Pharmaceutical Research, Vol. 24, No. 10, October 2007 (© 2007)
DOI: 10.1007/s11095-007-9285-5

Expert Review

A Cheminformatic Toolkit for Mining Biomedical Knowledge

Gus R. Rosania,[1,6] Gordon Crippen,[2] Peter Woolf,[3] David States,[4] and Kerby Shedden[5]

Received January 11, 2007; accepted February 27, 2007; published online March 24, 2007

Purpose. Cheminformatics can be broadly defined to encompass any activity related to the application of information technology to the study of properties, effects and uses of chemical agents. One of the most important current challenges in cheminformatics is to allow researchers to search databases of biomedical knowledge, using chemical structures as input.
Materials and Methods. An important step towards this goal was the establishment of PubChem, an open, centralized database of small molecules accessible through the World Wide Web. While PubChem is primarily intended to serve as a repository for high throughput screening data from federally-funded screening centers and academic research laboratories, the major impact of PubChem could also reside in its ability to serve as a chemical gateway to biomedical databases such as PubMed.
Conclusion. This article will review cheminformatic tools that can be applied to facilitate annotation of PubChem through links to the scientific literature; to integrate PubChem with transcriptomic, proteomic, and metabolomic datasets; to incorporate results of numerical simulations of physiological systems into PubChem annotation; and ultimately, to translate data of chemical genomics screening efforts into information that will benefit biomedical researchers and physician scientists across all therapeutic areas.

KEY WORDS: bioactivity fingerprints; bioinformatics; chemical genetics; chemical genomics; chemical space; cheminformatics; data mining; high throughput screening; mathematical modeling; QSAR.

THE CHALLENGE OF APPLYING CHEMOINFORMATICS TO BIOMEDICAL KNOWLEDGE

A biomedical research scientist has just discovered a potential link between a biochemical pathway and a disease-causing mechanism. Is there a small molecule that may be used to modulate that biochemical pathway, so as to alter the course of the disease? With limited budget and resources, he would like to purchase the largest possible number of molecules that could potentially be used to interfere with the activity of protein targets along that biochemical pathway, while at the same time having the least effects on other biochemical pathways leading to unwanted side effects. Because the biochemical pathway of interest is localized in mitochondria and the goal is to discover molecules of potential therapeutic relevance, it is important that the molecules have high solubility and cellular permeability, while at the same time accumulating specifically in mitochondria without disrupting mitochondrial function. What are those molecules? Are they commercially available? The investigator logs on to a computer terminal linked to the latest cheminformatic search engine on the web. Within a couple of hours, he is able to find a set of 1,000 relevant, candidate molecules, and orders them from 50 vendors, throughout the world.

Inspired by this vision, this review aims to highlight many of the new cheminformatic tools that are being developed in academic laboratories working in the fields of chemical genomics and drug discovery. These cheminformatic tools constitute a major contribution of cheminformatics research to the biomedical research community in general. As applied to drug discovery, traditional cheminformatic research has aimed to facilitate the design, analysis or interpretation of experiments, and to test specific models or experimental hypothesis. As a result, traditional cheminformatic tools can only be used by experts with highly specialized, quantitative skills, which are not shared by bench scientist lacking advanced cheminformatics training. As applied to mining biomedical knowledge, an emerging new direction in cheminformatics research aims to make chemical information more accessible to biomedical researchers. Cheminformatic tools are needed to disseminate chemical information to a wide audience without requiring a deep understanding of chemistry or statistics. Indeed, major challenges for integrating information emerge at the interface

[1] Department of Pharmaceutical Sciences, University of Michigan College of Pharmacy, 428 Church Street, Ann Arbor, MI 48109, USA.
[2] Department of Medicinal Chemistry, University of Michigan College of Pharmacy, Ann Arbor, MI, USA.
[3] Department of Chemical Engineering, University of Michigan School of Engineering, Ann Arbor, MI, USA.
[4] Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA.
[5] Department of Statistics, University of Michigan, Ann Arbor, MI, USA.
[6] To whom correspondence should be addressed. (e-mail: grosania@umich.edu)

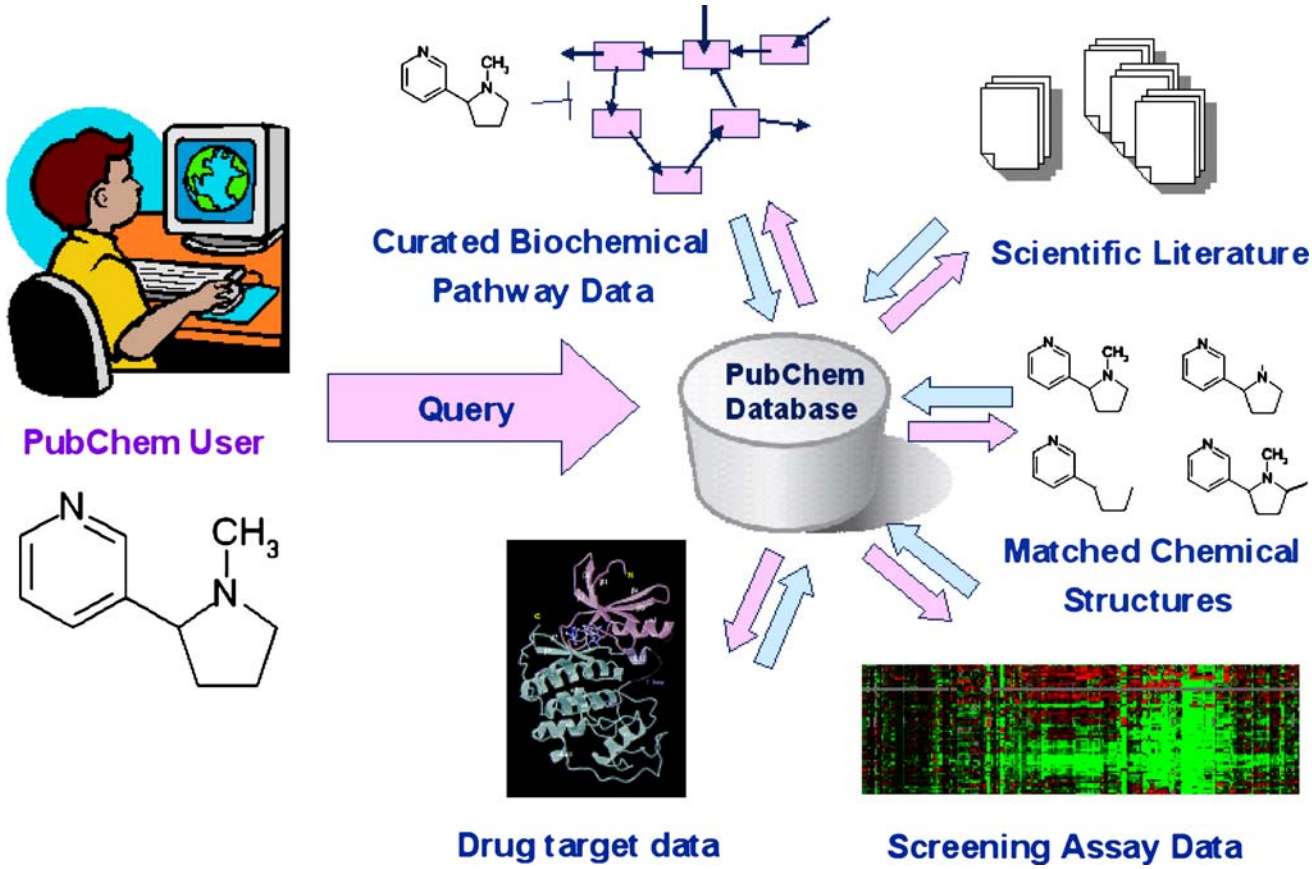

Fig. 1. A user can access the PubChem database to find not only compounds that are similar to a query structure, but also related research articles, protein structures, biochemical pathways, and screening assay data.

of chemistry, genomics, transcriptomics, proteomics, metabolomics and systems biology. Thus, instead of elaborating on cheminformatic tools that have been developed for drug discovery and are applicable to specific, project-oriented questions, this article will specifically concentrate on reviewing general purpose cheminformatic tools applicable to the integration of chemical and biological information.

TOOLS FOR MANAGING LARGE AMOUNTS OF DISPARATE DATA

Chemical genomics (also referred to as chemical genetics) aims to discover relationships between genes or proteins and cellular phenotypes by studying the influence of chemical agents on cell structure and function (1–10). This is unlike drug discovery, whose aim is to find chemical agents that achieve a therapeutic benefit in a patient population. Presently, cheminformatics is gaining momentum as a critical tool for chemical genomics research. There are various reasons for this: First, collections of diverse chemical agents can be readily synthesized and screened in academic research laboratories (4,11,12), creating a demand for methods to manage and disseminate screening results in a manner that can be readily accessed by the research community. Second, the development of computational tools to study activity data for small molecules held in large databases encompassing various assays and to identify relationships between chemical agents and their functional effects on physiological systems is technically feasible and timely (13–15). Third, the emerging field of metabolomics—comprising all endogenous small molecules occurring in living systems and their associated metabolic transformations—calls for the creation of computational tools to integrate, analyze and manage knowledge of biochemical reactions, pathways and networks, including the effect of exogenous chemical agents on those networks (13,16–25).

