1
Introduction to Chemoinformatics in Drug Discovery – A Personal View

Garland R. Marshall

1.1
Introduction

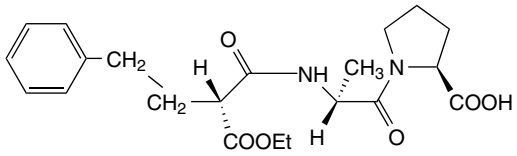
The first issue to be discussed is the definition of the topic. What is chemoinformatics and why should you care? There is no clear definition, although a consensus view appears to be emerging. ''Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization'' according to one view [1]. Hann and Green suggest that chemoinformatics is simply a new name for an old problem [2], a viewpoint I share. There are sufficient reviews [3–6] and even a book by Leach and Gillet [7] with the topic as their focus that there is little doubt what is meant, despite the absence of a precise definition that is generally accepted.

One aspect of a new emphasis is the sheer magnitude of chemical information that must be processed. For example, Chemical Abstracts Service adds over three-quarters of a million new compounds to its database annually, for which large amounts of physical and chemical property data are available. Some groups generate hundreds of thousands to millions of compounds on a regular basis through combinatorial chemistry that are screened for biological activity. Even more compounds are generated and screened *in silico* in the search for a magic bullet for a given disease. Either one of the two processes for generating information about chemistry has its own limitations. Experimental approaches have practical limitations despite automation; each *in vitro* bioassay utilizes a finite amount of reagents including valuable cloned and expressed receptors. Computational chemistry has to establish relevant criteria by which to select compounds of interest for synthesis and testing. The accuracy of prediction of affinities with current methodology is just now approaching sufficient accuracy to be of utility.

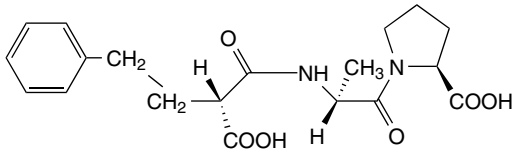
Let me emphasize the magnitude of the problem with a simple example. I was once asked to estimate the number of compounds covered by a typical issued patent for a drug of commercial interest. The patent that I selected to analyze was for enalapril, a prominent prodrug ACE inhibitor with a well-established commercial market. Given the parameters as outlined in the patent covering enalapril, an estimation of the total number of compounds included in the generic claim for enalaprilat, the active

Chemoinformatics in Drug Discovery. Edited by Tudor I. Oprea

ISBN: 3-527-30753-2

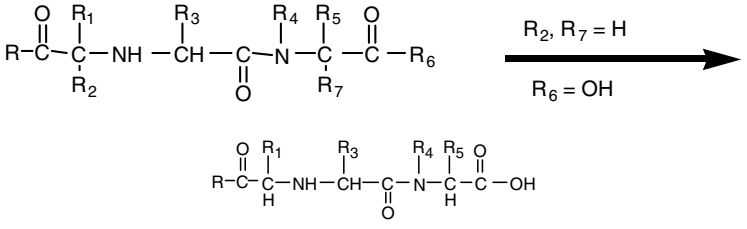


Enalapril



Enalaprilat

ingredient, was made. The following is the reference formula as described by the patent and simplified with $R_6 = OH$, and R_2 and $R_7 = H$:



Thus, one can simply enumerate the members of each class of substituent and combine them combinatorially. The following details the manner in which the number of each substituent was determined with the help of Chris Ho (Marshall and Ho, unpublished).

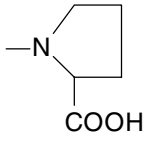
Substituent R: R is described as a lower alkoxy. The patent states that substituents are ''otherwise represented by any of the variables including straight and branched chain hydrocarbon radicals from one to six carbon atoms, for example, methyl, ethyl, isopentyl, hexyl or vinyl, allyl, butenyl and the like.'' DBMAKER [8] was used to generate a database of compounds containing any combination of one to six carbon atoms, interspersed with occasional double and triple bonds, as well as all possible branching patterns. Constraints were employed to forbid the generation of chemically impossible constructs. Concord 3.01 [9] was used to generate and validate the chemical integrity of all compounds. 290 unique substituents were generated as a minimal estimate.

Substituent R3: This substituent is identical to substituent R, only that it is an alkyl instead of an alkoxy. Again, 290 unique substituents of six or fewer carbon atoms were generated.

Substituent R1: R1 is described as a substituted lower alkyl wherein the substituent is a phenyl group. The patent is vague with regard to where this phenyl group should reside. If the phenyl group always resides at the carbon farthest away from the main chain, then again, 290 different substituents will result. However, if the phenyl group can reside anywhere along the 1- to 6-member chain, then approximately 1000 substituents are chemically and sterically possible.

Substituents R4 & R5: These two substituents are described by the patent as being lower alkyl groups, which may be linked to form a cyclic 4- to 6-membered ring in this position. This produces two scenarios: if these groups remain unlinked, then, as before, 290 substituents are found at *each* position.

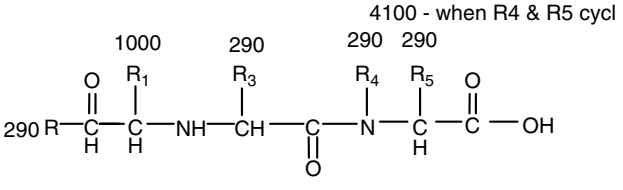
To determine the number of possible compounds when R4 and R5 are cyclized, a different approach was used. The patent states, ''R4 and R5 when joined through the carbon and nitrogen atoms to which they are attached form a 4- to 6-membered ring''. Preferred ring has the formula:



The patent is again vague in describing the generation of these cyclic systems. However, given that R4 and R5 are each 1–6 carbon alkyl groups with various branching patterns that are linked together, what results is a 4- to 6-membered ring system that may contain none, one or two side chains depending upon how R4 and R5 are connected. The overall requirement is that the total number of atoms comprising this ring system be less than or equal to 12.

To construct these ring systems, two databases were generated. The first database (''ring database'') contained three compounds – a 4-, 5- and 6-membered ring as specified by the patent. The second database (''side-chain database'') was constructed by cleaving each of the 290 alkyl compounds in half. One would assume that the first half of the alkyl chain would generate the ring, leaving the second half to dangle and form a side chain. A program DBCROSS (Ho, unpublished) was then used to join one compound from the ring database with up to two structures from the side-chain database at chemically appropriate substitution sites. Again, the overall requirement was that the number of atoms be less than or equal to 12. Approximately 4100 different cyclic systems were generated in this manner.

Total number of compounds



Summation	$(290)(1000)(290)(290)(290)$	$= 7.07 \cdot 10^{12}$	R4/R5 noncyclic
	$(290)(1000)(290)(4100)$	$= 3.44 \cdot 10^{11}$	R4/R5 cyclized

Sum $= 7.41 \cdot 10^{12} \rightarrow$ 3 chiral centers (carbons where R_1, R_3 and R_5 are attached to the backbone) in this molecule: X $8 = 5.93 \cdot 10^{13}$ or more than 59 *trillion* compounds included in the patent.
Note: If the phenyl group of substituent R1 is limited to the position farthest from the parent chain, then the number of compounds drops to $1.72 \cdot 10^{13}$ or more than 17 *trillion* compounds included in the patent.

