Journal of Computer-Aided Molecular Design, **14:** 215–232, 2000.
KLUWER/ESCOM
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

Review

Computational methods for the structural alignment of molecules

Christian Lemmen* & Thomas Lengauer
German National Research Center for Information Technology (GMD), Institute for Algorithms and Scientific Computing (SCAI), Schloß Birlinghoven, D-53754 Sankt Augustin, Germany

Received 27 January 1999; Accepted 22 July 1999

Key words: computational methods, review, structural alignment

Summary

In drug design, often enough, no structural information on a particular receptor protein is available. However, frequently a considerable number of different ligands is known together with their measured binding affinities towards a receptor under consideration. In such a situation, a set of plausible relative superpositions of different ligands, hopefully approximating their putative binding geometry, is usually the method of choice for preparing data for the subsequent application of 3D methods that analyze the similarity or diversity of the ligands. Examples are 3D-QSAR studies, pharmacophore elucidation, and receptor modeling. An aggravating fact is that ligands are usually quite flexible and a rigorous analysis has to incorporate molecular flexibility. We review the past six years of scientific publishing on molecular superposition. Our focus lies on automatic procedures to be performed on arbitrary molecular structures. Methodical aspects are our main concern here. Accordingly, plain application studies with few methodical elements are omitted in this presentation. While this review cannot mention every contribution to this actively developing field, we intend to provide pointers to the recent literature providing important contributions to computational methods for the structural alignment of molecules. Finally we provide a perspective on how superposition methods can effectively be used for the purpose of *virtual database screening*. In our opinion it is the ultimate goal to detect analogues in structure databases of nontrivial size in order to narrow down the search space for subsequent experiments.

Introduction

A major goal in pharmaceutical research is to design molecules that interfere with specific biochemical pathways in living systems. A corresponding area in drug design aims at developing small organic molecules with high affinity of binding toward a given protein receptor. If the three-dimensional structure of the receptor is known, standard, or *direct* rational drug design techniques, such as docking and all types of energy calculations, are applicable. However, frequently structural knowledge of the system under consideration is lacking. In many such cases, only a set of ligands, at best together with their measured biological activities toward the receptor is known. *Indirect* approaches try to infer in these situations, solely on the basis of the ligand data, about the process of binding to the target protein. If a proper superposition, approximating the binding geometry of a set of ligands, is available, the relevant chemical features of the ligands can be readily extracted in order to derive a pharmacophore model that, in turn, can be used to search for possible inhibitors in a ligand database. In addition, QSAR studies may provide an estimate of the binding affinity of a novel ligand toward the receptor under consideration. In fact, for 3D-QSAR methods like CoMFA, the molecular alignment may be considered the critical step in producing useful models. And finally, in some cases, it may be possible to take the negative imprint of the set of superimposed ligands as a crude description of the binding pocket under

*Present address: CombiChem Inc., 1804 Embarcadero Road, Palo Alto, CA 94303, U.S.A. E-mail: Christian.Lemmen@combichem.com consideration. Such a model can then be exploited by the more direct approaches mentioned above. However, with this kind of reasoning certainly several steps of experimental validation and refinement must be expected before, if any, reasonable hypothesis will result.

Indirect approaches incur serious limitations. First, the ligands must bind to the target protein at the same location and preferably adopt the same binding mode. If the former prerequisite is missing, there is nothing to infer from the superposition and, in fact, the models so generated will be misleading. Second, models generated on the basis of molecular superposition allow only to interpolate between the data. I.e., a region of space which is not occupied by any of the compounds cannot be judged. It might either be irrelevant if it relates to a volume not in contact with the protein, or it may zero out all activity if occupying this volume results in serious clashes. Also, and for the same reason as above, the data can only be meaningfully interpreted if a variation in physicochemical characteristics results in a variation in activity. Finally, 3D molecular models are usually restricted to low-energy conformations since the number of accessible conformers of a molecule increases dramatically with the conformational energy tolerated. Hence, bent conformations, such as observed as the transition state of a chemical reaction, will usually not be covered by such conformational models. While this limitation is frequently uncritical if potent drug molecules are considered, natural substrates may fail the low energy conformer assumption. Certainly, considering only the minimum energy state can be insufficient and the degree of conformational coverage will determine the success of a method.

