

# Retrospect and Prospect of Virtual Screening in Drug Discovery

Huafeng Xu\* and Dimitris K. Agrafiotis

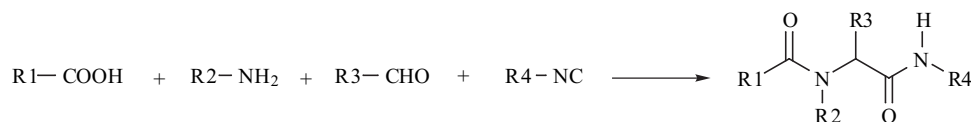
3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, PA 19341, USA

**Abstract:** We review the prominent technologies in virtual screening, and their applications in drug discovery.

## INTRODUCTION

Pharmaceutical companies are in pursuit of two goals: cure and profit. To achieve these goals, new drugs must be discovered at a fast pace to win the race against disease and expiring patents. Consequently the industry allocates a significant budget for research and development, and, as a result, enjoys the reward of being the second most profitable

compounds, and high-throughput screening enables expedient detection of those few that are biologically active. The numbers, however, are not in our favor. Consider, for example, the library based on the Ugi reaction illustrated in Fig. 1. If each R-group has 100 different variations,  $100^4 = 100,000,000$  distinct compounds can be assembled. In fact, it has been estimated that  $10^9$  compounds could, in theory, be synthesized using this chemistry from commercially



**Fig. (1).** Reaction scheme for the construction of the Ugi library.

industry of legitimate business, second only to tobacco, to which, ironically, is attributed a significant fraction of the pharmaceutical market. Discovery of new drugs, however, is becoming increasingly difficult, costly, and time consuming due to growing regulatory pressure from the FDA and counterpart agencies in other countries. In response, pharmaceutical companies have eagerly embraced a number of new technologies that promise to accelerate the pace and reduce the cost of discovering new drugs. Virtual screening of compound libraries in search of drug leads is one of these technologies.

In traditional medicinal chemistry, a lead compound is usually identified by screening collections of compounds consisting of natural products and other chemicals that organic chemists have toiled years to synthesize. A great leap in lead discovery came with the advent of combinatorial chemistry and high-throughput screening (HTS). Combinatorial chemistry provides access to millions of new

available starting materials alone [1]. Despite the ever-increasing capacity of synthetic and screening machinery, only a tiny fraction of these libraries can be materialized and tested in a biological assay. This situation presents some challenging opportunities and risks.

In fact, blind screening of large combinatorial libraries for drugs has had few conclusive success stories. This is because drugs are sparse among chemicals, and they are not uniformly distributed throughout the chemical space [2]. The screening capability of the pharmaceutical industry has increased by several orders of magnitude since the inception of modern medicine, but the speed at which new compounds enter advanced clinical trials and win FDA approval has accelerated very little by comparison. The number of new drugs is in no way proportional to the number of new compounds synthesized and screened. This is analogous to mathematicians' search for prime numbers [3]. There are 25 prime numbers smaller than 100 with a density of 0.4, 1229 prime numbers smaller than 10,000 with a density of 0.12, 78,498 prime numbers smaller than 1,000,000 with a density of 0.078, and 5,761,455 prime numbers smaller than 100,000,000 with a density of only 0.058. The prime number theorem states that the number of prime numbers smaller than  $n$  asymptotically approaches  $n/\ln n$ .

\*Address correspondence to this author at the 3-Dimensional Pharmaceuticals, Inc., 8 Clarke Drive, Cranbury, NJ 08512, USA; Tel: (609) 655-6978, Fax: (609) 655-6930, e-mail: huafeng.xu@3dp.com

Consequently the prime numbers are infinitely sparse among all integers. Likewise, drugs may be infinitely sparse among all accessible chemicals. Similar to the lack of a mathematical formula that will consistently generate prime numbers, medicinal chemists do not have a chemical formula to generate drugs.

This analogy, however, is not entirely valid. For many decades, the pharmaceutical industry has been successful in developing new medicines through a convergent SAR optimization process based on trial and error. This suggests that although the number of drugs may be vanishingly small compared to the size of chemistry space, the “fitness landscape” of drug-like matter exhibits relatively well-defined minima (drugs), surrounded by a large number of promising – albeit sub-optimal – molecules (hits and leads). Indeed, the process of finding drugs is considerably simpler than that of finding prime numbers, but it does require careful experimental design and good judgment. Virtual screening is a tool to aid such judgment. In addition to automating the selection process, virtual screening expands, by several orders of magnitude, the number of compounds that can be assessed as potential drugs. Utilizing the increasing power of computers, it is now possible to virtually “synthesize” billions of chemical compounds, analyze their properties, and select the ones that will most likely possess activity against the biological target(s) of interest [4].

## REALISTIC OBJECTIVES OF VIRTUAL SCREENING

Small molecule drugs act by binding to their respective protein (or nucleic acid) targets. To identify potential drug leads, we need to identify compounds that have specific affinity for the target of interest. In an ideal world, where molecular interactions and solvation effects are perfectly described, the binding free energy of each compound to the target can be computed, and the best compounds can be identified by sorting them according to their binding free energy. However, neither molecular interactions nor solvation effects can yet be fully described in a model that is amenable to computation, and simplified formalisms must be employed to enable the analysis of large virtual collections.

The use of simplified models means that the predicted binding affinities may differ substantially from their true values. Since in most cases drugs bind to their targets through reversible noncovalent interactions such as hydrogen bonds, hydrophobic interactions, and salt bridges, the binding free energy is small compared to covalent bonding. A small inaccuracy in the predicted binding affinity can alter the judgment between a strong and a weak binder. Therefore the output of virtual screening could include a significant number of compounds that are not truly bioactive (false positives), and omit others that are (false negatives). A good method must minimize the number of both false positives and false negatives. The ability to do so can be assessed through validation experiments, in which the method is used to screen a well-studied chemical library and protein target, and the results are compared to those obtained by actual

screening. Suppose that the library contains  $L$  compounds, of which  $D$  have been experimentally identified as bioactive. The virtual screening protocol picks  $N$  compounds as potential drug leads, among which  $d(N)$  compounds are truly bioactive (obviously  $d(N)$  is a function of  $N$ ). A random selection of  $N$  compounds from that library would result in  $DN/L$  bioactive entities. The quality of the method can thus be measured by the ratio of the percentage of truly bioactive compounds in its selection over that in a random selection, the so-called *enrichment factor*,  $f(N) = (d(N)/N) / (D/L)$ . For the method to be useful, the enrichment factor should be consistently greater than 1. The difference  $g(N) = f(N) - 1$  reflects the improvement of virtual screening over random selection. A common practice is to plot  $d(N)/D$  versus  $N/L$ , which should generate a curve well above the diagonal line.

Just as with many other promising technologies, it is important to have realistic expectations of the potential of virtual screening. It is impossible for any method to select all and only the compounds that are truly bioactive. From a practical point of view, the method should be employed as long as the corresponding enrichment factor  $f(N)$  is consistently greater than one and consistently better than that obtained by any alternative selection method, including experience and chemical intuition. Of course, for new chemical libraries and new targets the enrichment factor is not known *a priori*, and the decision as to whether the method should be employed must be based on prior experience and well-crafted, statistically sound retrospective validation studies.

## VIRTUAL SCREENING USING CHEMICAL DESCRIPTORS

Even though molecules interact with each other in 3-dimensional space and manifest their particular properties by adopting certain 3-dimensional conformations, chemists have long speculated that valuable information about molecular interactions and properties can be extracted directly from the connection table. Many different chemical descriptors based on invariants of the molecular graph have been employed in quantitative structure-activity relationship (QSAR) studies and virtual screening. The most popular ones – substructure keys, hashed fingerprints and topological indices – are described below. More extensive reviews can be found elsewhere [4-7].

### 1D/2D Chemical Descriptors

Substructure keys [8] are used to describe what substructures are present in a molecule. These keys are binary vectors, whose  $i$ -th field is set to 1 if the corresponding substructure is present in the molecule, 0 otherwise. The substructures are predefined, and range from simple atoms (carbon, nitrogen, *etc.*) to specific electronic configurations such as triple-bonded nitrogen, common functional groups such as alcohols or amines, and ring systems such as 5-member rings or specific heterocycles. Different substructures are used for different purposes. For drug screening, substructures of interest may include

hydrogen-bonding groups, hydrophobic fragments, or specific skeletal features such as steroids, *etc.*

Hashed fingerprints [9] also encode substructural information but, unlike substructure keys, the substructures to be encoded are not predefined. Instead, all possible patterns of a specific type (such as in a path of up to 3 bonds) are generated by systematic traversal of the molecular graph. Each pattern is converted to an integer number, which serves as a seed to a pseudorandom number generator, which turns on a small number of bits in the respective fingerprint (a process called hashing). Although, in theory, different molecules could share the same fingerprint, the probability of such a circumstance is extremely low if the hashing scheme is designed in a sensible manner. Both substructure keys and hashed fingerprints were originally designed for fast substructure search of chemical databases, but subsequently found widespread use in similarity searching, structure-activity correlation and virtual screening.