Actually, the number of compounds included in the patent is severalfold larger as esters of enalaprilat such as enalapril were also included. Of the 100 trillion or so compounds included in the patent, how many could be predicted to lack druglike properties (molecular weight too large? logP too high?)? How many would be predicted to be inactive on the basis of the known structure-activity data available on angiotensin-converting enzyme (ACE) inhibitors such as captopril? How many would be predicted to be inactive now that a crystal structure of a complex of ACE with an inhibitor has been published? Given the structure-activity relationships (SAR) available on the inhibitors, what could one determine regarding the active site of ACE? What novel classes of compound could be suggested on the basis of the SAR of inhibitors? On the basis of the new crystal structure of the complex? Do the most potent compounds share a set of properties that can be identified and used to optimize a novel lead structure? Can a predictive equation relating properties and affinity for the isolated enzyme be established? Can a similar equation relating properties and *in vitro* bioassay effectiveness be established? These are representative questions facing the current drug design community and one focus of chemoinformatics.

One significant tool that is employed is molecular modeling. Because I have been involved more directly with computational chemistry and molecular modeling, there is a certain bias in my perspective. This is the reason I have used ''A Personal View'' as part of the title. I have also chosen a historical presentation and focused largely on those contributions that significantly impacted my thinking. This approach, of course, has its own limitation, and I apologize to my colleagues for any distortions or omissions.

1.2 Historical Evolution

With the advent of computers and the ability to store and retrieve chemical information, serious efforts to compile relevant databases and construct information retrieval systems began. One of the first efforts to have a substantial long-term impact was to collect the crystal structure information for small molecules by Olga Kennard. The Cambridge Structural Database (CSD) stores crystal structures of small molecules and provides a fertile resource for geometrical data on molecular fragments for calibration of force fields and validation of results from computational chemistry [10, 11]. As protein crystallography gained momentum, the need for a common repository of

macromolecular structural data led to the Protein Data Base (PDB) originally located at Brookhaven National Laboratories [12]. These efforts focused on the accumulation and organization of experimental results on the three-dimensional structure of molecules, both large and small. Todd Wipke recognized the need for a chemical information system to handle the increasing numbers of small molecules generated in industry, and thus MDL and MACCS were born.

With the advent of computers and the availability of oscilloscopes, the idea of displaying a three-dimensional structure of the screen was obvious with rotation providing depth cueing. Cyrus Levinthal and colleagues utilized the primitive computer graphics facilities at MIT to generate rotating images of proteins and nucleic acids to provide insight into the three-dimensional aspects of these structures without having to build physical models. His paper in Scientific American in 1965 was sensational and inspired others (including myself [13]) to explore computer graphics (1966/1967) as a means of coping with the 3D nature of chemistry. Physical models (Dreiding stick figures, CPK models, etc.) were useful accepted tools for medicinal chemists, but physical overlap of two or more compounds was difficult and exploration of the potential energy surface hard to correlate with a given conformation of a physical model.

As more and more chemical data accumulated with its implicit information content, a multitude of approaches began to extract useful information. Certainly, the shape and variability in geometry of molecular fragments from CSD was mined to provide fragments of functional groups for a variety of purposes. As series of compounds were tested for biological activity in a given assay, the desire to distill the essence of the chemical requirements for such activity to guide optimization was generated. Initially, the efforts focused on congeneric series as the common scaffold presumably eliminated the molecular alignment problem with the assumption that all molecules bound with a common orientation of the scaffold. This was the intellectual basis of the Hansch approach (quantitative structure-activity relationships, QSAR), in which substituent parameters from physical chemistry were used to correlate chemical properties with biological activity for a series of compounds with the same substitution pattern on the congeneric scaffold [14, 15].

1.3
Known versus Unknown Targets

Intellectually, the application of molecular modeling has dichotomized into those methods dealing with biological systems where no structural information at the atomic level is known, the unknown receptor, and those systems that have become relatively common, where a three-dimensional structure is know from crystallography or NMR spectroscopy. The Washington University group has spent most of its efforts over the last three decades focused on the common problem encountered where one has little structural information. Others, such as Peter Goodford and Tak Kuntz, have taken the lead in developing approaches to therapeutic targets where the structure of the target was available at atomic resolution. The seminal work of Goodford and colleagues [16] on designing inhibitors of the 2,3-diphosphorylglycerate (DPG) binding site on hemoglobin

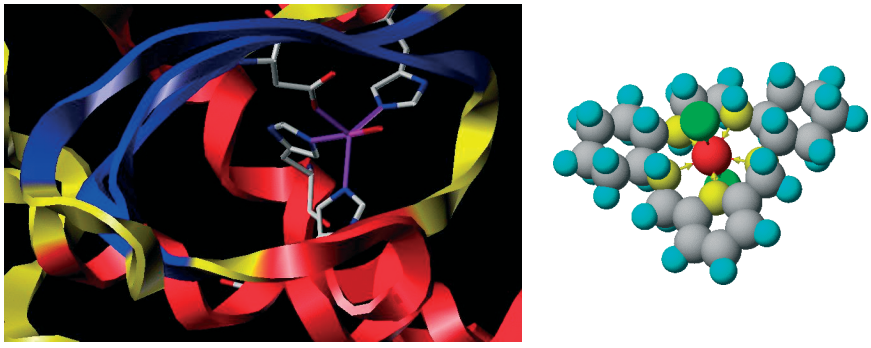


Fig. 1.1 **(a)** Active site of Mn superoxide dismutase (three histidine and one aspartic acid ligand to manganese) and **(b)** M40403, synthetic enzyme with 5 nitrogens (yellow) and two chloride (green) ligands.

for the treatment of sickle-cell disease certainly stimulated many others to obtain crystal structures of their therapeutic target. The most dramatic example of computer-aided drug design of which I am aware is the development of superoxide dismutase mimetics of below 500 molecular weight by Dennis Riley of Metaphore Pharmaceuticals. By understanding the redox chemistry of manganese superoxide reductase, Riley was able to design a totally novel pentaazacrown scaffold complexed with manganese (Figure 1.1) that catalyzes the conversion of superoxide to hydrogen peroxide at diffusion-controlled rates [17, 18]. This is the first example of a synthetic enzyme with a catalytic rate equal to or better than nature's best. The advances in molecular biology provided the means of cloning and expressing proteins in sufficient quantities to screen a variety of conditions for crystallization. Thus, it is almost expected that a crystal structure is available for any therapeutic target of interest. Unfortunately, many therapeutic targets such as G-protein-coupled receptors are still significant challenges to structural biology.

1.4 Graph Theory and Molecular Numerology

Considerable literature developed around the ability of numerical indices derived from graph theoretical considerations to correlate with SAR data. This was a source of mystery to me for some time. A colleague, Ioan Motoc, from Romania, with experience in this arena and a very strong intellect, helped me understand the ability of various indices to be useful parameters in QSAR equations [19–21]. Ioan correlated various indices with more physically relevant (at least to me) variables such as surface area and molecular volume. Since computational time was at a premium during the early days of QSAR and such indices could be calculated with minimal computations, they played a useful role and continue to be used. As a chemist, however, I am much more comfortable with parameters such as surface area or volume.