New cheminformatics tools are being developed so biomedical investigators are able to use chemical genomics data in a broad range of research projects, without the need for highly specialized resources or personnel. Cheminformatic tools will provide the guidance needed to link phenotypic outcomes to the chemical structures of molecules tested in biological assays (Fig. 1). Today, the search for small bioactive molecules is essentially a process of trial and error, whereby large collections of compounds are either chemically synthesized or isolated from natural sources, then experimentally tested for biochemical and cellular effects (9,10,26–28). In 2005, the Molecular Libraries Screening Center Network (MLSCN) was created as part of the NIH Roadmap Initiative (29). The MLSCN is a consortium of academic laboratories responsible for screening large libraries of compounds using biomedically-relevant cellular and biochemical assays (30). In parallel, the NIH also created a network of centers for cheminformatics research (31), to help develop software that will make chemical screening data more meaningful and relevant to the biomedical research community. Cheminformatics therefore will play an important role in integrating chemical structure and activity data

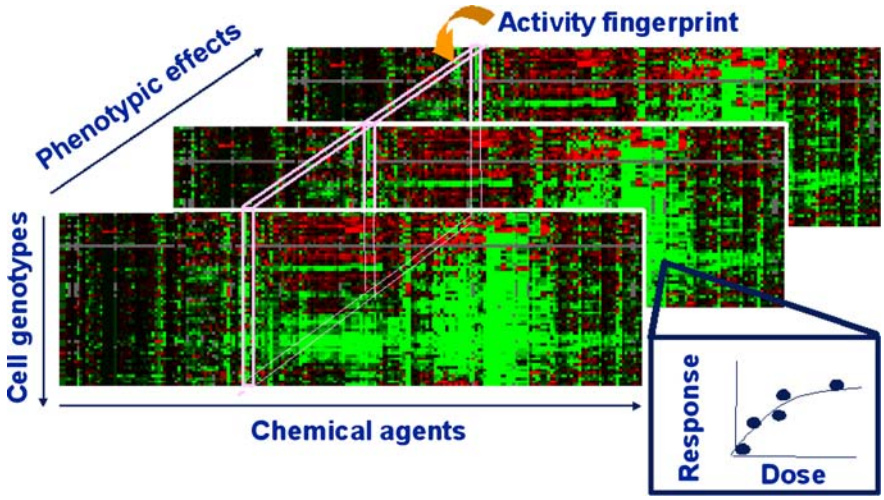

Fig. 2. Representation of activity data for many different compounds producing different phenotypic effects on a variety of cellular genotypes. Each dot represents the relative concentration at which a particular compound is active in a particular dose-response assay. Most active corresponds to *green*; least active is shown as *red*.

from the different screening centers, together with data deposited by individual investigators in academic laboratories or derived from published research articles.

Cheminformatics research centers will perform investigations whose results can be applied towards the (1) analysis, visualization, and interpretation of screening data (32,33); (2) efficient planning of new syntheses and assays (32,34–39); and (3) rapid advancement of screening results towards therapeutically relevant applications (40–47). Currently, the National Center of Biotechnology Information (NCBI) at NIH is developing the PubChem database of chemical structures and biological activities (48). PubChem will store all experimental screening results obtained by screening centers associated with the MLSCN. Data analysis methods and computational tools that can interface with PubChem as well as with other biomedically relevant databases such as PubMed will be important for linking screening results with the rest of scientific knowledge (15,49–51). Cheminformatic tools will be instrumental in identifying hits with useful activity patterns, as well as in the development of compounds with more potent and specific activities. While the 500,000–5,000,000 compounds to be screened may seem like a large number, this is a miniscule fraction of biologically relevant "chemical space" (52–54). However, by screening molecules possessing partially overlapping biological activities and substructural fragments, statistical analysis can be used to extract the features of the molecules that are most closely associated with different phenotypic activities and other annotated properties, thereby enhancing the information that can be gained from screening a limited set of molecules representing a small fraction of chemical space (33,53–59).

TOOLS FOR NAVIGATING BIOMEDICALLY-RELEVANT CHEMICAL SPACE

While cheminformatics has existed as an active field of study in medicinal chemistry, especially as practiced in the pharmaceutical industry, there are particular challenges posed by chemical genomics—such as the relationship between the chemical and functional diversity of molecules—that have little precedent in medicinal chemistry. In medicinal chemistry, quantitative structure-activity relationship (QSAR) analysis is applied to assays that have a single, well-defined endpoint and favored direction: improving the potency of a compound in an assay with a single phenotypic readout or increasing the binding affinity of a small molecule to its target (43,60–73). Building on the foundations of QSAR laid by C. Hansch and associates, thousands of QSAR models have been published over the past 40 years, and have been compiled into a single database (73). The QSAR approach is well-suited for pharmaceutical drug discovery, where candidate drug molecules need to be optimized so as to achieve a therapeutic benefit: for example, by increasing the intestinal permeability of a molecule that is otherwise not absorbed by the body. By contrast, when the intent is to study the effect of molecules on all biochemical pathways in a cell, the cheminformatic challenges center around the analysis of patterns of activity, also referred to as activity "fingerprints" or "profiles" (41,74–78) (Fig. 2). In this context, a useful cheminformatic tool should also be able to find links between genes, proteins or biochemical pathways differentially present in multiple cell lines with different genetic backgrounds, and relate these to the mechanism of action of small molecules (2,41,45–47,57,74–81).

Intended to be a single database of broad therapeutic relevance, PubChem will provide a framework for integrating experimental results obtained from molecules screened across multiple different assays. These assays will be relevant to various therapeutic applications, including but not limited to oncology, cardiovascular disease, infectious diseases, CNS disorders, immunity, metabolic disorders, and regenerative medicine. Accordingly, the chemical diversity encoded of molecules being screened by the MLSCN should be reflected in diverse activity fingerprints. Compounds should hit targets on biochemical pathways leading to different biological responses in different assays. One task of cheminformatic researchers therefore will be to analyze the diversity of activity fingerprints of MLSCN compounds and determine if the chemical diversity represented is well-suited for the intended

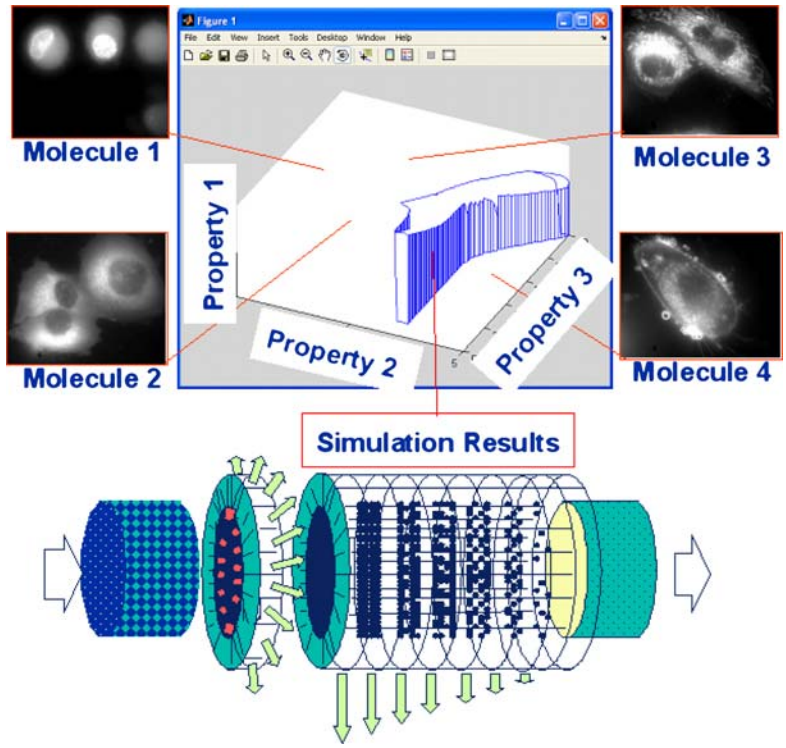

Fig. 3. The ability to visualize results of a mathematical simulation with image data obtained from screening experiments illustrates an important cheminformatic challenge of chemical genomics. In this example, image data acquired from cells incubated with four different fluorescent probes (Molecules 1, 2, 3 and 4) are graphed in a tridimensional, chemical property space where each axis represents a physicochemical characteristic of a molecule. Within this physicochemical property space, a computational simulation of intestinal absorption (represented by the cylinder at the bottom of the plot) is used to define combinations of physicochemical properties that should confer high permeability but low intracellular accumulation (indicated by the blue shaded region in the plot).

scope of the project. When the goal is to search for related compounds in a database or to cluster compounds with related chemical structures, chemical similarity and diversity can be thought of in terms of the shape of a molecule, the relative position of its constituent atoms, substructural motifs and topological descriptors (38,69,82–96). Such similarity metrics can be used for predicting the binding of small molecules to proteins based on the shape of the binding pocket (65,69,83,87,89,97–100). However, when the goal is to analyze the behavior of a collection of molecules on different cell based assays, the diversity of chemical agents can also be analyzed in terms of similarities or differences in the phenotypic responses of cells to small molecules across multiple assays (16,27,41,45–47,57–59,76–79,101–107) (Fig. 2). Thus, cheminformatic analysis of activity fingerprints will be important to determine if and how the collection of compounds being screened could be improved.

