PERSPECTIVE

Recent Advances in Chemoinformatics

Dimitris K. Agrafiotis,*,† Deepak Bandyopadhyay,† Jörg K. Wegner,‡ and Herman van Vlijmen‡

Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive, Exton, Pennsylvania 19341, and Tibotec BVBA, Gen De Wittelaan L 11B 3, 2800 Mechelen, Belgium

Received February 12, 2007

Chemoinformatics is a large scientific discipline that deals with the storage, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information. Chemoinformatics techniques are used extensively in drug discovery and development. Although many consider it a mature field, the advent of high-throughput experimental techniques and the need to analyze very large data sets have brought new life and challenges to it. Here, we review a selection of papers published in 2006 that caught our attention with regard to the novelty of the methodology that was presented. The field is seeing significant growth, which will be further catalyzed by the widespread availability of public databases to support the development and validation of new approaches.

INTRODUCTION

Chemoinformatics is a vast discipline, standing on the interface between chemistry, biology, and computer science. Despite being perceived by many as a mature field, it has seen considerable growth in 2006. This growth is evidenced by the fact that significant advances are no longer found in the pages of a few specialty publications but across a wide range of mainstream chemistry and general science journals such as JACS and PNAS.

In this review, we highlight a few papers that were published in 2006 and that we found intriguing for a variety of reasons. The review is not intended to be exhaustive or authoritative. It reflects strictly the views of the authors and their long-standing interest in new computational methodology. In order to manage the scope and length of the article, important studies describing primarily the application, validation, and comparison of various chemoinformatics techniques as well as incremental enhancements to established methodologies have not been included.

The remaining sections are organized in seven general areas: (1) advances in conformational analysis and pharmacophore development; (2) de novo and fragment-based design; (3) QSAR; (4) chemogenomics; (5) free energy and solvation; (6) geometric algorithms and combinatorial optimization; and (7) molecule mining. This is by no means an authoritative classification but rather an attempt to organize our thoughts into coherent themes and focus the readers' attention on topics pertinent to their own interests.

CONFORMATIONAL ANALYSIS AND PHARMACOPHORE DEVELOPMENT

Conformational sampling is a problem of central importance in computer-aided drug design. Several modeling techniques depend critically on the diversity of conformations sampled during the search, including protein docking, pharmacophore modeling, 3D database searching, and 3D-QSAR, to name a few. Recent analyses of crystal structures of protein−ligand complexes have shown that bioactive conformations tend to be more extended than random ones[1,2] and may lie several kcal/mol higher in energy than their respective global minima.[3] There have been a number of comparative studies of conformational analysis tools, focusing primarily on the ability to identify the bioactive conformation. While this is certainly a desired goal, our knowledge of pharmacologically relevant conformational space is very limited, and the ability to identify the bioactive conformation can only be guaranteed if the search method casts a wide net over the potential energy surface. Reproducing known ligand geometries is insufficient because these represent an extremely limited and biased sampling of all bound ligand conformations. Indeed, most ligands have never been crystallized in their own targets, even fewer have been crystallized in important countertargets, and many protein classes have never been crystallized at all.

While diversity is sometimes a goal in its own right, as in many approaches to library design, thoroughness of conformational sampling is usually not an end in itself, nor is it the sovereign virtue for a conformational search method. For example, Omega (OMEGA 1.8.1, distributed by Openeye Scientific Software (www.eyesopen.com)) is extremely fast, with sampling suitable for many applications. However, thorough sampling is an important means to many further ends, and any practicing computational chemists would want to know which methods sample the full ensemble of accessible conformations.

One study that is particularly indicative of the state of current conformational search techniques was recently published by Carta et al.[4] It is well-known that many stochastic 3D modeling techniques are very sensitive to starting configurations and random number effects, and the

* Corresponding author phone: (610)458-6045; fax: (610)458-8249; e-mail: dagrafio@prdus.jnj.com.
† Johnson & Johnson Pharmaceutical Research & Development, L.L.C.
‡ Tibotec BVBA.

10.1021/ci700059g CCC: $37.00
Published on Web 05/19/2007

results are often difficult to reproduce when the search is repeated under slightly different initial conditions. The paper by Carta et al. demonstrates that this reproducibility problem plagues systematic methods as well. More specifically, it examined how different permutations of the connection table affected the conformations generated by Corina, Omega, Catalyst, and Rubicon. The authors used Daylight and in-house utilities to generate different (noncanonical) variants of SMILES[5,6] (Simplified Molecular Input Line Entry System) and SD representations[7] for 17 bioactive ligands, effectively changing the order of the atoms and bonds while keeping the topology intact. Each variant was subjected to conformational search, using the same set of parameters for conformer generation. The results were evaluated, among other ways, by looking at the distribution of rmsds to the crystallographically determined bioactive conformation. Indeed, it was shown that Omega and Rubicon produced very different distributions of rmsds for the canonical and non-canonical SMILES/SD variants, suggesting that the methods exhibit an intrinsic bias and are highly dependent on the atom and bond ordering. On the contrary, Catalyst was found to be much less sensitive to permuted input. Principal component visualization of the conformational ensembles generated by each permuted input further revealed that the canonical and permuted SMILES sampled distinct regions of conformational space, and, in at least one case, the conformations generated by the permuted variants were much closer to the bioactive conformation.

Based on these findings, the authors recommend the use of multiple permuted inputs in order to improve the performance of methods such as Omega and Rubicon. Although this approach is symptomatic, it only requires a way to generate permuted connection tables and, therefore, can be used with any conformational search program to circumvent its intrinsic bias and enhance its sampling capacity.

Another approach aimed at expanding the range of geometries sampled during conformational search was presented by Izrailev et al.[8] The method is based on a self-organizing algorithm known as stochastic proximity embedding (SPE)[9] for producing coordinates in a low-dimensional space that best preserve a set of distance constraints. This algorithm was subsequently extended to the problem of conformational sampling using a distance geometry formalism.[10] SPE generates conformations that satisfy a set of interatomic distance constraints derived from the molecule's connection table and defined in the form of lower and upper bounds $\{l_{ij}\}$ and $\{u_{ij}\}$. While the method was originally shown to provide a good sampling of conformational space, it was observed that "extreme" conformations located near the periphery of conformational space were not as likely to be visited, and, therefore, important conformations could be missed.

To alleviate this problem, the authors introduced a boosting heuristic that can be used in conjunction with SPE or any other distance geometry algorithm to bias the search toward more extended or more compact geometries. The method generates increasingly extended (or compact) conformations through a series of embeddings, each seeded on the result of the previous one. In the first iteration, a normal SPE embedding is performed, generating a chemically sensible conformation c_1. The lower bounds of all atom pairs $\{l_{ij}\}$ are then replaced by the actual interatomic distances $\{d_{ij}\}$ in conformation c_1 and used along with the unchanged upper bounds $\{u_{ij}\}$ to perform a second embedding to generate another conformation, c_2. This process is repeated for a prescribed number of iterations. The lower bounds are then restored to their original default values, and a new sequence of embeddings is performed using a different random number seed. Because the distance constraints in any iteration are always equal to or greater than those in the previous iterations, successively more extended conformations are generated. This process will never yield a set of distance constraints that are impossible to satisfy, because there exists at least one conformation (i.e., the one generated in the preceding iteration) that satisfies them. An analogous procedure can be used to generate increasingly compact conformations.

Conformational boosting was subsequently validated against seven widely used conformational sampling techniques implemented in the Rubicon, Catalyst, Macromodel, Omega, and MOE software packages and was found, along with Catalyst, to be significantly more effective in sampling the full range of geometric sizes attainable by any given molecule compared to the other methods, which showed distinct preferences for either more extended or more compact geometries.[11,12] Since bioactive conformations tend to be extended and often fall outside the range sampled by an unbiased search, this heuristic significantly improves the chances of finding such conformations.

One important technique that benefits greatly from proper sampling of conformational space is pharmacophore modeling. A pharmacophore is the spatial arrangement of steric and electronic features that are necessary to confer the optimal interaction with a particular biomolecular target and to trigger (or block) its biological response. Ligand-based drug design methods that attempt to identify a pharmacophore from a set of active compounds have been known to fail when some of the compounds have different binding modes from the rest. Current approaches for pharmacophore identification typically use manual curation or consensus to remove actives that are presumed to bind with different binding modes from the majority that share a common mode. PharmID[13] is a new algorithm for pharmacophore detection that overcomes these problems by a statistical sampling approach, Gibbs Sampling,[14] that picks the most likely binding conformations and key binding features simultaneously and iteratively. PharmID's breakthrough lies in transforming the complex problem of matching N molecules with up to M conformations each into a simpler one of comparing each conformation of each molecule against a model of the active conformation and its key features. The method derives the probability that each feature is important and that each conformation is the active conformation, starting with no knowledge of important features or binding conformations. Each one of these probabilities iteratively determines the other one, and thus PharmID quickly converges to the correct answer for a large set of examples.

The algorithm begins with a set of distinct conformations for each molecule in the alignment, on which pharmacophore groups are defined using SMARTS. Pairs or triples of pharmacophore group types and their binned distances are defined as unique features. For each conformation of each

molecule, a bit string is generated to encode all the features that it contains. The program selects a conformation to align for each compound by sampling from a weight vector, which initially has equal weights for all conformations. One can estimate the occurrence/nonoccurrence probability of a feature, given the counts of that feature in the currently selected conformations and the *pseudocounts* or background probability of seeing the feature in a large data set. Probabilities are calculated for the bit strings of all conformations of a selected molecule, using the feature counts in the other *N*-1 compounds and using the background probabilities. The ratio of these two probability vectors defines the updated weights for each conformation of that molecule for the next iteration. The algorithm's convergence is determined by a scoring function that sums ratios of feature probabilities to background probabilities for each feature. After convergence, detected features are searched within selected conformations using a clique detection algorithm.[15] A postprocessing step ranks pharmacophore hypotheses based on how many molecules they fit and enumerates multiple binding modes if detected.

PharmID correctly detects two different binding modes in a mixed D2 and D4 ligand data set[16] on which Catalyst is known to fail. In a thrombin data set with highly flexible molecules having 1000 or more conformations each,[17] it converges to a pharmacophore hypothesis that is closer to the crystal structure conformation produced by Catalyst, DISCO, or GASP. This may be because PharmID can process thousands of conformations per molecule, while the other methods are limited to a smaller number. One weakness mentioned is the possibility of the sampling to converge to local minima if some speed optimizations are used; the authors suggest running several iterations that start with different binding conformations.