1.5
Pharmacophore

The success of QSAR led to efforts to extend the domain to noncongeneric series, where the structural similarity between molecules active in the same bioassay was not obvious. Certainly, the work of Beckett and Casey on opioids [22] to define those parts of the active molecules (pharmacophoric groups) essential for activity was seminal. Kier further developed the concept of pharmacophore and applied it to rationalize the SAR of several systems [23]. Peter Gund and Todd Wipke implemented the first *in silico* screening methodology with a program to screen a molecular database for pharmacophoric patterns in 1974 [24, 25]. Leslie Humber of Ayerst in Montreal exposed me to the wide variety of structures active in the dopaminergic system [26]. Overlaps of apomorphine, chlorpromazine and butaclamol to align the amines while maintaining the coplanarity of an aromatic ring led to a plausible hypothesis of a receptor-bound conformation. The least-squares fitting of atomic centers did not allow such an overlap, but the use of a centroid of the aromatic ring with normals to the plane for least-squares fitting accomplished the overlap. There still continues to be developments of methods to generate overlaps of hydrogen-bond donors and acceptors, aromatic rings, and so on, to generate a pharmacophore hypothesis from a set of compounds active at a given receptor/enzyme. One method developed early at Washington University was minimization of distances between groups in different molecules assigned by the

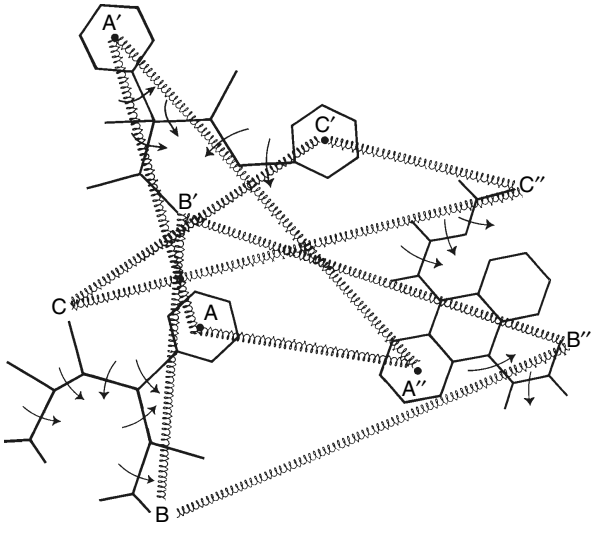


Fig. 1.2 Schematic diagram of minimization approach to the overlap of pharmacophoric groups (A with A′ with A″, B with B′ with B″, C with C′ with C″) by the introduction of constraints (springs) with intermolecular interactions ignored and only intramolecular interactions considered.

investigator with no intermolecular interactions. In effect, adding springs to cause the groups to overlap as the energy of the entire set of molecules was minimized excluding any interatomic interactions except those imposed by the springs (Figure 1.2). As a minimization procedure, the results were dependent on the starting conformation of the set of molecules being minimized, and multiple starting conformations were used to generate multiple pharmacophoric hypotheses.

1.6 Active-Analog Approach

The early work by medicinal chemists to try and rationalize their structure-activity relationships (SAR) with three-dimensional models as well as the success of Hansch and others in correlating SAR with physical properties led to exploration of molecular modeling as a means of combining the two approaches. Clearly, overall physical properties such as hydrophilicity, steric volume, charge and molar refractivity would be more meaningful in the context of a specific subsite within the receptor rather than when considered as an overall molecular average. One expected models with greater resolution and the ability to discriminate between stereoisomers, for example, as a result of the inclusion of geometrically sensitive parameters. By 1979, the group at Washington University had developed a systematic computational approach to the generation of pharmacophore hypotheses, the active-analog approach, which was disclosed at the ACS National Meeting that year [27].

The basic premise was that each compound tested against a biological assay was a three-dimensional question for the receptor. In effect, one was playing ''Twenty Questions'' in three dimensions. But each molecule was, in general, flexible and could present a plethora of possible three-dimensional arrays of interactive chemical groups. By computationally analyzing the sets of possible pharmacophoric patterns associated with each active molecule, one could find those pharmacophoric patterns common to a set of actives. In the simplest case, each inactive molecule would be geometrically precluded from presenting the given pharmacophoric pattern common to active molecules by steric or cyclic constraints. In practice, inactives that were capable of presenting the hypothetical pharmacophoric pattern were often found, so some other rationale for their inactivity was necessary to invoke. Aligning each active molecule to the candidate pharmacophoric pattern allowed determination of the volume requirements of the set of actives. One possible explanation for an inactive compound that could present the correct pharmacophoric pattern was a requirement for extra volume that was occupied by the receptor. When an inactive was aligned with the pharmacophore as scaffold, subtraction of the active volume space could identify such novel requirements. We had developed a Gaussian representation of molecular volume earlier [28], which readily allowed mathematical manipulation of atomic volumes. The best example of this rationalization of a data set occurred with a set of rigid bicyclic amino acids that inhibited the enzyme responsible for the synthesis of the active methyl donor in biology, S-adenosylmethionine, by Sufrin et al. [29]. In this case, the amino acid portion provided a common frame of reference that revealed that the compounds with

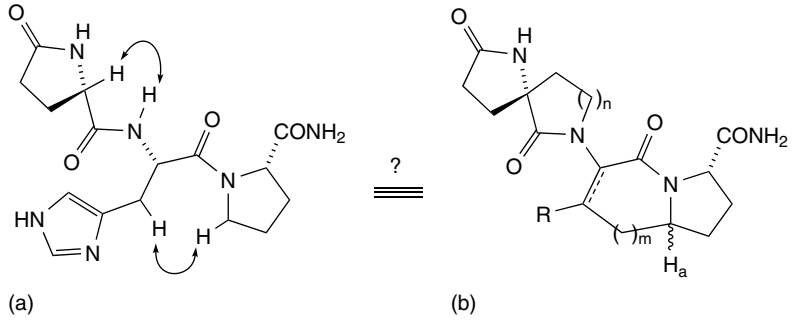


Fig. 1.3 Analysis of TRH **(a)** analogs by the active-analog approach by Font and Marshall led to a proposal for the receptor-bound conformation compatible with internal cyclization to generate polycyclic analogs **(b)**.

loss of the ability to inhibit the enzyme shared a small volume not required by active compounds, presumably required by an atom of the enzyme active site. Because the physical properties of the actives and inactives in this series were effectively identical, and the amino acid portion was clearly required for enzyme recognition, no other plausible suggestion for the data set has ever been suggested. Two other examples from analysis of SAR data were published on the glucose sensor [30] and on the GABA receptor [31].

One example of the early determination of the receptor-bound conformation of a biologically active peptide using the active-analog approach was the thesis work of Jose Font on the tripeptide TRH (thyrotropin releasing hormone), pyroglutamyl-histidyl-prolineamide. Only six torsional angles needed to be specified to determine the backbone conformation and the relative position of the imidazole ring of the bioactive conformation (Figure 1.3). Two alternative conformers were consistent with the conformational constraints required by the set of analogs analyzed. Font designed several polycyclic analogs, which were intractable for the synthetic procedures available at the time. In fact, these compounds served as a catalyst for the design of some novel electrochemical approaches by Prof. Kevin Moeller of Washington University [32–35]. Once the compounds could be prepared, their activity fully supported the receptor-bound conformation derived a decade before.

1.7 Active-Site Modeling

Any examination of crystal structures of complexes of a series of ligands binding to a protein (the set of complexes of thermolysin with a variety of inhibitors determined in the Brian Mathews lab, for example; see references in DePriest et al. [36]) shows clearly a major limitation of the pharmacophore assumption. Ligands do not optimize overlap of similar chemical functionality in complexes but find a way to maintain correct hydrogen-bonding geometry, for example, while accommodating other molecular interactions.

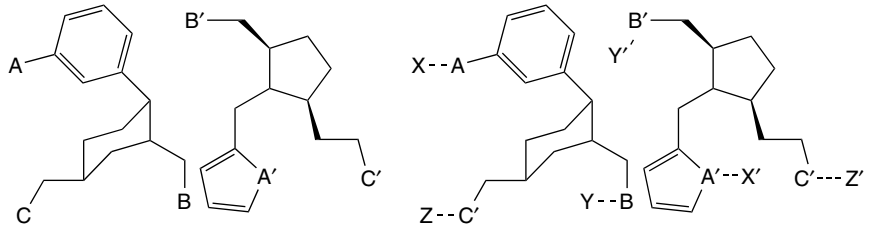


Fig. 1.4 Pharmacophore modeling with assumed ligand groups A = A′, B = B′ and C = C′. Active-site modeling with receptor groups X = X′, Y = Y′ and Z = Z′.

