ELSEVIER

Exploring the chemogenomic knowledge space with annotated chemical libraries

Nikolay P Savchuk*, Konstantin V Balakin and Sergey E Tkachenko

The recent human genome initiatives have led to the discovery of a multitude of genes that are potentially associated with various pathologic conditions and, thus, have opened new horizons in drug discovery. Simultaneously, annotated chemical libraries have emerged as information-rich databases to integrate biological and chemical data. They can be useful for the discovery of new pharmaceutical leads, the validation of new biotargets and the determination of the structural basis of ligand selectivity within target families. Annotated libraries provide a strong information basis for computational design of target-directed combinatorial libraries, which are a key component of modern drug discovery. Today, the rational design of chemical libraries enhanced with chemogenomics data is a new area of progressive research.

Addresses
Chemical Diversity Labs, Inc., 11558 Sorrento Valley Road, San Diego, California 92121, USA
*e-mail: ns@chemdiv.com

Current Opinion in Chemical Biology 2004, **8**:412–417

This review comes from a themed section on
Next-generation therapeutics
Edited by Tudor Oprea and John Tallarico

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2004.06.003

Abbreviations

ADME	absorption, distribution, metabolism, excretion
GPCR	G-protein-coupled receptor
NCI	National Cancer Institute

Introduction

Today, we are beginning to witness a clear trend in drug discovery to focus research efforts on gene families, which, because of their interrelationships, provide an opportunity for parallel processing of multiple targets. The effective identification of high quality hits and leads across diverse classes of therapeutic targets is based on the systematic analysis of structural genomics data [1,2,3•]. Chemogenomics refers to the determination and practical application of the relationships between chemical and genomic spaces. The key element of chemogenomics is the ligand–target space, in which all ligands are annotated according to their targets (Figure 1). The effective exploration of this ligand–target space requires application of special data-mining tools for both ligand and target domains, as well as algorithms to combine the results. The purpose of the ligand–target classification is to allow annotation-based exploration of this information-rich space. Using this classification, the genome sequence information of the target can be directly associated with the ligand, thus allowing gene homology-based identification of ligands to closely related targets. On the other hand, annotated ligands can serve as comprehensive reference sets for chemoinformatics-based similarity searches and for discovery of novel therapeutically relevant biotargets. In this review, we elucidate the reported examples of annotated chemical libraries and discuss their possibilities within contemporary drug discovery.

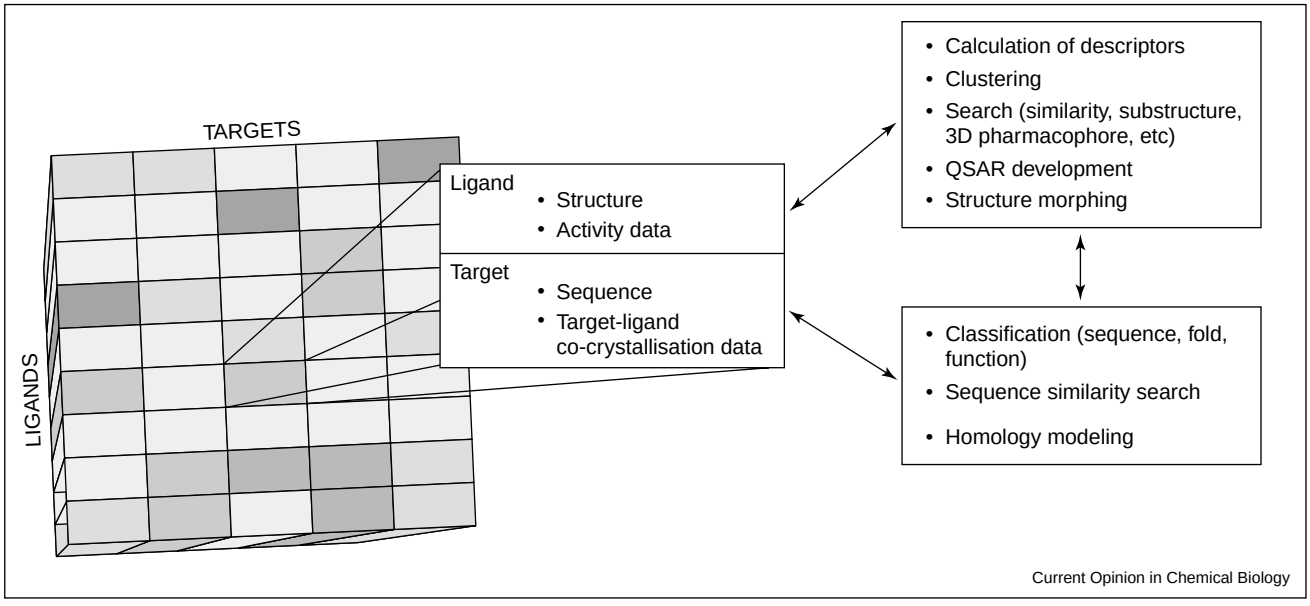
Annotated chemical libraries in drug design