To relate phenotypic diversity to chemical diversity, cheminformatic analysis can incorporate genomic, transcriptomic, proteomic, metabolomic data sets—as well as systems biology and mathematical modeling components (2,44–46,74–79,108). The -omics and mathematical modeling components can lead to hypotheses about the effects of interfering with specific molecular targets in different biochemical pathways. Systems biology can provide the conceptual framework for understanding the relationship between the different phenotypic states of the cell at a molecular level (e.g. through mathematical modeling of mechanistic physiological phenomena, or through Bayesian, statistical analysis of causal networks of gene or protein expression patterns (109–117)). Mathematical modeling can also be used to predict important cellular variables related to transport (42,117–125) and distribution of small molecules (117,126,127) as well as the global effects of metabolic perturbations (16,20,128).

TOOLS FOR PREDICTING PHENOTYPIC ACTIVITY AND SPECIFICITY

Computational studies predicting drug activity and specificity have focused on studying and analyzing how the structure of a molecule determines the binding to a specific cellular target, and in turn, how the binding to the specific target promotes or inhibits its biological activity. Such cheminformatic tools are useful to “dock” a small molecule to the active site of a molecular target, for example, to predict the binding affinity. However, cells are structurally and functionally organized into organelles delimited by membranes, and all organelles are associated with specific physiological and biochemical signaling functions. Therefore, a molecule may exert phenotypic effects simply by accumulating in an organelle so as to perturb the function of that organelle. In melanocytes, for example, accumulation of small molecules in mitochondria can interfere with ATP generation and trigger a protective signal transduction response whereby cells increment the levels of pigment production (129–132). In contrast, accumulation of small molecules in mitochondria of cancer cells generally induces

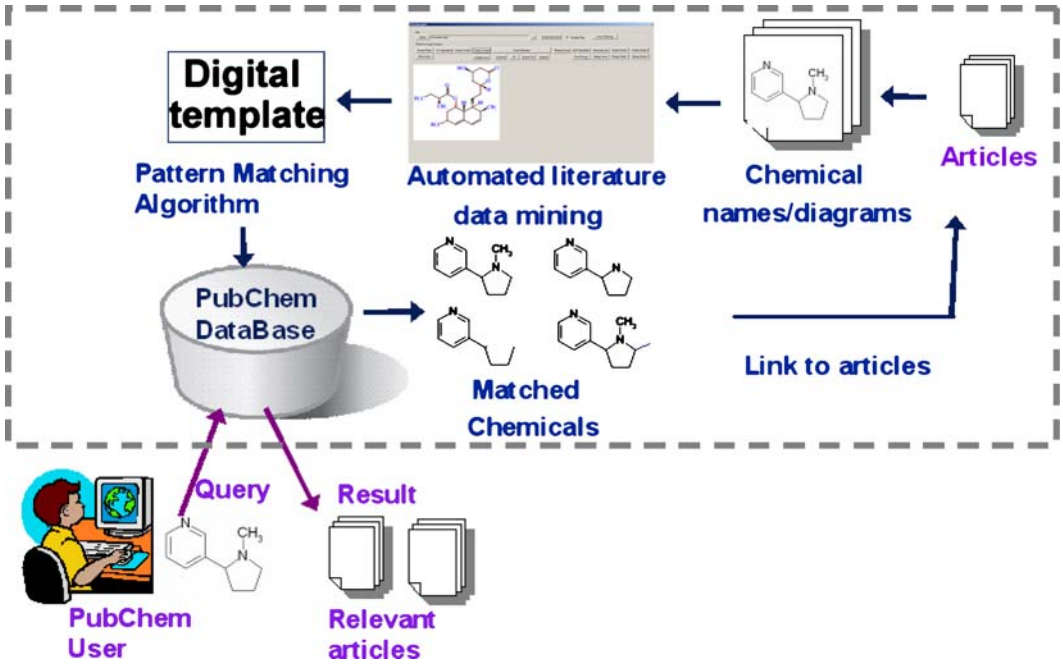

Fig. 4. Development of automated methods for linking the scientific literature to the Pubchem database can increase Pubchem's usefulness in biomedical research. From published scientific research articles, chemical names and structure diagrams can be extracted and converted into a digital template representing the structure of a chemical agent (or a partial substructure). In turn, the digital template can be linked to one or more Pubchem entries using digital pattern matching algorithms. Matched Pubchem entries can be linked to published research articles via html links to article abstracts in PubMed. A user would be able to query these articles simply by drawing a chemical structure in the PubChem search engine, without being aware of the sophisticated cheminformatics infrastructure that makes such a query possible.

apoptosis (103,133–138). Accordingly, knowledge of biochemical pathways associated with different organelles in different cell types is relevant to understanding the activity and specificity of chemical agents (3,7,10,81,129,139).

Statistical and computational approaches to predict how the structure of small molecules influences their intracellular distribution have a precedent in the study of dyes and fluorescent probes for histochemical staining. Quantitative structure-property relationships and mechanistic models have been developed for studying the intracellular distribution of molecules in different organelles inside cells (117,126,127,140–145), and for predicting the ability of small molecules to traverse cell monolayers and intracellular membranes (42,64,118–125,146–149). Many QSAR models for predicting membrane passage, intracellular accumulation and absorption have been published over the years, and can be compiled into a comprehensive database (73). Because of differences in pH in different intracellular compartments and transmembrane electrochemical gradients, membrane permeant hydrophobic molecules can reach different equilibrium concentrations at different intracellular locations, in a cell-type dependent manner (7,126,127,150,151). In addition, molecules can be substrates of active transport mechanisms, such as ATP-dependent xenobiotic transporters, driving the accumulation of small molecules in specific organelles or promoting their expulsion from the cell in a manner that affects the differential response of cells to small molecules (7,104,152–157). New types of data mining tools developed for chemical genomics can combine predictions about intracellular distribution and permeability of small molecules together with other types of data such as images or gene expression profiles to reveal associations between intracellular accumulation and bioactivity (Fig. 3) (81,158,159).

TOOLS FOR LINKING TO THE SCIENTIFIC LITERATURE

Recently, interest in automated, computational tools for organizing and mining the scientific literature has increased (49,160–168). Automated, natural language processing algorithms may be used for extracting information from PubMed (160,164–171) (Fig. 4). A similar approach can help link PubChem molecules to biochemical pathways and small molecule agonists and antagonists whose activities have been characterized and published by individual investigators. There have already been attempts to automate incorporation of small molecule activity data into biochemical pathway maps, and tools are being developed to expand, visualize and mine data from maps of chemical reaction networks through the internet (21–25,172–177). Cheminformatic efforts can build upon these first steps at integrating biochemical knowledge, particularly in the direction of linking PubChem molecules with metabolites and molecular targets.

Virtually all current knowledge about chemistry and biology is contained as natural language text in scientific research articles. In this context, development of machine vision and natural language processing algorithms will be relevant to cheminformatic efforts aiming to link PubChem molecules with biochemical pathway mapping resources (21,22,51,176,177). It is within the capabilities of current machine vision technology to convert analog, chemical structure diagrams into digital representations that can be