Thus, a transformation from overlapping chemical groups to the use of optimal sites of interaction with a stable active site of the receptor by extending hydrogen-bond donors to include an acceptor with optimal geometry, or adding a zinc to zinc-binding ligands was utilized to map the enzyme binding site (Figure 1.4). This was first emphasized for me in a study of the binding of inhibitors of angiotensin-converting enzyme (ACE) by Andrews et al. [37] in 1985. At the time, ACE was an object of intense interest in the pharmaceutical industry as captopril and enalapril, the first two approved drugs inhibiting ACE, were being extensively used to treat hypertension. Thus, each company was endeavoring to design novel chemical structures that inhibited ACE to gain a piece of the market (an activity that still occurs, witness the plethora of HMGCoA reductase and COX-2 inhibitors on the market, or in clinical studies). Analysis of the minimum energy conformations of eight ACE inhibitors revealed a common low-energy conformation of the Ala-Pro segment. Appending the pendant carboxy of enalapril or the sulfhydryl group of captopril determined a plausible site for the zinc atom involved in the enzymatic activity in the active site of ACE.

We had initiated a similar investigation of ACE inhibitors with the active-analog approach. By including additional geometrical parameters, a carboxyl group could include the zinc atom with optimal geometry from crystal structures of zinc-carboxyl complexes. Similarly, the sulfhydryl group could be expanded to include the zinc site as well with additional parameters to allow for appropriate geometrical variation. It seemed much more reasonable to assume that the groups involved in chemical catalysis and substrate recognition in the enzyme must have a relatively stable geometrical relationship, in contrast to chemical groups in a set of diverse ligands. Mayer et al.[38] analyzed 28 ACE inhibitors of diverse chemical structure available by 1987 and two inactive compounds with appropriate chemical functionality. On the basis of this data, a unique conformation for the core portion of each molecule interacting with a hypothetical ACE active site was deduced; the two inactive compounds were geometrically incapable of appropriate interaction.

1.8
Validation of the Active-Analog Approach and Active-Site Modeling

Inhibitors of the angiotensin-converting enzyme (ACE) served as a test bed for the active-analog approach in which one tries to deduce the receptor-bound conformation

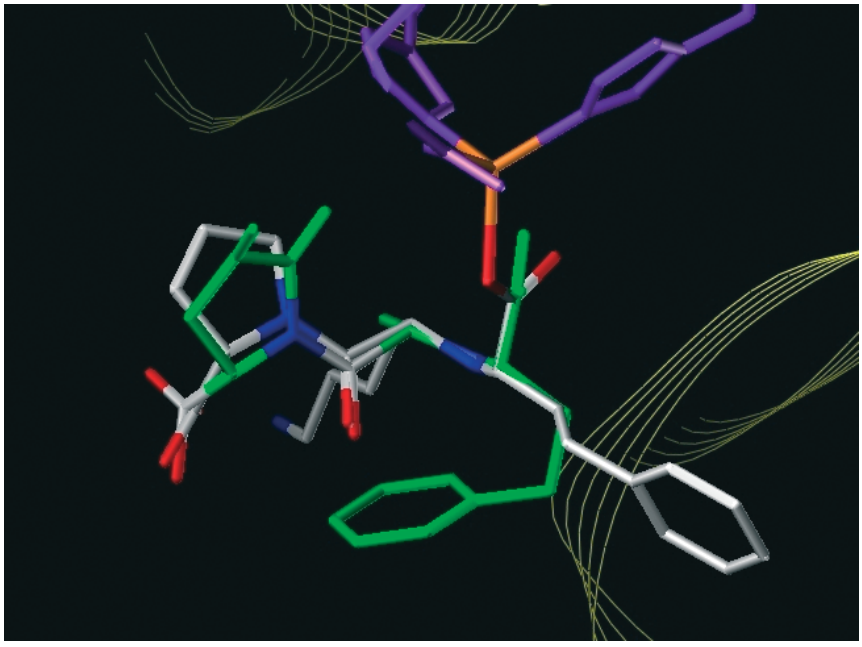


Fig. 1.5 Overlap of crystal structure of complex of the inhibitor lysinopril with angiotensin-converting enzyme and the predicted enzyme-bound conformation of ACE inhibitors by Mayer et al. [38]. Note the overlap between positions of pharmacophoric groups interacting with zinc (orange), C-terminal carboxyl and carbonyl oxygen of amide, the groups targeted by active-site modeling. The phenyl group common to enalapril analogs such as lisinopril (white ring) was not constrained (green ring) by analogs available at the time of the analysis in 1987.

of a series of active analogs based on the assumption of a common binding site. After a long delay, the crystal structure of the complex of lisinopril with ACE was finally determined [39] (not for lack of trying over two decades). The common backbone conformation of ACE inhibitors and the location of the zinc atom, hydrogen-bond donor and cationic site of the enzyme determined by Mayer et al. [38] essentially overlaps that seen in the crystal structure of the complex (Figure 1.5) arguing that, at least for this case, the assumption regarding the relative stability of groups important in catalysis or recognition is valid.

1.9 PLS/CoMFA

Dick Cramer provided insight and inspiration that led to my interest in 3D QSAR methodology [40] and was the impetus (the precursor of CoMFA was a lattice model [41] developed by Cramer and Milne at SKF) behind the development of CoMFA (Comparative Molecular Field Analysis) by Tripos [42]. The success of CoMFA in

generating predictive models rested entirely on the shoulders of a new statistical approach, Partial Least Squares of Latent Variables (PLS) [43], applied to chemistry by Prof. Svante Wold of the University of Umeå, Sweden. What was revolutionary at the time was the concept that you could extract useful correlations from situations where there were more variables than observations. Traditional linear regression analysis protects the user from chance correlations when too many variables are used. PLS recognized and corrects for cross-correlation between variables, and avoids chance correlations in models by systematically determining the sensitivity of the predictability of a model to omission of training data [44].

One seminal paper [45] by Cramer examined the principal components derived from examining the physical property data of a large set of chemicals from the Handbook of Chemistry and Physics. In effect, only two principal components were responsible for a significant amount of the variance of the data in the model derived. The nature of these two properties has always intrigued me, as well as the impact/possible simplification that derivation of chemical principles in terms of these two properties might have had. My viewpoint on the potential impact of the frame of reference is analogous to the impact that transformation of variables of coordinate systems can have on simplification of mathematical equations. A good example is the simplification that arises from using internal coordinates, that is, distances between atoms rather than coordinates, in structural comparisons so that the global orientation of each molecule is eliminated from consideration.

1.10 Prediction of Affinity

In order to prioritize synthesis and testing of a compound, an accurate estimate of the binding affinity to compare with the synthetic effort is of practical utility. Unfortunately, even for systems where the crystal structure of a complex of the compound with the receptor/enzyme is already available, accurate prediction of affinity *de novo* is still challenging and relates, among other limitations of current methodology, to difficulties in estimating changes in entropy as well as lack of inclusion of multipoles and polarizability in the electrostatics used by force fields currently employed. While interpolation based on experimental data can be effective as with CoMFA and other predictive models, and simulation techniques that mutate a related compound with known activity into the compound of interest often give useful predictions of affinity, accurate *de novo* prediction of the activity is still elusive. Oprea and Marshall have recently reviewed this topic [46].