Ligand–target annotation schemes have become a useful tool for analyzing the inhibition of tumor cell proliferation based on the National Cancer Institute's (NCI) screening panel. One of the first attempts to analyze a complex ligand–target space was the work published by Weinstein in 1997 [4]. In another complementary study, the gene expression patterns were related not just to the drugs as entities but to approximately 27 000 substructures and other chemical features within the drugs [5]. Using a systematic substructure analysis coupled with statistical correlations of compound activity with differential gene expression, two subclasses of quinones were identified whose patterns of activity in the NCI's 60-cell line screening panel correlate strongly with the expression patterns of particular genes.

An extensive study of ca. 20 000 compounds tested against 80 of NCI's tumor cell lines was recently performed using Kohonen self-organizing maps [6]. The developed classifications can serve as hypotheses for the rational discovery of antitumor agents, and their usefulness was experimentally confirmed by the same group of researchers [7]. The expression profiles of 400 genes were used to locate similar activity profiles of synthetic agents screened against 60 of NCI's tumor cell lines. A correspondence was found between mRNA expression patterns and 50% growth inhibition response patterns of screened agents for 11 cases. This work supports the idea that similarities between expression patterns and chemical responses for the NCI tumor panel can be related to known aspects of molecular structure and putative cellular function.

A collection of properly characterized ligands covering a diverse set of mechanisms of action can be an extremely useful tool to probe disease pathways and identify new

Figure 1



The elements of the annotated ligand-target space.