Conformational analysis is also important in the study of macromolecules. The analysis of protein folding pathways is very complex due to the large number of conformational variables involved. To study important folding descriptors such as the transition state (TS), reaction coordinates need to be defined with a dimensionality that is as low as possible. Linear dimensionality reduction techniques such as principal component analysis (PCA) are widely used but have not been very useful in the study of protein folding because of the inherent nonlinearity of the event. Das et al.[18] used a nonlinear dimensionality reduction technique based on the recently proposed ISOMAP algorithm.[19] This algorithm attempts to preserve as best as possible the geodesic distances between all pairs of data points in a low-dimensional embedding. The geodesic distance is defined as the shortest path between a pair of points, and the path is confined to lie of the low-dimensional manifold. This manifold is not known a priori, and the distance is approximated by taking the shortest path between two points obtained by adding all subpaths between neighboring points. This approach works only if there is a good sampling of points on the manifold. The ISOMAP algorithm was modified to allow calculations on the large protein folding data sets (molecular dynamics trajectories of an SH3 protein model) by using landmark points in the geodesic distance calculations. The results showed that this approach was much more accurate than PCA at describing the original data, and the free energy profile in a one- or two-dimensional projection space showed a consistent transition state location. Validation of the TS region was done by showing that 50% of the dynamics trajectories that started from the TS went to a folded state and 50% went to an unfolded state. ISOMAP scales to the third power of the number of data points and becomes prohibitive for large data sets. In contrast, the stochastic proximity embedding (SPE) algorithm described above[9] obviates the need to estimate the geodesic distances, scales linearly with the size of the data set, and has been shown to be equally effective in extracting the intrinsic dimensionality and nonlinear structure of the underlying manifold.

DE NOVO AND FRAGMENT-BASED DESIGN

The generation of novel chemical entities that are clearly distinguished from competitive products has become one of the pharmaceutical industry's most pressing needs. Modern medicinal chemists find themselves navigating through a maze of patents covering not only the active pharmaceutical ingredient but also a myriad of related variants. These variants are protected long before their true safety and efficacy profile is understood, in order to increase the likelihood that a commercially useful compound will emerge from that patent space. The threshold of innovation has increased dramatically in recent years, and there is a great need for methods that can provide access to uncharted chemical space. One such method is de novo drug design.

De novo design is the automated, computer-assisted construction of new molecules that satisfy a set of desired constraints, such as shape and electrostatic complementarity to a protein binding site. A number of algorithms have been developed since the early 1990s, such as Sprout, GrowMol, Ludi, Legend, Smog, MCDNLG, and Synopsis. De novo design algorithms consist of three major components: (1) a way of "growing" molecules from smaller fragments, (2) a way of scoring those molecules against the desired objectives, and (3) a way of directing the algorithm toward the most productive areas of the search space in order to tackle the vast combinatorial complexity of the problem.

De novo design programs are notoriously difficult to validate. They tend to generate a large number of candidate ligands, and only a small fraction of them can be physically synthesized and tested in a biological context. Although several de novo design algorithms have yielded active structures, there are no systematic studies on the likelihood of finding actives in the raw outputs of these algorithms. In most cases,[20,21] de novo design has produced molecular frameworks that were subsequently converted into active compounds through conventional medicinal chemistry, and it is often unclear whether the final success is the result of the de novo design algorithm or the imagination of the medicinal chemist.

The most common criticism of de novo design approaches is the synthetic accessibility of the resulting molecules. Depending on the approach that is taken, these molecules can be either too simple or too complex. Some programs employ simplistic molecule-growing techniques that yield unimaginative structures, while others lead to overly complex topologies that are very difficult, if not impossible, to synthesize.

A novel heuristic for evaluating the complexity of the ligands generated in the course of a de novo design run was presented by Boda and Johnson.[22] The authors propose a scoring function based on the statistical distribution of various cyclic and acyclic fragments and atom substitution patterns in existing drugs or commercially available starting materials. The method starts by systematically enumerating all possible acyclic and cyclic patterns found in a database of reference molecules (the authors used MDDR and the union of the Aldrich, Maybridge, and Lancaster catalogs as representatives of druglike and synthetically accessible chemical space, respectively). These patterns include 1-, 2-, 3-, and 4-centered chain fragments composed of nonterminal chain atoms along with their nearest heavy atom neighbors as well as ring and ring substitution patterns comprised of fused, bridged, and spiro ring systems with and without their immediate bonded neighbors. The patterns are canonicalized and stored in a hierarchical "complexity" database. A unique aspect of the algorithm is that the patterns are represented at multiple levels of abstraction. At the highest level, only the topology of the pattern is preserved, taking into account only connectivity, hybridization, and bond order. Each such generic pattern is linked to all of its specific variants, which represent specific atom substitutions anywhere on the generic structure. The patterns are organized in a hierarchical structure, with the generic structure at the root, single-atom substitutions at the first level, two-atom substitutions at the second level, and so on. Each node in the tree also maintains the frequency of that pattern in the reference database.

This database can be used to score the complexity of structures generated during the de novo design cycle. The complexity score penalizes patterns that are absent or rare in the reference database and consists of two subscores, one based on the generic topology and one on the exact atom substitution. The method was validated on a set of 50 top-selling drugs, whose complexity scores were found to correlate reasonably well with the synthetic accessibility indices computed by the retrosynthetic analysis program CAESA.[20] The method was also used to analyze the structures generated by Sprout on an enzyme target (dihydroorotate dehydrogenase complexed with brequinar, PDB code 1D3G), and it was found that only a small fraction of them (<10%) had all their structural generic patterns matched in the complexity database, and an even smaller fraction (<1%) had all their atom-substituted patterns matched as well. Still, striking the right balance between novelty and complexity in de novo design remains largely an art.

Fragment-based drug design is a closely related technique that has become increasingly popular since the first applications published by the Fesik group.[23] In this approach, X-ray crystallography and/or NMR spectroscopy is used to determine the binding mode of low molecular weight (<250), usually weakly active ligands (high micromolar to low millimolar binding affinities). The small ligands (fragments) are subsequently translated into high affinity ligands by combination, extension, or other methods.[24] Although the logic of the approach is appealing, it is still unpredictable whether a particular application will be successful. Especially the step of linking different fragments into a single molecule has proven to be difficult. Hajduk[25] presented a detailed study on 15 different internal discovery programs at Abbott to see whether certain rules could be extracted that increase the odds for successful fragment-based design. From every discovery program, the most optimized compound (potency, safety, bioavailability) was systematically reduced in size, and the corporate database was queried for the resulting substructures. On average, 4.1 substructures existed per optimized inhibitor. Interestingly, a linear relationship between potency and molecular weight (MW) in every compound subseries was observed, leading to an average increase of 1 pK_D unit for every MW increase by 64 mass units. This observation suggests that along the optimal optimization path the ligand efficiency (binding free energy per mass unit) remains relatively constant. Several conclusions are drawn from this work. First, if the ligand efficiency drops significantly when the size of a fragment is increased, it is likely that the site or nature of the modification is not ideal. Second, it is important to start with the most efficient fragment lead if one wants to obtain a highly potent molecule with a relatively low MW. The results also enable a prediction of the expected size of the final optimized inhibitor, given the initial fragment lead/hit and the required potency. It should be noted that the predictions resulting from the analysis can only be applied to fragment leads with maximal binding efficiency, i.e., a significant amount of SAR data is required. The analysis also highlights binding efficiency differences between targets, which are consistent with the notion that certain targets require larger molecules to achieve high potency (e.g., HIV protease, protein−protein interactions).

QUANTITATIVE STRUCTURE−ACTIVITY RELATIONSHIPS (QSAR)

QSAR is one of the oldest and most widely used methods in computational drug design. It employs statistical methods (typically regression, pattern recognition and machine learning techniques) to derive quantitative mathematical relationships linking chemical structure and biological activity. Most conventional QSAR approaches construct "global" models that can provide insight into general structural trends that affect bioactivity. Such methods, however, often fail to identify relationships that are evident only within a particular subset of molecules, as the rules for activity tend to be dominated by the most active and most inactive compounds.

Consider, for example, the SAR data illustrated in Figure 1. Most QSAR methods will conclude that having an imidazole substituent at a certain position (molecules 1 and 2) is good for activity, whereas having a carboxylate group is bad (molecules 10, 11, and 12). In contrast, since Cl and Br are present in compounds spanning the entire activity range, such methods will lose sight of the fact that changing a Br to a Cl in otherwise identical compounds will always result in improved potency. These "local" and "global" perspectives on the SAR data are equally valuable and complementary.

Sheridan et al. recently presented a method for automating the exploration of "local" QSAR based on similarity comparisons between pairs of molecules.[26] The authors examine the effect of "transformations", i.e., how small changes in structure from molecule A to B change the activity of the compound regardless of how active it is. Two methods were presented, T-ANALYZE and T-MORPH.