Head et al. developed a PLS-based model VALIDATE [47] to scale the relative contributions of entropy and enthalpy to binding affinity for a variety of complexes whose crystal structures had been determined. Molecular mechanics were used to calculate several parameters most correlated with enthalpy of binding, while changes in surface area, number of rotatable bonds fixed upon binding and other parameters more related to the entropy of binding were also included in the model. Of interest was that the principal components of the model were dominated by two terms (ΔH and ΔS,

hopefully), several other terms had significant weight for the relative accurate model derived. Of course, doing the statistical mechanics right with a next-generation force field is an obvious solution, but a scoring function to quickly discard compounds with low affinity is still desired as witnessed by the amount of effort expended.

1.11 Protein Structure Prediction

Despite the Levinthal paradox [48, 49] that suggests the futility of attempting to predict the structure of a small protein based solely on sequence, efforts continue with increasing evidence that predictions are becoming more reliable [50]. This is based on the realization that all combinations of dihedral angles are not systematically explored by a protein in solution; rather, a funnel-like potential energy landscape guides the process. In general, the process of prediction can be dichotomized into conformer generation and conformer scoring. Obviously, one must generate a set of folds as candidates that contain the correct fold at some level of resolution to be identified and refined. Homology modeling, where one has a crystal structure of a homologous sequence, has proven a powerful approach that can generate useful models. Other approaches assemble models from homologous peptide segments [51]. The David Baker group has had considerable success in the recent CASP competitions with this approach [52], and recently designed, expressed and crystallized a small protein with a novel fold [53]. Prof. Stan Galaktionov developed a novel *ab initio* approach to fold prediction based on constraints from the contact matrix of predicted folds (Figure 1.6) that restricted possible folds to those with the correct density seen in experimental structures [54, 55]. By eliminating folds that were extended or overly compact, the fold space could be efficiently explored to generate sets of carbon alphas for further consideration. To generate a low-resolution structural model, a polyalanine chain was threaded through the carbon alphas and the orientation of the peptide planes optimized for hydrogen bonding (Welsh and Marshall, unpublished). What was needed to discriminate the candidate folds was a scoring function to evaluate the polyalanine trace utilizing the amino acid side chain information from the sequence to determine those folds worthy of full atomic representation and refinement.

We have recently published a low-resolution scoring function ProVal [56] developed with PLS that uses a multipole representation of side chains centered on the carbon alphas and betas that can distinguish the correct structure in the midst of plausible decoy folds in a large percentage of the 28 test cases studied (Figure 1.7). For 18 of the protein sets (~64%), the crystal structure scored best. In 24 sets (~86% of the cases), including the previous 18, the crystal structure ranked in the top 5%, and the crystal structure was ranked in the top 10% in all 28 cases. A second objective was to obtain a favorable correlation between the root-mean-square values for the α-carbons of the amino acids (CRMS value) of decoys and the experimental structure and the calculated score that was obtained for many of the test sets (Figure 1.7). In effect, ProVal can eliminate approximately 90% of ''good'' fold predictions from further consideration without specifying the coordinate position of side chains past the carbon beta. The

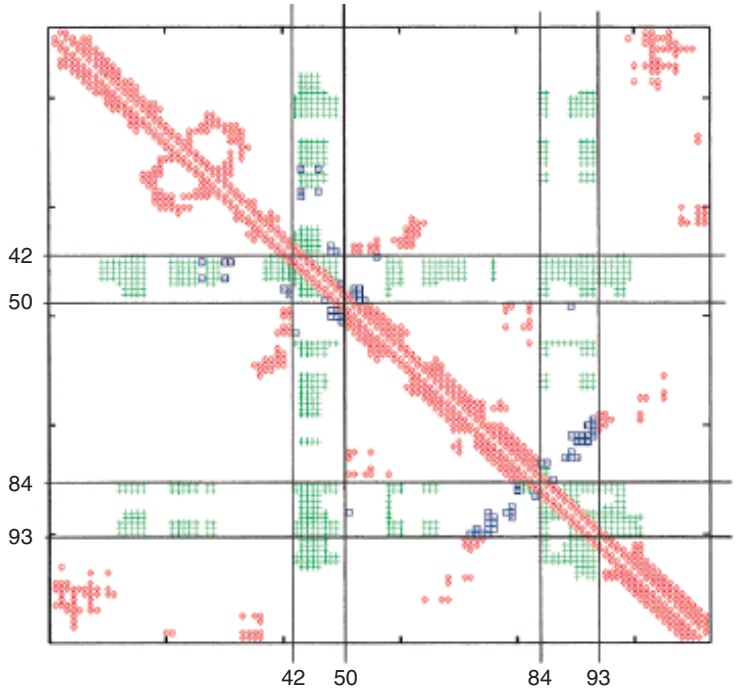


Fig. 1.6 Residue–residue contact matrix for predicted 3D structure of 3c2c (blue and green lines). The constant part *Ac* is shown in red, the "noncontact" matrix *An* is shown in green, and predicted variable contacts *Ax* are shown in blue. Numbers correspond to the predicted loops.

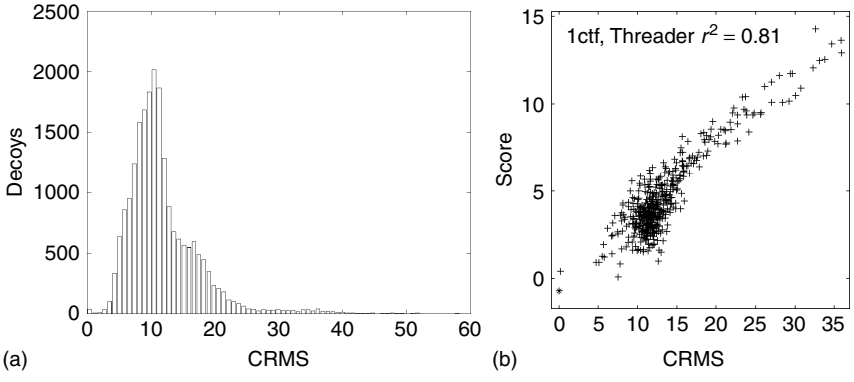


Fig. 1.7 Distribution of alpha carbon rmsd (CRMS) of 28 sets of decoy folds from crystal structures **(a)**. ProVal score versus CRMS for one decoy set generated by Threader 2.5 [57] for 1 ctf (Brookhaven PDB code) **(b)**.

details of atomic resolution are avoided because of the precision required to pack side chains efficiently with van der Waals overlap. As the quality of fold predictions increase, however, the ability to discriminate between alternatives becomes more difficult and requires a high resolution force field including multipole electrostatics and polarization (for example, the next-generation AMOEBA, Atomic Multipole Optimized Energetics for Biomolecular Applications) force field now being validated in Prof. Jay Ponder's group at Washington University, http://dasher.wustl.edu/ponder/.

1.12 Structure-Based Drug Design

As determination of crystal structures has become more commonplace, the efforts to design ligands to compliment a cavity on a molecular surface have become more sophisticated. Certainly, the pioneering effects [16] of Goodford and colleagues to design compounds binding to the DPG site on hemoglobin were dominated by chemical intuition, physical models and very primitive modeling systems. The development of DOCK by Kuntz et al. [58] was a major innovation, and sons of DOCK (AutoDock [59, 60], DREAM [61], etc.) are readily available over the web for exploring possible complex formation that include solvation approximations [62] and flexible ligands [63, 64]. Another major innovation from UCSF was the application of distance geometry as a means of generating three-dimensional coordinates from a set of distance constraints (bond lengths, sum of van der Waals radii, experimental distance constraints from NMR, etc.) [65]. Goodford has developed the use of probe atoms and chemical groups in GRID to map the binding site and identify optimal binding subsites [66]. This was certainly a prelude to experimentally determining subsite binding with subsequent assembly of fragments either by crystallography [67, 68] or by NMR [69]. The Washington University group led by Chris Ho has developed its share of structure-based design tools [8, 70–72]. Recently, a new generation software package RACHEL developed by Chris has been commercialized by Tripos, as discussed in Chapter 8 [73]. The ability to include synthetic feasibility and generate candidates with druglike properties has become a dominant theme in structure-based drug design.