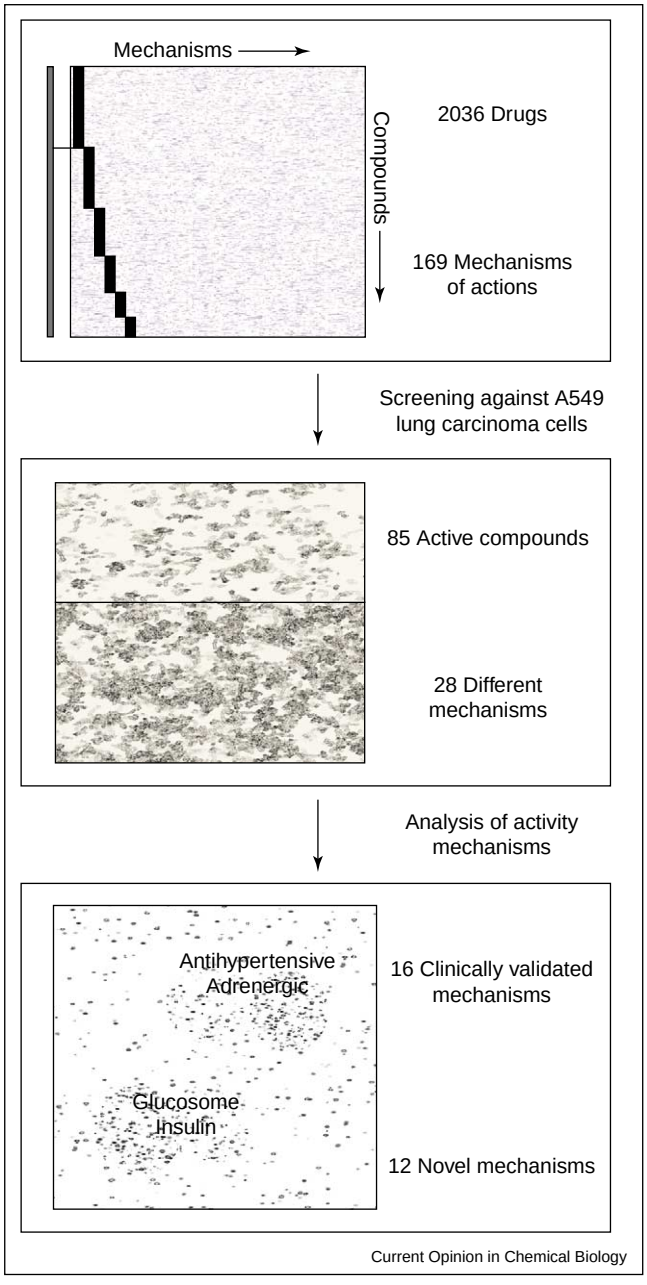
disease-associated targets belonging to well validated target families within these pathways. A method was recently reported for testing many biological mechanisms and related biotargets in cellular assays using an annotated compound library [8••]. This library represents a collection of 2036 biologically active compounds with 169 diverse, experimentally confirmed biological mechanisms and effects. These compounds were experimentally tested against A549 lung carcinoma cells, and subsequent analysis of the screen results allowed the determination of 12 previously unknown mechanisms associated with the proliferation of the studied carcinoma cells (Figure 2).

The structure–activity relationship homology principle (SARAH) formulated by Frye in 1999 [9] uses the knowledge obtained in the screening experiments for one target to discover leads for another one. There are several recent works demonstrating the practical utility of this strategy [10–12,13•,14]. This potential can be most effectively realized using annotated databases suitable for effective data mining. Thus, researchers at the French biotechnology company Cerep reported an analysis of properties of more than 500 drugs screened against 42 targets [11]. Based on this target–ligand database, they derived similarity metrics for both targets and ligands based on Fuzzy Bipolar Pharmacophore Fingerprints, which are the conformation-dependent (3D) two-point pharmacophore vectors. It was observed that ligands for subtypes of one target or closely related target families usually have similar ligand binding profiles, whereas homologically distant targets (such as 5HT-binding G-protein-coupled receptors [GPCRs] and ion channels), despite the common endogenous ligand, have highly distinct ligand binding profiles from each other.

Based on the concept of homology-based similarity searching, researchers at Novartis developed an annotation scheme for the ligands of four major target classes — enzymes, GPCRs, nuclear receptors and ligand-gated ion channels — for *in silico* screening and combinatorial design of targeted libraries [12]. According to their approach, the homology-based targeted library design consists of several principal steps (Figure 3). In the initial steps, gene sequences for targets that have been identified by genomics approaches are cloned and expressed as target proteins that are suitable for screening. Using the annotated ligand–target database, at least one target with known ligands is selected that is homologous to this new target. Then, the known ligands of the selected target are combined into a reference set. Finally, the potential ligands for the new target are searched on the basis of their similarity to the reference set. Retrospective *in silico* screening experiments have shown that such reference sets can be useful for the identification of ligands that bind to receptors closely related to the reference system. The same group of scientists reported a modified [13•] homology-based similarity searching based on special molecular representations, Similog keys.