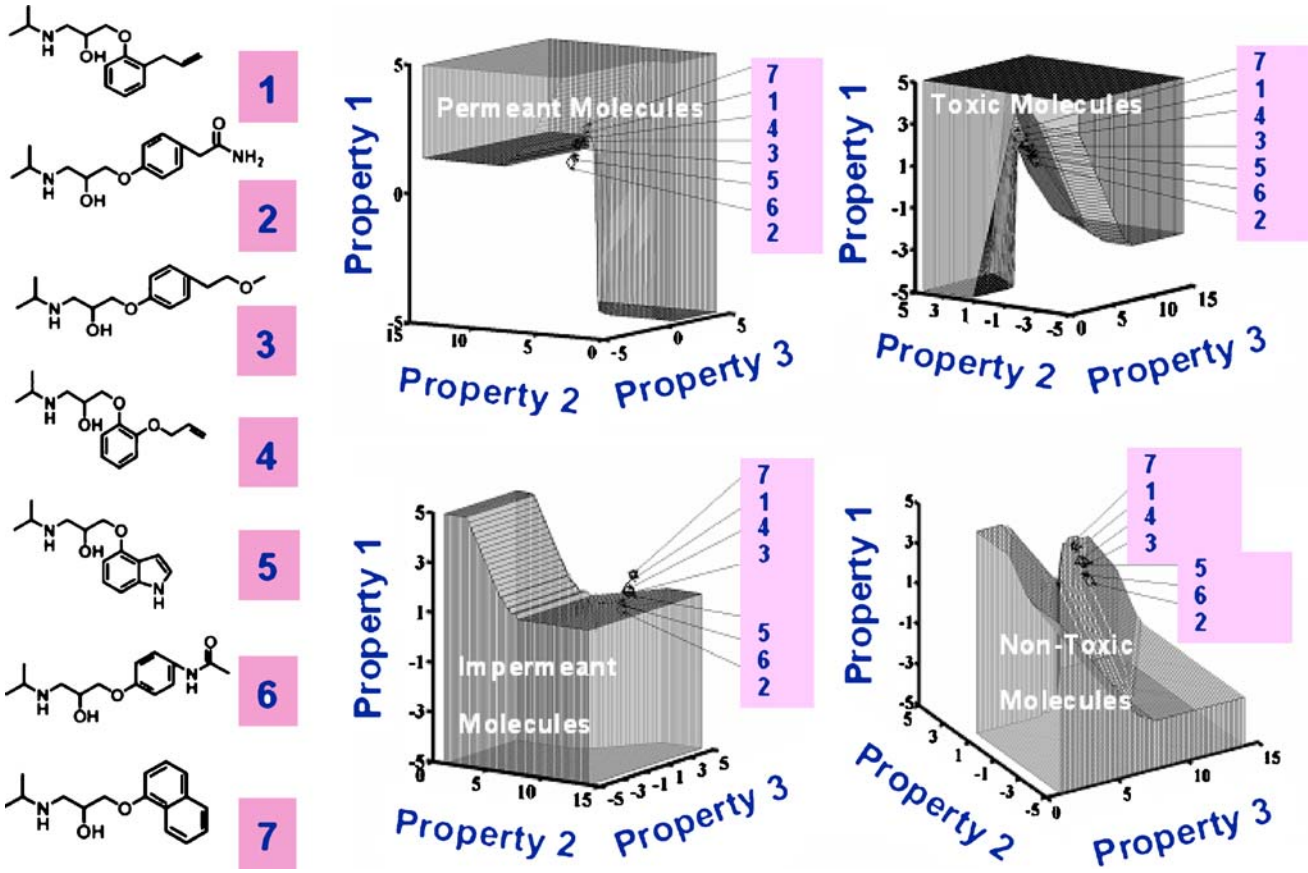

Fig. 5. Cheinformatic-assisted visualization tools can be used to facilitate simultaneous exploration of chemical features and related phenotypic effects of small molecules. Using mathematical simulations or QSAR models, it is possible to define regions of physicochemical property space occupied by cell permeant, cell impermeant, toxic and non-toxic molecules. The ability to see the chemical structures of a set of molecules (1–7) within this property space can facilitate selection of molecules with the most desirable features for specific experiments.

used for database annotation (162,178–180). Using mechanistic insights obtained from published research articles, it is possible to connect cellular responses to small molecules, gene expression patterns and drug mechanism of action (181). Natural language processing can be used to link molecular targets, biochemical pathways, cellular phenotypes to diseases (170). Graph-similarity analysis (102) can be explored to make predictions of molecule/phenotype interactions based on the small molecule/protein interactions by drawing inferences in the context of a catalogue of protein/protein assertions and biochemical pathways.

Cheminformatics researchers are also developing new, structured formats for sharing chemical information (50,95,161,182). Such structured formats would facilitate automated integration of Pubchem with other structured repositories of scientific knowlege. At present, PubMed can be linked to PubChem through manually-curated MeSH headings of PubMed articles (183,184). Future PubChem links that can be envisioned include The Protein Data Bank (PDB) and manually-curated protein-ligand interaction databases (26,79,185–187). Currently, molecular docking and pharmacophore similarity searching algorithms can be used to mine potentially-relevant interactions between small molecule ligands and cellular macromolecules (69,79,83,87–94,186–192). Thousands of molecular structures including proteins, nucleic acids and macromolecular complexes are already publicly available (193). Together with data on the binding and functional activity of small molecules on those targets, powerful cheminformatics tools may be developed to assess structural features of a molecule that determine differential affinity to different targets, and identify the topological features of a molecule that preclude or promote binding to a target or a set of targets (35,186–191,194,195).

TOOLS FOR VISUALIZING MULTIDIMENSIONAL RELATIONSHIPS

Many of the ongoing chemical genomic MLSCN screens are being performed with phenotypic cell-based assays run on "high content screening" instruments (196–200). These instruments generate large amounts of microscopic image data, capturing how the intensity and spatial distribution of a fluorescent biosensor changes in response to the activity of small molecules. High content screening therefore creates a need for computationally-intensive data management and analysis strategies (201–204). And, just like image data is multidimensional, the chemical space of the molecules being screened is also multidimensional: any collection of molecules can be sorted based on molecular weight, logP, pK_a, permeability, toxicity, activity, subcellular distribution, stability, solubility, or degree of similarity to other molecules (13,27). Within this multidimensional chemical-image-assay space, molecules with certain desirable characteristics may exist in regions with complex shapes (117) (Figs. 3, 5). These

regions may form a continuous chunk of chemical space, but they could also be present as discontinuous islands or pockets. The ability to navigate multidimensional chemical space and search for high content data associated with specific molecules, as well as the ability to visualize phenotypic data and its relationship to chemical structures of the molecules being screened, is yet another challenge posed by chemical genomics for which cheminformatics tools are being developed (117) (Fig. 5).

CONCLUSION

To conclude, this review highlights some unique cheminformatic challenges associated with mining biomedical knowledge, and various tools that are being developed to solve them. As cheminformatics evolves in tandem with chemical genomics research, more scientists will become fluent with the new cheminformatic tools; this will lead, in turn, to even more powerful tools. An effective cheminformatics infrastructure must be able to manage the vast amount of present knowledge, and to disseminate this knowledge in a manner that is biomedically-relevant and useful to researchers in different fields. Thus, there is an important educational component to cheminformatics research (205,206). Cheminformatics research centers can help build an adaptable infrastructure that can accommodate new generations of investigators looking at problems from very different angles (50,173,207). Therefore, the ultimate solution to the cheminformatic challenges associated with mining biomedical knowledge will depend on new types of computer hardware, software, network systems, and immersive and interactive data management and visualization technology, as much as it will depend on incremental improvements to the more familiar cheminformatic methodologies such as QSAR and molecular modeling

ACKNOWLEDGEMENTS

This work was supported by NIH grants RO1-GM078200 and P20-HG003890 to G.R.R and K.S. We would like to thank Kazu Saitou, Jungkap Park and Xinyuan Zhang for help with the illustrations and graphics.

REFERENCES

1. B. K. Wagner, S. J. Haggarty, and P. A. Clemons Chemical genomics: probing protein function using small molecules. *Am. J. Pharmacogenomics.* **4**:313–320 (2004).
2. R. A. Butcher and S. L. Schreiber Using genome-wide transcriptional profiling to elucidate small-molecule mechanism. *Curr. Opin. Chem. Biol.* **9**:25–30 (2005).
3. H. S. Moon, E. M. Jacobson, S. M. Khersonsky, M. R. Luzung, D. P. Walsh, W. Xiong, J. W. Lee, P.B. Parikh, J. C. Lam, T. W. Kang, G. R. Rosania, A. F. Schier, and Y. T. Chang A novel microtubule destabilizing entity from orthogonal synthesis of triazine library and zebrafish embryo screening. *J. Am. Chem. Soc.* **124**:11608–11609 (2002).
4. B. R. Stockwell Exploring biology with small organic molecules. *Nature* **432**:846–854 (2004).
5. G. R. Rosania, J. Merlie, Jr., N. Gray, Y. T. Chang, P. G. Schultz, and R. Heald A cyclin-dependent kinase inhibitor inducing cancer cell differentiation: biochemical identification using Xenopus egg extracts. *Proc. Natl. Acad. Sci. U. S. A.* **96**:4797–4802 (1999).
6. G. R. Rosania, Y. T. Chang, O. Perez, D. Sutherlin, H. Dong, D. J. Lockhart, and P. G. Schultz Myoseverin, a microtubule-binding molecule with novel cellular effects. *Nat. Biotechnol.* **18**:304–308 (2000).
7. G. R. Rosania Supertargeted chemistry: identifying relationships between molecular structures and their sub-cellular distribution. *Curr. Top Med. Chem.* **3**:659–685 (2003).
8. O. D. Perez, Y. T. Chang, G. Rosania, D. Sutherlin, and P. G. Schultz Inhibition and reversal of myogenic differentiation by purine-based microtubule assembly inhibitors. *Chem. Biol.* **9**:475–483 (2002).
9. Y. T. Chang, S. M. Wignall, G. R. Rosania, N. S. Gray, S. R. Hanson, A. I. Su, J. Merlie, Jr., H. S. Moon, S. B. Sangankar, O. Perez, R. Heald, and P. G. Schultz Synthesis and biological evaluation of myoseverin derivatives: microtubule assembly inhibitors. *J. Med. Chem.* **44**:4497–4500 (2001).
10. Y. T. Chang, N. S. Gray, G. R. Rosania, D. P. Sutherlin, S. Kwon, T. C. Norman, R. Sarohia, M. Leost, L. Meijer, and P. G. Schultz Synthesis and application of functionally diverse 2,6,9-trisubstituted purine libraries as CDK inhibitors. *Chem. Biol.* **6**:361–375 (1999).
11. I. Smukste and B. R. Stockwell Advances in chemical genetics. *Annu. Rev. Genomics. Hum. Genet.* **6**:261–286 (2005).
12. S. L. Schreiber, K. C. Nicolaou, and K. Davies Diversity-oriented organic synthesis and proteomics. New frontiers for chemistry & biology. *Chem. Biol.* **9**:1–2 (2002).
13. Y. K. Kim, M. A. Arai, T. Arai, J. O. Lamenzo, E. F. Dean, , 3rd, N. Patterson, P. A. Clemons, and S. L. Schreiber Relationship of stereochemical and skeletal diversity of small molecules to cellular measurement space. *J. Am. Chem. Soc.* **126**:14740–14745 (2004).
14. J. Klekota, E. Brauner, F. P. Roth, and S. L. Schreiber Using high-throughput screening data to discriminate compounds with single-target effects from those with side effects. *J. Chem. Inf. Model* **46**:1549–1562 (2006).
15. C. P. Austin The completed human genome: implications for chemical biology. *Curr. Opin. Chem. Biol.* **7**:511–515 (2003).
16. A. Ramanathan, C. Wang, and S. L. Schreiber Perturbational profiling of a cell-line model of tumorigenesis by using metabolic measurements. *Proc. Natl. Acad. Sci. U. S. A.* **102**:5992–5997 (2005).
17. K. Dettmer, P. A. Aronov, and B. D. Hammock Mass spectrometry-based metabolomics. *Mass. Spectrom. Rev.* **26** (1): 51–78 (2007).
18. K. Hollywood, D. R. Brison, and R. Goodacre Metabolomics: current technologies and future trends. *Proteomics* **6**:4716–4723 (2006).
19. N. Schauer and A. R. Fernie Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci.* **11**:508–516 (2006).
20. O. Ebenhoh, T. Handorf, and R. Heinrich Structural analysis of expanding metabolic networks. *Genome. Inform. Ser. Workshop Genome. Inform.* **15**:35–45 (2004).
21. S. Goto, T. Nishioka, and M. Kanehisa LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* **28**:380–382 (2000).
22. S. Goto, Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30**:402–404 (2002).
23. D. C. McShan, M. Updadhayaya, and I. Shah. Symbolic inference of xenobiotic metabolism. *Pac. Symp. Biocomput.* 545–556 (2004).
24. M. Reitz, A. Homeyer, and J. Gasteiger Query generation to search for inhibitors of enzymatic reactions. *J. Chem. Inf. Model* **46**:2333–2341 (2006).
25. E. Selkov, Jr., Y. Grechkin, N. Mikhailova, and E. Selkov MPW: the Metabolic Pathways Database. *Nucleic Acids Res.* **26**:43–45 (1998).
26. J. Liu, X. Wu, B. Mitchell, C. Kintner, S. Ding, and P. G. Schultz A Small-Molecule Agonist of the Wnt Signaling Pathway. *Angew Chem. Int. Ed. Engl.* **44** (13): 1987–1990 (2005).
27. J. S. Melnick, J. Janes, S. Kim, J. Y. Chang, D. G. Sipes, D. Gunderson, L. Jarnes, J. T. Matzen, M. E. Garcia, T. L. Hood,