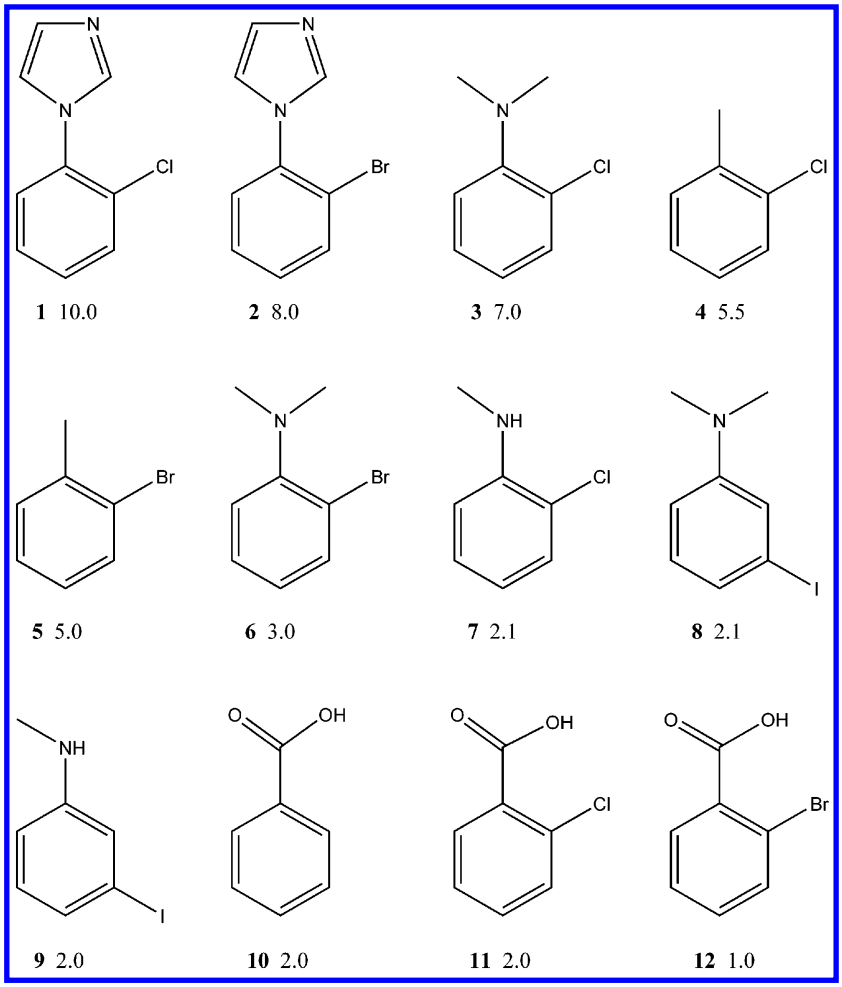

Figure 1. A hypothetical SAR data set.

T-ANALYZE takes a data set, finds the structural differences between similar molecules by means of maximum common substructures, and records the minimum transformation that is required to convert one molecule into the other. These are then compared and clustered into groups of related transformations, and the clusters are ranked based on the magnitude and nature of the activity range of their constituent transformations (i.e., positive or negative). In the example of Figure 1, pairs of related molecules with Br to Cl transformations at the same position would be grouped under the same cluster, and an assertion would be made as to whether converting a Br to a Cl always increases activity, always decreases it, or leads to mixed results. This is not a "fit" to any kind of model, it is merely a way of reorganizing observed data and presenting it to the user. The idea is to do what a medicinal chemist would normally do by hand but do it in a systematic and nonbiased way on a large data set. T-ANALYZE scales to the fourth power of the number of molecules, and thus the use of the method with data sets exceeding a few hundred records is computationally prohibitive.

The second technique, T-MORPH, takes a particular molecule and compares it to all the A to B transformations in a data set to see which transformations could transform that molecule into something more active. For instance, if the target contains a Br, T-MORPH would find all the transformations that could change a Br to something more active like Cl. It then lists the transformations in decreasing order of "goodness". This technique is much faster than T-ANALYZE but must be repeated for each molecule to be analyzed.

T-ANALYZE and T-MORPH were compared to traditional QSAR methods using three data sets: one containing 116 dopamine agonists, a second containing 397 DHFR inhibitors, and a third containing 114 ACE inhibitors. Although the global methods, such as PLS, random forests, and SVMs, provided more accurate summaries of the global properties of the data, T-ANALYZE and T-MORPH were able to identify a greater number of chemical transformations that were relevant for distinct subsets of molecules.

Another QSAR-related advance is ENTess,[27] a program for ligand binding affinity prediction using quantitative structure-binding affinity relationship (QSBR). Geometric and statistical functions for scoring and analyzing protein structures based on interactions made by neighboring residues in Delaunay tetrahedra are a longtime research focus of the Tropsha group at UNC.[28] The innovation of ENTess is to assign atom types as their Pauling electronegativity values,[29] with the score for a quadruplet being the sum of their atomic electronegativities, thus the name EN (electronegativity) and Tess (Delaunay tessellation). The choice of electronegativity follows from its presence as the first-order term of the energy function of molecules given by the equation below (μ_a is electronegativity, Q_a is partial charge of atom a):

$$E(Q_a) = E_0 + {}_a\Sigma\mu_a Q_a + {}^1/_2\ {}_a\Sigma\eta_a Q_a{}^2 + ...$$

The authors trained on 517 ligand−protein complexes, 264 of which have known binding affinities from literature. A Delaunay tessellation of the complex was done, and only tetrahedra on the receptor−ligand interface were kept and were classified into three types (RRRL, RRLL, RLLL) based on how many atoms came from the ligand and the receptor. The atomic alphabet was reduced to four (C, N, O, S) for ligands and six (C, N, O, S, halogen/P, metal) for proteins, and a lookup table was created with electronegativity sums for 554 possible quadruplet compositions. The 264 complexes were divided into training, test, and validation subsets, and a QSBR model relating the ENTess score to known binding affinity was derived using k-nearest neighbor search with variable selection, with the leave-one-out correlation optimized over the training set. The QSBR model was validated using Y-randomization and refined after matching predictions of binding affinity for the test set with known values. Since matching during refinement destroys the independence of training and test sets from each other, prediction was also done on an external validation set of 24 complexes, achieving similar or better R2 values than earlier scoring functions, such as BLEEP, PMF, SMoG96, SCORE, XSCORE, LUDI, VALIDATE, and ChemScore. ENTess was used to predict the binding affinities of several docked structures constructed from four protein−ligand pairs and always distinguished the crystal or lowest energy pose from other poses. One limitation is that prediction sometimes fails when a structure's ENTess representation is too far from all the ones used for training, perhaps because of an infrequently occurring interaction.

CHEMOGENOMICS

The National Institutes of Health has recently funded the formation of a network of high throughput screening centers[30] that aim to identify and further develop modulators and chemical probes of gene, pathway, and cell functions. The network has begun to provide publicly available data on a wide variety of biological assays tested against a common small molecule library, with all assay descriptions, chemical structures, and data made available through the NIH's Pubchem database.[31]

Advances in automation, liquid handling, and data analysis have led to high-capacity integrated technologies enabling the assay and analysis of more than 1 million biological reactions per day. Perhaps more important than output is the potential of these technologies to improve data quality generated from HTS campaigns. Currently, the industry standard for the primary screen is to test compounds at a single concentration. However, for the purposes of creating a chemical genomic map of the cross-section between chemical space and biological activity, a more thorough analysis of each compound is required. One of the major depositors of Pubchem assay data, the NIH Chemical Genomics Center, has developed a paradigm (quantitative HTS or qHTS) to efficiently generate pharmacological data from the primary screen.[32] The high quality and in-depth nature of this data has provided the chemoinformatics community a wealth of biological activities that can be mined for the elucidation of structure−activity relationships.[33]

In this recent study, Inglese et al. tested their qHTS paradigm with the enzyme pyruvate kinase to generate activation and inhibition concentration response curves for more than 60 000 compounds in a single experiment. In addition to highly reproducible activity profiles, their method provided efficacy data on all compounds. To simplify the analysis of the diversity of concentration response curves, the authors created an automated curve classification scheme to group compounds into full titration inhibitors and activators, partial modulators, incomplete curves, and inactive compounds. Using these curve classifications, they devised a method for the systematic generation of SAR. First, all compounds with apparent concentration-dependent effects were identified using select classifications. This set of actives was used as an initial seed for compound clustering, which was performed using Leadscope[34] fingerprints and agglomerative hierarchical clustering. Maximal common substructures were extracted from each cluster and were subsequently abridged to create SAR scaffolds. A canonical list of scaffolds across all clusters was generated and was used as substructure queries to capture all remaining inactive data for each scaffold. The authors observed that qHTS data show varying effects of R-groups, including a distribution of potencies, efficacies, and modes of activity (activation and inhibition). While the qHTS method was applied to a simple biochemical system in the study, its utility in cell-based assays has demonstrated its versatility.[35]

This new high throughput pharmacology paradigm promises to provide a platform for chemical genomics and for the discovery of probes for gene products in uncharted areas of the human genome. The public availability of this data through Pubchem is providing modeling communities a plethora of high quality profiles of small molecules.

The work by the NIH group described above is one of several attempts to understand the relationships between chemical scaffolds and protein families by mining large volumes of HTS data. Most of these studies have focused on "druggable" targets, mostly GPCRs, kinases, and proteases.[36−39] Yan et al.[40] used a related methodology to analyze the Novartis HTS corporate database, but their primary aim was to identify chemotypes that appear to give artifacts related to the underlying screening technology or demonstrate activities that are specific to target families. Indeed, it is well-known that general toxic compounds tend to show consistently high activity in many cell-based assays, while compounds that form aggregates also exhibit misleadingly high activities in enzyme inhibition studies. These artifacts are difficult to detect by looking at a single HTS data set. However, by analyzing the activities of large libraries of compounds across multiple assays, one can more easily identify promiscuous or unwanted chemotypes and scaffolds with target-specific activities.

The authors utilized a statistical method called ontology-based pattern identification (OPI) that was previously applied to the analysis of microarray gene expression profiles. OPI attempts to identify subsets of structurally similar compounds that also exhibit similar biological profiles. The method can be used with both single and multiresponse data. Given a panel of relevant assays, the algorithm begins by identifying compounds that have significant activity in at least one of these assays and have been tested on the majority of the

remaining assays (the authors use 80% as a threshold). These compounds are clustered into a set of structural families using Daylight fingerprints and a Tanimoto similarity cutoff of 0.85. Given the statistical nature of the analysis, singletons are eliminated from further consideration. For each remaining cluster, C, the algorithm constructs a representative biological profile, $\boldsymbol{Q}_C$, based on the activities of all the members in C. $\boldsymbol{Q}_C$ can be either the average profile of all the members of C or the profile of the compound that is most similar to the profiles of all the other members of C. The algorithm then computes the similarity S_i of each compound in $\boldsymbol{C}$ to $\boldsymbol{Q}_C$, $\boldsymbol{S}_i = \boldsymbol{sim}(\boldsymbol{Q}_C, \boldsymbol{Q}_i)$, ranks all the compounds in C in descending order of their similarities, S_i, and identifies the similarity cutoff S that minimizes the likelihood that the compounds with similarity $S_i \geq S$ have a similar biological profile by chance. In essence, the OPI algorithm identifies the optimal subset of compounds that best demonstrates neighborhood behavior by minimizing the probability of random enrichment.