An annotated library of inhibitors for a group of closely related enzymes, the papain family of cysteine proteases, was used to classify the reference group of proteases into subfamilies based on their small-molecule affinity fingerprints [14]. This classification scheme was used to identify

Figure 2



Identifying new cancer-associated mechanisms using an annotated compound library [8••].

cysteine protease targets in complex proteomes and predicts their small-molecule inhibitors on the basis of reported crystal structures.

Annotated chemical libraries can be useful for determination of the structural basis of ligand specificity within target families. In another recent publication, publicly available selectivity data were employed by researchers from Lilly to define chemically relevant chemogenomic kinase space [15•]. These data were used to create a chemogenomic kinase dendrogram for 43 kinases. Conservation of the ATP binding site between known kinase structures, together with knowledge of the human kinase genome and an abundance of selectivity data, enabled significant advances in understanding the relationship between kinase targets and inhibitors. The developed chemogenomic classification of kinase space can be used as an effective guideline for making decisions about target selection, panel selection, and use of existing inhibitors for chemogenomically related targets. Recently, chemogenomics technologies at Lilly have aided in several successful lead generation strategies [16].

Commercial annotated databases

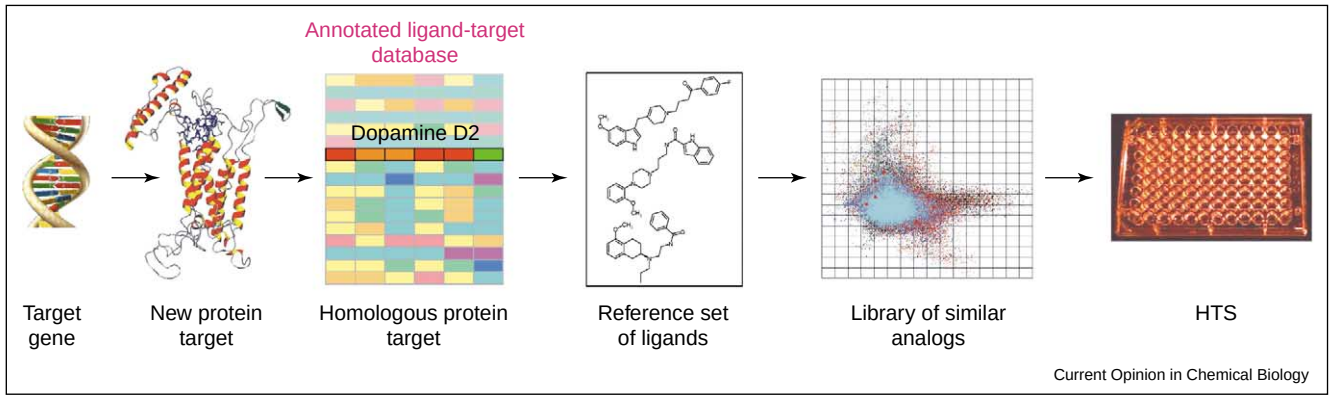
During the past two years, there has been a large increase in the number of special databases that aim to provide useful information for analysis of the ligand–target knowledge space (Table 1). These information repositories, typically focused on well-established target families or known drugs, can be used for homology-based search, validation of putative targets, automatic generation of quantitative structure–activity relationship models and prediction of ADME (absorption, distribution, metabolism, excretion) properties. WOMBAT 2004.1, available from Sunset Molecular Discovery LLC, represents a good example of an annotated database integrated in a multifunctional chemoinformatics platform. It contains 76 165 entries (68 543 unique SMILES; Simplified Molecular Input Line Entry System) covering 3039 series (over 3000 papers) and ~143 000 activities on ~630 targets. Activities include 635 inactives, 8916 'less than', 259 'greater than', and 578 single-dose results. Activity types include 37.1% K_i, 55.9% IC_{50}, 4.44% EC_{50} (agonists and substrates), 0.9% K_b and K_d (binding data), 0.1% MIC and 0.04% ED_{50}. WOMBAT 2004.1 covers a decade of the *Journal of Medicinal Chemistry* (1991–2000), and papers from five other journals. It is fully integrated in FEDORA (Metaphorics LLC).