The third class of graph theoretic descriptors are connectivity indices [10]. The generalized molecular connectivities  $\chi$  are defined as:

$${}^l\chi = \sum_{\text{all paths of length } l} \frac{1}{\sqrt{\prod_i \delta_i}}$$

where the summation is taken over all the paths of length  $l$  in the molecule, the product is taken over all the atoms in each path, and  $\delta_i$  is the number of bonds that atom  $i$  forms with non-hydrogen atoms. The expression has been modified to include different atomic species, by replacing  $\delta_i$  with  $\delta_i^Z = Z_i - n_i^H$ , where  $Z_i$  and  $n_i^H$  represent the number of valence electrons and the number of attached hydrogen atoms of the  $i$ -th atom, respectively. The  $\chi$  indices have been used extensively in QSAR studies, and have been shown to correlate well with a number of molecular properties, such as anesthetic potencies [11], toxicities [12], and inhibitory potencies [13]. Similar to the  $\chi$  indices, the  $\kappa$  indices were devised to encode molecular shape, and are based on the number of paths of a specific length. A full description can be found in a review article by Hall and Kier [10].

### Similarity Metrics

Molecular descriptors can be combined using a suitable distance function to provide a numerical estimate of the similarity between two compounds. A number of distance functions have been proposed, including the well-known Manhattan and Euclidean metrics for real-valued descriptors, and the Tanimoto coefficient for binary descriptors. In general, the diversity of chemical structure encourages the use of large descriptor sets to provide adequate structural and/or biological discrimination. However, the more descriptors are used to describe the molecules, the greater the likelihood that they are correlated. Dimensionality reduction attempts to minimize this redundancy by eliminating features that add very little to the overall picture. Dimensionality reduction techniques fall into two broad categories: 1) methods that preserve some of the original descriptors (such as fast random elimination of descriptors

[14], cluster significance analysis [15], and information theory [16]), and 2) methods that generate alternative *latent* features based on the original descriptors. The latter may be accomplished using linear methods such as principal component analysis [17], singular value decomposition [18,19] and factor analysis [20], and nonlinear methods such as multidimensional scaling [21] (MDS) and nonlinear mapping [22] (NLM).

The most popular linear dimensionality reduction method is principal component analysis (PCA). Let  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$  and  $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$  be the vectors of descriptors of two compounds, and  $d^2 = \sum_k (x_k - y_k)^2$  the (squared) Euclidean distance between them. If the descriptors are highly correlated, the Euclidean metric will result in amplifying the contribution of some descriptors at the expense of others. The correlation between the descriptors can be measured by the covariance matrix  $\mathbf{C} = \{\sigma_{ij}\}$ , where  $\sigma_{ij} = \langle (x_i - \bar{x}_i)(x_j - \bar{x}_j) \rangle$  is the covariance between the descriptors  $x_i$  and  $x_j$ , as measured by averaging over all the compounds in the library. A more unbiased estimate of the dissimilarity between the compounds can be obtained using the Mahalanobis distance [23]:

$$d^2 = (\mathbf{x} - \mathbf{y})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{y})$$

where  $\mathbf{C}^{-1}$  is the inverse of the covariance matrix  $\mathbf{C}$ , and  $\mathbf{x}^T$  denotes the transpose of  $\mathbf{x}$ .

Since  $\mathbf{C}$  is a symmetric matrix, we can solve the eigenvector problem  $\mathbf{C}\mathbf{v}_k = \lambda_k \mathbf{v}_k$ , and order the  $n$  real eigenvalues in descending order. The eigenvectors  $\{\mathbf{v}_k\}$  form an orthonormal basis for the  $n$ -dimensional space. In this basis, the covariance matrix is diagonal:

$$\mathbf{C}' = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix},$$

and its inverse is simply:

$$\mathbf{C}'^{-1} = \begin{pmatrix} \lambda_1^{-1} & & & \\ & \lambda_2^{-1} & & \\ & & \ddots & \\ & & & \lambda_n^{-1} \end{pmatrix}.$$

The descriptor vectors in the new basis are  $\mathbf{x}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n\}$  and  $\mathbf{y}' = \{\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_n\}$ , and the Mahalanobis distance is given by:

$$d^2 = \sum_i (x'_i - y'_i)^2 / \lambda_i.$$

The eigenvectors  $\{\mathbf{v}_k\}$  are usually called the principal components (PC). Frequently a few eigenvalues will dominate the variance in the dataset. A common practice is

to select only the largest few eigenvalues and the corresponding eigenvectors in the analysis, and treat the variance corresponding to the remaining eigenvalues as statistical noise.

Chemical space, however, is very unlikely to be linear. For this reason, it is preferable to reduce the dimensionality using a nonlinear method such as MDS or NLM. These methods attempt to embed the compounds in a lower-dimensional Euclidean space in a way that preserves their distances in the original high-dimensional manifold. These methods are particularly valuable because they can also be used to produce Cartesian coordinate vectors from data supplied directly in the form of proximities, which simplifies their analysis using conventional statistical and data mining techniques. Although MDS is a computationally intensive procedure, our group has developed a number of methods involving the use of neural networks that allow the expedient mapping of very large data sets such as those encountered in combinatorial chemistry and virtual screening [24]. This general strategy was subsequently extended to employ local learning techniques [25], generalized to handle complex distance functions and input data supplied in non-vectorial form [26], and modified to allow the scaling of combinatorial libraries in a way that circumvents explicit enumeration [27].

### Virtual Screening by Similarity Search

Similarity searching is one of the oldest and most trusted methods for computer-assisted lead discovery and optimization [7]. Once a set of active inhibitors has been identified, computers can be used to search large chemical libraries for molecules that closely resemble these compounds. The search is not restricted to compounds that have already been synthesized, but may include synthetically accessible compounds that exist only in virtual form.

In a very interesting study, Schneider *et al.* used a reference molecule to guide the *de novo* design of K<sup>+</sup>-channel blocking agents through an evolutionary algorithm [28]. The reference molecule was first simplified to an abstract representation consisting of generalized atoms (hydrogen bond donors/acceptors, positive/negative charges, lipophilic groups) connected by bonds. Correlation vectors were then used to encode this abstract representation [29], and the distance between these vectors was used as a measure of the similarity between the corresponding compounds. New candidate molecules were designed by an evolutionary protocol, guided by a fitness function based on the distance of the correlation vector from that of the reference structure. Using this approach, the researchers were able to design a novel potent K<sup>+</sup>-channel blocker.

The ability to search large virtual libraries has become increasingly important with combinatorial chemistry, which provides access to billions of chemical analogs from a relatively limited number of building blocks. These virtual libraries are orders of magnitude larger than a typical chemical bank, and several sophisticated algorithms have been devised to simplify their search and manipulation.

Our group's efforts have been based on the concept of "lazy enumeration", that is, delaying virtual synthesis and storage of a particular product until it is needed [4]. These algorithms offer enumeration speeds in excess of 20,000 products/second, and can compress a billion-member library into a 1-5 Mb data stream in few seconds on a modern PC. In addition, two complementary similarity searching algorithms have been devised that capitalize on the massive redundancy and constrained nature of combinatorial libraries. The first is based on the observation that the structural diversity of a combinatorial library stems from a limited number of building blocks, so that it is possible, through random sampling, to identify reagents leading to the products that are most closely related to the query structure [30]. The second makes use of an optimization heuristic that allows the independent evaluation and ranking of candidate reagents in each variation site in the library, thus breaking the task into a set of well-defined, computationally tractable sub-problems [31].

However, in our experience the most important advance is the ability to circumvent the enumeration process altogether, and perform similarity searching and other related calculations using approximate descriptors calculated on-the-fly as needed by the application. This is accomplished using feed-forward neural networks, which are trained to predict descriptors of products from pertinent features of their respective building blocks [32]. The training set consists of the descriptors of all the reagents that make up the combinatorial library, along with the descriptors of a relatively small number of randomly chosen products. The latter are computed in a conventional way, i.e. by running a computer program or subroutine on the fully enumerated structures. Once the networks are trained, screening the virtual library (or any subset thereof) becomes a matter of retrieving the precomputed descriptors of the reagents, feeding them through the neural networks in order to compute the descriptors of the products, and using these descriptors for any subsequent analysis, searching, or classification task. Unlike previous attempts, which focused exclusively on decomposable descriptors [33-36], this method is general and can be applied to a wide variety of molecular properties, regardless of origin and complexity. The algorithm is fast, and permits the screening of virtual libraries at a rate of tens of thousands of compounds per second on a modern personal computer.

2D descriptors are also gaining popularity in designing diverse chemical libraries for general screening. Diversity analysis has been reviewed in several recent texts, and will not be addressed in this work. We would like to note, however, that it is now becoming increasingly apparent that many other factors need to be considered besides diversity in order to increase the quality and information content of a screening deck. Successful drug development requires the optimization of many additional properties such as solubility, uptake and distribution, metabolism, pharmacokinetics, toxicity, and chemical and metabolic stability. Lipinski's rule-of-5 [2] is an example of how 2D descriptors coupled with simple statistical reasoning can be used to flag compounds with potential development liabilities. Lipinski's work, along with the successful history of QSAR, has prompted several authors to develop

more rigorous statistical models to predict drug-likeness in a more general sense [37-40]. This approach is based on the thesis that drugs have certain common properties that differentiate them from ordinary chemicals, and the belief that it should be possible to extract that information through statistical analysis of large databases of marketed drugs and advanced clinical candidates. In practice, these methods are used either as "hard" filters that eliminate problematic compounds from further consideration [1], or "soft" biases that enrich the selection with compounds with a more favorable ADME profile [41,42].