This paper gives an overview of recent methodical contributions to structural alignment. We attempt to provide pointers to the recent literature covering all major methodical developments in molecular superpositioning. These pointers should enable the interested reader to thoroughly explore most facets of this steadily emerging field of research. Where applicable we provide also links to prominent approaches presented in the earlier literature. Good coverage of the publications predating 1990 can be found in [1, 2]. An outstanding collection of articles and reviews on this subject can be found in [3].

Frequently, the terms molecular superposition and pharmacophore elucidation appear interchangeably in the literature. However, we consider these two separate tasks where the former is usually a prerequisite of the latter. Given the molecular superposition, determining the key features essential for binding is a difficult problem on its own which is restricted by the inherent limitations of any indirect approach (cf. above). In this review we focus on the methodical aspect of the structural alignment problem. Reviews focusing on pharmacophore elucidation can be found in [4, 5].

The presentation is structured as follows. The following section reviews the important algorithmic innovations that have taken place over the past few years. These algorithms recur in several places in the literature. Therefore, it is helpful to describe them in general terms, in the beginning. The third section summarizes contributions towards assessing the similarity of molecules. Most of the methods discussed here do not actually tackle the superpositioning problem. However, since most of the alignment algorithms described later are based on molecular similarity, this section is in our opinion of general interest. Next follow three sections on different variants of the structural alignment problem grouped by application domain. First, methods for screening large ligand databases for candidates that are similar to a given molecule are discussed. Second, achievements in comparing protein structures are described. This field is quite different from the superposition of small molecules. Nevertheless, there is significant methodical exchange between these two fields which is the reason why we include this section. Third, we concentrate on the structural superposition of small rigid or flexible molecules. The final sections provide a comparison of the approaches, and conclude the paper with an outlook.

Table 1 classifies the literature that we have considered with respect to the algorithms used (rows) and the applications considered (columns). Note that the entries of the table are populated to widely different degrees.

Algorithmic methods for aligning molecules

In this section we briefly summarize the basic methods that appear in different variants in many of the approaches described in the sections to follow. Later we refer to these techniques and comment on variations and combinations that are applied.

If the correspondences between at least three pairs of reference points of two rigid-body objects are available, *rms-fitting* minimizes the sum of the squared

Table 1. Literature classification scheme

Method	Problem		
	Database screening	Protein structural alignment	Small molecule superposition
Matching-based techniques	[39] [47]	[57] [71] [72]	[77] [81] [88] [109] [113] [130] [136] [141]
Optimization-based techniques	[8] [41] [42]	[66] [69]	[75] [78] [79] [82] [85] [86] [87] [91] [93] [115] [117] [134]
Grid-based techniques	[45]	[63]	[83] [94] [102] [105]
Graph-based techniques	[49]		[13] [121] [131]

distances of corresponding points. In 1976, Kabsch provided an analytical solution to this problem based on eigenvector calculation [6]. Later computational improvements followed [7]. An important extension is the so-called *directed tweak* technique, first introduced by Hurst [8]. Directed tweak allows for an rms-fit considering molecular flexibility. By the use of local coordinates for the handling of rotatable bonds it is possible to formulate analytical derivatives of the objective function. With a gradient-based local optimizer flexible rms-fits are obtained extremely fast. However, no torsional preferences may be introduced. Therefore, directed tweak may result in energetically unfavorable solutions.

Volume overlap optimization is another basic technique quite often employed in molecular superposition. The approach is usually decomposed into three steps that are subject to variation. First, each molecule is represented by a set of spheres or Gaussians displaying some property which is frequently the *molecular electrostatic potential* (*MEP*). The MEP may be derived from any kind of charge distribution. Usually the MEP is first calculated on a grid and subsequently transformed to the sphere or Gaussian representation. Also the use of multiple properties is quite common. Based on such a representation, a sample of starting configurations is generated in a second step. Depending on the degrees of freedom considered, rotational, translational, and conformational sample points are generated. Finally, local optimizations are carried out using any variant of the classical *similarity measures* provided by Carbó in [9] and Hodgkin in [10] as the objective function. Often also the corresponding normalized *similarity indices* are used.

Database screening is frequently performed applying a stepwise filtering protocol. The sequence of steps is ordered by increasing computational demands and the filtering is adjusted such that the amount of retained structures can be handled by the subsequent step. Typically, the screening starts with a descriptor-based comparison. Such a descriptor represents an encoding of a molecule as a vector (*1D approach*). This may range from a single figure (e.g. the molecular weight) up to a complicated *fingerprint* consisting of numerous components. Topology-based approaches such as *substructure comparison* or purely distance-based filters (*2D approaches*) are applied next. Subsequently, 3D rigid-body alignments are calculated. Finally, flexible fitting may be applied to a limited amount of candidate structures.