The rapid growth of such databases reflects the high level of interest of the scientific and business communities in chemogenomics approaches. It suggests that the development of annotated library-related technologies is considered a significant competitive advantage in the industry. Nevertheless, there are several key questions about these products. What is the quality of the data? How frequently is the information in them updated? Is the user interface convenient? Are the databases compatible with other industry-standard chemoinformatics platforms? Questions such as these will be taken into consideration should one implement these resources for target-directed research.

Design of target-directed combinatorial libraries

According to the described examples, annotated libraries provide a strong information basis for computational

Figure 3



Homology-based targeted library design.

design of target-directed combinatorial libraries, which are a key component of modern chemogenomic drug discovery platforms. Many examples of successful lead compounds discovered from target-directed libraries can be found in the recent literature (2002–2004) [18••,19]. Advanced chemoinformatics approaches are required to tailor the design of such libraries to address both target potency and other properties necessary for further pharmaceutical development [1,20••,21,22].

Thus, 3D structural data from a representative member of a drugable target family can be used to design chemical screening libraries using modern docking technologies [23]. For rapid identification of ligands to putative targets, the design based on privileged structures seems to be a very useful approach [24•]. This concept is inherently consistent with the homology-based similarity principle. For example, using a series of docking experiments for a set of class A GPCRs, a good correlation has been found between conservation patterns of residues in the ligand-binding pocket and the privileged structure fragments in class A GPCR ligands [25•]. As a result, target-directed ligand libraries can be effectively designed without any foreknowledge of the structure of the endogenous ligand, which in turn means that even orphan receptors can be addressed as potential drug targets.

An inherent concern associated with this approach is the restricted availability of privileged substructures for known target families. The resulting issues concerning intellectual property clearly limit the scope of this approach. This problem can be addressed using special methods of rational transformation (morphing) of active scaffolds to generate analogs with enhanced IP position. The goal of morphing is to select structures with similar steric and electrostatic parameters but different chemistry. The bioisosteric approach, which is one of the key concepts in drug design, represents a useful morphing strategy [26]. Scientists at Chemical Diversity have applied the concept of bioisosterism to build a software tool for designing target-directed libraries. Using this software, a series of potent submicromolar inhibitors of *abl* tyrosine kinase have been identified using structures of initial high-throughput screening hits as queries (Tkachenko SE, Okun I, Balakin KV, Petersen CE, Ivanenkov YA, Savchuk NP, Ivashchenko AA, unpublished data).

Table 1

Annotated databases

Database	Description	Web-site
AUR-STORE	Structure–activity information focused on different targets and activities	http://www.aureus-pharma.com
Bioprint	Cerep's proprietary ligand profiling data including target-specific activity and ADME-related properties	http://www.cerep.fr
ChemBank [17]	Structures and biological activity data for over 2000 compounds	http://chembank.med.harvard.edu
Drugmatrix	Pharmacological, pathology and gene expression profiles for benchmark drugs	http://www.iconixpharm.com
GPCR Annotator	Structural and functional information related to GPCRs and their ligands	http://www.jubilantbiosys.com
Kinase ChemBioBase	Annotated database of over 170 000 molecules covering over 350 kinases	
Sertanty	Annotated database of 22 000 compounds focused on protein kinases	http://www.sertanty.com
WOMBAT	63 000 structures focused on over 500 targets and more than 100 000 biological activities	http://www.sunsetmolecular.com

Optimization of the initial hits can also proceed via calculating the similarity of 2D molecular fingerprints, a conceptually simple yet practically useful methodology extensively reviewed in the scientific literature [27]. Another straightforward approach is related to a 2D substructure search for analogs of known ligands [28]. Going beyond analysis of 2D structural representations, the optimization libraries can be generated using 3D pharmacophore hypotheses [29]. A patented Tripos technology of molecular shape-based similarity search [30] can successfully complement the more traditional approaches. Combinatorial library design can be carried out at either the product or the reagent level, and both strategies can be useful approaches in chemogenomic applications [31,32].