## PHARMACOPHORES

Because a drug interacts with its protein target in 3D space, 2D chemical descriptors are incapable of fully describing these interactions. Thus, it is often desirable to search for compounds that share common 3-dimensional structural elements with known inhibitors. This is usually accomplished by detecting common spatial arrangements of specific functional groups among a series of known inhibitors that correlate well with biological activity. These functional groups with their spatial configuration are commonly called pharmacophores [43]. Once the pharmacophores have been extracted by superimposition of known inhibitor structures, they can be used to query a 3-dimensional structural database of small molecules, such as the Cambridge Structure Database (CSD), or some other physical or virtual collection. Molecules that contain similar pharmacophoric elements will be retrieved as potential leads for further study. Chen *et al.* used this approach to discover a novel scaffold for inhibitors of rat 5 $\alpha$ -reductase [44].

To simplify search and retrieval, pharmacophores are often encoded in the form of long binary vectors referred to as pharmacophore fingerprints. This mechanism allows pharmacophore perception and screening through simple and fast binary operations. The pharmacophore itself is typically represented as a set of 3 or 4 pharmacophoric centers forming a triangle or tetrahedron. These centers include macromolecular recognition sites, such as charged centers, hydrogen bond donors and acceptors, hydrophobic centers, and aromatic ring centers. To generate the fingerprint, the pharmacophores (i.e. the pharmacophoric centers and their respective distances) exposed by a particular conformation are mapped onto specific bits in the bit string. Individual fingerprints can be combined into "molecular fingerprints" which represent the union across all conformations of a particular molecule, and "library fingerprints" which represent the union across all molecules in a library. The latter has been the method of choice for defining the pharmacophoric diversity of a collection of compounds. These descriptors are available from a number of commercial molecular modeling packages and have been widely adopted by the computational chemistry community.

The problem with pharmacophoric descriptors is that they collapse on a massive scale. The large number of possible products and the large number of possible low-energy conformations per product limit the application of pharmacophoric techniques in four important ways: 1) to relatively small virtual libraries, 2) to only 3- and 4-point

pharmacophores, 3) to inadequate conformational sampling, and 4) to simplistic similarity and/or diversity measures. Perhaps the biggest source of error comes from inadequate conformational sampling. To be meaningful in a biological context, a conformational search algorithm must be able to produce an accurate and thorough account of every molecular conformation that is accessible in a particular biological setting. Most conformational search algorithms perform calculations either in vacuum or in a simple polarizable continuum. Moreover, as the flexibility of the molecule increases, thorough conformational sampling becomes impossible, and drastic measures are taken to truncate the size of the search space. Missing or erroneous conformations can introduce significant errors in the resulting fingerprint.

Another problem comes from aggregation. When the pharmacophore keys of the individual conformations are combined to produce the molecular fingerprint, higher-order relationships are lost. That is, the relative geometric arrangements of, for example, 4 or more pharmacophoric centers cannot be inferred from a set of potentially overlapping 3-point pharmacophores, because these pharmacophores may have originated from different conformations, and could never be simultaneously present in a single conformation. As Kahn has recently demonstrated, the use of higher order pharmacophores results in increasing ability to separate active from inactive molecules, but the complexity of the calculations increases accordingly.

These problems, and the instinctive desire to somehow account for the third dimension, has prompted several groups to look at substituent-based pharmacophoric descriptors. The most prominent of these methods are Cramer's topomerically aligned conformer fields [45], and Martin's oriented substituent pharmacophores [46].

The former are based on the steric fields of single side chain conformers topomerically aligned around a common combinatorial core. This alignment tries to find a representative conformation for each side chain attached to a particular variation site on a combinatorial template. First, a model-building routine generates a low energy conformation, which is then fitted as a rigid body onto the template using least-squares minimization. After that, the bond torsions are adjusted one at a time starting from the bond closest to the template, according to a simple set of topological precedence rules. Once the alignment is complete, the steric field of the side chain is calculated using a CoMFA-like approach. These fields can be used to compute a similarity index between two compounds from the root of the squares of the differences in steric field values summed over all lattice points in the CoMFA region, or another equivalent distance function.

In the special case of combinatorial libraries, Martin's approach is to employ pharmacophore keys that represent the substituents rather than the final products [46]. To ensure that the relative orientation of the substituents is accounted for, two additional points and the corresponding distances are added to each substituent pharmacophore. The first is the point of attachment to the template, and the second is an orienting point placed at some distance away from the attachment point along the bond attaching the substituent to

the scaffold. In theory, by dissecting the molecule into its constituent parts, one can construct a richer, more informative representation of the pharmacophoric space accessible to each compound. In practice, the results are only as valid as the underlying combinatorial conformer and template alignment assumptions (i.e. the premise that the conformations adopted by a particular substituent do not depend on the conformations adopted by the other substituents or the orientation of the scaffold). However, the most important limitation of these types of approaches is that the descriptors are inherently specific to a particular combinatorial library, and cannot be used to compare that library with other combinatorial libraries, non-combinatorial collections, or even individual compounds.

Of course, pharmacophore detection is not the only approach for introducing 3-dimensional information in virtual screening. A variety of alternative descriptors have been explored, including geometric atom pairs and topological torsions, spatial autocorrelation vectors, WHIM indices, molecular hashkeys, and BCUTs, to name a few.

## DOCKING

Despite its long, successful history in drug design, similarity searching suffers from a number of limitations. The method cannot be used with targets for which no active compounds are known, and, when 2D descriptors are used, has a tendency to discover molecules that closely resemble the input queries, thus limiting the ability to discover novel, patentable scaffolds. A promising alternative is to model the 3-dimensional structure of the receptor-ligand complex and assess its stability, a process known as small molecule docking.

Experimentally, the structures of receptor-ligand complexes can be determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. This approach, however, is excruciatingly slow. The Protein Data Bank (PDB) currently contains about 18,000 structures, and grows at a pace of 7 additional structures per day [47]. This represents the collective effort of thousands of crystallographers and NMR spectroscopists working in the field. In practice, experimental structure determination can only be used for a select number of ligands, and serves as a source of ideas for analog design.

Small molecule docking attempts to model the key interactions involved in ligand binding, and extract plausible 3-dimensional models and binding affinities that can be subsequently used to prioritize synthetic candidates. The basic assumption is that if a small molecule binds to a protein with good affinity, the geometric shape of that molecule has to fit with that of the cavity (or binding pocket) in the protein. The fit between the small molecule and the binding pocket is then used to evaluate the affinity of the small molecule for the respective target. If the fit is poor, steric hindrance will prevent the ligand from binding to the protein, and the molecule will be rejected as an improbable lead.

As a protein interacts with a ligand, both may undergo conformational changes [48-52]. Theoretically, such conformational flexibility should be considered in the model, but its effective treatment is computationally very expensive. Consequently, several levels of simplification are introduced to reduce the complexity of the task, which differ in what portion of the protein-ligand complex is treated as flexible.

## Rigid Docking

When both the ligand and the binding pocket of the protein are held rigid, the docking problem is reduced to finding a match between characteristic features of the two molecules in 3-dimensional space. Usually a distance compatibility graph is constructed in order to identify potential matches. Each pair of matching features between the protein and the ligand is a node in the graph. If two pairs of matching features have the same distance within certain tolerance, the two nodes that represent the two pairs are joined by an edge. For example, consider a binding pocket that consists of two hydrophobic cavities (R1 and R2) and one hydrogen bond donor site (D), and a candidate ligand that consists of three lipophilic sites (r1, r2, and r3) and one hydrogen bond acceptor (a). Then the nodes in the graph are R1r1, R2r1, R1r2, R2r2, R1r3, R2r3, and Da (see Fig. 2). If the distance between R1 and R2 is the same as that between r1 and r3, the nodes R1r1 and R2r3 are joined by an edge. A 3-dimensional match between the two molecules is thus equivalent to a fully connected subgraph in the distance compatibility graph, and finding the correct docking orientation is reduced to a search for the largest fully connected subgraph. In graph theory, such a fully connected subgraph is called a clique, and many algorithms have been developed to find the maximum cliques in a graph [53,54]. Once a clique is found, the matching features in the clique are superimposed to give an initial orientation of the ligand with respect to the binding pocket.

DOCK [55,56], one of the most widely used docking programs, evaluates the orientation of the ligand using a scoring function based on the AMBER force field. Since the starting orientation is quite crude, the ligand may have bad van der Waals (VDW) contacts with the receptor and thus a high energy. Normally, this would lead to rejection of nearly every docking conformation. To circumvent this difficulty, the program uses a soft VDW potential for the ligand, where the VDW well depth ( $\epsilon$ ) is scaled down by a factor of 1000, corresponding to about 1 Å extra displacement into the VDW repulsive core. The rationale behind this approach is to allow favorable electrostatic interactions to dominate the orientation of the ligand, and resolve bad VDW contacts at a later stage in the optimization.