From our point of view, the following techniques are important methodical contributions. Directed tweak [8], as stated above, comprises a major advance over the previous restriction of rms-fitting to rigid structures.

Geometric hashing originates from computer vision [11] and comprises a major invention in the field of generating matches between corresponding elements in the two structures to be superimposed. In [12] this technique has first been applied to structural biology data. Geometric hashing is a two-step method. In the first step, a highly redundant representation of one of the molecules is generated, that is invariant under rotation and translation. This information is stored in a hash table. In the second step, the hash table is queried with structural features from the second molecule. Each hit in the table identifies a transformation between the two molecules. Transformations that receive many hits are those that are likely to superimpose essential structural features of both molecules.

Another prominent technique to solve the matching problem considering multiple molecules is based on *clique detection* [13]. Structures are again represented by point sets. Depending on a distance tolerance δ, the algorithm generates a so-called *distance compatibility graph*. This graph contains a node for each type and length compatible distance in a reference structure and one conformer of every other molecule. Two nodes are connected if they share a common point in each of the structures. The matching procedure utilizes clique detection [14] to determine overall valid distance constraints.

Distance geometry is an elegant way to describe molecules in a translation- and rotation-invariant fashion. Conformational flexibility can be addressed, as well, by providing distance intervals rather than fixed distances, for all atom pairs. The resulting underdetermined constraint system which is assembled in a so-called distance-bound matrix, in principle, provides a means to generate all sets of 3D coordinates for the molecules obeying the restrictions. Picking a legal solution from the set of all possible solutions, which is in addition low-energy, requires to solve the difficult *embedding problem* which appears to be the bottleneck in computational speed. Crippen and co-workers laid the foundation for this mathematical vehicle to numerous applications in molecular structural biology. The monograph by Crippen and Havel [15] provides an excellent and comprehensive overview on this subject.

Genetic algorithms are a general purpose, global optimization technique that provides promising results in the entire area of computational structural biology [16]. Genetic algorithms mimic the process of evolution. A generation within this process comprises a set of configurations that are coded via chromosomes. Chromosomes are subjected to manipulation by some genetic operators such as crossover and mutation. The information content of the chromosomes varies depending on the application. Typically, it comprises the intramolecular matches or a coding of the orientational degrees of freedom and a coding of the torsional degrees of freedom in the case of considered molecular flexibility. The fitness function used to enable the process of selection typically comprises an efficiently computable similarity function.

Gaussian molecular representation and *Gaussian overlap optimization* [17], as described above, comprise a major advance over grid-based techniques which were most often used previously. This kind of modeling provides a high information content, avoids the dependency from additional parameters as for example the grid spacing, and allows for efficient optimization.

Molecular similarity

Molecular similarity is a fundamental concept in chemical research. Here we review recent approaches to molecular similarity that have a strong relation to structural alignment or structure comparison. In 1990 Johnson and Maggiora provided in [18] overviews of the definition, computation, and application of molecular similarity. In 1995 Dean presented in [19] a collection of contributions to molecular similarity used in drug design from researchers in academia as well as in industry. The overview deals with both molecular similarity and molecular complementarity. In his 1995 paper [20], Good reviewed molecular similarity measures that provide a mathematical footing for the sometimes vague and intuitive notion of similarity. Also in 1995 Rouvray surveyed similarity in chemistry in a broader scientific framework. Discussing various kinds of similarity, the author points out that ‘all scientific concepts and classifications have a basis in similarity’. We now briefly discuss specific contributions.

- Klopman [22] presents the CASE approach to molecular similarity based on molecular fragments, in order to distinguish between active and inactive molecules. From a training set of 30–50 molecules of known activity a linear dependency between the composition of fragments and activity is deduced by regression analysis. For this purpose fragmentation is performed automatically in a first step. In an extension called MULTICASE [23] the approach is iterated in a hierarchical fashion. Those fragments that are largely responsible for activity (or inactivity) are determined first. Subsequently, the training set is grouped by those fragments and, within each group, the method is applied again in order to improve the classification.
- Petke [24] introduces a new similarity index suitable for molecular comparison on the basis of scalar as well as vector fields. His aim is to arrive at an index that reacts sensitively to moderate or

small variations of these fields. Discrete variants of the proposed similarity indices are established, considering localized vector (respectively scalar) properties that are subjected to statistical analysis.