Finally, the knowledge-based data mining methods used for correlation of molecular properties with specific activities represents an alternative promising strategy of targeted library design [33]. For example, recursive partitioning is a prominent statistical method that uses decision trees to identify specific partitions enriched with active molecules. The method has the potential to leverage copious data from an older, well-studied target while beginning to study a newer target for which only a small amount of data are available [34•]. Several groups have reported the application of artificial neural networks, Kohonen self-organizing maps and Support Vector Machines [35,36,37•,38,39] to the design of target-specific libraries. The ability to identify compounds with desired target-specific activity and optimize a large number of other molecular parameters (such as ADME/Tox-related properties, lead- and drug-likeness) in a parallel fashion is a characteristic feature of these methods, which favorably complement other strategies of targeted library design. The works by Oprea [40,41] demonstrate that issues beyond binding affinity, such as lead-likeness and pharmacokinetic behaviour, need to be considered in the early stages of drug design. Using a special VolSurf program [42], both ligand–receptor binding and the ligand's pharmacokinetic behavior can be modeled simultaneously.

Conclusions

A modern chemogenomics platform for drug discovery comprises a variety of sophisticated technologies for the discovery and optimization of multiple compounds interacting with different members of a pharmaceutically relevant target family. A key component of such a chemogenomics platform is a combinatorial library focused on specific protein targets. The knowledge-based design of chemical libraries can be enhanced using a systematic exploration of the annotated chemical libraries for selected target families. Such a meaningful integration of chemogenomics data with advanced virtual screening technologies holds great promise for more efficient discovery of leads across diverse classes of biological targets. However, the technologies related to annotated databases are still in their infancy, and future practical works will highlight their role in contemporary drug discovery.

Update

Recently, scientists from Vertex reviewed the application of a chemogenomics approach to the protein kinases of the human genome with emphasis upon the synergies and efficiencies to be gained [43]. Specific examples from the SAPK-family were discussed. Scientists from the University of Barcelona demonstrated that making optimal use of chemogenomics information represents an efficient knowledge-based strategy for improving binding affinity estimations, ligand binding-mode predictions, and virtual screening enrichments obtained from protein–ligand docking [44].

Acknowledgements

We would like to thank Dr Yan A Ivanenkov for his assistance in searching the literature and preparation of the manuscript.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Agrafiotis DK, Lobanov VS, Salemme FR: **Combinatorial informatics in the post-genomics era**. *Nat Rev Drug Discov* 2002, **1**:337-346.

2. Salemme FR: **Chemical genomics as an emerging paradigm for postgenomic drug discovery**. *Pharmacogenomics* 2003, **4**:1-11.

3. • Jacoby E, Schuffenhauer A, Floersheim P: **Chemogenomics knowledge-based strategies in drug discovery**. *Drug News Perspect* 2003, **16**:93-102.
A useful survey of current chemogenomics knowledge-based strategies for drug discovery.

4. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL *et al.*: **An information-intensive approach to the molecular pharmacology of cancer**. *Science* 1997, **275**:343-349.

5. Blower PE, Yang C, Fligner MA, Verducci JS, Yu L, Richman S, Weinstein JN: **Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data**. *Pharmacogenomics J* 2002, **2**:259-271.

6. Rabow AA, Shoemaker RH, Sausville EA, Covell DG: **Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities**. *J Med Chem* 2002, **45**:818-840.

7. Wallqvist A, Rabow AA, Shoemaker RH, Sausville EA, Covell DG: **Establishing connections between microarray expression data and chemotherapeutic cancer pharmacology**. *Mol Cancer Ther* 2002, **1**:311-320.

8. •• Root DE, Flaherty SP, Kelley BP, Stockwell BR: **Biological mechanism profiling using an annotated compound library**. *Chem Biol* 2003, **10**:881-892.
The authors demonstrate an approach to identification of new disease-associated targets using an annotated library.