This analysis resulted in ~1500 scaffolds (out of ~33 000 molecules tested on a panel of 74 assays) with statistically significant structure-profile relationships. Further ANOVA and Kruskal-Wallis analysis was used to identify four broad categories of scaffolds, that include tumor cytotoxic, generally toxic, potential reporter gene artifacts, and target family specific. These annotated scaffolds can be used as substructure search queries to flag compounds in a corporate collection, select meaningful diversity libraries to enhance a corporate deck, and identify reliable SAR from HTS campaigns.

Several recent studies have investigated the targets in currently marketed drugs[41] and the proteins in the genome that are theoretically amenable to small molecule drugs, also known as the druggable genome.[42] Cleves and Jain used an interesting approach to analyze ligand−ligand (drug−drug) and target−target similarities in the set of currently marketed drugs.[43] Both types of similarities were calculated by determining 3D ligand−ligand similarities using previously published methods such as morphological similarity (shape similarity) and molecular imprinting. The level of similarity between models of cognate ligands of two targets was used as a measure of the target−target similarity. An important goal of the study was to see if this approach would make it possible to rationalize the off-target activities of drugs. A nontrivial task in this analysis was to define the exact molecular targets of all drugs in the list that was used. Two-way hierarchical clustering on the drug and target similarities resulted in intuitively correct target clustering, even though target similarity was evaluated indirectly through their ligands. Ligand similarity could be used very effectively to enrich cognate ligands within a large set of screening compounds or a set of other drugs. Enrichment levels were comparable to the currently best performing docking methods. The approach also led to several previously nonannotated activity overlaps. Especially interesting are cases where two ligands do not share a common primary target but share a common secondary target. In most cases these target similarities would not be detected by protein sequence or structure similarity. Examples of this are the cyclooxygenase inhibitors and nucleoside antivirals. It is suggested that both classes share a common secondary target, possibly Bcl-2 or Bcl-XL.

While most genomic and proteomic data reside in public databases, most pharmacological data (i.e., SAR of bioactive chemicals) exist in proprietary databases and journals and are not accessible to data mining efforts. The Global Map of Pharmacological Space[44] is a large scale ligand-target matrix constructed to relate and link the chemical structure and biological target spaces and enable genomic data and mining techniques to be applied in drug design (chemogenomics). The data warehouse behind the global map contains 4.8 million nonredundant molecules of which 275 000 are known to be biologically active and includes 600 000 SARs of molecular binding from publications and proprietary screening data.

Using the global map, several important questions can be answered at a glance. Only 836 human genes are associated with known small molecule binding, of which 141 have been targeted by known oral small-molecule drugs. One can also study polypharmacology, i.e., specific binding of a compound to two or more protein targets, which are then said to interact in chemical space. 35% of the bioactive compounds in the database show polypharmacology, and their interrelationships are visualized in the human polypharmacology interaction network. The majority of promiscuous compounds are active against targets within the same gene family, but a quarter of them are also active across different gene families. Molecular properties such as molecular weight and lipophilicity were found to cluster together by target class and vary widely across target classes. A steady rise in the number of targets screened and molecular weight of reported compounds over the years was inferred from the data, but so was a fall in the molecular weight of drugs surviving clinical trials. The global map allows answers to questions on the druggability of parts of target space and probabilistic modeling of properties of drugs and targets. It may also allow, as desired, either the prediction of side effects from polypharmacology or the rational design of polypharmaceutical drugs.

The value of large data sets has also been exploited at the 3D structural level. The popularity of structure-based design to facilitate drug discovery has led to a wealth of structural information on proteins bound to small molecules. There exist a large number of X-ray/NMR small-molecule−protein complexes in the public protein databanks and an even greater number of structures are being generated from virtual screening efforts. Advances in high-throughput crystallography/NMR methods will likely lead to a dramatic increase in the number of experimental structures available in the near future. Given the wealth of such data, there is a critical need for improved methods to organize and analyze this information and apply it to virtual screening. Current approaches for examining protein−ligand interactions such as LIGPLOT[45] or various computer graphics programs are not useful for the analysis of these large data sets. Scoring functions are popular for filtering large virtual data sets of protein−small-molecule complexes but are poor at distinguishing correct binding modes from incorrect ones, leading to a large number of high-scoring false positives that result in lower enrichment rates in virtual screening.[46]

Singh et al. have developed a series of fingerprint-based approaches for analyzing protein−small-molecule complexes.[47] The method called SIFt (Structural Interaction

Fingerprint) translates the 3D interaction pattern of a protein−small-molecule complex into a 1D binary fingerprint. The concept underlying all of the SIFt approaches is the generation of a binary fingerprint that encodes the interactions between the ligand and its receptor.[48] The first step is identifying the residues that are in contact between a small molecule(s) and a protein(s). The resulting set of ligand binding site residues is used as a common reference frame to align the fingerprints across multiple ligands and/or multiple proteins. The fingerprints can be generated for multiple ligands bound to a single protein (e.g., docking results) or to multiple proteins when a common binding site can be defined, as in the case of protein kinase X-ray structures. Each binding site residue is described by a number of bits that describe the nature of the interaction with the ligand(s). The length of the bit-string is flexible, ranging from a single bit denoting the presence or absence of an interaction, to longer strings encoding the specific nature of the interaction, e.g., side chain, main chain, hydrogen bonding, polarity, etc. The complete interaction fingerprint of the complex is constructed by sequentially concatenating the bit-strings corresponding to each binding site residue according to their numbering order. (When multiple protein structures are analyzed, the order of the binding site residues is determined by multiple sequence alignment.) Thus, for a given family of related receptor−ligand complexes, the interaction fingerprints are of the same length, and each bit represents the presence or absence of a particular interaction at a particular binding site. This fingerprint representation enables the rapid clustering and analysis of large numbers of protein complexes.

The basic SIFt methodology has been extended to generate interaction profiles for groups of structures in order to describe the conservation of interactions present in a set of receptor−ligand complexes (p-SIFt).[49] The interaction profiles are constructed by averaging each interaction and generating a residue-by-residue measure of the degree to which these interactions are observed in the reference structures. This approach was used to obtain important insights into the similarities and differences in the binding of inhibitors to the kinase family. The utility of the method was further demonstrated by scoring each docked pose against an interaction profile generated from a training set of target X-ray complexes, thus using the interaction profile as an effective knowledge-based filter eliminating promising poses.

The SIFt methodology was further extended to combinatorial libraries by redefining the interaction bits to encode whether or not a particular R-group or core fragment of the compound satisfies a contact interaction with a particular protein residue (r-SIFt).[50] In effect, the modified fingerprint tags what interactions are satisfied by the variable R-groups in a combinatorial library. The r-SIFts can therefore form the basis for generating classification models for R-group selection based on the likelihood of satisfying the targeted binding mode. The resulting predictive models for each R-group are independent of each other, enabling filtering without the need to fully enumerate the combinatorial library.

Another technique to expand the drug and target space using the information from genomics is to search for similar protein structures to known targets, with the hope of retargeting known drugs or combinatorial libraries to new uses. Protein Structure Similarity Clustering (PSSC)[51] aims to group protein targets into clusters, which can enable libraries of compounds designed to bind to one target to be tried on close structural relatives that might have similar binding sites. The clustering algorithm compares atoms within a sphere defined by the ligand or centered on a binding pocket, dubbed the "ligand-sensing core". Protein structural similarity search programs CE, DALI, and VAST were employed to find similar proteins, but in many cases they were unsuccessful or needed several iterations or manual lookup of the SCOP classification to find all similar proteins. To address these deficiencies, the authors introduced molecular dynamics into the structural homology search, guessing that the crystallographic structure was a snapshot of a number of possible conformations and trying a range of conformations for each protein might yield a better clustering result. They demonstrate, with a 1 ns simulation on the Cdc25A protein phosphatase, that conformations sampled at different time-steps find different and overlapping clusters of VAST neighbors, with some conformations finding the entire known cluster. They also determine clusters for the M6P/IGF2 Receptor, for which conventional VAST searches show very few structural neighbors due to its uniqueness in the PDB, while VAST searches of representative MD conformations converged on two other protein families with structural homology near the active site. Thus, the use of MD simulations is shown to expose new structural space for VAST neighbor searches beyond what is accessible from static crystal structures, and thus broaden the list of potential targets for drugs with known targets.

FREE ENERGY AND SOLVATION

Protein−ligand binding depends strongly on shape complementarity and electrostatic interactions. Solvation effects have a very large impact on electrostatic interactions, and it is essential to include them for accurate prediction of binding free energies. Two commonly used computational methods to model solvation effects are the Generalized Born (GB) method and the Poisson−Boltzmann (PB) method. These methods are still too slow to be used in large scale virtual screening, but they are often employed to rerank docking hits with a more accurate binding energy estimate.

An interesting method that uses PB was earlier developed by Kangas and Tidor.[52] For a molecule with a given shape bound to a protein, their method calculates the most optimal charge distribution for binding over all atoms in the molecule. Unfortunately the optimal charges are often physically unrealistic, and suggestions for improved molecules usually require visual inspection and recalculation. A recent method that allows PB binding energy calculations on a large set of isosteric compounds was developed by Sayle and Nicholls.[53] Their method addresses the commonly occurring situation of having the structure of a bound small molecule that requires further optimization, preferably by making small changes. The calculation of the electrostatic interaction can be written as

$$\Delta E = \mathbf{q}^{\mathrm{T}}\,\mathbf{M}\,\mathbf{q} + \mathbf{I}\,\mathbf{q} + \mathrm{C}$$

Here **M**, **I**, and C are a precalculated matrix, vector, and constant, respectively, that are constant for every molecular

variation because the shape is assumed to be constant. The vectors **q** contain the atomic charges of the small molecule that is being evaluated. Because in usual PB calculations most time is spent evaluating the **M** and **I** matrices, their method is able to evaluate 10−100 000 binding energies per second. To optimize a bound molecule, the authors generated a large number of isosteres using chemical patterns observed in a training set of drug molecules. Several protein−ligand complexes were used as test cases for the ability of the method to find better binding variants. Where possible, the results were compared to known analogs and their binding energies. The results were variable; in some cases good correlations with experiment were found, but in others the agreement was poor. Their conclusion was that the method performs well for truly isosteric series (no atomic size variations) in systems where there are significant electrostatic interactions. It performs less well when there is significant flexibility in the active site or when the resolution of the crystal structure is low.

