is known, docking calculations can be applied to virtually screen compound databases and select molecules for experimental testing. Frequently, these databases are large virtual compound libraries (for example, obtained by diversity design), from which candidate molecules can be selected for synthesis. It is also possible to directly focus the design of libraries on the 3D structure of a target. For example, molecular scaffolds (core structures) could be selected for combinatorial exploration that are found to spatially fit the binding (or active) site of the targeted receptor (or enzyme), based on docking calculations or other computational studies. Such selections are attempted to substantially increase the probability of generating library compounds with desired specific activity. Once hits have been identified by experimental testing of library compounds (diverse or focused), structure-based design methods are again employed to convert hits into more potent leads and further optimize these leads. In such cases, 3D structures of complexes of target proteins and active compounds are determined experimentally or modeled and serve as templates for iterative design and evaluation of analogs with increased potency.

Protein structure prediction, target-focused libraries and structure-based drug design

The prediction and analysis of 3D structures of therapeutic targets, together with extended concepts of structure-based drug and library design, is another area of high complementarity and overlap between bio- and chemo-informatics, as well as experimental drug discovery. Figure 1 illustrates the interplay between structure-based and informatics methodologies. In drug discovery settings, it is increasingly the attempt to combine structure-based design and combinatorial chemistry[48–50]. In this context, combinatorial library design and development is shifting from large and diverse screening libraries to smaller and more specialized libraries (again departing from the 'numbers game') that are focused on either specific therapeutic targets or hits and leads with desired initial activity[48]. A popular way to achieve such target-focus is to combine docking simulations with library design[48,49], for example, by identifying template molecules for combinatorial exploration that 'fit' the binding (or active site) of a target protein (or enzyme). In fact, several successful structure-based library designs have been reported[49]. These developments also influence structure-based drug design techniques themselves. For example, rather than predicting compound modifications in a step-wise manner and improving the binding characteristics of synthetic derivatives by iteration, it

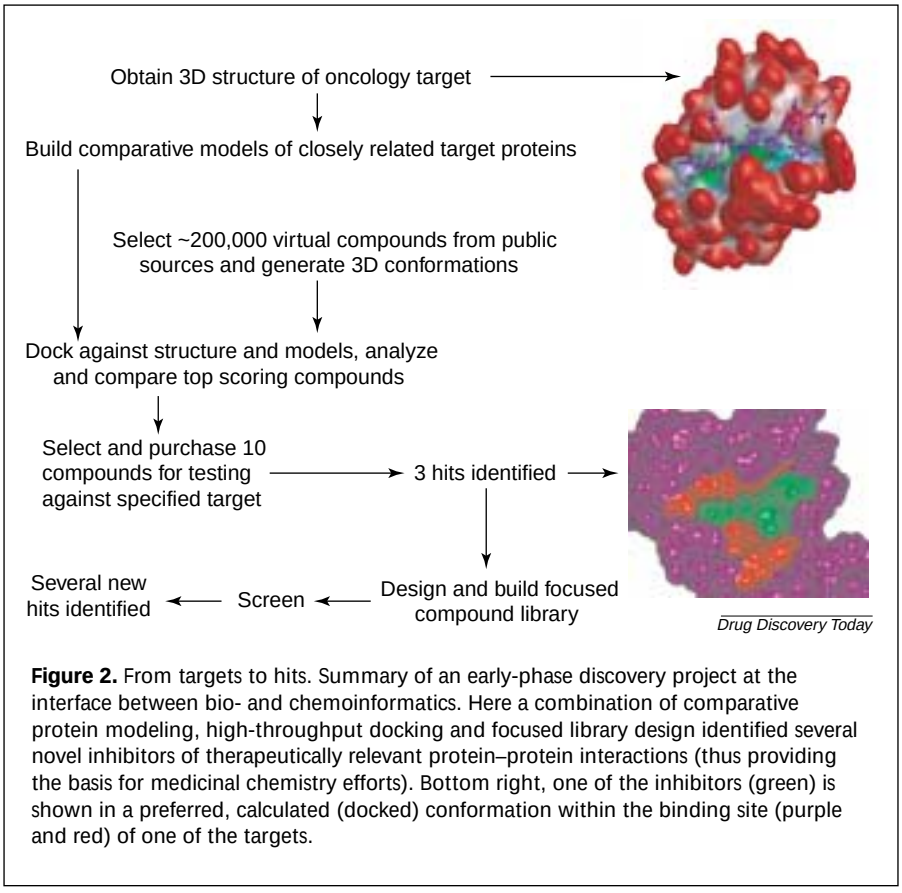

Figure 2. From targets to hits. Summary of an early-phase discovery project at the interface between bio- and chemoinformatics. Here a combination of comparative protein modeling, high-throughput docking and focused library design identified several novel inhibitors of therapeutically relevant protein–protein interactions (thus providing the basis for medicinal chemistry efforts). Bottom right, one of the inhibitors (green) is shown in a preferred, calculated (docked) conformation within the binding site (purple and red) of one of the targets.