9. Frye SV: **Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era**. *Chem Biol* 1999, **6**:R3-R7.

10. Koch MA, Breinbauer R, Waldmann H: **Protein structure similarity as guiding principle for combinatorial library design**. *Biol Chem* 2003, **384**:1265-1272.

11. Horvath D, Jeandenans C: **Neighborhood behavior of *in silico* structural spaces with respect to *in vitro* activity spaces - a**

novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J Chem Inf Comput Sci* 2003, **43**:680-690.

12. Schuffenhauer A, Zimmermann J, Stoop R, van der Vyver JJ, Lecchini S, Jacoby E: **An ontology for pharmaceutical ligands and its application for *in silico* screening and library design**. *J Chem Inf Comput Sci* 2002, **42**:947-955.

13. • Schuffenhauer A, Floersheim P, Acklin P, Jacoby E: **Similarity metrics for ligands reflecting the similarity of the target proteins**. *J Chem Inf Comput Sci* 2003, **43**:391-405.
A good example illustrating the practical approach developed by scientists at Novartis for the homology-based design of target-specific libraries.

14. Greenbaum DC, Arnold WD, Lu F, Hayrapetian L, Baruch A, Krumrine J, Toba S, Chehade K, Bromme D, Kuntz ID, Bogyo M: **Small molecule affinity fingerprinting. A tool for enzyme family subclassification, target identification, and inhibitor design**. *Chem Biol* 2002, **9**:1085-1094.

15. • Vieth M, Higgs RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H: **Kinomics – structural biology and chemogenomics of kinase inhibitors and targets**. *Biochim Biophys Acta* 2004, **1697**:243-257.
An article describing a practical approach developed by scientists at Lilly for collection and exploration of chemical and biological data about kinase inhibitors and targets.

16. Vieth M, Brooks HB, Hamdouchi C, McMillen W, Sawyer JS, Yingling JM, Zhang F: **Combining medicinal chemistry with chemogenomic and computer-aided structure-based design in development of novel kinase inhibitors**. *Cell Mol Biol Lett* 2003, **8**:566-567.

17. Strausberg RL, Schreiber SL: **From knowing to controlling: a path from genomics to drugs using small molecule probes**. *Science* 2003, **300**:294-295.

18. •• Golebiowski A, Klopfenstein SR, Portlock DE: **Lead compounds discovered from libraries: part 2**. *Curr Opin Chem Biol* 2003, **7**:308-325.
A comprehensive review of lead compounds with the potential to progress to viable drug candidates identified from combinatorial libraries.

19. Dolle RE: **Comprehensive survey of combinatorial library synthesis: 2002**. *J Comb Chem* 2003, **5**:693-753.

20. •• Bleicher KH, Bohm HJ, Muller K, Alanine AI: **Hit and lead generation: beyond high-throughput screening**. *Nat Rev Drug Discov* 2003, **2**:369-378.
A comprehensive review describing the current methods of transformation of initial hits into high-content lead series.

21. Rose S, Stevens A: **Computational design strategies for combinatorial libraries**. *Curr Opin Chem Biol* 2003, **7**:331-339.

22. Lengauer T, Lemmen C, Rarey M, Zimmermann M: **Novel technologies for virtual screening**. *Drug Discov Today* 2004, **9**:27-34.

23. Krumrine J, Raubacher F, Brooijmans N, Kuntz I: **Principles and methods of docking and ligand design**. *Methods Biochem Anal* 2003, **44**:443-476.

24. • Muller G: **Medicinal chemistry of target family-directed masterkeys**. *Drug Discov Today* 2003, **8**:681-691.
This review focuses on multipurpose privileged structures that address a variety of targets from a gene family of interest, irrespective of therapeutic area.