R. Beigi, G. Xia, R. A. Harig, H. Asatryan, S. F. Yan, Y. Zhou, X. J. Gu, A. Saadat, V. Zhou, F. J. King, C. M. Shaw, A. I. Su, R. Downs, N. S. Gray, P. G. Schultz, M. Warmuth, and J. S. Caldwell An efficient rapid system for profiling the cellular activities of molecular libraries. *Proc. Natl. Acad. Sci. U. S. A.* **103**:3153–3158 (2006).
28. M. Warashina, K. H. Min, T. Kuwabara, A. Huynh, F. H. Gage, P. G. Schultz, and S. Ding A synthetic small molecule that induces neuronal differentiation of adult hippocampal neural progenitor cells. *Angew Chem. Int. Ed. Engl.* **45**:591–593 (2006).
29. E. Zerhouni Medicine. The NIH Roadmap. *Science* **302**:63–72 (2003).
30. http://grants1.nih.gov/grants/guide/rfa-files/RFA-RM-04-017.html. MLSCN.
31. http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-05-012.html. RFA ECCR.
32. T. I. Oprea, J. Gottfries, V. Sherbukhin, P. Svensson, and T. C. Kuhler Chemical information management in drug discovery: optimizing the computational and combinatorial chemistry interfaces. *J. Mol. Graph. Model* **18**:512–524, 541 (2000).
33. P. Japertas, R. Didziapetris, and A. Petrauskas Fragmental methods in the analysis of biological activities of diverse compound sets. *Mini. Rev. Med. Chem.* **3**:797–808 (2003).
34. J. Bajorath Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries. *J. Comput. Aided. Mol. Des.* **16**:431–439 (2002).
35. J. M. Shin and D. H. Cho PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res.* **33**:D238–D241 (2005).
36. B. A. Tounge, L. B. Pfahler, and C. H. Reynolds Chemical information based scaling of molecular descriptors: a universal chemical scale for library design and analysis. *J. Chem. Inf. Comput. Sci.* **42**:879–884 (2002).
37. A. Tropsha and W. Zheng Rational principles of compound selection for combinatorial library design. *Comb. Chem. High. Throughput. Screen* **5**:111–123 (2002).
38. P. Willett Chemoinformatics—similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* **11**:85–88 (2000).
39. C. H. Reynolds, A. Tropsha, L. B. Pfahler, R. Druker, S. Chakravorty, G. Ethiraj, and W. Zheng Diversity and coverage of structural sublibraries selected using the SAGE and SCA algorithms. *J. Chem. Inf. Comput. Sci.* **41**:1470–1477 (2001).
40. M. Alvarez, R. Robey, V. Sandor, K. Nishiyama, Y. Matsumoto, K. Paull, S. Bates, and T. Fojo Using the national cancer institute anticancer drug screen to assess the effect of MRP expression on drug sensitivity profiles. *Mol. Pharmacol.* **54**:802–814 (1998).
41. S. A. Amundson, T. G. Myers, D. Scudiero, S. Kitada, J. C. Reed, and A. J. Fornace, Jr. An informatics approach identifying markers of chemosensitivity in human cancer cell lines. *Cancer Res.* **60**:6101–6110 (2000).
42. S. Tavelin, J. Taipalensuu, L. Soderberg, R. Morrison, S. Chong, and P. Artursson Prediction of the oral absorption of low-permeability drugs using small intestine-like 2/4/A1 cell monolayers. *Pharm. Res.* **20**:397–405 (2003).
43. W. J. Egan and G. Lauri Prediction of intestinal permeability. *Adv. Drug Deliv. Rev.* **54**:273–289 (2002).
44. R. Huang, A. Wallqvist, and D. G. Covell Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. *Genomics* **87**:315–328 (2006).
45. R. Huang, A. Wallqvist, N. Thanki, and D. G. Covell Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute's anticancer screen and drug mechanism of action. *Pharmacogenomics J.* **5**:381–399 (2005).
46. A. D. Koutsoukos, L. V. Rubinstein, D. Faraggi, R. M. Simon, S. Kalyandrug, J. N. Weinstein, K. W. Kohn, and K. D. Paull Discrimination techniques applied to the NCI *in vitro* antitumour drug screen: predicting biochemical mechanism of action. *Stat. Med.* **13**:719–730 (1994).
47. J. N. Weinstein, T. G. Myers, P. M. O'Connor, S. H. Friend, A. J. Fornace, K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. van Osdol, A. P. Monks, D. A. Scudiero, E. A. Sausville, D. W. Zaharevitz, B. Bunow, V. N. Viswanadhan, G. S. Johnson, R. E. Wittes, and K. D. Paull An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**:343–349 (1997).
48. http://pubchem.ncbi.nlm.nih.gov. PubChem.
49. H. Shatkay and R. Feldman Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.* **10**:821–855 (2003).
50. R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E. L. Willighagen The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model* **46**:991–998 (2006).
51. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**:D668–D672 (2006).
52. T. I. Oprea, I. Zamora, and A. L. Ungell Pharmacokinetically based mapping device for chemical space navigation. *J. Comb. Chem.* **4**:258–266 (2002).
53. T. I. Oprea and J. Gottfries Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **3**:157–166 (2001).
54. N. P. Savchuk, K. V. Balakin, and S. E. Tkachenko Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr. Opin. Chem. Biol.* **8**:412–417 (2004).
55. R. D. Cramer, R. J. Jilek, and K. M. Andrews Dbtop: topomer similarity searching of conventional structure databases. *J. Mol. Graph. Model* **20**:447–462 (2002).
56. C. Merlot, D. Domine, and D. J. Church Fragment analysis in small molecule discovery. *Curr. Opin. Drug Discov. Devel.* **5**:391–399 (2002).
57. O. Keskin, I. Bahar, R. L. Jernigan, J. A. Beutler, R. H. Shoemaker, E. A. Sausville, and D. G. Covell Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure. *Anticancer Drug Des.* **15**:79–98 (2000).
58. W. G. Rice, J. A. Turpin, C. A. Schaeffer, L. Graham, D. Clanton, R. W. Buckheit, Jr., D. Zaharevitz, M. F. Summers, A. Wallqvist, and D. G. Covell Evaluation of selected chemotypes in coupled cellular and molecular target-based screens identifies novel HIV-1 zinc finger inhibitors. *J. Med. Chem.* **39**:3606–3616 (1996).
59. A. Wallqvist, R. Huang, N. Thanki, and D. G. Covell Evaluating chemical structure similarity as an indicator of cellular growth inhibition. *J. Chem. Inf. Model* **46**:430–437 (2006).
60. G. Cianchetta, Y. Li, J. Kang, D. Rampe, A. Fravolini, G. Cruciani, and R. J. Vaz Predictive models for hERG potassium channel blockers. *Bioorg. Med. Chem. Lett.* **15**:3637–3642 (2005).
61. G. Cianchetta, R. W. Singleton, M. Zhang, M. Wildgoose, D. Giesing, A. Fravolini, G. Cruciani, and R. J. Vaz A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J. Med. Chem.* **48**:2927–2935 (2005).
62. P. de Cerqueira Lima, A. Golbraikh, S. Oloff, Y. Xiao, and A. Tropsha Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model* **46**:1245–1254 (2006).
63. C. Hansch, A. Leo, and D. H. Hoekman. Exploring QSAR, American Chemical Society, Washington, DC, 1995.
64. T. J. Hou, W. Zhang, K. Xia, X. B. Qiao, and X. J. Xu ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *J. Chem. Inf. Comput. Sci.* **44**:1585–1600 (2004).
65. S. Oloff, R. B. Mailman, and A. Tropsha Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J. Med. Chem.* **48**:7322–7332 (2005).
66. C. M. Breneman, C. M. Sundling, N. Sukumar, L. Shen, W. P. Katt, and M. J. Embrechts New developments in PEST shape/property hybrid descriptors. *J. Comput. Aided Mol. Des.* **17**:231–240 (2003).
67. J. R. Votano, M. Parham, L. M. Hall, L. H. Hall, L. B. Kier, S. Oloff, and A. Tropsha QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing

Structure-Information Representation. *J. Med. Chem.* **49**:7169–7181 (2006).
68. A. Golbraikh, M. Shen, Z. Xiao, Y. D. Xiao, K. H. Lee, and A. Tropsha Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* **17**:241–253 (2003).
69. S. Oloff, S. Zhang, N. Sukumar, C. Breneman, and A. Tropsha Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). *J. Chem. Inf. Model.* **46**:844–851 (2006).
70. R. Guha, J. R. Serra, and P. C. Jurs Generation of QSAR sets with a self-organizing map. *J. Mol. Graph. Model* **23**:1–14 (2004).
71. R. Guha and P. C. Jurs Interpreting computational neural network QSAR models: a measure of descriptor importance. *J. Chem. Inf. Model* **45**:800–806 (2005).
72. R. Guha and P. C. Jurs Determining the validity of a QSAR model—a classification approach. *J. Chem. Inf. Model* **45**:65–73 (2005).
73. A. Kurup C-QSAR: a database of 18,000 QSARs and associated biological and physical data. *J. Comput. Aided Mol. Des.* **17**:187–196 (2003).
74. U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **24**:236–244 (2000).
75. K. Shedden and G. R. Rosania Exploratory chemoinformatic analysis of cell type-selective anticancer drug targeting. *Mol. Pharm.* **1**:267–280 (2004).
76. J. E. Staunton, D. K. Slonim, H. A. Coller, P. Tamayo, M. J. Angelo, J. Park, U. Scherf, J. K. Lee, W. O. Reinhold, J. N. Weinstein, J. P. Mesirov, E. S. Lander, and T. R. Golub Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. U. S. A.* **98**:10787–10792 (2001).
77. A. Wallqvist, A. A. Rabow, R. H. Shoemaker, E. A. Sausville, and D. G. Covell Establishing connections between microarray expression data and chemotherapeutic cancer pharmacology. *Mol. Cancer Ther.* **1**:311–320 (2002).
78. A. Wallqvist, A. A. Rabow, R. H. Shoemaker, E. A. Sausville, and D. G. Covell Linking the growth inhibition response from the National Cancer Institute's anticancer screen to gene expression levels and other molecular target data. *Bioinformatics* **19**:2212–2224 (2003).
79. D. G. Covell, A. Wallqvist, R. Huang, N. Thanki, A. A. Rabow, and X. J. Lu Linking tumor cell cytotoxicity to mechanism of drug action: an integrated analysis of gene expression, small-molecule screening and structural databases. *Proteins* **59**:403–433 (2005).
80. M. Monga and E. A. Sausville Developmental therapeutics program at the NCI: molecular target and drug discovery process. *Leukemia* **16**:520–526 (2002).
81. K. Shedden, J. Brumer, Y. T. Chang, and G. R. Rosania Chemoinformatic analysis of a supertargeted combinatorial library of styryl molecules. *J. Chem. Inf. Comput. Sci.* **43**:2068–2080 (2003).
82. P. J. Artymiuk, P. A. Bath, H. M. Grindley, C. A. Pepperrell, A. R. Poirrette, D. W. Rice, D. A. Thorner, D. J. Wild, P. Willett, and F. H. Allen, *et al*. Similarity searching in databases of three-dimensional molecules and macromolecules. *J. Chem. Inf. Comput. Sci.* **32**:617–630 (1992).
83. M. Bradley, W. Richardson, and G. M. Crippen Deducing molecular similarity using Voronoi binding sites. *J. Chem. Inf. Comput. Sci.* **33**:750–755 (1993).
84. R. Bruschweiler Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation. *Proteins* **50**:26–34 (2003).
85. G. Cruciani, M. Pastor, and R. Mannhold Suitability of molecular descriptors for database mining. A comparative analysis. *J. Med. Chem.* **45**:2685–2694 (2002).
86. G. M. Maggiora and V. Shanmugasundaram Molecular similarity measures. *Methods Mol. Biol.* **275**:1–50 (2004).
87. P. Willett Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**:1046–1053 (2006).
88. P. Willett Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **48**:4183–4199 (2005).
89. P. Willett Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **31**:603–606 (2003).
90. A. Schuffenhauer, V. J. Gillet, and P. Willett Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **40**:295–307 (2000).
91. N. J. Richmond, P. Willett, and R. D. Clark Alignment of three-dimensional molecules using an image recognition algorithm. *J. Mol. Graph. Model* **23**:199–209 (2004).
92. N. J. Richmond, C. A. Abrams, P. R. Wolohan, E. Abrahamian, P. Willett, and R. D. Clark GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput. Aided Mol. Des.* **20**:567–587 (2006).
93. J. W. Raymond, C. J. Blankley, and P. Willett Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *J. Mol. Graph. Model* **21**:421–433 (2003).
94. N. E. Jewell, D. B. Turner, P. Willett, and G. J. Sexton Automatic generation of alignments for 3D QSAR analyses. *J. Mol. Graph. Model* **20**:111–121 (2001).
95. S. J. Edgar, J. D. Holliday, and P. Willett Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *J. Mol. Graph. Model* **18**:343–357 (2000).
96. W. Deng, C. Breneman, and M. J. Embrechts Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J. Chem. Inf. Comput. Sci.* **44**:699–703 (2004).
97. A. K. Ghose and G. M. Crippen Modeling the benzodiazepine receptor binding site by the general three-dimensional structure-directed quantitative structure-activity relationship method REMOTEDISC. *Mol. Pharmacol.* **37**:725–734 (1990).
98. L. Boulu and G. M. Crippen Voronoi receptor site models. *Prog. Clin. Biol. Res.* **289**:267–277 (1989).
99. A. S. Smellie, G. M. Crippen, and W. G. Richards Fast drug-receptor mapping by site-directed distances: a novel method of predicting new pharmacological leads. *J. Chem. Inf. Comput. Sci.* **31**:386–392 (1991).
100. C. A. Parks, G. M. Crippen, and J. G. Topliss The measurement of molecular diversity by receptor site interaction simulation. *J. Comput. Aided Mol. Des.* **12**:441–449 (1998).
101. L. Hodes, K. Paull, A. Koutsoukos, and L. Rubinstein Exploratory data analytic techniques to evaluate anticancer agents screened in a cell culture panel. *J. Biopharm. Stat.* **2**:31–48 (1992).
102. Y. Tian, R. C. McEachin, C. Santos, D. J. States, and J. M. Patel. SAGA: a subgraph matching tool for biological graphs. *Bioinformatics* **23**(2):232–239 (2007).
103. M. Kawakami, K. Koya, T. Ukai, N. Tatsuta, A. Ikegawa, K. Ogawa, T. Shishido, and L. B. Chen Synthesis and evaluation of novel rhodacyanine dyes that exhibit antitumor activity. *J. Med. Chem.* **40**:3151–3160 (1997).
104. Y. Huang, P. Anderle, K. J. Bussey, C. Barbacioru, U. Shankavaram, Z. Dai, W. C. Reinhold, A. Papp, J. N. Weinstein, and W. Sadee Membrane transporters and channels: role of the transportome in cancer chemosensitivity and chemoresistance. *Cancer Res.* **64**:4294–4301 (2004).
105. T. J. Mitchison Small-molecule screening and profiling by using automated microscopy. *Chembiochem* **6**:33–39 (2005).
106. Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler Multidimensional drug profiling by automated microscopy. *Science* **306**:1194–1198 (2004).
107. J. N. Weinstein Integromic Analysis of the NCI-60 Cancer Cell Lines. *Breast Dis.* **19**:11–22 (2004).
108. T. Yamori. Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. *Cancer Chemother. Pharmacol.* **52**(Suppl 1):S74–S79 (2003).
109. S. Ekins, Y. Nikolsky, A. Bugrim, E. Kirillov, and T. Nikolskaya Pathway mapping tools for analysis of high content data. *Methods Mol. Biol.* **356**:319–350 (2007).
110. A. Rzhetsky, T. Koike, S. Kalachikov, S. M. Gomez, M. Krauthammer, S. H. Kaplan, P. Kra, J. J. Russo, and C.