Although PB calculations are inherently complex and therefore slow, improvements in computational methodology have made significant impacts on the computational speed. Initially most methods used finite difference or finite element approximations, which scale proportionally to the volume of the molecular system. This limited the application to small to medium-sized molecular systems, and protein−protein association or dissociation calculations were precluded. More recently, boundary element, boundary integral equation, and fast multipole (FMM) methods have significantly improved the speed of PB calculations. The latest improvement is described by Lu et al.[54] They combined an optimized FMM method into the boundary element/boundary integral equation formulation and obtained very significant increases in calculation speed and reductions in memory usage without sacrificing accuracy. In addition, the approach enables fast calculations of forces and torques, which are necessary for application to molecular or Brownian dynamics calculations. They suggest additional numerical techniques to increase the speed of PB calculations another order of magnitude, which could allow the routine application of PB in molecular dynamics calculations.

The Generalized Born (GB) solvent model[55] attempts to approximate the solvation free energy computed from integration of the Poisson−Boltzmann equation by an analytical function of the atomic coordinates. The new Gaussian Surface-Generalized Born (GSGB) method[56] replaces the model of an atom as a hard sphere of size equal to the van der Waals radius by a smooth model where each atomic surface is an isocontour of the electron density represented by a Gaussian function. Using a smooth model for electrostatic solvation force enables computation of analytical gradients at every point in space; with a hard sphere model the forces are discontinuous, so gradients are unavailable at the intersections of spheres.

The new model directly constructs the Gaussian surface around a molecular system using the electron density of each atom, with the density at a point in space calculated by summing contributions from all the atoms. The Gaussian radius parameter a controls the smoothness of the surface and is set to 2.5, and the density contour value to cut off the surface is set to match the hard-sphere surface for a single atom. The Gaussian surface is then computed as a closed triangulation using the Marching Cubes algorithm from computer graphics,[57] with rules for placing triangles in cubic grid cells whose vertices change from being inside to outside the Gaussian surface; some rules were modified to ensure the surface is well-defined and closed. Each surface triangle is characterized by its center, area and unit normal, and analytical gradients of these quantities are also computed and plugged into the equation for the GB energy.

The method was validated on short loop sequences and peptides, and was able to correctly predict the energies of long loop sequences and their decoys that previous hard-sphere surface Generalized Born models could not. The calculation also better matched the Poisson−Boltzmann energies than when using the van der Waals hard sphere model, and the tendency of such a model to introduce spurious cavities that are inaccessible to solvent was also eliminated.

Accurate calculations of the free energy of binding (ΔG_b) of ligands to targets is a very important goal in computational chemistry. There is a wide range of methodologies for calculating ΔG_b to varying levels of accuracy. Tirado-Rives and Jorgensen[58] explore the accuracy and consistency of one of the terms contributing to ΔG_b, named the conformer focusing term ΔG_{cf}. This term consists of (a) the energy difference between the bound ligand conformation and most optimal unbound conformation and (b) the entropy loss resulting from the reduction of ligand conformational freedom in the bound state. Four non-nucleoside reverse transcriptase inhibitors were used as test systems in this analysis. The main result of the analysis was that the greatest uncertainty in the ΔG_{cf} calculation arises from the evaluation of the bound-unbound conformational ligand energy difference, amounting to an uncertainty of approximately 5 kcal/mol for ΔG_{cf}. This uncertainty can be improved in principle by application of ab initio or DFT calculations, but this is not practical due to computational limitations. They suggest that the inaccuracy of the force field methods in calculating conformational energies is caused by the fact that molecules much smaller than drugs are used in parametrizations of the force fields. To improve this situation, the authors suggest that the only practical solution is to develop very accurate force fields. Until that has been accomplished the most realistic approach is to focus on relatively rigid molecules or to compare the ΔG_b of very similar molecules.

On the other end of the ΔG_b calculation methods are the knowledge-based approaches, in which statistical interaction potentials are derived from observed atomic contacts in experimental crystal structures. The method assumes Boltzmann-like energetics, which means that the pair potentials $u_{ij}(r)$ between two atom types i (ligand) and j (protein) can be written as

$$u_{ij}(r) = -k_B T \ln(\rho_{ij}(r)/\rho_{ij}^*(r))$$

where k_B is the Boltzmann constant, T is the temperature, and ρ_{ij} is the number density of pairs of types i and j at a distance r in the observed crystal structures. The quantity ρ_{ij}^* is the pair density of the so-called reference state, in which there are no interactions between the atoms i and j. The determination of the reference state values has been approached in different ways but continues to be a source of debate. Huang and Zou[59] developed a method to determine accurate knowledge-based potentials in an iterative way

without the need to define an a priori reference state. The pair potentials $u_{ij}(r)$ are iteratively modified until the best-scored binding mode is the correct binding mode in more than 99% of the protein−ligand complexes tested (786 in total). If the prediction accuracy is less than 99%, $u_{ij}(r)$ is adjusted by comparing the experimentally observed i,j pair distribution with the predicted i,j pair distribution from the current form of $u_{ij}(r)$, thereby reducing the difference between experimental and predicted pair distributions. The potentials usually converged within 20 iterations, and the resulting pair potentials had functional forms that agreed with physical insights. Three independent test sets of protein−ligand complex structures and associated binding affinities were used to validate the performance of the method in an associated article.[60]

GEOMETRIC ALGORITHMS AND COMBINATORIAL OPTIMIZATION

Drug discovery informatics often involves analysis of the 3D geometry of drug and protein structures, and techniques from computational geometry and allied application areas such as geometric modeling and computer graphics form the basis of several drug discovery informatics solutions. Since exploring chemical and biological space to find drug leads is a needle-in-the-haystack combinatorial search problem of staggering complexity, it also needs advanced combinatorial optimization methods. Geometric data structures and combinatorial methods are extensively studied and well established for analyzing protein structures in computational biology and to a lesser extent in chemoinformatics and drug discovery. Several recent and comprehensive reviews have been written on the subject;[61,62] compendia of geometric applications such as Geometry in Action[63] are available; and research initiatives such as BioGeometry[64] have pioneered interdisciplinary research and training. In this section we review two recent papers describing geometric and combinatorial data structures and methods applicable to small molecule informatics: the Voronoi S-network for nearest neighbor analysis[65] and the Multiple Common Point Set problem applied to binding patterns.[66]

Geometric and combinatorial structures such as Voronoi diagrams and Delaunay tessellations that capture the nearest neighboring relationships among a set of points in 3D have been used in structural bioinformatics and chemoinformatics. Though very useful, they have their limitations; for example, the Voronoi/Delaunay had earlier been shown to be unstable in the face of small coordinate perturbations, and a solution was proposed by one of the authors.[67] Another limitation of these diagrams is that they model atoms as geometric points and thus do not capture the nonuniform radii of the interacting atoms or the anisotropy of the interaction. Voronoi S-networks, also known as 3D additively weighted Voronoi diagrams, are a generalization of Voronoi diagrams for potentially overlapping spheres with varying radii, as seen in models of molecular systems. So far, these and other generalizations have been studied mostly in 2D,[68] and fast 3D implementation introduces major complications. Medvedev et al.[65] define Voronoi S-networks, prove some of their properties, and give an algorithm to compute them that takes linear time in practice.

The computed Voronoi S-network for a set of spheres consists of S-surfaces that are hyperboloids, S-channels that are open or closed curves, S-vertices that are centers of interstitial spheres, and S-regions that are nearer to one sphere than to any others. Pairs, triples, and quadruples of spheres generate zero, one, or two of these S-entities, depending on their size, relative location, and occlusion. Whenever adjacent sphere radii are equal, the S-network and its parts degenerate into the Voronoi diagram for points. The Voronoi S-network is computationally represented by the 3D coordinates of its S-vertices, a connectivity matrix with indices of the (up to four) neighbors of each S-vertex; and an index matrix that specifies which four input spheres contribute to each site. The algorithm sequentially determines the S-network sites (vertices, channels, and surfaces), starting with an initial S-vertex and traversing its S-channels to find new S-vertices by sorting the point coordinates. The algorithm's main step is to find inscribed spheres between quadruples of points, and there are some elegant substeps to sort approximate distances along curved S-channels, determine the direction to seek a new site, and accelerate the algorithm for large systems. Radii of interstitial spheres of the S-vertices and clearance (bottleneck width) around each S-channel are computed and stored as the algorithm proceeds. Pairs of adjacent spheres can be connected by straight lines to yield the dual Delaunay S-simplices, which differ from the Delaunay tessellation, and potentially overlap.

Applications of the Voronoi S-network to molecular systems in computational chemistry and biology are reviewed in earlier publications.[69,70] The implementation is well described in this paper, and a demo is readily available.[71] We believe that this method will enable computational chemists to more accurately calculate molecular properties, binding affinity, and other quantities that are derived from nearest neighboring interactions. They may also re-evaluate computations that use Voronoi cells as atomic volumes to ensure that the qualitative results of calculations done this way are still valid when a more accurate representation of atomic volume and boundaries is used.

Another celebrated problem in computational geometry and combinatorial optimization with applications in drug discovery informatics is the largest common point set problem (LCP), which involves finding a transformation that maximizes the overlap between two point sets. If the larger of the two sets is of size n, it is known that LCP has a high polynomial degree[72] of $O(n^{32.5})$ and can be approximated[73] to $O(n^{8.5})$. Shatsky et al.[66] introduce the related multiple common point set problem, with k point sets and ϵ-overlap between points allowed. They give an elegant and rigorous proof that this variation of the problem, called k-partite-ϵ-3D matching, is NP-hard and also hard to approximate. This is an important theoretical result and implies that trying to design an asymptotically optimal exact or approximate algorithm for this class of problems is futile. However, a practical solution to this hard problem would be extremely useful in bio/chemoinformatics to find the largest set of aligned atoms/residues among binding sites of multiple related proteins, ignoring the sequence order. The determination of this multiple common point set helps close in on the smallest set of features responsible for a biological effect (i.e., a receptor-based pharmacophore) and facilitates more sensitive database searches. High-scoring pairwise alignments are not always enough to determine a sensitive multiple alignment.