1.13 Real World Pharmaceutical Issues

In reality, the approaches discussed above are all focused on the relatively simple part of developing a therapeutic, namely, lead generation. The reality of drug development is that there are many ways to interact with a given active site on a macromolecule. For example, look at the diversity of the structures capable of inhibiting HIV protease or ACE. The difficulty is in predicting adsorption, distribution, metabolism and elimination (ADME), which determine the pharmacokinetics, dosage regime and quantity of drug required. Even more problematic is prediction of toxicity, the ultimate filter that eliminates many compounds from clinical studies, and the major determinate of therapeutic ratio. At one of the first QSAR Gordon Conferences I attended, I paraphrased

Elizabeth Barrett Browning after hearing a discussion of toxicity prediction, ''How can I kill thee, let me count the ways''. A recent article by Stouch et al. [74] presents a thoughtful analysis of the validation effort for four such ADME/Tox models. Oprea et al. [75, 76] have compared drugs leads with compounds in development and in the marketplace and shown that compounds increase in molecular weight and logP as they progress to the bedside. *In silico* approaches certainly have their place in the pharmaceutical industry as one more tool to increase the probability of success [77].

1.14 Combinatorial Chemistry and High-throughput Screens

Development of automation and *in vitro* high-throughput biological screens has had a dramatic impact on lead discovery. Molecular biology has provided the tools for identification and validation of therapeutic targets, cloning and expression of sufficient protein to accommodate high-throughput screening, and determining the impact of elimination of the therapeutic gene by knock-out mutations.

Once the ability to screen libraries developed, the pressure on medicinal chemists increased to generate large quantities of compounds for screening. Ironically, combinatorial chemistry developed utilizing the technology of solid-phase organic chemistry. Solid-phase chemistry [78] was developed by Prof. R. Bruce Merrifield of Rockefeller University, my thesis advisor, as an automated method to assemble polypeptides, and later adapted by Prof. Marvin Caruthers of the University of Colorado for automated DNA synthesis [79]. Pioneering applications of this approach to synthetic organic chemistry in general was pioneered by a Canadian chemist, Charles Leznoff, who received little academic support, despite elegant applications including synthesis of some juvenile hormone analogs [80]. A paper on the solid-phase synthesis of benzodiazepine libraries [81] was a clarion call to medicinal chemists in industry due to the known pharmacological activity of benzodiazepines. Much of the reactions utilized in modern synthetic chemistry have been adapted to solid-phase organic chemistry for the synthesis of combinatorial libraries for high-throughput screening in the pharmaceutical industry.

Once the initial diversity fetish had run its course and management realized that it was inefficient to attempt to span the entirety of chemical space in search of drugs, more rational approaches based on chemoinformatics were developed to design combinatorial libraries and select candidates for screening on the basis of properties that have proven to be associated with successful therapeutics in the past.

1.15 Diversity and Similarity

Molecular recognition is an essential process in biological processes. One assumes that similar molecules are more likely to interact with a given receptor site than molecules that differ dramatically in size, shape or electronic distribution. This has led to the desire to compare molecules computationally prior to biological testing in order to prioritize

them and test only those molecules most likely to have the desired activity. For example, in a random-screening program, a compound that generates the desired biological effect may be found. One would like to examine the company's compound library of 500 000 compounds to select those 20 compounds most likely to show the desired activity. Often, one wishes to transcend a congeneric series (not choose 20 analogs of the same basic structure) and so comparisons must be done in some three-dimensional representation. Alternatively, if one is using combinatorial chemistry in a lead discovery effort, then one may want to explore as diverse a set of potential ligands as possible with a given number of assays. This leads to the concept of chemical diversity space and, if one is not careful, to a diversity fetish.

One relevant concern has been to prioritize the order of screening, or to decide which compound libraries to purchase for screening. One approach that has been used relies on the complementary concepts of diversity and similarity. Given two compounds, how do you quantitate how divergent the two structures are? One major problem is the choice of a relevant metric, what parameters are considered, how are the parameters scaled, and so on. Similarity, like beauty, is clearly in the eye of the beholder. There is no generally relevant set of parameters to explain all observations and one should expect that a given subset of parameters will be more relevant to one problem than to another. It should be pointed out that one is focused on properties of molecules in the absence of the receptor in contrast to the detailed focus on the complex in drug design studies. Many approaches to similarity fail to even consider chirality, a common discriminator of receptors. For a recent overview of the current status of virtual screening in lead discovery, see the review by Oprea and Matter [82].

1.16
Prediction of ADME

Methods for estimating the molecular properties that correlate with ADME problems have also been a very active arena for chemoinformatics. By studying the isozymes of cytochrome P450 enzymes, for example, certain molecular signatures for metabolic stability can be discerned. In a similar way, properties such as lipohilicity, pKa, number of hydrogen-bond donors and acceptors, and so on, correlate with oral bioavailability. For any drug development effort, oral bioavailability is often a requirement to compete in the marketplace with drugs already available. For a bioactive compound to succeed as a drug, it must pass many selective filters during development (toxicity, etc.) as well as in the body including metabolism, uptake, excretion, and so on. The most potent HIV protease inhibitor prepared in my lab was also an excellent substrate for excretion by the liver in first pass; thus high levels were found in the portal circulation, but nowhere else.

1.17
Failures to Accurately Predict

Why do we still have difficulty in developing useful predictive models? Where are the sources of noise? From the point of view of molecular modeling and computational

chemistry, the potential functions in common use have intrinsic significant errors in electrostatics. Estimating the entropy of binding is complex unless one is willing to sample solvent configurations sufficiently to adequately represent the partition function. Solvation models such as GB/SA [83] are certainly better than ignoring the significant impact that desolvation has on energetics. Multiple binding modes are not uncommon, but often too difficult to handle while modeling. The normal assumption of rigid receptor sites or, at the very least, limited exploration of the dynamics of the structure seen in the crystal is inherently dangerous. An excellent demonstration of the fallacy of assuming stability of receptor structure comes from the work of Don Abraham, where a compound designed to bind to an allosteric site on hemoglobin actually displaces core hydrophobic residues to optimize its interactions [84], a story which is detailed in Chapter 17 [85]. Receptors, at least GPCRs, have multiple conformations and probably different modes of activation and coupling with different G-proteins. The role of dimerization of GPCRs has only recently been shown to be important for a variety of receptors. In summary, we routinely apply Occam's razor for convenience, or as a rough approximation where we hope the results will withstand scrutiny by comparison with experimental results. The reality of biological systems is that Mother Nature never shaved with Occam's razor, and we cannot expect significant signal-to-noise from systems that have not been calibrated with the tools we are applying. If one is not using accurate *ab initio* methods, then one must remember the old dictum; to extrapolate is human, to interpolate is correct, but only within a relevant data set.

1.18 Summary

Chemoinformatics is the science of determining those important aspects of molecular structures related to desirable properties for some given function. One can contrast the atomic level concerns of drug design where interaction with another molecule is of primary importance with the set of physical attributes related to ADME, for example. In the latter case, interaction with a variety of macromolecules provides a set of molecular filters that can average out specific geometrical details and allows significant models developed by consideration of molecular properties alone.

Acknowledgments

It should be clear from the text that the author has benefited from the efforts of excellent (no, exceptional) collaborators and students. He has benefited immensely from a variety of mentors over the years, who have been both encouraging and critical regarding the development and application of computer-aided drug design. Sitting at the interface between the revolution in microelectronics, where CPUs are now a commodity, and the revolution in molecular biology and genomics, which provides a plethora of therapeutic targets and interesting conundrums to consider, has been both exciting and humbling. If the past is a preface to the future, get set for an exhilarating ride.