- Crippen [25] does not provide precisely a similarity method, but still addresses the general drug design problem of correlating binding affinities with structural properties of the ligand molecules in the absence of a receptor structure. Also he provides a mapping of pharmacophoric features across the set of molecules and thus implicitly allows for aligning the molecules structurally. The method presented combines the earlier distance geometry work of Crippen and co-workers [26] with their research in Voronoi binding site models [27]. The distance intervals, usually used in distance geometry, are extended to intervals of intervals in order to obtain more accurate constraints. The hypothetical receptor is modeled as a Voronoi tesselation of the 3D space. The complexity of the combinatorial problem of mapping ligand features across a set of molecules to the prospective portion of the putative binding pocket is greatly reduced by the restriction of all possible assignments to those that represent a convex partition of each of the molecules. A decision tree method is used to hierarchically refine the receptor model starting from a single region containing all parts of all molecules, and gradually partitioning this region into distinct sub-pockets. A single application to a set of 20 cocaine analogues is provided in this study.
- Gasteiger et al. [28] describe an approach to molecular similarity that is based on neural networks. The rigid-body comparison is carried out by mapping the MEP from the van der Waals surface onto the surface of a torus (resulting from periodically extending a square). The mapping process is performed using self-organizing Kohonen neural networks [29]. The quadratic map that results from unfolding the surface of the torus allows for qualitative visual inspection and comparison of the corresponding molecules. Recently an application of the approach to QSAR has been presented [30].
- Apaya et al. [31] provide a similarity approach on the basis of the matching of local extrema of the MEP. Accurate MEPs, derived from a distributed multipole analysis of ab initio charge densities, afford an appropriate representation of effects such as lone pairs or π-electron densities. The local extrema are then determined using probes of appropriate sizes and charges. Subsequently, a few (three to four) such locations are superimposed using rms-fitting. The resulting matches are extended to introduce further pairs of locations. The paper reports on a single specific test case (PDE III ligands) which was analyzed involving manual intervention.
- Blaney et al. [32] introduce an interactive similarity approach on the basis of spherical projection. The MEP calculated on a regular grid is mapped to the surface of a sphere by gnomic projection. Two rigid structures can be compared by calculating differences of MEP-values at points on the sphere. Computational enhancements allow for carrying out such comparisons on-line as the user rotates one of the molecules. The authors compare various empirical, semiempirical, and ab initio calculations of the molecular electrostatic potential. Recently, another similarity method has been presented that utilizes the reduction of computational complexity by series of spherical projections [33].
- Mezey [34] presents a similarity approach on the basis of molecular shape descriptors. The shape of a molecule is mapped to a number in three steps. First, the molecule is decomposed into rigid fragments. Second, electron densities of each fragment are taken from a database generated with ab initio methods. The resulting densities are added. Third, an isodensity contour is generated and submitted to *numerical shape code* analysis. These concepts have been extended to shape complementarity in order to predict molecular recognition. The paper contains no serious validation of these methods.

Database screening

3D database screening approaches based on either structure alignment or structure comparison provide rapid processing usually at the expense of limited accuracy. We do not consider 1D or 2D approaches based on, e.g., fingerprints or substructure matching, since these do not provide a means of actually aligning different molecules. A recent review on descriptor-based methods can be found in [35]. In 1993 Humblet and Dunbar [36] provided a broad overview of different 3D database approaches including substructure comparison, similarity searching, and docking. In 1995 Willett [37] reviewed pharmacophore matching methods considering rigid-body as well as flexible alignment during database screening. In 1996 Brown and Martin [38] surveyed different screening approaches in a comparative study. Since screening approaches aim at processing large numbers of molecules in order to detect analogs rather than providing a detailed comparison of a few compounds, these methods need to be validated by different means. Typically *enrichment factors, throughput, and hit rates* are used to demonstrate the performance of an approach. However, since the focus of this survey is much more on the detailed comparison we will not compare approaches in this respect. Again we discuss specific contributions.