The method presented, MultiBind,[66] efficiently and practically solves the NP-hard problem of finding multiple common protein binding patterns for small but useful problem sizes (of up to about 100 proteins). Potential matching points from each point set are generated by a geometric hashing preprocess.[74] A molecule is picked as the pivot, and coordinates of each atomic pseudocenter in the other molecules are transformed into reference frames given by all triplets of pseudocenters and stored in a geometric hash table along with a chemical property such as the atom type. Pseudocenters from the pivot structure are used to query the hash table for entries with coordinates within an ϵ-sphere and with the same chemical property. The resulting hits are counted for each reference frame of the pivot and matching structures, storing the transformation between the two reference frames if the points that overlap within ϵ equal or exceed the maximum number of points matched so far. From the multiple candidate transformations of each structure other than the pivot, a set is picked that optimizes the number and chemical similarity of the points matched over all the molecules, using a branch-and-bound algorithm that converges quickly in practice.

The authors validate the biological correctness of the multiple common point set by considering the structurally related protein kinase family and the structurally diverse proteins sharing transition state analogues and estradiol binding sites. MultiBind was able to pinpoint 6−14 shared pseudocenters in these cases, starting with 30−70 pseudocenters around each binding site and taking a few minutes to an hour. The common atoms were fewer than those found by pairwise alignment and overlapped well with the literature. The geometric binding patterns found were also highly specific, typically occurring in less than 1% of the Protein Data Bank, and could be used much like family specific fingerprints[75] to seek proteins with similar biological functions or drugs with similar binding activities.

MOLECULE MINING

"Small molecule" similarity measures are one of the most important topics in rational drug design, QSAR, and combinatorial library design as well as in analyzing HTS libraries. With their growing relevance and use, several questions often arise:

1. Is this a reasonable similarity for the problem at hand?

2. Why does this similarity measure not show the "obvious" similarity to "this" molecule, which has exactly the same activity?

3. Is it possible to design a similarity measure for in-house data?

An overview of several descriptor vector codings and metrics can be found elsewhere.[76,77] Several reviews have shown that specific data sets require specific similarity measures. The basic assumption behind this is that a small change in structure should cause a relatively small change in activity. The problem of defining the term "small change" is sometimes referred to as the similarity paradox.[78,79]

An interesting development in 2006 is that the machine learning community has become interested in mining "structured data", such as strings, trees, and also molecules represented as graphs.[80−82] As a result, the number of "molecule mining" publications has increased dramatically over the last few months. Due to the strong interplay between chemistry and informatics, it is crucial that mined chemical information flows back from the informatics side to support chemistry and rational drug design. Many classical descriptor calculation methods are not amenable to this approach, but some of the newly developed methods show considerable promise. This section focuses on a few important studies published in 2006 and provides a more thorough tabular overview covering some older molecule mining techniques with links to the source code, where available. It is hoped that the public availability of source code, encoding chemistry in the form of hypothesis languages, will help improve available methods and allow proper benchmarking of new methods.

Molecule mining approaches can be separated into two main categories:[83]

1. Single molecule coding (coordinate-based space), where each molecule is described by a substructure or fragment vector, and therefore has an absolute position in a multidimensional space.

2. Pairwise molecule coding (coordinate-free space), where only the distances between two molecules are computed, using an explicit or implicit (maybe infinite) similarity measure. The absolute position of molecules in this space can only be calculated by measuring all pairwise distances, and the dimensionality of the space may be unknown.

SMIREP is a single-molecule coding technique that combines the SMILES molecular representation with an algorithm for concept learning by association known as IREP (Incremental Reduced Error Pruning).[84,85] IREP is faster than earlier inductive logic programming (ILP) approaches due to a modified pruning strategy and is suitable for noisy data domains since it avoids overfitting.[86] In the context of SMILES, an overfitted SMILES representation would function just as a lookup table, without generalizing to unseen molecules. The authors mention in this context the weakness of the similar MULTICASE approach, which has also limitations in the overall fragment length.[87] SMIREP is based on an incremental divide-and-conquer algorithm searching for a single rule that covers many of the active compounds and none (or only very few) of the inactive ones. An example for a typical rule might be a combination of SMILES expressions, e.g., "c1ccccc1 AND Nccc". As new rules are being discovered, they are appended to the rule set until a scoring function that measures the classification accuracy of the rule base shows no further improvement. The incremental generation step uses weighted information gain in order to rank the individual rules, which is a well-known principle in information theory.[88] Although it is not mentioned by the authors, we must point out that rules with minimum support might cause difficulties.[89] Additional numerical constraints, e.g., logP > -1.11, are allowed to restrict the substructure space and provide expert chemical knowledge.

The method was tested on several data sets containing less than 500 molecules: binding to the estrogen receptor based on EPA's DSSTox NCTRER Database, carcinogenicity based on the carcinogen potency database CPDB, and biodegradability based on a subset of the Environmental Fate Database (EFDB). The quality of the results is comparable to other state-of-the-art learning methods.

Fragment co-occurrence mining is a novel method for analyzing a large chemical compound collection.[90] Not only fragments that occur frequently but also pairs of nonoverlapping fragments are determined. The fragment co-occurrence detection method discusses two fragmentation approaches called "graph splitting" and "virtual retrosynthesis". Graph splitting breaks molecules at topologically interesting points, such as the bond between a substituent and a ring, while virtual retrosynthesis uses rules based on chemical reactivity (for example, breaking ester bonds). Since both methods could produce hundreds of thousands of fragments on a large database, full combinatorial evaluation of all co-occurring fragments is not possible. For the NCI database with 250 000 compounds, the co-occurrence of the fragments was estimated by a stochastic sampling procedure. After repeating the experiment 1000 times, the authors report several typical co-occurrences, like ring/linker ratios, branch/attachment point ratios, metals/nonmetals, and a few other typical patterns. A detailed list of the fragments is available upon request in SD format. The authors also demonstrate the use of co-occurrence information in de novo design. Several strategies can be envisioned: combining pairs of fragments that are known to occur together in many synthesized compounds; detecting and avoiding pairs that do not co-occur for reasons such as toxicity or synthetic tractability; or finding new areas of chemical space by combining fragments that could be combined but have not yet been.

A typical problem of all pairwise molecule-coding approaches is that they can only be used for either small data sets or reduced feature graphs. Two new algorithms in this class are MCS-EditDistance-GA[91] and the T-ANALYZE method[26] described in an earlier section. While the first approach uses reduced graphs, the second uses only similar subsets for a detailed analysis of the maximum common substructure (MCS).

Gillet et al. use the "graph edit distance" between two molecules as a similarity measure.[91] Though computationally demanding, this approach belongs to the class of "inexact graph matching" algorithms, which make it more robust than methods based only on strict graph matching. The similarity metric is based on a reduced feature graph applying a cost matrix for editing. The goal is to calculate the edit costs when transforming a source feature graph into a target feature graph. The edit operations are solved by dynamic programming, whereas the cost matrix is established using genetic algorithms. The method was tested on an MDDR subset containing ~5000 compounds of the following classes: 5HT1 agonists (5HT1A), 5HT reuptake inhibitors (5HTRT), renin inhibitors (Renin), cyclooxygenase inhibitors (COX), 5HT3 antagonists (5HT3 Ant), dopamine D2 antagonists (D2 Ant), angiotensin II AT1 antagonists (AT1 Ant), thrombin inhibitors (Thrombin), substance P antagonists (Sub P Ant), HIV 1 protease inhibitors (HIV 1 Prot), and protein kinase C inhibitors (PKC). The authors conclude that this similarity measure is able to cover atom insertions and deletions due to the flexible cost matrix optimization.

Kernel-based similarity metrics and learning methods have already been reviewed elsewhere.[92] Similar to MCS approaches, their time and space complexity is quite demanding. Nonetheless, these methods have a big advantage. As in many other kernel approaches, a similarity value might be calculated recursively and implicitly, perhaps within infinite and highly nonlinear spaces. In other words, kernel-based similarity measures can be completely data driven, allowing multiple chemical codings without making many assumptions or reducing the problem to "atom reduced graphs" or "feature reduced graphs". Besides, the familiar branch-and-bound approach to executing algorithms on small and similar compound sets is always an option. In 2006, the *Journal of Machine Learning Research* published a special issue entitled "JMLR Special Topic on Machine Learning and Large Scale Optimization", which contains several articles for large-scale kernel calculations. This should be of interest to the readers of this review, as large-scale optimization problems abound in the pharmaceutical industry.

The optimal assignment kernel[93−96] first calculates local optimal atom similarities, allowing multiple chemical, physicochemical, and other atom properties, such as electrotoplogical states. This is quite different from most other similarity indices, which usually work on only one atom type or physicochemical property (atom reduced graphs). Here, multiple atom and bond properties are combined to an optimal local atom environment (LAE) kernel, which is then extended by obtaining two other similarity measures. The first measure is an atom-based similarity using an optimal assignment between two molecules. The second measure is a feature-reduced similarity computed by executing the graph similarity algorithm recursively on the atom level and then on the feature level. The difference compared to atom-pairs[97,98] or feature-trees[99] is that the reduced features graph already includes multiple atom properties, which allows it to cover more chemical/biological information. The method was evaluated on multiple ADME and other bioactivity data sets with less than 2000 molecules. The authors showed that the best similarity measures use multiple atom and bond properties and can be further improved by adding problem-specific expert knowledge, such as (in this instance) topological polar surface areas. No attempt was made to optimize the internal parameters for the data sets under investigation.

The pharmacophore kernel[100,101] is a three-point pharmacophore implementation allowing multiple codings including reduced feature graphs and physicochemical properties, such as partial charges. The authors provide an in-depth discussion on optimizing the runtime of multiple variants of their similarity measures. In addition to the full three-point method, they also present a fast approximation called 2D and 3D spectrum, which involves reducing the number of pharmacophoric labels. The method was evaluated on data sets with less than 1000 compounds containing biological activity information for benzodiazepine (BZR), cyclooxygenase-2 (COX), dihydrofolate reductase (DHFR), and estrogen (ER). The authors demonstrated that the best similarity coding contains multiple types of information covering pharmacophore types and partial charges. This is consistent with results reported in other publications and shows that it is advisable to include additional expert knowledge in the form of user constraints (as in SMIREP), multiple atom and bond properties (as in the LAE kernel), or multiple atom and feature information. In general, this algorithm is not restricted to linear fingerprints, spherical fingerprints, or any other 2D-based molecule mining method. This innovative kernel concept for three-dimensional (3D) molecular structures might therefore open a new class of algorithms for structure-based drug design and complement

methods like CavBase, SuMo, ConSurf, Rate4Site, APROPOS, CAST, GASPS, and many more.[102]

CONCLUSION

The preceding discussion makes it abundantly clear that chemoinformatics spans a very broad range of problems and approaches, which are often inter-related and sometimes difficult to categorize. As high-throughput technologies continue to advance, informatics techniques will become indispensable in managing and analyzing the exploding volumes of data. We believe that this will be the defining theme over the next few years. Hopefully, the availability of publicly available data will catalyze further advancements and open new possibilities.