Another technique applicable to rigid molecular docking is geometric hashing [57-60]. This technique was originally developed in the field of computer vision to recognize partially occluded objects. Geometric hashing is a model-based method. The program first preprocesses all the known scenes (models) and stores them into "memory". When presented with a new scene, the program attempts to find a model in memory that best matches the new scene. A scene

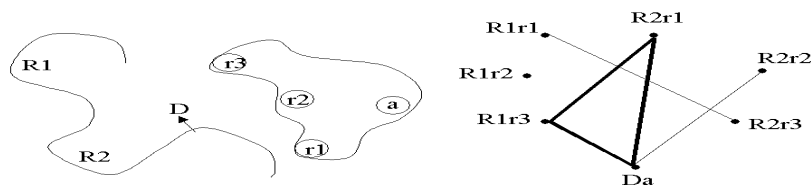


Fig. (2). Distance compatibility graph.

is represented by a set of points and their coordinates. In the case of protein docking, a scene consists of characteristic structural features and their spatial coordinates. Every triplet of non-collinear points in a scene can uniquely specify a body-fixed reference frame (for instance, for a triplet of points  $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ , the reference frame can be determined by placing  $\mathbf{p}_1$  at the origin, setting  $\mathbf{e}_1 = \mathbf{p}_2 - \mathbf{p}_1$  as the x-axis,  $\mathbf{e}_3 = \mathbf{e}_1 \times (\mathbf{p}_3 - \mathbf{p}_1)$  as the z-axis, and  $\mathbf{e}_2 = \mathbf{e}_3 \times \mathbf{e}_1$  as the y-axis). The position of each point in the scene can then be specified by its coordinates with respect to this reference frame. Geometric hashing compares two scenes using these body-fixed coordinates. If two scenes are identical, for every triplet in one scene, there exists a matching triplet in the other, such that in the respective body-fixed reference frame, the set of coordinates of all the points in each scene are identical to each another. If two scenes are similar but not identical, such a one-to-one correspondence no longer exists between the triplets. However, there should still be some pairs of matching triplets in each scene, such that the set of coordinates in the respective body-fixed reference frame are similar. In the preprocessing stage, every triplet of non-collinear points in a model scene is used to establish a body-fixed reference frame, and the coordinates in the reference frame of every point in the scene are binned and used as a key in a hash table. The hash entry consists of the model's identity, the triplet used to determine the reference frame, and the type of the point. Apparently both the time and space complexity of preprocessing scales as  $O(Mn^4)$ , where  $M$  is the number of models, and  $n$  is the number of points in each model (by using a two dimensional coordinate system, the complexity can be reduced to  $O(Mn^3)$  [61]). In the recognition stage, a triplet of non-collinear points in the unknown scene is used to establish a body-fixed reference frame, and the coordinates of all the points in the scene are used to generate hash keys as in the preprocessing stage. For each hash key, the method retrieves all the stored hash entries corresponding to that key, and each model and triplet in the retrieved hash entry receives a vote. If the new scene matches a model in memory, that model and the corresponding triplet will receive a large number of votes. The method can therefore decide whether the new scene matches a memorized model by looking at whether the model receives a large number of votes (more than a predefined threshold). The hashing technique enables recognition to occur in constant time, if the number of bins

in the hash table is chosen to be on the same order as the number of models [62].

To apply geometric hashing to molecular docking, the ligands are simply represented by feature sites, such as lipophilic centers, hydrogen bond donor sites, etc. All the conformations of all the ligands serve as models. Every low-energy conformation of every ligand is preprocessed and hashed in a database. Complementary feature sites are then extracted from the binding pocket of the protein target. The extracted representation is the scene to be recognized. This allows ligand conformations that fit into the binding pocket to be easily identified.

In theory, geometric hashing has the advantage that the ligand library can be processed once and for all, and protein targets can then be recognized at constant time. At this moment, however, such an approach is impractical because of the enormous storage requirements of preprocessing.

### Flexible Ligand / Rigid Receptor Docking

In a small molecule, most of the conformational degrees of freedom stem primarily from rotatable bonds. Most drug-like molecules possess between two and eight rotatable bonds [63]. Because the energetic differences between different conformations are often small relative to the binding energies, a ligand may bind to the receptor in a conformation different from the one with the lowest energy *in vacuo*. Consequently, the flexibility of the ligand must be taken into account explicitly by the docking method, if it is to be successful in discriminating active from inactive compounds.

### Conformation Ensembles

The most obvious way to include ligand flexibility is to individually dock every reasonable conformation of the target molecule. Under this paradigm, all low energy conformations of the ligands in the library are identified and stored in a database, and a rigid docking program docks each of these conformations into the protein target. An exemplary system that employs such conformational ensembles is the Flexibas/FLOG suite [64] developed by Kearsley *et al.* For each compound in the library, the program generates a

number of conformations using distance geometry coupled with a crude energy minimization scheme, selects up to 25 diverse conformations per molecule using an RMSD dissimilarity score, and stores these conformations in a database. The rigid docking program FLOG [65] is then used to dock each conformation in the receptor of interest.

### **Fragmentation**

Another approach is to “cut” the rotatable bonds and divide the molecule into a set of rigid fragments, and are then docked individually using the methods developed for rigid docking. In the “place-and-join” approach, the fragments are docked independently in the receptor, and then joined together by reconnecting the broken bonds. In the “incremental construction” approach, a core fragment is first docked, and the remaining fragments are grown from the core in energetically favorable orientations.

The place-and-join approach is the method of choice in *de novo* drug design [66,67], and was originally applied to flexible ligand docking by DesJarlais *et al.* [68]. The ligand was manually divided into two fragments with one common atom, and they were reconnected in their docked positions based on the proximity of that common atom. The reconnected ligand was then energy minimized and evaluated.

Sandak *et al.* [69,70] developed another place-and-join algorithm within the geometric hashing framework. The ligand is divided into rigid fragments with one overlapping atom, called the hinge atom, and the hash entries are modified to include the location of that hinge atom. The method casts a vote for the hinge location for every match of a fragment to the receptor, and the hinge locations with the greatest number of votes are selected for rejoining the fragments.

In practice, the place-and-join algorithms are best suited to ligands comprised of a small set of medium-sized rigid fragments. When the fragments are too small, the number of potential docking orientations for these fragments is very large, resulting in an intractably large search space. Another disadvantage is that the bond lengths and bond angles around the hinge atoms are not preserved during the docking of individual fragments. Energy minimization is usually required to restore them to acceptable values.

In contrast, incremental construction naturally preserves the bond lengths and bond angles near the hinge atoms. The most critical step is placing the core fragment in the binding site. If an appropriate core fragment is selected, the potential docking configurations can be relatively small, rendering the problem more tractable compared to the docking of small fragments.

Leach and Kuntz [71] developed the first flexible docking program based on incremental construction. In their approach, a single anchoring fragment was manually selected and docked into the receptor, and the rest of the ligand was grown onto the anchoring fragment. In the recently released DOCK 4.0 [72] the approach was further developed and automated. The ligand is divided into fragments by cutting

the rotatable bonds. The largest fragment is selected as the anchor, and docked into the receptor. The other fragments are organized into layers, where a fragment in one layer is joined to a fragment in the immediately preceding layer by one rotatable bond. Then a breadth-first search is carried out for the torsion angles with which the fragments in each layer can be joined with the fragments in the preceding layer. As each fragment is grown onto the structure, the partial configuration is evaluated based on its score and its dissimilarity from the top-ranking configuration discovered at that stage. Partial configurations with low score or high similarity to the top-ranking configurations are discarded in order to truncate the size of the search space.

A number of other flexible ligand docking programs similar to DOCK 4.0 exist. Among the most prominent ones is FlexX [73-75]. FlexX also utilizes incremental construction, but it uses the pose clustering algorithm [76,77] to dock the rigid components into the receptor. Pose clustering matches each triangle of points in one object to the corresponding triangles in another object. For each such match, a transformation that superimposes the two objects can be determined (*i.e.* a pose). The poses are clustered, and the largest cluster suggests the best transformation to superimpose the two objects, while the size of the cluster tells whether the two objects match. FlexX uses the pose clustering technique to match the interaction surface of the receptor to the interaction centers on the ligand, and the interaction surface of the ligand to the interaction centers on the receptor.

Another program based on incremental construction is Hammerhead [78]. It differs from FlexX in that it divides the ligand into a small set of large fragments, while FlexX divides the ligand into a large core fragment and many small fragments. It adds the next fragment by overlapping the hinge atom and forming maximum interaction with the protein.