– Lauri et al. [39] present an approach to 3D database screening of rigid structures based on matching pairs of bonds represented as vectors. The screening software, embedded in the CAVEAT system [40], is designed to facilitate the search for a given pattern of bonds in large databases in order to allow for the identification of connecting fragments (e.g., for *de novo* ligand design). The screening comprises three separate steps. A preprocessing compiles the source database into a CAVEAT database, representing bond pairs in bins indexed by distance and angles. A subsequent screening step seeks matching bond pairs, one in the search pattern and the other in the compiled database. Finally, the results of the query are clustered and representative hits are handed to the user.
– Hurst [8] provides flexible 3D database searching based on fitting of predefined pharmacophoric points. Using directed tweak, point sets of known correspondence are superimposed simultaneously considering the torsional degrees of freedom. At the expense of significantly increased computational costs, van der Waals penetrations may be checked and avoided during optimization, in order to enhance the generation of energetically accessible conformers.
– Moock et al. [41] describe flexible 3D database searching following the usual multi-step protocol. The so-called *conformationally flexible search* (CFS) starts with a topological *key screening* step. Subsequently, in a matching step, distance bounds of pairs of matched atoms in rigid database structures are estimated and checked against the query. Finally, fitting of flexible structures is performed applying two alternative optimization techniques.
– Thorner et al. [42] perform flexible field-based database searching using genetic algorithms to optimize the alignment of MEP fields. This is an extension of related work concerning rigid-body alignment [43, 44]. Molecular fields are represented by sets of Gaussian functions. The intermolecular overlap of the Gaussians is used as the fitness function. The chromosome contains the six orientational degrees of freedom and the torsional degrees of freedom of one of the molecules.
– Hahn [45] describes shape-based 3D database screening with a rigid so-called *receptor surface model* [46]. To take molecular flexibility into account a multi-conformer database is created as a preprocessing step. Another preprocessing step consists of the generation of a so-called *shape filter database* containing seven shape indices based on molecular volume and principal axes. In a first screening step, the shape filter is screened for candidate molecules with similar characteristics. Subsequent steps comprise principal axes alignment, rigid volume overlap optimization, and flexible fitting.
– Wang et al. [47] provide the traditional stepwise 1D/2D/3D screening approach to large structural databases. Initially, molecules are discarded that do not fit a query according to the requested number of different functional groups. The subsequent 2D screening comprises substructure comparison [48] and rigid 3D screening checks distance and angular constraints. A final optional flexible fitting phase enables searching for a conformation of a database structure to the template, on the basis of the 2D mapping, via directed tweak. Atomic as well as general functional queries are supported. Tests have been performed with databases of up to 200 000 structures.
– Rarey et al. [49] present a novel approach to analyzing molecular similarity in large databases based on tree mapping. A so-called *feature tree* represents hydrophobic fragments and functional groups of a molecule while preserving its topology. The nodes of the tree are labeled with the physicochemical properties of the corresponding fragments. The comparison of feature trees is carried out efficiently using a recursive subtree matching procedure. An advantage of the method is that, in addition to the similarity score, a matching of fragments is produced, thus enabling subsequent molecular superposition.

Protein structural alignment

The protein structural alignment task has much in common with rigid-body superpositioning. However, upon closer inspection there are several important differences. First, the size of the compared molecules, in terms of *number of atoms*, differs by two to three orders of magnitude. Second, the chemical diversity of proteins is somewhat reduced by the restriction to only 20 building blocks, namely the amino acids. Finally, a hierarchy of structural entities – atoms, amino acids, secondary structure elements, domains – may be utilized for comparison purposes. Therefore, the methodical overlap is relatively small and justifies a separate section. An 1994 overview on protein structural alignment by Orengo can be found in [50]. In the same year Holm and Sander elucidated the screening aspects of this task in a review [51]. In 1996 Godzik [52] performed a comparative study utilizing a variety of approaches on a set of examples to demonstrate the ambiguity of the results. Here we omit methods that inherently rely on the fact that the structures being compared are proteins, since our focus is chemical structure comparison in general and not techniques that borrow from the specific topology of proteins. This concerns the simplification to specific backbone atoms, the consideration of secondary structure elements, and the implicit assumption about a linear concatenation of the building blocks. Prominent examples of such primarily sequence-based approaches are those software tools that have been employed for evaluation purposes of the CASP competitions [53], VAST [54], DALI [55], and SSAP [56]. Most of the missing approaches should be found in the reviews cited above. Specific contributions in this area are the following.