APPENDIX − OVERVIEW OF MOLECULE MINING METHODS

In Table 1 is shown an overview of molecule mining methods.

Table 1. Mining Methods

Classical Mining Methods Working on Single Molecules Mi Generating a Feature Vector

	molecular descriptor transformations		
algorithm	source code	references	dimension
RDF	Java[96,103,104]	105	2D/3D
BCUT	Java[96,103,104]	77	2D
1001 other descriptors		77	2D/3D/4D/xD

Mining Methods Working on Single Molecules Mi Generating Substructures or Fragments

	substructure or fragment based		
algorithm	source code	references	dimension
MoFa/MoSS	Java[106]	107,108,109	2D
ParMol	Java[110]	111	2D
PolyFARM	Haskell[112]	113	2D
SMIREP	C++[84]	85	2D
Warmr	Prolog[114]	115,116	2D
AGM	none	117,118	2D
Dmax	none	119	2D
Gaston	Java[110]	120	2D
optimized gSpan	C++[121]	122,123	2D
MolFea	none	124	2D
Sam/AIm/RHC	none	125	2D
LAZAR	C++/Java[126]	82	2D
co-occurrence	none	90	2D

Mining Methods Working on Molecule Pairs s(Mi,Mj≠i) Working on Atom Type or Pharmacophore Type Reduced Graphs

	atom pair based		
algorithm	source code	references	dimension
atom pair	Java[96]	97	2D
CATS	none	98	2D/3D
feature trees	none	99	2D/3D

Mining Methods Working on Molecule Pairs s(Mi,Mj≠i) Working with Single Atom or Pharmacophoric Properties

	maximum common substructure based		
algorithm	source code	references	dimension
MCS−HSCS	Java[96]	127	2D
Gillet et al.	none	91	2D
T-Analyze	none	26	2D

Mining Methods Working on Molecule Pairs s(Mi,Mj≠i) Working with Multiple Atom or Pharmacophoric Properties

	kernel based		
algorithm	source code	references	dimension
marginalized graph	Java[96]	128	2D
optimal assignment	Java[96]	93,94,95	2D
pharmacophore	C++[100]	101	2D/3D

REFERENCES AND NOTES

(1) Baker, M. Open-access chemistry databases evolving slowly but not surely. *Nat. Rev. Drug Discovery* **2006**, *5*, 707−708.
(2) Diller, D. J.; Merz, K. M., Jr. Can we separate active from inactive conformations? *J. Med. Chem*. **2002**, *16*, 105−112.
(3) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model*. **2005**, *45*, 422.
(4) Carta, G.; Onnis, V.; Knox, A. J. S.; Fayne, D.; Lloyd, D. G. Permuting input for more effective sampling of 3D conformer space. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 179−190.
(5) Weininger, D. SMILES, A Chemical Language for Information Systems. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(6) Weininger, D. SMILES 2: Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(7) Structured Data File format, Inc. MDL Information System.
(8) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *J. Comput. Chem*. **2006**, *27* (16), 1962−1969.
(9) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869.
(10) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational sampling by self-organization. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1186−1191.
(11) Agrafiotis, D. K.; Gibbs, A.; Zhu, F.; Izrailev, S.; Martin, E. Conformational boosting. *Aust. J. Chem*. **2006**, *59*, 874−878.
(12) Agrafiotis, D. K.; Gibbs, A.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model*. **2007**, in press.
(13) Feng, J.; Sanil, A.; Young, S. S. PharmID: Pharmacophore Identification Using Gibbs Sampling. *J. Chem. Inf. Model.* **2006**, *46* (3), 1352−1359.
(14) Rouchka, E. C. A Brief Overview of Gibbs Sampling. http://sapiens.wustl.edu/∼ecr/PAPERS/gibbs.pdf (accessed online 2007-01-30).
(15) Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575−577.
(16) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.
(17) Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 653−681.
(18) Das, P. D.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *PNAS* **2006**, *103*, 9885−9890.
(19) Tenenbaum, J.; de Silva, V.; Langford, J. *Science* **2000**, *290*, 2319−2323.
(20) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34−50.
(21) Bohacek, R. S.; McMartin, C. Multiple highly diverse structures complementary to enzyme binding sites: results of extensive application of a de novo design method incorporating combinatorial growth. *J. Am. Chem. Soc*. **1994**, *116*, 5560.
(22) Boda, K.; Johnson, A. P. Molecular Complexity Analysis of de Novo Designed Ligands. *J. Med. Chem.* **2006**, *49* (20), 5869−5879.
(23) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. *Science* **1996**, *274*, 1531−1534.
(24) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 660−672.
(25) Hajduk, P. J. Fragment-based drug design: how big is too big? *J. Med. Chem.* **2006**, *49*, 6972−6976.
(26) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180−192.
(27) Zhang, S.; Golbraikh, A.; Tropsha, A. Development of Quantitative Structure-Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein-Ligand Interfaces. *J. Med. Chem.* **2006**, *49* (9), 2713−2724.
(28) Tropsha, A.; Carter, C. W., Jr.; Cammer, S. A.; Vaisman, I. I. Simplicial Neighborhood Analysis of Protein Packing (SNAPP): A Computational Geometry Approach to Studying Proteins. In *Methods in Enzymology*; Carter, C. W., Jr., Sweet, R. M., Eds.; Elsevier: 2003; Vol. 374, pp 509−544.
(29) Pauling, L. The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. *J. Am. Chem. Soc.* **1932**, *54*, 3570−3582.

(30) Austin, C.; Brady, L. S.; Insel, T. R.; Collins F. S. *Science* **2004**, *306* (5699), 1138−1139.
(31) The Pubchem Project. http://pubchem.ncbi.nlm.nih.gov/ (accessed month year).
(32) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11473−11478.
(33) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model* **2007**, *47* (1), 47−58.
(34) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On Combining Recursive Partitioning and Simulated Annealing To Detect Groups of Biologically Active Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (2), 393−404.
(35) Davis, et al. *Assay Drug Dev. Technol*. **2007**, in press.
(36) Bredel, M.; Jacoby, E. Chemogenomics: an emergent strategy for rapid target and drug discovery. *Nat. Rev. Genet*. **2004**, *5*, 262−275.
(37) Fischer, H. P.; Heyse, S. From targets to leads: the importance of advanced data analysis for decision support in drug discovery. *Curr. Opin. Drug. Discovery Dev.* **2005**, *8*, 334−346.
(38) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Campbell, R. M. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10*, 839−846.
(39) Root, D. E.; Flaherty, S. P.; Kelley, B. P.; Stockwell, B. R. Biological mechanism profiling using an annotated compound library. *Chem. Biol*. **2003**, *10*, 881−892.
(40) Yan, S. F.; King, F. J.; He, Y.; Caldwell, J. S.; Zhou, Y. Learning from the data: mining of large high-throughput screening databases. *J. Chem. Inf. Model* **2006**, *46*, 2381−2395.
(41) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993−996.
(42) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727−730.
(43) Cleves, A. E.; Jain, A. N. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* **2006**, *49*, 2921−2938.
(44) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.
(45) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*. **1995**, *8*, 127−134.
(46) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(47) Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem. Biol. Drug Des*. **2006**, *67*, 5−12.
(48) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem*. **2004**, *47*, 337−344.
(49) Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem*. **2005**, *48*, 121−133.
(50) Deng, Z.; Chuaqui, C.; Singh, J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J. Med. Chem*. **2006**, *49*, 490−500.
(51) Charette, B. D.; MacDonald, R. G.; Wetzel, S.; Berkowitz, D. B.; Waldmann, H. Protein Structure Similarity Clustering: Dynamic Treatment of PDB Structures Facilitates Clustering. *Angew. Chem., Int. Ed.* **2006**, *45*, 7766−7770.
(52) Kangas, E.; Tidor, B. Charge optimization leads to favorable electrostatic binding free energy. *Phys. Rev. E* **1999**, *59*, 5958.
(53) Sayle, R.; Nicholls, A. Electrostatic evaluation of isosteric analogues. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 191−208.
(54) Lu, B.; Cheng, X.; Huang, J.; McCammon J. A. Order N algorithm for computation of electrostatic interactions in biomolecular systems. *PNAS* **2006**, *103*, 19314−19319.
(55) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
(56) Yu, Z.; Jacobson, M. P.; Friesner, R. A. What Role Do Surfaces Play in GB Models? A New-Generation of Surface-Generalized Born Model Based on a Novel Gaussian Surface for Biomolecules. *J. Comput. Chem.* **2006**, *27* (1), 72−89.
(57) Lorensen, W. E.; Cline, H. E. Marching cubes: A high resolution 3D surface construction algorithm. *Comput. Graphics **1987**, 21, 163.*
(58) Tirado-Rives, J.; Jorgensen W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **2006**, *49*, 5880−5884.
(59) Huang, S. Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866−1875.
(60) Huang, S. Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876−1882.
(61) Wolters, H. J. Geometric modeling applications in rational drug design: a survey. *Comput. Aided Geom. Des.* **2006**, *23* (6), 482−494.
(62) Greenberg, H. J.; Hart, W. E.; Lancia, G. Opportunities for Combinatorial Optimization in Computational Biology. *INFORMS J. Comput.* **2004**, *16* (3), 211−231.
(63) Eppstein, D. Geometry in Action. http://www.ics.uci.edu/∼eppstein/geom.html (accessed month year).
(64) UNC−Duke−Stanford BioGeometry project. http://biogeom.duke.edu (accessed month year).
(65) Medvedev, N. N.; Voloshin, V. P.; Luchnikov, V. A.; Gavrilova, M. L. An Algorithm for Three-Dimensional Voronoi S-Network. *J. Comput. Chem.* **2006**, *27* (14), 1676−1692.
(66) Shatsky, M.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. The multiple common point set problem and its application to molecule binding pattern detection. *J. Comput. Biol.* **2006**, *13* (2), 407−28.
(67) Bandyopadhyay, D.; Snoeyink, J. Almost-Delaunay simplices: nearest neighbor relations for imprecise points. In Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Jan 11−14, 2004, New Orleans, pp 410−419.
(68) Okabe, A.; Boots, B.; Sugihara, K.; Chiu, S. N. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd ed.; Wiley: New York, 2000.
(69) Alinchenko, M. G.; Anikeenko, A. V.; Medvedev, N. N.; Voloshin, V. P.; Mezei, M.; Jedlovszky, P. Morphology of voids in molecular systems. A Voronoi-Delaunay analysis of a simulated DMPC membrane. *J. Phys. Chem. B* **2004**, *108*, 19056−19057.
(70) Anikeenko, A. V.; Alinchenko, M. G.; Voloshin, V. P.; Medvedev, N. N.; Gavrilova, M. L.; Jedlovszky, P. Implementation of the Voronoi-Delaunay Method for Analysis of Intermolecular Voids. *Lect. Notes Comput Sci.* **2004**, *3045*, 217−226.
(71) Medvedev, N. N. Voronoi S-network demo. http://www.kinetics.nsc.ru/mvd/SOFTS/softs_eng.html (accessed month year).
(72) Ambuhl, C.; Chakraborty, S.; Gartner, B. Computing largest common point sets under approximate congruence. *Proc. 8th Ann. European Symp. Alg.* **2000**, 52−63.
(73) Akutsu, T.; Halldorson, M. M. On the approximation of largest common trees and largest common point sets. *Theor. Comput. Sci.* **2000**, *233*, 33−50.
(74) Wolfson, H. J. Model-based object recognition by geometric hashing. *Proc. 1st Eur. Conf. Comput. Vision* **2000**, LNCS, 526−536.
(75) Bandyopadhyay, D.; Huan, J.; Liu, J.; Prins, J.; Snoeyink, J.; Wang, W.; Tropsha, A. Structure-based function inference using protein family-specific fingerprints. *Protein Sci.* **2006**, *15* (6), 1537−1543.
(76) *Chemoinformatics - A Textbook*; Gasteiger, J. G., Engel, T., Eds.; Wiley-VCH: Weinheim, Germany, 2003; ISBN 3-527-30681-1.
(77) *Handbook of Molecular Descriptors*; Todeschini, R., Consonni, V., Eds.; Wiley-VCH: Weinheim, Germany, 2000; ISBN 3-52-29913-0.
(78) MacCuish, J.; Nicolaou, C.; MacCuish, N. E. Ties in Proximity and Clustering Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 134−146.
(79) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.
(80) Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*; Cambridge University Press: 1997; ISBN 0-521-58519-8.
(81) Schölkopf, B.; Tsuda, K.; Vert, J. P. *Kernel Methods in Computational Biology*; MIT Press: Cambridge, MA, 2004.
(82) *Predictive Toxicology*; Helma, C., Ed.; CRC: 2005; ISBN 0-8247-2397-X.
(83) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. In *Chemoinformatics*; Bajorath, J., Ed.; Humana Press: 2004; Vol. 275, pp 1−50, ISBN 1-58829-261-4.
(84) SMIREP. http://www.karwath.org/systems/smirep.html (accessed Jan 1, 2007).
(85) Karwath, A.; Raedt L. D. SMIREP: predicting chemical activity from SMILES. *J. Chem. Inf. Model.* **2006**, *46*, 2432−2444.
(86) Furnkranz, J. Incremental Reduced Error Pruning. *Int. Conf. Machine Learning* **1994**, 70−77.
(87) Klopman, G. MultiCASE: A hierarchical computer automated structure evaluation program - Quantitative Structure-Activity Relationships. *Quant. Struct.−Act. Relat.* **1992**, *11*, 176−184.
(88) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; John Wiley and Sons, Inc.: 1991; ISBN 0-471-06259-6.