References

1 Brown, F. Chemoinformatics: What is it and How does it impact drug discovery. *Annu. Rep. Med. Chem.* **1998**, *33*, 375–384.

2 Hann, M., Green, R. Chemo-informatics – a new name for an old problem? *Curr. Opin. Chem. Biol.* **1999**, *3*, 379–383.

3 Ritchie, T. Chemoinformatics; manipulating chemical information to facilitate decision-making in drug discovery. *Drug Discovery Today* **2001**, *6*(16), 813–814.

4 Johnson, D.E., Blower, P.E., Jr, Myatt, G.J., Wolfgang, G.H. Chem-tox informatics: data mining using a medicinal chemistry building block approach. *Curr. Opin. Drug Disc. Dev.* **2001**, *4*(1), 92–101.

5 Oprea, T.I., Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157–166.

6 Oprea, T.I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **2002**, *6*(3), 384–389.

7 Leach, A.R., Gillet, V.J. *An Introduction to Chemoinformatics*. Kluwer Academic Publishers, Dordrecht, Netherlands, **2003**, 259.

8 Ho, C.M., Marshall, G.R. DBMAKER: a set of programs to generate three-dimensional databases based upon user-specified criteria. *J. Comput.-Aided Mol. Des.* **1995**, *9*(1), 65–86.

9 Pearlman, R.S. *CONCORD User's Manual*. Tripos Associates, St. Louis, MO, **1992**.

10 Allen, F.H., Motherwell, W.D. Applications of the Cambridge structural database in organic chemistry and crystal chemistry. *Acta Crystallogr, Sect B* **2002**, *58*(Pt 3 Pt 1), 407–422.

11 Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrea, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M., Watson, D.G. The developments of versions 3 and 4 of the Cambridge database system. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187–204.

12 Abola, E.E., Bernstein, F.C., Koetzle, T.F. The protein data bank. In *The Role of Data in Scientific Progress*, Glaeser, P.S. (ed.) Elsevier, New York, **1985**.

13 Barry, C.D., Ellis, R.A., Graesser, S., Marshall, G.R. Display and manipulation in three dimensions. In *Pertinent Concepts in Computer Graphics*, Faiman, M. Nievergelt, J. (Eds). University of Illinois Press, Chicago, IL, **1969**, 104–153.

14 Fujita, T., Iwasa, J., Hansch, C. A new substituent constant, π, derived from partition coefficients. *J. Am. Chem. Soc.* **1964**, *86*(December 5), 5175–5180.

15 Hansch, C., Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Wiley & Sons, New York, **1979**.

16 Goodford, P.J. Drug design by the method of receptor fit. *J. Med. Chem.* **1984**, *27*(5), 557–564.

17 Riley, D.P. Rational design of synthetic enzymes and their potential utility as human pharmaceuticals: development of manganese(II)-based superoxide dismutase mimics. *Adv. Supramol. Chem.* **2000**, *6*, 217–244.

18 Riley, D.P., Henke, S.L., Lennon, P.J., Aston, K. Computer-aided design (CAD) of synzymes: use of molecular mechanics (MM) for the rational design of super-oxide dismutase mimics. *Inorg. Chem.* **1999**, *38*(8), 1908–1917.

19 Motoc, I., Marshall, G.R., Labanowski, J. Molecular shape descriptors. 3. Steric mapping of biological receptor. *Z. Naturforsch.* **1985**, *40a*, 1121–1127.

20 Motoc, I., Marshall, G.R. Molecular shape descriptors. 2. Quantitative structure-activity relationships based upon three-dimensional molecular shape descriptor. *Z. Naturforsch.* **1985**, *40a*, 1114–1120.

21 Motoc, I., Marshall, G.R., Dammkoehler, R.A., Labanowski, J. Molecular shape descriptors. 1. Three-dimensional molecular shape descriptor. *Z. Naturforsch.* **1985**, *40a*, 1108–1113.

22 Beckett, A.H., Casey, A.F. Synthetic analgesics: stereochemical consider-ations. *J. Pharm. Pharmacol.* **1954**, *6*, 986–999.

23 KIER, L.B., ALDRICH, H.S. A theoretical study of receptor site models for trimethylammonium group interactions. *J. Theor. Biol.* **1974**, *46*, 529–541.

24 GUND, P., WIPKE, W.T., LANGRIDGE, R. Computer searching of a molecular structure file for pharmacophoric patterns. *Comput. Chem. Res. Educ. Technol.* **1974**, *3*, 5–21.

25 GUND, P. Three-dimensional pharmacophoric pattern searching. *Prog. Mol. Subcell. Biol.* **1977**, *11*, 117–143.

26 HUMBER, L.G., BRUDERLIN, F.T., PHILIPP, A.H., GOTZ, M., VOITH, K. Mapping the dopamine receptor. 1. Features derived from modifications in ring E of the neuroleptic butaclamol. *J. Med. Chem.* **1979**, *22*, 761–767.

27 MARSHALL, G.R., BARRY, C.D., BOSSHARD, H.E., DAMMKOEHLER, R.A., DUNN, D.A. The conformational parameter in drug design: the active analog approach. In *Computer-Assisted Drug Design*, OLSON, E.C. CHRISTOFFERSEN, R.E. (eds). American Chemical Society, Washington, DC, **1979**, 205–226.

28 MARSHALL, G.R., BARRY, C.D. Functional representation of molecular volume for computer-aided drug design. *Abstr. Amer. Cryst. Assoc., Honolulu, Hawaii* **1979**.

29 SUFRIN, J.R., DUNN, D.A., MARSHALL, G.R. Steric mapping of the L-methionine binding site of ATP: L-methionine S-adenosyltransferase. *Mol. Pharmacol.* **1981**, *l9*, 307–313.

30 WEAVER, D.C., BARRY, C.D., MCDANIEL, M.L., MARSHALL, G.R., LACY, P.E. Molecular requirements for recognition at glucoreceptor for insulin release. *Mol. Pharmacol.* **1979**, *16*(2), 361–368.

31 KLUNK, W.E., KALMAN, B.L., FERRENDELLI, J.A., COVEY, D.F. Computer-assisted modeling of the picrotoxinin and γ-butyrolactone receptor site. *Mol. Pharmacol.* **1982**, *23*, 511–518.

32 RUTLEDGE, L.D., PERLMAN, J.H., GERSHENGORN, M.C., MARSHALL, G.R., MOELLER, K.D. Conformationally restricted TRH analogs: a probe for the pyroglutamate region. *J. Med. Chem.* **1996**, *39*(8), 1571–1574.

33 SIMPSON, J.C., HO, C.M.C., SHANDS, E.F.B., GERSHENGORN, M.C., MARSHALL, G.R., MOELLER, K.D. Conformationally restricted TRH analogs: constraining the pyroglutamate region. *Bioorg. Med. Chem.* **2002**, *10*, 291–302.

34 SLOMCZYNSKA, U., CHALMERS, D.K., CORNILLE, F., SMYTHE, M.L., BEUSEN, D.D., MOELLER, K.D., MARSHALL, G.R. Electrochemical cyclization of dipeptides to form novel bicyclic, reverse-turn peptidomimetics. 2. Synthesis and conformational analysis of 6,5-bicyclic systems. *J. Org. Chem.* **1996**, *61*(4), 1198–1204.

35 TONG, Y.S., OLCZAK, J., ZABROCKI, J., GERSHENGORN, M.C., MARSHALL, G.R., MOELLER, K.D. Constrained peptidomimetics for TRH: cis-peptide bond analogs. *Tetrahedron* **2000**, *56*(50), 9791–9800.