### **Stochastic Methods**

The above algorithms represent greedy optimization heuristics that attempt to truncate the size of the search space and quickly produce good, though not necessarily optimal, solutions. Docking is, of course, a global optimization problem where the goal is to identify the ligand conformation, position and orientation relative to the receptor where the binding energy is at its minimum. For flexible ligand docking, the free parameters of the system are the position and orientation of the ligand in the binding pocket and the torsions of its rotatable bonds. While global optimization is a very difficult problem for which often no deterministic algorithm with polynomial time complexity exists [79] several efficient stochastic techniques have been devised and successfully applied to a wide variety of problems, including flexible docking.

Olson and coworkers developed the AUTODOCK program [80,81] based on the simulated annealing algorithm [82]. To reduce the computational cost of fully evaluating the binding energy at each Monte Carlo step, molecular affinity potentials are calculated once and for all on a grid surrounding the binding site [83]. The standard Metropolis

scheme [84] is used to update the conformation and location of the ligand, causing it to move along a continuous and stochastic trajectory in its configuration space. If the annealing protocol is long and slow enough, the ligand will eventually settle in the state with the lowest binding energy.

Another Monte Carlo-based global optimization method, very efficient and robust in the authors' opinion, is Monte Carlo minimization [85] or basin hopping [86]. In this approach, the potential energy surface is transformed, so that the energy of each configuration represents the energy of the local minimum that corresponds to that particular configuration. A regular Monte Carlo simulation is then carried out on this transformed potential energy surface. In other words, before the energy of a state is evaluated after a Monte-Carlo step, the state is locally minimized, and the energy of the respective local minimum is used as a measure of fitness. This approach greatly reduces the search space, as well as the energy barriers between local energy minima, and has proven itself in a number of very difficult optimization problems. Trosset *et al.* developed PRODOCK [87-89] using Monte Carlo minimization. To accurately evaluate the energy and its gradient with the precomputed grid potential, the authors used a Bezier spline, which is continuous and differentiable everywhere on the grid.

Another widely applicable optimization technique is the genetic algorithm (GA) [90]. In common genetic algorithms, a state of the system is encoded as a binary string, called a chromosome. A pool of candidate states go through sexual and asexual reproductions to produce the next generation of states. In sexual reproductions, two chromosomes encoding the respective states swap bits with each other. In asexual reproduction, a chromosome goes through point mutations, in which some of its bits are flipped. After each reproduction cycle, the binary strings are translated back to candidate states, whose fitness is evaluated by an appropriate function. The candidates with better fitness are given a higher probability to participate in the next reproduction cycle. In the end, only the states with the highest fitness will survive.

Jones *et al.* developed one of the most widely used flexible docking programs, GOLD [91,92] based on a genetic algorithm. GOLD represents a configuration of the ligand by the torsional angle of each rotatable bond, and the mapping between the hydrogen bond partners of the protein and the ligand. This representation is then encoded into a binary string, and the system is allowed to evolve guided by a scoring function comprised of the internal strain energy of the ligand, and the van der Waals and hydrogen bonding energies of the protein-ligand complex.

A few other docking programs based on genetic algorithms [93] or their closely related evolutionary programming are also available [94,95]. Other global optimization methods have also been applied to the molecular docking problem, such as tabu search [96].

### Flexible Ligand / Flexible Protein Docking

For many protein-ligand complexes, the protein conformation does not differ significantly from its unbound

state. In such situations, holding the protein rigid during docking is a valid approximation. However, conformational changes in proteins induced by ligand binding are not uncommon. In particular, when forming hydrophobic interactions, the receptor usually undergoes some conformational changes that allow for a more tight hydrophobic interaction with the ligand. This is usually termed an induced fit. For instance, trifluoperazine (TFP) induces a drastic conformational change in  $\text{Ca}^{2+}$ -calmodulin ( $\text{Ca}^{2+}$ -CaM) from an elongated dumbbell shape into a compact form [97]. Because of this large conformational change, none of the predicted binding modes between TFP and  $\text{Ca}^{2+}$ -CaM resembled the observed structure of the complex obtained by X-ray crystallography.

Receptor flexibility can be naturally handled by molecular dynamics and Monte Carlo-based methods. To retain the advantages of grid-based energy calculation, the receptor is divided into a flexible part consisting of atoms on or near the active site, and a rigid part comprised of the remaining atoms in the biopolymer. The affinity potential is precomputed on a grid for the rigid part, and used to compute the corresponding energy contributions [98]. The flexible part and the ligand are allowed to move in the ensuing simulation [99] to search for the lowest binding energy.

Fragment-based approaches can also be adapted to include receptor flexibility. In this case, it is necessary to consider the flexible part of the receptor as well as the rotatable bonds in the ligand during incremental construction. The most computationally efficient way to handle flexible side-chains on the receptor is through the use of rotamers, which represent the most favorable amino acid conformations as determined by statistical analysis of protein structural databases [100,101]. Leach adapted the A\*-algorithm [102] from artificial intelligence to search through the ligand orientations and conformations along with a discrete set of rotamers for the flexible part of the protein [103].

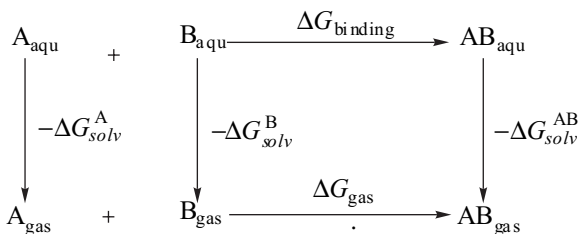
Another straightforward approach to handle receptor flexibility is to use an ensemble of protein conformations. A number of programs have been developed along these lines [104,105]. By placing hinges on the receptors, Sandak *et al.* extended their flexible ligand docking program to flexible receptors as well, and docked the flexible  $\text{Ca}^{2+}$ -CaM to a rigid ligand [69]. However, it is not obvious how this approach can be further extended to handle flexible ligand / flexible protein docking.

### Scoring Functions

To distinguish good from bad binding configurations, and to select the ligands with the best binding affinities from a pool of candidates, a reliable scoring function is required to evaluate any protein-ligand complex in a specific bound conformation. Indeed, without a good scoring function, even the most effective search algorithm will be rendered useless.

The ideal scoring function is, of course, the binding free energy between the protein and the ligand. This is the free

energy difference between the protein-ligand complex in solution and the unbound free protein and ligand in solution. Theoretically, the binding free energy can be calculated by considering the following thermodynamic cycle:



The binding free energy can be computed by the following formula:

$$\Delta G_{\text{binding}} = \Delta G_{\text{gas}} - \Delta G_{\text{solv}}^A - \Delta G_{\text{solv}}^B + \Delta G_{\text{solv}}^{AB}$$

where  $\Delta G_{\text{gas}} = \Delta H - T\Delta S$  is the free energy of binding in the gas phase. The binding enthalpy,  $\Delta H$ , can be estimated by molecular dynamics simulations. Several approaches exist for estimating the entropy, such as normal-mode analysis [106], quasi-harmonic analysis [80-82], and the quasi-Gaussian approach [107].

There have been some really exciting developments in computing the absolute free energies of binding [84, 85]. A relatively mature technique, thermodynamic integration, has been routinely employed to compute the difference in binding free energy between two different ligands [108]. These methods provide a way to select the best candidate from a small set of candidates with sufficient confidence. However, they all involve expensive molecular dynamics or Monte Carlo simulations, and are not suitable for virtual screening. If docking is to be used for prioritizing a large number of synthetic or screening candidates, simple, fast scoring functions have to be used. Designing such scoring functions is the holy grail of molecular docking. Protein-ligand binding involves complex interactions that are difficult to capture with simple terms that are amenable to computation. Although there has been tremendous effort in developing such scoring functions and significant progress has been made, the problem is still unsolved as manifested by the variety of scoring functions and the lack of consensus.

### Potential Energy as the Scoring Function

In predicting the native structure of a protein, we seek the structure that globally minimizes the free energy  $G$  of the protein in solution. According to the energy landscape theory [109], the native structures of most proteins reside in a deep and narrow well on the potential energy surface. Thus the potential energy contributes dominantly to the free energy and overshadows entropy. As a result, the structure at the global free energy minimum coincides with that at the global potential energy minimum, and we can minimize the potential energy to predict the native structure, sidestepping the difficulty of computing the free energy. Substituting the potential energy for the free energy is also practiced in

molecular docking, and scoring functions based on nonbonded interactions between protein and ligand have been utilized in docking experiments. Handling the electrostatic interactions requires some care in these calculations. Electrostatic interactions in vacuum differ greatly from those in the dielectric continuum of water. In vacuum, the Coulomb interactions between bare ions and dipoles usually dominate the nonbonded interaction energies. In contrast, highly polar water molecules cloud the ions and dipoles in aqueous solutions, and the electrostatic interactions are shielded to a large extent. For homogeneous solutions, the electrostatic interactions between ions can be computed by scaling the Coulomb interaction by the dielectric constant,  $\epsilon$ . But this approach is hardly appropriate at the protein-ligand interface. The heterogeneity at the interface precludes the use of a uniform dielectric constant. In this environment, the electrostatic interaction can be computed by solving the Poisson-Boltzmann equation [110,111], but the computational cost is prohibitively high. Therefore, in most potential energy-like scoring functions, the Coulomb energy is simply scaled down by the dielectric constant  $D$  of the protein environment.