is possible to design analog libraries with the aid of 3D structures of targets, select the most attractive compounds and thereby shortcut at least some optimization cycles.

These and other tasks in chemistry further emphasize the need for experimental structures of target proteins or at least reliable molecular models. This is in addition to the more biological applications of structures in assigning protein functions[20], as discussed earlier, studying genetic polymorphisms in three dimensions[51] or identifying ligand binding sites, typically in combination with mutagenesis studies[52]. Thus, an important question is how to obtain a sufficient number of template structures for these applications? Genome projects have already revealed many more potential targets at the sequence level than could be studied experimentally, without considering the determination of their 3D structures. This is one reason why comparative modeling, where a molecular model of a novel protein is built based on the structure of a related one, is experiencing a renaissance[53]. Supported by an array of sequence and structure analysis and comparison tools[20,54] as well as fold recognition (or threading) calculations[55], comparative modeling is usually more accurate than *ab initio* methods and also amenable to automated high-throughput structure prediction[56]. Accordingly, one of the major goals of the structural genomics initiative[16] is the determination of at least one representative structure of each known protein family (or superfamily) so that other family members or similar proteins can be modeled comparatively. Thus, similar to the situation in HTS, proteomics or structure-based design, as discussed herein, experimental structure-determination and protein structure prediction will be highly complementary disciplines in the post-genomics era. For drug discovery, the close interplay between complementary small molecular and macro molecular approaches, both theoretical and experimental, offers many opportunities. As an example, Figure 2 shows a summary of a discovery project (in-house) at the interface between bio- and chemo-informatics.

The future?

It seems certain that the impact of informatics approaches on drug discovery will increase steadily. This will not only include the generation of further improved data processing and management infrastructure, but also the introduction of novel research concepts and predictive methods; for example, for virtual screening or the prediction of molecular transport and metabolic parameters. As the drug discovery information and knowledge base grows, it is also anticipated that computational analyses will reduce the magnitude of experimental programs in areas such as compound synthesis and screening. For example, much information concerning *in vitro* assays, pharmacological profiles or *in vivo* models can be made available in context in relational or object-oriented databases. Mining such databases is likely to aid the rational pre-selection of subsets of compound decks for testing on selected target classes. It is also expected that biological and chemical informatics research will continue to merge, at least in drug discovery settings. As indicated here, distinguishing between bio- and chemo-informatics might, in some instances, already be artificial because underlying scientific concepts and goals are often similar or difficult to separate. This is, in part, reflected by the fact that terms such as 'research informatics' or 'drug discovery informatics' are used more frequently in the field. In any event, the design and implementation of information-rich R&D concepts will be increasingly important for effective drug discovery.