25. • Bondensgaard K, Ankersen M, Thogersen H, Hansen BS, Wulff BS, Bywater RP: **Recognition of privileged structures by g-protein coupled receptors**. *J Med Chem* 2004, **47**:888-899.
An important work demonstrating the correlation between conservation patterns of residues in GPCR ligand binding pocket and the privileged structural fragments.

26. Olesen PH: **The use of bioisosteric groups in lead optimization**. *Curr Opin Drug Discov Devel* 2001, **4**:471-478.

27. Xue L, Godden JW, Bajorath J: **Mini-fingerprints for virtual screening: design principles and generation of novel prototypes based on information theory**. *SAR QSAR Environ Res* 2003, **14**:27-40.

28. Merlot C, Domine D, Cleva C, Church DJ: **Chemical substructures in drug discovery**. *Drug Discov Today* 2003, **8**:594-602.

29. van Drie JH: **Pharmacophore discovery–lessons learned**. *Curr Pharm Des* 2003, **9**:1649-1664.

30. Cramer RD, Jilek RJ, Andrews KM: **Dbtop: topomer similarity searching of conventional structure databases**. *J Mol Graph Model* 2002, **20**:447-462.

31. Makara GM, Nash H, Zheng Z, Orminati JP, Wintner EA: **A reagent-based strategy for the design of large combinatorial libraries: a preliminary experimental validation**. *Mol Divers* 2003, **7**:3-14.

32. Jamois EA: **Reagent-based and product-based computational approaches in library design**. *Curr Opin Chem Biol* 2003, **7**:326-330.

33. Viswanadhan VN, Balan C, Hulme C, Cheetham JC, Sun Y: **Knowledge-based approaches in the design and selection of compound libraries for drug discovery**. *Curr Opin Drug Discov Devel* 2002, **5**:400-406.

34. • Stockfisch TP: **Partially unified multiple property recursive partitioning (PUMP-RP): a new method for predicting and understanding drug selectivity**. *J Chem Inf Comput Sci* 2003, **43**:1608-1613.
The authors demonstrate how data from an older, well-studied target can be used to study a newer target for which only a small amount of data are available.

35. Balakin KV, Lang SA, Skorenko AV, Tkachenko SE, Ivashchenko AA, Savchuk NP: **Structure-based versus property-based approaches in the design of G-protein-coupled receptor-targeted libraries**. *J Chem Inf Comput Sci* 2003, **43**:1553-1562.

36. Manallack DT, Pitt WR, Gancia E, Montana JG, Livingstone DJ, Ford MG, Whitley DC: **Selecting screening candidates for kinase and G-protein coupled receptor targets using neural networks**. *J Chem Inf Comput Sci* 2002, **42**:1256-1262.

37. • Schneider G, Nettekoven M: **Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps**. *J Comb Chem* 2003, **5**:233-237.
An interesting example demonstrating the possibility to design a small, activity-enriched focused library with an improved property profile using a virtual screening approach based on Kohonen maps.

38. Balakin K: **Pharma ex machina**. *Mod Drug Disc* 2003, **8**:45-47.

39. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV: **Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions**. *J Chem Inf Comput Sci* 2003, **43**:2048-2056.

40. Oprea TI: **Current trends in lead discovery: Are we looking for the appropriate properties?** *J Comp-Aid Drug Des* 2002, **16**:325-334.

41. Oprea TI: **Virtual screening in lead discovery: a viewpoint**. *Mol* 2002, **7**:51-62.

42. Zamora I, Oprea T, Cruciani G, Pastor M, Ungell A-L: **Surface descriptors for protein-ligand affinity prediction**. *J Med Chem* 2003, **46**:25-33.

43. ter Haar E, Walters WP, Pazhanisamy S, Taslimi P, Pierce AC, Bemis GW, Salituro FG, Harbeson S: **Kinase chemogenomics: targeting the human kinome for target validation and drug discovery**. *Mini Rev Med Chem* 2004, **4**:235-253.

44. Fradera X, Mestres J: **Guided docking approaches to structure-based design and screening**. *Curr Top Med Chem* 2004, **4**:687-700.