Since ligand desolvation is often an important driving force for protein-ligand binding, it is important to include the solvation energy in the scoring function. The complete scoring function should be expressed as [112,113].

$$E_{\text{bind}} = E_{\text{nonbond}} - E_{\text{solv,elec}} - E_{\text{solv,np}}$$

where  $E_{\text{solv,elec}}$  is the solvation energy of bringing the electric charges from vacuum into solution, and  $E_{\text{solv,np}}$  is the solvation energy of the nonpolar groups. These two terms can be conveniently calculated using the generalized Born/surface area (GB/SA) model [114,115]. The nonpolar solvation energy is associated with the solvent accessible surface area by the simple linear relationship  $E_{\text{solv,np}} = \sum_k \sigma_k A_k$ , where  $A_k$  is the total solvent accessible surface area [116,117] of atom type  $k$ , and  $\sigma_k$  is an empirical atomic solvation parameter for the atom type. The work needed to create a distribution of charged spheres with charges  $q_i$  and radii  $a_i$  in a dielectric continuum of permittivity  $\epsilon$  is trivially computed by considering the process of bringing the charges from infinity to their corresponding sphere. The energy of the final system is given by:

$$E(\epsilon) = \frac{332}{\epsilon} \left( \sum_{i<j} \frac{q_i q_j}{r_{ij}} + \frac{1}{2} \sum_i \frac{q_i^2}{a_i} \right)$$

The constant 332 comes from unit conversions. Here, the charges  $q_i$  are specified in electrons, the radii  $a_i$  and the charge separation  $r_{ij}$  in Å, the permittivity  $\epsilon$  is 1 in vacuum, and the energy is in kcal/mol. Taking the difference between the charges in vacuum and in solution, the solvation energy  $E_{\text{solv,elec}}$  is given by:

$$E_{\text{solv,elec}} = E(\epsilon) - E(1) = 332 \left( \frac{1}{\epsilon} - 1 \right) \left( \sum_{i<j} \frac{q_i q_j}{r_{ij}} + \frac{1}{2} \sum_i \frac{q_i^2}{a_i} \right)$$

To accommodate overlapping atoms, the above equation is modified to

$$E_{\text{solv,elec}} = E(\epsilon) - E(1) = 332 \left( \frac{1}{\epsilon} - 1 \right) \sum_{i < j} \frac{q_i q_j}{\left( r_{ij}^2 + a_{ij}^2 e^{-r_{ij}^2 / (2a_{ij}^2)} \right)^{1/2}}$$

This expression reduces to the original equation when  $r_{ij} \gg a_{ij}$  and  $r_{ij} \ll a_{ij}$ , and approximates the Onsager reaction field energy when  $r_{ij} \sim a_{ij}$ . It avoids the singularity in case of small charge separations, but introduces an exponential function, and thus is more expensive to compute.

### Empirical Scoring Functions

An alternative to calculating the binding free energies from first-principles, as in the potential energy approach, is to use available binding complexes and their binding free energies, and create a simple model that best reproduces these data. A variety of empirical scoring functions have been developed this way. Many of them resemble the potential energy function in that they decompose the binding free energy into individual contributions from physically intuitive interaction types, such as hydrogen bond interaction, ion-pair interactions, etc. They can be viewed as simplified energy functions, sometimes with additional terms for describing contributions from entropy and solvation. In LUDI [118] and FlexX [75], for example, the free energy of binding is expressed as the sum of free energy contributions from the rotatable bonds in the ligand, hydrogen bonding, ion-pair interactions, hydrophobic and  $\delta$ -stacking interactions between aromatic groups, and lipophilic interactions. Each rotatable bond contributes a constant amount to the free energy, which is intended to represent the entropy loss due to the freezing of one rotational degree of freedom upon binding, though in practice it serves to diminish the dependence of the score on molecular size [119]. Every other term in the free energy expression imposes a penalty for deviating from the ideal interacting geometry as specified in the training set. Because the scoring function in FlexX and LUDI is based on specific pair interactions such as hydrogen bonding, it performs very well for binding complexes where hydrogen bonds are the dominant binding force, but encounters problems when hydrophobic interactions play the major role [119].

Another class of empirical scoring functions is constructed from the statistical preferences of atom pairs. It discards the first-principle approach altogether, and attempts to capture the preferential separation of different types of atoms by statistical analysis of known complexes. These so-called knowledge-based potentials [120-124] stem from the field of protein structure prediction, where the pairwise pseudo-energy between amino acids has been derived from a protein database [125,126]. The pairwise potential of mean force is computed from

$$A(r) = -kT \ln g(r)$$

where  $A(r)$  is the Holmoltz energy for bringing the pair into proximity of  $r$ ,  $k$  is the Boltzmann constant,  $T$  is body

temperature, and  $g(r)$  is the pair distribution function that describes how much more the pair prefers the distance  $r$  over randomly distributed points.  $g(r)$  can be computed from all the known structures of protein-ligand complexes. Muegge *et al.* developed such a knowledge-based scoring function, PMF\_SCORE [124], which is given by the sum of the pairwise energy of mean force. Knowledge-based scoring functions have demonstrated some success in finding the correct binding structure. However, directional interactions (such as hydrogen bonds) and unidirectional ones (such as hydrophobic packing) are averaged together, and consequently, these methods often perform poorly in when directional interactions dominate the binding energy or when binding involves a very tight fit [119].

### Consensus Scoring

Studies have shown that no scoring function consistently outperforms all other scoring functions [119]. In fact, most scoring functions have a domain of problems where they perform better than others [119,121,124,127-129]. Consequently, consensus scoring emerged to take advantage of as many scoring functions as possible. Sometimes different scoring functions are employed to dock the ligand onto the receptor and to score the resulting ligand-protein complex [121,130,131]. A more promising technique is to select compounds that consistently score high with a number of different scoring functions. Charifson *et al.* selected compounds that are common in the top-300 lists of various combinations of scoring functions, and found that it greatly reduced the number of false hits, while retaining the genuinely bioactive ones reasonably well [132].

Different scoring functions often emphasize different aspects of the binding interactions, even though each one of them makes an attempt to capture the comprehensive picture of binding. If we compute a series of scoring functions,  $\{f^a\}$ ,  $a = 1, 2, \dots, n$ , we can then form a vector for each compound  $i$  whose components are the scoring functions for the compound,  $f_i = (f_i^1, f_i^2, \dots, f_i^n)$ . We will henceforth call this vector the scoring vector. The components of the scoring vector represent different aspects of binding, and are complementary to each other. If a good number of sensible scoring functions are used, the scoring vector will contain all the essential information as regards the true binding affinity. Apparently the components  $f_i^a$ 's are not independent. If all the scoring functions are perfect, they should move in unison. We need to maximally utilize the binding information contained in the scoring vector, and predict the binding affinity of the ligand. Mathematically, we hope to find a mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}$  that produces an accurate estimate of the binding affinity from the scoring vector. The simplest mapping is a linear transform, where the combined score is given by  $F = \sum_i \lambda_i f_i$ . Stahl *et al.* [119] created a new scoring function by linear combination of terms from two different scoring functions: FlexX [75] and PLP [133]. This new scoring function showed comparable performance to each of its parent functions. More complicated, nonlinear techniques can be conceived to calculate the binding free energy from the scoring vector.

## Pressing Issues

Molecular docking has to overcome several hurdles before it can reliably predict binding mode and affinity. Some of the challenges are algorithmic, where a little creativity is needed to solve the problem. One of the examples is how to handle receptor flexibility. Other challenges are theoretical – a fundamental understanding of the phenomenon holds the key.

Theoretical challenges abound. The weakest link in the current scoring functions is hydrophobic interactions. In other interactions such as hydrogen bonds and polar interactions, the interactions are pairwise, and can be easily enumerated explicitly. Hydrophobic interactions, in contrast, are a collective phenomenon that originates from the perturbation of the water network by the immersed alien molecules. Water molecules, unfortunately, are too many to be explicitly included in docking calculations. Therefore we need to model the water ambience implicitly. Simple solvation models approximate the hydrophobic solvation energy as being linearly proportional to the solvent-accessible surface area. This is hardly an accurate approximation for the length scale of drug-protein interaction [134]. Significant progress has been made on the theory of hydrophobicity [134,135]. If we can formulate these new theories in a computationally efficient way and incorporate them in the docking programs, we will be able to predict the hydrophobic contribution to binding affinity more accurately, and greatly increase the reliability of virtual screening.

## Docking Applications

Despite many unsolved problems and the need for a reliable scoring function, docking has been widely applied in virtual screening and has achieved significant success. Perola *et al.* used docking to search the Available Chemicals Directory (ACD) and discovered a number of Farnesyltransferase inhibitors [136]. Horvath applied rigid body docking and found a few compounds that exhibited inhibitory activities against trypanothione reductase [137]. More examples can be found in Muegge and Rarey's review [112] and references within.