Friedman A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics* **16**:1120–1128 (2000).
111. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 422–433 (2001).
112. S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac. Symp. Biocomput.* 175–186 (2002).
113. S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *J. Bioinform. Comput. Biol.* **2**:77–98 (2004).
114. K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger. Bayesian network approach to cell signaling pathway modeling. *Sci. STKE* **2002**:PE38 (2002).
115. P. J. Woolf, W. Prudhomme, L. Daheron, G. Q. Daley, and D. A. Lauffenburger. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics* **21**(6):741–753 (2005).
116. Y. Xia, H. Yu, R. Jansen, M. Seringhaus, S. Baxter, D. Greenbaum, H. Zhao, and M. Gerstein Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.* **73**:1051–1087 (2004).
117. X. Zhang, K. Shedden, and G. R. Rosania A cell-based molecular transport simulator for pharmacokinetic prediction and cheminformatic exploration. *Mol. Pharm.* **3**:704–716 (2006).
118. V. Tantishaiyakul Prediction of Caco-2 cell permeability using partial least squares multivariate analysis. *Pharmazie* **56**:407–411 (2001).
119. H. H. Refsgaard, B. F. Jensen, P. B. Brockhoff, S. B. Padkjaer, M. Guldbrandt, and M. S. Christensen In silico prediction of membrane permeability from calculated molecular parameters. *J. Med. Chem.* **48**:805–811 (2005).
120. J. T. Penniston, L. Beckett, D. L. Bentley, and C. Hansch Passive permeation of organic compounds through biological tissue: a non-steady-state theory. *Mol. Pharmacol.* **5**:333–341 (1969).
121. K. Palm, K. Luthman, J. Ros, J. Grasjo, and P. Artursson Effect of molecular charge on intestinal epithelial drug transport: pH-dependent transport of cationic drugs. *J. Pharmacol. Exp. Ther.* **291**:435–443 (1999).
122. D. A. Norris, G. D. Leesman, P. J. Sinko, and G. M. Grass Development of predictive pharmacokinetic simulation models for drug discovery. *J. Control. Release* **65**:55–62 (2000).
123. Y. Marrero Ponce, M. A. Cabrera Perez, V. Romero Zaldivar, H. Gonzalez Diaz, and F. Torrens A new topological descriptors based model for predicting intestinal epithelial transport of drugs in Caco-2 cell culture. *J. Pharm. Pharm. Sci.* **7**:186–199 (2004).
124. S. Fujiwara, F. Yamashita, and M. Hashida Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. *Int. J. Pharm.* **237**:95–105 (2002).
125. G. Camenisch, J. Alsenz, H. van de Waterbeemd, and G. Folkers Estimation of permeability by passive diffusion through Caco-2 cell monolayers using the drugs' lipophilicity and molecular weight. *Eur. J. Pharm. Sci.* **6**:317–324 (1998).
126. S. Trapp and R. W. Horobin A predictive model for the selective accumulation of chemicals in tumor cells. *European Biophysics Journal With Biophysics Letters* **34**:959–966 (2005).
127. S. Trapp Plant uptake and transport models for neutral and ionic chemicals. *Environ. Sci. Pollut. Res.* **11**:33–39 (2004).
128. M. E. Andersen Toxicokinetic modeling and its applications in chemical risk assessment. *Toxicol. Lett.* **138**:9–27 (2003).
129. G. R. Rosania Mitochondria give cells a tan. *Chem. Biol.* **12**:412–413 (2005).
130. D. Williams, D. W. Jung, S. M. Khersonsky, N. Heidary, Y. T. Chang, and S. J. Orlow Identification of compounds that bind mitochondrial F1F0 ATPase by screening a triazine library for correction of albinism. *Chem. Biol.* **11**:1251–1259 (2004).
131. D. W. Jung, D. Williams, S. M. Khersonsky, T. W. Kang, N. Heidary, Y. T. Chang, and J. S. Orlow Identification of the F1F0 mitochondrial ATPase as a target for modulating skin pigmentation by screening a tagged triazine library in zebrafish. *Mol. Biosyst.* **1**:85–92 (2005).
132. J. R. Snyder, A. Hall, L. Ni-Komatsu, S. M. Khersonsky, Y. T. Chang, and S. J. Orlow Dissection of melanogenesis with small molecules identifies prohibitin as a regulator. *Chem. Biol.* **12**:477–484 (2005).
133. V. R. Fantin and P. Leder F16, a mitochondriotoxic compound, triggers apoptosis or necrosis depending on the genetic background of the target carcinoma cell. *Cancer Res.* **64**:329–336 (2004).
134. V. R. Fantin, M. J. Berardi, L. Scorrano, S. J. Korsmeyer, and P. Leder A novel mitochondriotoxic small molecule that selectively inhibits tumor cell growth. *Cancer Cell* **2**:29–42 (2002).
135. J. S. Modica-Napolitano and J. R. Aprille Delocalized lipophilic cations selectively target the mitochondria of carcinoma cells. *Adv. Drug Deliv. Rev.* **49**:63–70 (2001).
136. A. Manetta, G. Gamboa, A. Nasseri, Y. D. Podnos, D. Emma, G. Dorion, L. Rawlings, P. M. Carpenter, A. Bustamante, J. Patel, and D. Rideout Novel phosphonium salts display *in vitro* and *in vivo* cytotoxic activity against human ovarian cancer cell lines. *Gynecol. Oncol.* **60**:203–212 (1996).
137. T. J. Lampidis, S. D. Bernal, I. C. Summerhayes, and L. B. Chen Selective toxicity of rhodamine 123 in carcinoma cells in vitro. *Cancer Res.* **43**:716–720 (1983).
138. P. Costantini, E. Jacotot, D. Decaudin, and G. Kroemer Mitochondrion as a novel target of anticancer chemotherapy. *J. Natl. Cancer Inst.* **92**:1042–1053 (2000).
139. Q. Li, Y. Kim, J. Namm, A. Kulkarni, G. R. Rosania, Y. H. Ahn, and Y. T. Chang RNA-selective, live cell imaging probes for studying nuclear structure and function. *Chem. Biol.* **13**:615–623 (2006).
140. F. Rashid, R. W. Horobin, and M. A. Williams Predicting the behaviour and selectivity of fluorescent probes for lysosomes and related structures by means of structure-activity models. *Histochem. J.* **23**:450–459 (1991).
141. F. Rashid and R. W. Horobin. Accumulation of fluorescent non-cationic probes in mitochondria of cultured cells: observations, a proposed mechanism, and some implications. *J. Microsc.* **163**(Pt 2):233–241 (1991).
142. F. Rashid and R. W. Horobin Interaction of molecular probes with living cells and tissues. Part 2. A structure-activity analysis of mitochondrial staining by cationic probes, and a discussion of the synergistic nature of image-based and biochemical approaches. *Histochemistry* **94**:303–308 (1990).
143. R. W. Horobin and F. Rashid Interactions of molecular probes with living cells and tissues. Part 1. Some general mechanistic proposals, making use of a simplistic Chinese box model. *Histochemistry* **94**:205–209 (1990).
144. R. W. Horobin Structure-staining relationships in histochemistry and biological staining. I. Theoretical background and a general account of correlation of histochemical staining with the chemical structure of the reagents used. *J. Microsc.* **119**:345–355 (1980).
145. J. Colston, R. W. Horobin, F. Rashid-Doubell, J. Pediani, and K. K. Johal Why fluorescent probes for endoplasmic reticulum are selective: an experimental and QSAR-modelling study. *Biotech. Histochem.* **78**:323–332 (2003).
146. E. Walter, S. Janich, B. J. Roessler, J. M. Hilfinger, and G. L. Amidon HT29-MTX/Caco-2 cocultures as an *in vitro* model for the intestinal epithelium: *in vitro-in vivo* correlation with permeability data from rats and humans. *J. Pharm. Sci.* **85**:1070–1076 (1996).
147. M. V. Varma, K. Sateesh, and R. Panchagnula Functional role of P-glycoprotein in limiting intestinal absorption of drugs: contribution of passive permeability to P-glycoprotein mediated efflux transport. *Mol. Pharm.* **2**:12–21 (2005).
148. M. Sugawara, Y. Takekuma, H. Yamada, M. Kobayashi, K. Iseki, and K. Miyazaki A general approach for the prediction of the intestinal absorption of drugs: regression analysis using the physicochemical properties and drug-membrane electrostatic interaction. *J. Pharm. Sci.* **87**:960–966 (1998).
149. A. Malkia, L. Murtomaki, A. Urtti, and K. Kontturi Drug permeation in biomembranes: *in vitro* and in silico prediction and influence of physicochemical properties. *Eur. J. Pharm. Sci.* **23**:13–47 (2004).