36 DEPRIEST, S.A., MAYER, D., NAYLOR, C.B., MARSHALL, G.R. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determned active site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372–5384.

37 ANDREWS, P.R., CARSON, J.M., CASELLI, A., SPARK, M.J., WOODS, R. Conformational analysis and active site modelling of angiotensin-converting enzyme inhibitors. *J. Med. Chem.* **1985**, *28*(3), 393–399.

38 MAYER, D., NAYLOR, C.B., MOTOC, I., MARSHALL, G.R. A unique geometry of the active site of angiotensin-converting enzyme consistent with structure-activity studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*(1), 3–16.

39 NATESH, R., SCHWAGER, S.L., STURROCK, E.D., ACHARYA, K.R. Crystal structure of the human angiotensin-converting enzyme-lisinopril complex. *Nature* **2003**, *421*(6922), 551–554.

40 MARSHALL, G.R., CRAMER III, R.D. Three-dimensional structure-activity relationships. *Trends Pharmacol. Sci.* **1988**, *9*, 285–289.

41 CRAMER III, R.D., MILNE, M. The lattice model: a general paradigm for shape-related structure/activity correlation. In *19th National Meeting of the American Chemical Society*. American Chemical Society, Washington, DC, **1979**, COMP 44.

42 Cramer III, R.D., Patterson, D.E., Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*(18), 5959–5967.

43 Wold, S., Abano, C., Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E., Lindberg, W., Sjostrom, M. Modelling data tables by principal components and PLS: class patterns and quantitative predictive relations. *Analysis*, **1984**, *12*(10), 477–485.

44 Cramer III, R.D., Bunce, J., Patterson, D., Frank, I. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quantum Struct. Act. Relat.* **1988**, *7*, 18–25.

45 Cramer, I., Richard, D. BC(DEF) parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Am. Chem. Soc.* **1980**, *10*(6), 1837–1849.

46 Marshall, G.R., Arimoto, R., Ragno, R., Head, R.D. Predicting affinity: the sina qua non of activity. *Abstr. Pap. Am. Chem. Soc.* **2000**, *219*, 056–COMP.

47 Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M., Marshall, G.R. Validate – a new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*(16), 3959–3969.

48 Zwanzig, R., Szabo, A., Bagchi, B. Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20–22.

49 Karplus, M. The Levinthal paradox: yesterday and today. *Fold. Des.* **1997**, *2*(4), S69–S75.

50 Simons, K.T., Strauss, C., Baker, D. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **2001**, 1191–1199.

51 Bystroff, C., Baker, D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* **1998**, *281*(3), 565–577.

52 Simons, K.T., Bonneau, R., Ruczinski, I., Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **1999**, 171–176.

53 Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*(5649), 1364–1368.

54 Galaktionov, S., Nikiforovich, G.V., Marshall, G.R. Ab initio modeling of small, medium, and large loops in proteins. *Biopolymers* **2001**, *60*(2), 153–168.

55 Galaktionov, S.G., Marshall, G.R. Properties of intraglobular contacts in proteins: an approach to prediction of tertiary structure. *Proceedings of the 27th Hawaii International Conference on System Sciences*. IEEE Computational Society, Washington, DC, **1994**, 326–335.

56 Berglund, A., Head, R.D., Welsh, E., Marshall, G.R. ProVal: a protein scoring function for the selection of native and near-native folds. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 289–302.

57 Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. *Nature* **1992**, *358*, 86–89.

58 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269.

59 Goodsell, D.S., Lauble, H., Stout, C.D., Olson, A.J. Automated docking in crystallography: analysis of the substrates of aconitase. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 1–10.

60 Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*(14), 1639–1662.

61 Makino, S., Ewing, T.J.A., Kuntz, I.D. DREAM++: Flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*(5), 513–532.

62 Shoichet, B.K., Leach, A.R., Kuntz, I.D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*(1), 4–16.

63 Lamb, M.L., Burdick, K.W., Toba, S., Young, M.M., Skillman, K.G., Zou, X.Q., Arnold, J.R., Kuntz, I.D. Design, docking, and evaluation of multiple libraries against multiple targets.

Proteins: Struct., Funct., Genet. **2001**, *42*(3), 296–318.

64 Abagyan, R., Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*(4), 375–382.

65 Crippen, G.M. Distance geometry and conformational calculations. In *Chemometrics Research Studies*, Vol. 1, Bawden, D. (ed.) John Wiley, Chichester, UK, **1981**.

66 Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Am. Chem. Soc.* **1985**, *28*(7), 849–856.

67 Ringe, D. Structure-aided drug design: crystallography and computational approaches. *J. Nucl. Med.* **1995**, *36* (6 Suppl), 28S–30S.

68 Ringe, D., Mattos, C. Analysis of the binding surfaces of proteins. *Med. Res. Rev.* **1999**, *19*(4), 321–331.

69 Shuker, S.B., Hajduk, P.J., Meadows, R.P., Fesik, S.W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*(5292), 1531–1534.

70 Ho, C.M., Marshall, G.R. Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J. Comput.-Aided Mol. Des.* **1990**, *4*(4), 337–354.

71 Ho, C.M.W., Marshall, G.R. SPLICE – a program to assemble partial query solutions from three-dimensional database searches into novel ligands. *J. Comput.-Aided Mol. Des.* **1993**, *7*(6), 623–647.

72 Ho, C.M., Marshall, G.R. FOUNDATION: a program to retrieve all possible structures containing a user-defined minimum number of matching query elements from three-dimensional databases. *J. Comput.-Aided Mol. Des.* **1993**, *7*(1), 3–22.

73 Ho, C.M.W. In silico lead optimization. In *Chemoinformatics in Drug Discovery*, Oprea, T.I. (ed.) Wiley-VCH, Weinheim, **2004**, 199–219.

74 Stouch, T.R., Kenyon, J.R., Johnson, S.R., Chen, X.-Q., Dowyko, A., Li, Y. In silico ADME/Tox: why models fail. *J. Comput.-Aided Drug Des.* **2003**, *17*, 83–92.

75 Oprea, T.I., Davis, A.M., Teague, S.J., Leeson, P.D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, 1308–1315.

76 Oprea, T.I. Current trends in lead discovery: Are we looking for the apporpriate properties? *J. Comput.-Aided Drug Des.* **2002**, *16*, 325–334.

77 Koppal, T. Pharmacology's resurrection. *Drug Disc. Dev.* **2003**, *6*(11), 28–32.

78 Merrifield, R.B. Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. *J. Am. Chem. Soc.* **1963**, 2149–2154.

79 Caruthers, M.H. Gene synthesis machines: DNA chemistry and its uses. *Science* **1985**, *230*(4723), 281–285.

80 Leznoff, C.C. The use of insoluble polymer supports on general organic synthesis. *Acc. Chem. Res.* **1978**, *11*, 327–333.

81 Bunin, B.A., Ellman, J.A. A general and expedient method for the solid-phase synthesis of 1,4-benzodiazepine derivatives. *J. Am. Chem. Soc.* **1992**, *114*, 10 997–10 998.

82 Oprea, T.I., Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **2004**, *8*, 349–358.

83 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*(16), 6127–6129.

84 Wireko, F.C., Kellogg, G.E., Abraham, D.J. Allosteric modifiers of hemoglobin. 2. Crystallographically determined binding-sites and hydrophobic binding interaction analysis of novel hemoglobin oxygen effectors. *J. Med. Chem.* **1991**, *34*(2), 758–767.

85 Abraham, D.J. Drug discovery in academia. In *Chemoinformatics in Drug Discovery*, Oprea, T.I. (ed.) Wiley-VCH, Weinheim, **2004**, 457–484.