## CONCLUDING REMARKS

We now possess a variety of tools for analyzing and manipulating molecules in the computer, and searching for optimal drug leads. A challenging task currently facing scientists is to assemble all the tools into an integrated, easy-to-use system, so that virtual screening can be carried out automatically, achieve true high-throughput, and become a routine desktop tool for medicinal chemists. The latter, however, may prove to be a challenging task. Virtual screening techniques often require careful preparation of training sets, judicious selection of parameters, and careful analysis of the results. It is difficult for a person without extensive knowledge of the internal workings of the software and the underlying theory to use it effectively and obtain sensible results. Considering that there are even specialized

consultants for Microsoft Windows, we can safely say that, at least in the foreseeable future, well-trained computational chemists will be the main operators of such systems. Our experience suggests that pharmaceutical companies will benefit greatly by creating an environment of free and constant communication between computational and medicinal chemists.

## REFERENCES

- [1] Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening - an overview. *Drug Discovery Today* **1998**, *3*, 160-178.
- [2] Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharm. Toxic. Methods* **2000**, *44*, 235-249.
- [3] Zheng, Y. Personal Communications, **2002**.
- [4] Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nature Reviews Drug Discovery* **2002**, *1*, 337-346.
- [5] Agrafiotis, D. K. The diversity of chemical libraries. *Encyclopedia of computational chemistry*; John Wiley & Sons: Chichester, **1998**; pp 742-761.
- [6] Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Mol. Divers.* **1999**, *4*, 1-22.
- [7] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
- [8] MDL Information Systems, <http://www.mdli.com>.
- [9] James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual; Daylight Chemical Information Systems, Inc., **2000**.
- [10] Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Reviews in Computational Chemistry*; VCH Publishers: New York, **1991**; pp 367-422.
- [11] Dipaolo, T. Molecular connectivity in quantitative structure-activity relationship study of anesthetic and toxic activity of aliphatic hydrocarbons, ethers, and ketones. *J. Pharm. Sci.* **1978**, *67*, 566.
- [12] Hall, L. H.; Kier, L. B. Molecular connectivity of phenols and their toxicity to fish. *Bulletin of Environmental Contamination and Toxicology* **1984**, *32*, 354.
- [13] Sabljic, A.; Protic-Sabljić, M. Quantitative structure-activity study on the mechanism of inhibition of microsomal p-hydroxylation of aniline by alcohols. *Mol. Pharmacol.* **1982**, *23*, 213.
- [14] Waller, C. L.; Bradley, M. P. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345-355.

- [15] Rose, V. S.; Wood, J. Generalized cluster significance analysis with conditional probabilities. *Quant. Struc. Activ. Rel.* **1998**, *17*, 348-356.
- [16] Godden, J. W.; Bajorath, J. Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060-1066.
- [17] Cooley, W.; Lohnes, P. *Multivariate Data Analysis*; Wiley, **1971**.
- [18] Hull, R. D.; Singh, S. B.; Nachbar, R. B.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent semantic structure indexing (LASSI) for defining chemical similarity. *J. Med. Chem.* **2001**, *44*, 1177-1184.
- [19] Xie, D.; Tropsha, A.; Schlick, T. An efficient projection protocol for chemical databases: singular value decomposition combined with truncated Newton minimization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 167-177.
- [20] Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750-763.
- [21] Kruskal, J. B. Non-metric multidimensional scaling: a numerical method. *Psychometrika* **1964**, *29*, 115-129.
- [22] Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comp.* **1969**, *18*, 401-409.
- [23] Mahalanobis, P. C. *Proc. Natl. Inst. Sci. India* **1936**, *2*, 49.
- [24] Agrafiotis, D. K.; Lobanov, V. S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356-1362.
- [25] Rassokhin, D. N.; Lobanov, V. S.; Agrafiotis, D. K. Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. *J. Comput. Chem.* **2001**, *22*, 373-386.
- [26] Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **2001**, *22*, 488-500.
- [27] Agrafiotis, D. K.; Lobanov, V. S. Multidimensional scaling of combinatorial libraries without explicit enumeration. *J. Comput. Chem.* **2001**, *22*, 1712-1722.
- [28] Schneider, G.; Clément-Chomienne, O.; Hilfiger, L.; Schneider, P.; Kirsch, S.; Boehm, H. J.; Neidhart, W. Virtual screening for bioactive molecules by evolutionary *de novo* design. *Angew. Chem. Int. Ed.* **2000**, *39*, 4130-4133.
- [29] Schneider, G.; Lee, M.-L.; Stahl, M.; Schneider, P. *De novo* design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Design* **2000**, *14*, 487-497.
- [30] Lobanov, V. S.; Agrafiotis, D. K. Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 460-470.
- [31] Agrafiotis, D. K.; Lobanov, V. S. Ultrafast algorithm for designing focused combinatorial arrays. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1030-1038.
- [32] Lobanov, V. S.; Agrafiotis, D. K. Combinatorial networks. *J. Mol. Graphics Modell.* **2001**, *19*, 571-578.
- [33] Downs, G. M.; Barnard, J. M. Techniques for generating descriptive fingerprints in combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 59-61.
- [34] Ivanciuc, O.; Klein, D. J. Computing Weiner-type indices for virtual combinatorial libraries generated from heteroatom-containing building blocks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 8-22.
- [35] Shi, S.; Peng, Z.; Kostrowicki, J.; Paderes, G.J.; Kuki, A. Efficient combinatorial filtering for desired molecular properties of reaction products. *J. Mol. Graphics Modell.* **2000**, *18*, 478-496.
- [36] Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual compounds libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010-1023.
- [37] Sadowski, J.; Kubinyi, H. A. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325-3329.
- [38] Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between drug-like and nondrug-like molecules? *J. Med. Chem.* **1998**, *41*, 3314-3324.
- [39] Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841-1846.
- [40] Wang, J.; Ramnarayan, K. Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* **1999**, *1*, 524-533.
- [41] Rassokhin, D. N.; Agrafiotis, D. K. Kolmogorov-Smirnov statistic and its application in library design. *J. Mol. Graphics Modell.* **2000**, *18*, 368-382.
- [42] Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169-177.
- [43] *Pharmacophore perception, development and use in drug design*; International University Line: La Jolla, CA, **2000**.
- [44] Chen, G. S.; Chang, C.-S.; Kan, W. M.; Chang, C.-L.; Wang, K. C.; Chern, J.-W. Novel lead generation through hypothetical pharmacophore three-dimensional database searching: Discovery of isoflavonoids as nonsteroidal inhibitors of rat 5 $\alpha$ -reductase. *J. Med. Chem.* **2001**, *44*, 3759-3763.
- [45] Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers. *J. Med. Chem.* **1996**, *39*, 3060-3069.
- [46] Martin, E. J.; Hoeffel, T. J. Oriented substituent pharmacophore property space (OSPPREYS): A

- substituent-based calculation that describes combinatorial library products better than the corresponding product-based selection. *J. Mol. Graphics Modell.* **2000**, *18*, 383-403.
- [47] PDB Protein Data Bank, <http://www.rcsb.org/pdb>.
- [48] Joseph, D.; Petsko, G. A.; Karplus, M. Anatomy of conformational change: Hinged "lid" motion of the triphosphate isomerase loop. *Science* **1990**, *249*, 1425-1428.
- [49] Weber, C.; Wilder, G.; von Freyberg, B.; Traber, R.; Braun, W.; Widmer, H.; Wüthrich, K. The NMR structure of cyclosporin A bound to cyclophilin in aqueous solution. *Biochemistry* **1991**, *30*, 6563-6574.
- [50] Rini, J. M.; Schulze-Gahmen, U.; Wilson, I. A. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science* **1992**, *255*, 959-965.
- [51] Stanfield, R. L.; Fieser, T. M.; Lerner, R. A.; Wilson, I. A. Crystal structure of an antibody to a peptide and its complex with peptide antigen at 2.8Å. *Science* **1990**, *248*, 712-719.
- [52] Wedemayer, G. E.; Patten, P. A.; Wang, L. E.; Schultz, P. G.; Stevens, R. C. Structural insights into the evolution of an antibody combining site. *Science* **1997**, *276*, 1665-1669.
- [53] Bron, C.; Kerbosch, J. Finding all cliques of an undirected graph. *Comm. ACM* **1973**, *16*, 575-577.
- [54] Ullman, J. R. An algorithm for subgraph isomorphism. *J. ACM* **1976**, *23*, 31-42.
- [55] Gschwend, D. A.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking, revisited: On-the-fly optimization and degeneracy removal. *J. Comput.-Aided Mol. Design* **1996**, *10*, 123-132.
- [56] Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. Orientational sampling and rigid body minimization in molecular docking. *Proteins: Struct. Funct. Gene.* **1993**, *17*, 266-278.
- [57] Lamdan, Y.; Schwartz, J. T.; Wolfson, H. J. Affine invariant model-based object recognition. *IEEE Trans. Robot. Automat.* **1990**, *6*, 578-589.
- [58] Rigoutsos, I.; Wolfson, H. J. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.* **1997**, *4*, 10-21.
- [59] Fischer, D.; Norel, R.; Wolfson, H. J.; Nussinov, R. Surface motifs by a computer vision technique: Searches, detection and implications for protein-ligand recognition. *Proteins: Struct. Funct. Gene.* **1993**, *16*, 278-292.
- [60] Fischer, D.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. A geometry-based suite of molecular docking processes. *J. Mol. Biol.* **1995**, *248*, 459-477.
- [61] Nussinov, R.; Wolfson, H. J. Efficient detection of three-dimensional motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 10495-10499.
- [62] Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; The MIT Press McGraw-Hill Book Company: Cambridge, 1990.
- [63] Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Design* **2000**, *14*, 251-264.
- [64] Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: a way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Design* **1994**, *8*, 565-582.
- [65] Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select quasi-flexible ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Design* **1994**, *8*, 153-174.
- [66] Clark, D. E.; Murray, C. W.; Li, J. Current issues in *de novo* molecular design. *Reviews in Computational Chemistry*; Wiley-VCH: New York, **1997**; pp 67-125.
- [67] Murcko, M. A. Recent advances in ligand design methods. *Reviews in Computational Chemistry*; Wiley-VCH: New York, **1997**; pp 1-66.
- [68] DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **1986**, *29*, 2149-2153.
- [69] Sandak, B.; Nussinov, R.; Wolfson, H. J. A method for biomolecular structural recognition and docking allowing conformational flexibility. *J. Comput. Biol.* **1998**, *5*, 631-654.
- [70] Sandak, B.; Nussinov, R.; Wolfson, H. J. An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching. *Comp. Appl. Biosci.* **1995**, *11*, 87-99.
- [71] Leach, A. R.; Kuntz, I. D. Conformation analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.* **1992**, *13*, 730-748.
- [72] Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Design* **2001**, *15*, 411-428.
- [73] Rarey, M.; Kramer, B.; Lengauer, T. Multiple automatic base selection: Protein-ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Design* **1997**, *11*, 369-384.
- [74] Rarey, M.; Kramer, B.; Lengauer, T. Docking of hydrophobic ligands with interaction-based matching algorithms. *Bioinformatics* **1999**, *15*, 243-250.
- [75] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.
- [76] Linnainmaa, S.; Harwood, D.; Davis, L. S. Pose determination of a three-dimensional object using triangle pairs. *IEEE Trans. Pattern Anal. Machine Intell.* **1988**, *10*, 634-647.