References

1 Searls, D.B. (2000) Using bioinformatics in gene and drug discovery. *Drug Discov. Today* 5, 135–143
2 Brown, F.K. (1998) Chemoinformatics: what is it and how does it impact drug discovery. *Annu. Rep. Med. Chem.* 33, 375–384
3 Hansch, C. *et al.* (2001) Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophopic terms. *Chem. Rev.* 101, 619–672
4 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
5 Pendley, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* 405, 837–846
6 Gaasterland, T. and Oprea, M. (2001) Whole-genome analysis: annotations and updates. *Curr. Opin. Struct. Biol.* 11, 377–381
7 Martin, Y.C. (2001) Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* 3, 1–20
8 Xue, L. and Bajorath, J. (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combi. Chem. High Throughput Screen.* 3, 363–372
9 Gasteiger, J. *et al.* (2000) Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Perspect. Drug Des. Discovery* 20, 1–21
10 Blake, J.F. (2000) Chemoinformatics – predicting the physicochemical properties of 'drug-like' molecules. *Curr. Opin. Biotechnol.* 11, 104–107
11 Edwards, A.M. *et al.* (2000) Proteomics: new tools for a new era. *Mod. Drug Discov.* 3, 34–40
12 Adam, D. (2001) Genetics group targets disease markers in the human sequence. *Nature* 412, 105
13 Drews, J. (2000) Drug discovery: a historical perspective. *Science* 287, 1960–1964
14 Murcko, M.A. *et al.* (1999) Structure-based drug design. *Annu. Rep. Med. Chem.* 34, 297–306
15 Kim, S.H. (1998) Shining a light on structural genomics. *Nat. Struct. Biol.* 5 (Suppl.), 643–645
16 Burley, S.K. (2000) An overview of structural genomics. *Nat. Struct. Biol.* 7 (Suppl.), 932–934
17 Eddershaw, P.J. *et al.* (2000) ADME/PK as part of a rational approach to drug discovery. *Drug Discov. Today* 5, 409–414
18 Rawlings, C.J. and Searls, D.B. (1997) Computational gene discovery and human disease. *Curr. Opin. Genet. Dev.* 7, 416–423
19 Hoersch, S. *et al.* (2000) The GeneQuiz server: protein functional analysis through the Web. *Trends Biochem. Sci.* 25, 33–35
20 Thornton, J.M. (2001) From genome to function. *Science* 292, 2095–2097
21 Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression, and DNS arrays. *Nature* 405, 827–836
22 Hartwell, L.H. *et al.* (1997) Integrating genetic approaches into the discovery of anticancer drugs. *Science* 278, 1064–1068
23 Weinstein, J.N. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349
24 Altman, R.B. and Raychaudhuri, S. (2001) Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.* 11, 340–347
25 Carlson, C.S. *et al.* (2001) SNPing the human genome. *Curr. Opin. Chem. Biol.* 5, 78–85
26 Evans, W.E. and Relling, M.V. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 286, 487–491
27 MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763
28 Emili, A.Q. and Cagney, G. (2000) Large-scale functional analysis using peptide or protein arrays. *Nat. Biotechnol.* 18, 393–397
29 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
30 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753
31 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
32 Gane, P.J. and Dean, P.M. (2000) Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* 10, 401–404
33 Charifson, P.S. *et al.* (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* 42, 5100–5109
34 Bajorath, J. (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* 41, 233–245
35 Good, A.C. and Mason, J. (1996) Three-dimensional structure database searches. *Rev. Comput. Chem.* 7, 67–117
36 Hopfinger, A.J. *et al.* (1999) Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* 39, 1151–1160
37 Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of drug-likeness. *Drug Discov. Today* 5, 49–58
38 Walters, W.P. *et al.* (1999) Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* 3, 384–387
39 Fox, A.W. (2001) Clinical development – look before you leap. *Nat. Biotechnol.* 19 (Suppl.), BE27–BE29
40 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
41 Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Design* 14, 251–264
42 Muegge, I. *et al.* (2001) Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* 44, 1841–1846
43 Sadowski, J. (2000) Optimization of chemical libraries by neural networks. *Curr. Opin. Chem. Biol.* 4, 280–282
44 Boguslavsky, J. (2001) Creating knowledge from HTS data. *Drug Discov. Devel.* 4 (6), 34–38
45 Rusinko, A., III *et al.* (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* 39, 1017–1026
46 Labute, P. (1999) Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* 4, 444–455
47 van Rhee, A.M. *et al.* (2001) Retrospective analysis of an experimental high-throughput screening data set by recursive partitioning. *J. Comb. Chem.* 3, 267–277
48 Antel, J. (1999) Integration of combinatorial chemistry and structure-based drug design. *Curr. Opin. Drug Discov. Devel.* 2, 223–224
49 Böhm, H-J. and Stahl, M. (2000) Structure-based library design: molecular modeling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* 4, 283–286
50 Sawyer, T. (2001) Drug discovery in 3D. *Curr. Drug Discov.* 1 (3), 33–37
51 Bajorath, J. *et al.* (1996) Classification of mutants in the human CD40 ligand, gp39, that are associated with X-linked hyper IgM syndrome. *Protein Sci.* 5, 531–534
52 Bowen, M.A. *et al.* (2000) Cell surface receptors and their ligands: *in vitro* analysis of CD6–CD166 interactions. *Proteins* 40, 420–428
53 Marti-Renom, M. *et al.* (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291–325
54 Teichmann, S.A. *et al.* (2000) Advances in structural genomics. *Curr. Opin. Struct. Biol.* 9, 390–399
55 Miller, R.T. *et al.* (1996) Protein fold recognition by sequence threading – tools and assessment techniques. *FASEB J.* 10, 171–181
56 Sanchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13597–13602