150. M. Duvvuri, W. Feng, A. Mathis, and J.P. Krise A cell fractionation approach for the quantitative analysis of subcellular drug disposition. *Pharm. Res.* **21**:26–32 (2004).
151. Y. Gong, M. Duvvuri, and J. P. Krise Separate roles for the Golgi apparatus and lysosomes in the sequestration of drugs in the multidrug-resistant human leukemic cell line HL-60. *J. Biol. Chem.* **278**:50234–50239 (2003).
152. Y. Lai, C. M. Tse, and J. D. Unadkat Mitochondrial expression of the human equilibrative nucleoside transporter 1 (hENT1) results in enhanced mitochondrial toxicity of antiviral drugs. *J. Biol. Chem.* **279**:4490–4497 (2004).
153. R. Lill and G. Kispal Mitochondrial ABC transporters. *Res. Microbiol.* **152**:331–340 (2001).
154. L. M. Mangravite, I. Badagnani, and K. M. Giacomini Nucleoside transporters in the disposition and targeting of nucleoside analogs in the kidney. *Eur. J. Pharmacol.* **479**:269–281 (2003).
155. G. Szakacs, J. P. Annereau, S. Lababidi, U. Shankavaram, A. Arciello, K. J. Bussey, W. Reinhold, Y. Guo, G. D. Kruh, M. Reimers, J. N. Weinstein, and M. M. Gottesman Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. *Cancer Cell* **6**:129–137 (2004).
156. V. Y. Chen and G. R. Rosania The great multidrug-resistance paradox. *ACS Chem. Biol.* **1**:271–273 (2006).
157. V. Y. Chen, M. M. Posada, L. L. Blazer, T. Zhao, and G. R. Rosania The role of the VPS4A-exosome pathway in the intrinsic egress route of a DNA-binding anticancer drug. *Pharm. Res.* **23**:1687–1695 (2006).
158. Q. Li, Y. K. Kim, J. Namm, A. Kulkarni, G. Rosania, Y. H. Ahn, and Y. T. Chang. RNA-selective, live cell imaging probes for studying nuclear structure and function. *Chem. Biol.* in press. (2006).
159. V. Y. Chen, S. M. Khersonsky, K. Shedden, Y. T. Chang, and G. R. Rosania System dynamics of subcellular transport. *Mol. Pharm.* **1**:414–425 (2004).
160. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* 17–21 (2001).
161. H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**:D452–D455 (2004).
162. P. Ibison, M. Jacquot, F. Kam, A. G. Neville, R. W. Simpson, C. Tonnelier, T. Venczel, and A. P. Johnson Chemical Literature Data Extraction: The CLiDE Project. *J. Chem. Inf. Comput. Sci.* **33**:338–344 (1993).
163. R. N. Kostoff and R. A. DeMarco Extracting information from the literature by text mining. *Anal. Chem.* **73**:370A–378A (2001).
164. M. Krauthammer, P. Kra, I. Iossifov, S. M. Gomez, G. Hripcsak, V. Hatzivassiloglou, C. Friedman, and A. Rzhetsky. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* **18**(Suppl 1):S249–S257 (2002).
165. S. Raychaudhuri, J. T. Chang, P. D. Sutphin, and R. B. Altman Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome. Res.* **12**:203–214 (2002).
166. T. C. Rindflesch, L. Hunter, and A. R. Aronson. Mining molecular binding terminology from biomedical text. *Proc. AMIA Symp.* 127–131 (1999).
167. T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* 517–528 (2000).
168. M. Weeber, H. Klein, A. R. Aronson, J. G. Mork, L. T. de Jong-van den Berg, and R. Vos. Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc. AMIA Symp.* 903–907 (2000).
169. K. Baclawski, J. Cigna, M. M. Kokar, P. Mager, and B. Indurkhya. Knowledge representation and indexing using the unified medical language system. *Pac. Symp. Biocomput.* 493–504 (2000).
170. C. Santos, D. Eggle, and D. J. States. Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics* **21**(8):1653–1658 (2005).
171. M. D. Yandell and W. H. Majoros Genomics and natural language processing. *Nat. Rev. Genet.* **3**:601–610 (2002).
172. M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **125**:11853–11865 (2003).
173. W. D. Ihlenfeldt and J. Gasteiger. Beyond the hyperactive molecule: search, salvage and visualization of chemical information from the Internet. *Pac. Symp. Biocomput.* 384–395 (1996).
174. W. D. Ihlenfeldt and J. Gasteiger Augmenting connectivity information by compound name parsing: automatic assignment of stereochemistry and isotope labeling. *J. Chem. Inf. Comput. Sci.* **35**:663–674 (1995).
175. M. Reitz, O. Sacher, A. Tarkhov, D. Trumbach, and J. Gasteiger Enabling the exploration of biochemical pathways. *Org. Biomol. Chem.* **2**:3226–3237 (2004).
176. M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**:42–46 (2002).
177. M. Kanehisa and S. Goto KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**:27–30 (2000).
178. J. R. McDaniel and J. R. Balmuth Kekule: OCR—Optical chemical (structure) recognition. *J. Chem. Inf. Comput. Sci.* **32**:373–378 (1992).
179. M. L. Contreras, C. Allendes, L. T. Alvarez, and R. Rozas Computational perception and recognition of digitized molecular structures. *J. Chem. Inf. Comput. Sci.* **30**:302–307 (1990).
180. R. Casey, S. Boyer, P. Healey, A. Miller, B. Oudot, and K. Zilles. Optical recognition of chemical graphics. *Proceedings of the 2nd International Conference on Document Analysis and Recognition* 627–631 (1993).
181. K. Shedden, L. B. Townsend, J. C. Drach, and G. R. Rosania A rational approach to personalized anticancer therapy: chemoinformatic analysis reveals mechanistic gene-drug associations. *Pharm. Res.* **20**:843–847 (2003).
182. G. V. Gkoutos, P. R. Kenway, and H. S. Rzepa JChemTidy: a tool for converting chemical Web document collections to an XHTML representation. *J. Chem. Inf. Comput. Sci.* **41**:253–258 (2001).
183. P. Srinivasan. MeSHmap: a text mining tool for MEDLINE. *Proc. AMIA Symp.* 642–646 (2001).
184. P. Srinivasan and T. Rindflesch. Exploring text mining from MEDLINE. *Proc. AMIA Symp.* 722–726 (2002).
185. R. D. Smith, L. Hu, J. A. Falkner, M. L. Benson, J. P. Nerothin, and H. A. Carlson Exploring protein-ligand recognition with Binding MOAD. *J. Mol. Graph. Model* **24**:414–425 (2006).
186. R. Wang, X. Fang, Y. Lu, and S. Wang The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**:2977–2980 (2004).
187. R. Wang, Y. Lu, X. Fang, and S. Wang An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.* **44**:2114–2125 (2004).
188. Y. Z. Chen and C. Y. Ung Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *J. Mol. Graph. Model* **20**:199–218 (2001).
189. Y. Z. Chen and D. G. Zhi Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **43**:217–226 (2001).
190. A. Lindstrom, F. Pettersson, F. Almqvist, A. Berglund, J. Kihlberg, and A. Linusson Hierarchical PLS modeling for predicting the binding of a comprehensive set of structurally diverse protein-ligand complexes. *J. Chem. Inf. Model* **46**:1154–1167 (2006).
191. N. Paul, E. Kellenberger, G. Bret, P. Muller, and D. Rognan Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins* **54**:671–680 (2004).
192. C. P. Mpamhanga, B. Chen, I. M. McLay, and P. Willett Knowledge-based interaction fingerprint scoring: a simple

method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model* **46**:686–698 (2006).
193. H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*. **35**(Database issue):D301–303 (2007).
194. E. Kellenberger, P. Muller, C. Schalon, G. Bret, N. Foata, and D. Rognan sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model* **46**:717–727 (2006).
195. W. C. Byrem, S. C. Armstead, S. Kobayashi, R. G. Eckenhoff, and D. M. Eckmann A guest molecule-host cavity fitting algorithm to mine PDB for small molecule targets. *Biochim. Biophys. Acta* **1764**:1320–1324 (2006).
196. Z. E. Perlman, T. J. Mitchison, and T. U. Mayer High-content screening and profiling of drug activity in an automated centrosome-duplication assay. *Chembiochem* **6**:145–151 (2005).
197. J. C. Yarrow, Z. E. Perlman, N. J. Westwood, and T. J. Mitchison. A high-throughput cell migration assay using scratch wound healing, a comparison of image-based readout methods. *BMC Biotechnol*. **4**:21 (2004).
198. J. C. Yarrow, Y. Feng, Z. E. Perlman, T. Kirchhausen, and T. J. Mitchison Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb. Chem. High Throughput Screen.* **6**:279–286 (2003).
199. J. C. Yarrow, G. Totsukawa, G. T. Charras, and T. J. Mitchison Screening for cell migration inhibitors via automated microscopy reveals a Rho-kinase inhibitor. *Chem. Biol.* **12**:385–395 (2005).
200. V. C. Abraham, D. L. Taylor, and J. R. Haskins High content screening applied to large-scale cell biology. *Trends Biotechnol.* **22**:15–22 (2004).
201. D. L. Taylor Past, present, and future of high content screening and the field of cellomics. *Methods Mol. Biol.* **356**:3–18 (2007).
202. A. H. Gough and P. A. Johnston Requirements, features, and performance of high content screening platforms. *Methods Mol. Biol.* **356**:41–61 (2007).
203. K. A. Giuliano, J. R. Haskins, and D. L. Taylor Advances in high content screening for drug discovery. *Assay Drug Dev. Technol.* **1**:565–577 (2003).
204. R. T. Dunlay, W. J. Czekalski, and M. A. Collins Overview of informatics for high content screening. *Methods Mol. Biol.* **356**:269–280 (2007).
205. D. J. Wild and G. D. Wiggins Challenges for chemoinformatics education in drug discovery. *Drug Discov. Today* **11**:436–439 (2006).
206. D. J. Wild and G. D. Wiggins Videoconferencing and other distance education techniques in chemoinformatics teaching and research at Indiana University. *J. Chem. Inf. Model* **46**:495–502 (2006).
207. C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo—and bioinformatics. *Curr. Pharm. Des.* **12**:2111–2120 (2006).