- [77] Olson, C. F. Time and space efficient pose clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **1994**; pp 251-258.
- [78] Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449-462.
- [79] Gary, M. R.; Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*; Freeman: San Francisco, **1979**.
- [80] Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct. Funct. Gene.* **1990**, *8*, 195-202.
- [81] Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Design* **1996**, *10*, 293-304.
- [82] Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671.
- [83] Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.
- [84] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.
- [85] Li, Z.; Scheraga, H. A. Monte Carlo minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611-6615.
- [86] Wales, D. J.; Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Chem. Phys.* **1997**, *101*, 5111-5116.
- [87] Trosset, J.-Y.; Scheraga, H. A. Flexible docking simulations: scaled collective variable Monte Carlo minimization approach using Bezier splines. *J. Comput. Chem.* **1999**, *20*, 244.
- [88] Trosset, J.-Y.; Scheraga, H. A. Reaching the global minimum in docking simulations: A Monte Carlo energy minimization approach using Bezier splines. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 8011-8015.
- [89] Trosset, J.-Y.; Scheraga, H. A. PRODOCK: Software package for protein modeling and docking. *J. Comput. Chem.* **1999**, *20*, 412.
- [90] Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, **1989**.
- [91] Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Design* **1995**, *9*, 532-549.
- [92] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727-748.
- [93] Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided Mol. Design* **1995**, *9*, 113-130.
- [94] Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317-324.
- [95] Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Fogel, D. B.; Fogel, L. J.; freer, S. T. Docking conformationally flexible small molecules into a protein binding site through evolutionary programming. *The Fourth Annual Conference on Evolutionary Programming*; MIT Press, **1995**; pp 615-627.
- [96] Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins: Struct. Funct. Gene.* **1998**, *33*, 367-382.
- [97] Vandonselaar, M.; Hickie, R. A.; Quail, J. W.; Delbaere, L. T. J. Trifluoperazine-induced conformational change in Ca<sup>2+</sup>-calmodulin. *Nature Struct. Biol.* **1994**, *1*, 795-801.
- [98] Luty, B. A.; Wasserman, Z. R.; Stouten, P. F. W.; Hodge, C. N.; Zacharias, M.; McCammon, J. A. A molecular mechanics / grid method for evaluation of ligand-receptor interactions. *J. Comput. Chem.* **1995**, *16*, 454-464.
- [99] Wasserman, Z. R.; Hodge, C. N. Fitting an inhibitor into the active site of thermolysin: A molecular dynamics case study. *Proteins: Struct. Funct. Gene.* **1996**, *24*, 227.
- [100] Ponder, J. W.; Richards, F. M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **1987**, *193*, 775-791.
- [101] Dunbrack, R. L.; Karplus, M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature Struct. Biol.* **1994**, *1*, 334-340.
- [102] Russell, S.; Norvig, P. *Artificial Intelligence: a Modern Approach*; Prentice Hall: Upper Saddle River, **1995**.
- [103] Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345.
- [104] Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377.
- [105] Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424.
- [106] Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; McGraw-Hill: New York, **1955**.
- [107] Roccatano, D.; Amadei, A.; Apol, M. E. F.; Nola, A. D.; Berendsen, H. J. C. Application of the quasi-Gaussian entropy theory to molecular dynamics simulations of Lennard-Jones fluids. *J. Chem. Phys.* **1998**, *109*, 6358-6363.
- [108] Price, M. L. P.; Jorgensen, W. L. Analysis of binding affinities for Celecoxib analogues with COX-1 and COX-

- 2 from combined docking and Monte Carlo simulations and insight into the COX-2/COX-1 selectivity. *J. Am. Chem. Soc.* **2000**, *122*, 9455-9466.
- [109] Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545-600.
- [110] Sharp, J. A.; Honig, B. Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301-332.
- [111] Davis, M. E.; MCCammon, J. A. Electrostatics in biomolecular structure and dynamics. *Chemical Reviews* **1990**, *90*, 509-521.
- [112] Muegge, I.; Rarey, M. Small molecule docking and scoring. *Reviews in Computational Chemistry* **2001**, *17*, 1-60.
- [113] Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins: Struct. Funct. Gene.* **1999**, *34*, 4-16.
- [114] 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*, 6127-6129.
- [115] Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A.* **1997**, *101*, 3005-3014.
- [116] Lee, B.; Richards, F. M. Interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379-400.
- [117] Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 709-713.
- [118] Boehm, H. J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Design* **1992**, *6*, 593-606.
- [119] Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035-1042.
- [120] DeWitte, R. S.; Shakhnovich, E. I. SMOG: *de novo* design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 1733.
- [121] Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337.
- [122] Mitchell, J. B. O.; Laskowski, A. A.; Thornton, J. M. BLEEP-Potential of mean force describing potential ligand interactions. I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165.
- [123] Wallqvist, A.; Jernigan, R. L.; Covell, D. G. A preference-based free energy parameterization of enzyme-inhibitor binding. Applications to HIV1-protease inhibitor design. *Protein Sci.* **1995**, *4*, 1881-1903.
- [124] Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791-804.
- [125] Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures of globular proteins. *J. Mol. Biol.* **1990**, *213*, 859-883.
- [126] Jones, D. T.; Taylor, W. R.; Thornton, J. M. A new approach to protein fold recognition. *Nature* **1992**, *358*, 86-89.
- [127] Knegtel, R. M. A.; Bayada, D. M.; Engh, R. A.; von der Saal, W.; van Geerestein, V. J.; Grootenhuis, P. D. J. Comparison of two implementations of the incremental construction algorithm in flexible docking of thrombin inhibitors. *J. Comput.-Aided Mol. Design* **1999**, *13*, 167.
- [128] Ha, S.; Andreani, R.; Robbins, A.; Muegge, I. Evaluation of docking/scoring approaches: A comparative study based on MMP-3 inhibitors. *J. Comput.-Aided Mol. Design* **2000**, *14*, 435.
- [129] Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **1999**, *42*, 2498.
- [130] Stahl, M.; Boehm, H. J. Development of filter functions for protein-ligand docking. *J. Mol. Graphics Modell.* **1998**, *16*, 121.
- [131] Wallqvist, A.; Covell, D. G. Docking enzyme-inhibitor complexes using a preference-based free-energy surface. *Proteins: Struct. Funct. Gene.* **1996**, *25*, 403.
- [132] Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100-5109.
- [133] Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Design* **2000**, *14*, 731-751.
- [134] Lum, K.; Chandler, D.; Weeks, J. D. Hydrophobicity at small and large length scales. *J. Phys. Chem. B.* **1999**, *103*, 4570-4577.
- [135] Hummer, G.; Garde, S.; Garcia, A. E.; Pohorille, A.; Pratt, L. R. An information theory model of hydrophobic interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8951-8955.
- [136] Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y.-P. Successful virtual screening of a chemical database for farnesyltransferase inhibitors leads. *J. Med. Chem.* **2000**, *43*, 401-408.
- [137] Horvath, D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J. Med. Chem.* **1997**, *40*, 2412-2423.