(89) Xiong, H.; Tan, P.; Kumar, V. Hyperclique pattern discovery. *Data Min. Knowledge Discovery* **2006**, *13*, 219−242.
(90) Lameijer, E.; Kok, J. N.; Bäck, T.; IJzerman, A. P. Mining a chemical database for fragment co-occurrence: discovery of chemical clichés. *J. Chem. Inf. Model.* **2006**, *46*, 553−562.
(91) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Training similarity measures for specific activities: application to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 577−586.
(92) Schölkopf, B.; Smola A. J. *Learning with Kernels-Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: 2002; ISBN 0-262-19475-9.
(93) Fröhlich, H.; Wegner, J. K.; Zell, A. Assignment Kernels For Chemical Compounds. *International Joint Conference on Neural Networks 2005 (IJCNN'05)*; 2005; pp 913−918.
(94) Fröhlich, H.; Wegner, J. K.; Zell, A. Optimal Assignment Kernels For Attributed Molecular Graphs. *The 22nd International Conference on Machine Learning (ICML 2005*); Omnipress: Madison, WI, U.S.A., 2005; pp 225−232.
(95) Fröhlich, H.; Wegner, J. K.; Zell, A. Kernel Functions for Attributed Molecular Graphs - A New Similarity Based Approach To ADME Prediction in Classification and Regression. *QSAR Comb. Sci.* **2006**, *25*, 317−326.
(96) JOELib2. http://joelib.sf.net (accessed Jan 1, 2007).
(97) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
(98) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. 'Scaffold-Hopping' by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.
(99) Rarey, M.; Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided. Mol. Des.* **1998**, *12*, 471−490.
(100) Pharmacophore kernel. http://chemcpp.sourceforge.net/html/index.html (accessed Jan 1, 2007).
(101) Mahé, P.; Ralaivola, L.; Stoven, V.; Vert, J. The pharmacophore kernel for virtual screening with support vector machines. *J. Chem. Inf. Model.* **2006**, *46*, 2003−2014.
(102) *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective*; Kubinyi, H., Müller, G., Mannhold, R., Folkers, G., Eds.; Wiley-VCH: 2004; ISBN 3-527-30987-X.
(103) Guha, R.; Howard, M.; Hutchison, G.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J. K.; Willighagen, E. L. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991−998.
(104) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open−Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500. http://cdk.sf.net (accessed Jan 1, 2007).
(105) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.
(106) MoSS. http://fuzzy.cs.uni-magdeburg.de/∼borgelt/moss.html (accessed Jan 1, 2007).
(107) Meinl, T.; Berthold M. R. *Hybrid Fragment Mining with MoFa and FSG*, Proceedings of the 2004 IEEE Conference on Systems, Man & Cybernetics (SMC2004), 2004.
(108) Meinl, T.; Borgelt, C.; Berthold, M. R.; Philippsen, M. *Mining Fragments with Fuzzy Chains in Molecular Databases*, Proceedings of the 2004 IEEE Conference on Systems, Man & Cybernetics (SMC2004), 2004.
(109) Meinl, T.; Borgelt, C.; Berthold M. R. *Discriminative Closed Fragment Mining and Pefect Extensions in MoFa*, Proceedings of the Second Starting AI Researchers Symposium (STAIRS 2004), 2004.
(110) ParMol. http://www2.informatik.uni-erlangen.de/Forschung/Projekte/ParMol/ (accessed Jan 1, 2007).
(111) Wörlein, M. Extension and parallelization of a graph-mining-algorithm, Master Thesis, Friedrich-Alexander-Universitaet, 2006.
(112) PolyFARM. http://www.aber.ac.uk/compsci/Research/bio/dss/polyfarm/ (accessed Jan 1, 2007).
(113) Clare, A.; King R. D. *Data mining the yeast genome in a lazy functional language*, Practical Aspects of Declarative Languages (PADL2003), 2003.
(114) Warmr. http://www.cs.kuleuven.be/%7Edtai/ACE/ (accessed Jan 1, 2007).
(115) King, R. D.; Srinivasan, A.; Dehaspe, L. Warmr: a data mining tool for chemical data. *J. Comput.-Aid. Mol. Des.* **2001**, *15*, 173−181.
(116) Dehaspe, L.; Toivonen, H.; King R. D. Finding frequent substructures in chemical compounds. *4th International Conference on Knowledge Discovery and Data Mining*; AAAI Press: 1998; pp 30−36.
(117) Inokuchi, A.; Washio, T.; Nishimura, K.; Motoda, H. A *Fast Algorithm for Mining Frequent Connected Subgraphs*; IBM Research, Tokyo Research Laboratory: 2002.
(118) Inokuchi, A.; Washio, T.; Okada, T.; Motoda, H. Applying the Apriori−based Graph Mining Method to Mutagenesis Data Analysis. *J. Comput.-Aided Chem.* **2001**, *2*, 87−92.
(119) Ando, H.; Dehaspe, L.; Luyten, W.; Craenenbroeck, E.; Vandecasteele, E.; Meervelt, L. Discovering H-Bonding Rules in Crystals with Inductive Logic Programming. *Mol. Pharm.* **2006**, *3*, 665−674.
(120) Nijssen, S.; Kok J. N. *Frequent Graph Mining and its Application to Molecular Databases*, Proceedings of the 2004 IEEE Conference on Systems, Man & Cybernetics (SMC2004), 2004.
(121) optimized gSpan. http://wwwkramer.in.tum.de/projects/gSpan.tgz (accessed Jan 1, 2007).
(122) Jahn, K.; Kramer, S. *Optimizing gSpan for Molecular Data Sets*, Proceedings of the Third International Workshop on Mining Graphs, Trees and Sequences (MGTS-2005), 2005.
(123) Yan, X.; Han, J. *gSpan: Graph-Based Substructure Pattern Mining*, Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), IEEE Computer Society: 2002; pp 721−724.
(124) Helma, C.; Cramer, T.; Kramer, S.; de Raedt, L. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402−1411.
(125) Mazzatorta, P.; Tran, L.; Schilter, B.; Grigorov, M. Integration of Structure-Activity Relationship and Artificial Intelligence Systems To Improve in Silico Prediction of Ames Test Mutagenicity. *J. Chem. Inf. Model.* **2007**, *47*, 34−38.
(126) LAZAR. http://www.predictive-toxicology.org/lazar/ (accessed Jan 1, 2007).
(127) Wegner, J. K.; Froehlich, H.; Mielenz, H.; Zell, A. Data and Graph Mining in Chemical Space for ADME and Activity Data Sets. *QSAR Comb. Sci.* **2006**, *25*, 205−220.
(128) Kashima, H.; Tsuda, K.; Inokuchi, A. Marginalized Kernels Between Labeled Graphs. *The 20th International Conference on Machine Learning (ICML2003*); 2003.

CI700059G