- Fischer et al. [57] describe several enhancements of the geometric hashing technique described previously [12] that primarily applies to protein-protein comparison. However, the authors point out that this method also applies to the comparison of protein and ligand surfaces. The structural features on which the hashing is based are points on surfaces, backbones, or in active sites, respectively. In related studies Nussinov and co-workers provide methods for secondary structure comparison [58], protein-protein docking [59] and docking of ligands with limited flexibility [60]. Rarey et al. have modified the geometric hashing technique in order to place molecular fragments into the active site of a protein [61] which is used as the first step in a flexible docking procedure [62].
- Diederichs [63] presents protein structural alignment based on a brute-force grid search in orientational space. The global strategy, adapted from [64], is to sample rotational space and to optimize the remaining three translational degrees of freedom in the inner loop efficiently. Adequate rotational sampling [65] in combination with a novel translational optimization technique results in this highly efficient rigid-body superposition technique. Interatomic distances between the molecules to be aligned are determined and multiple occurrences of the same distance vector are counted in a cubic array. Subsequently, this array can be rapidly searched for its maxima. Various filters enable additional pruning of the searches.
- May et al. [66] use genetic algorithms to do protein rigid-body superpositioning. The algorithmic engine has been described previously [67]. The chromosome comprises the six rotational degrees of freedom and a gap penalty. The fitness function is based on an alignment score obtained by a dynamic programming approach [68]. The algorithmic enhancements comprise the evaluation of alternative alignments based on varying gap penalties, and the generalization to multiple structure comparisons.
- Poirette et al. [69] describe a rigid-body superposition approach based on genetic algorithms for the purpose of macromolecular surface comparison. Connolly surfaces [70] of the structures to be compared are transferred to a rectangular grid coding the type of shape and the hydrogen bonding potential at each point. The genetic representation allows for six-dimensional optimization on a fixed grid of rotational and translational increments. The fitness function comprises the number of surface points with similar characteristics that come close to each other.
- Wallace et al. [71] present a variant of geometric hashing to screen protein structure databases. In a preprocessing step the database is converted according to a specific key amino acid to be searched for. Each occurrence of this amino acid including all atoms in a 36 Å cube around it is transformed according to a predefined *frame of reference* defined for the specific amino acid of interest. Atom positions are stored in a hash table, according to a coarse grid within the frame of reference. The query pattern undergoes the same transformations and, for each query atom, a matching atom is searched for in the hash tables. Sufficiently large matches with a tolerable rms distance are declared hits of such a search.

- Escalier et al. [72] provide a protein structure alignment method based on the recursive assembly of lists of simultaneously superimposable intermolecular matches, satisfying a certain similarity criterion. The authors point out that the application of their methods to small molecule rigid-body superposition is straightforward. Two sets of atoms of size *k* taken from either molecule are considered a list of *k matches* if all intramolecular distances between these *k* atoms agree up to a certain tolerance. Two lists of matches of the same length are considered combinable if they differ only in a single pair and joining them does not violate the distance threshold. Note that this is similar to detecting cliques in a distance compatibility graph [13]. Since the method becomes intractable for large matching lists in real applications on protein structures, a second algorithmic phase which combines local matching lists applying a branch & bound technique is appended. An extension to multi-molecule alignment is provided as well.

Small molecule superposition

For a long time, rigid-body structural alignment of ligand molecules has been the method of choice for structure comparison. A great variety of techniques has been invented to tackle the problem, with an emphasis on optimization methods (cf. Table 1). Some of the approaches can be fed with multi-conformer input that is processed sequentially (*semiflexible*). Today, flexible ligand superposition is the method of choice for structure comparison. It defines a difficult problem that requires either a matching of points in space or a rigorous optimization. Given the matching which is sometimes provided by combinatorial approaches, determining the alignment is straightforward. On the other hand, given the global (or any convenient local) optimum, deriving the matching is straightforward as well. The dilemma is that frequently neither of them is available. Various techniques are provided in the literature. In 1993, a review of different alignment techniques was provided by Klebe [73] and an overview by Bures [74] dates from 1997. We present recent approaches starting with rigid-body superpositioning.

- Good et al. [75] present rigid-body superpositioning based on a simplex optimization of the molecular electrostatic fields. The authors were the first to replace grid-based evaluation techniques by analytical evaluation of a number of Gaussian functions, and thus gained a substantial increase in speed. No sampling of alternative starting positions is provided by the method. An essential ingredient of this technique is the so-called *Hodgkin similarity index* which was previously developed for the purpose of molecular comparison [10, 76].
- Feuilleaubois et al. [77] describe a pattern-matching approach to pharmacophore identification and rigid-body superpositioning. The method is an implementation of the *Boltzmann-machine* which combines the optimization property of the Hopfield neural network with the stochastic simulated annealing method. It searches for an optimal set of *m* interatomic matches and minimizes the respective distances. The method can also handle suboptimal and partial solutions. Tests have been carried out on a specific pair of Na^+-channel blockers.
- Petitjean [78] approaches the rigid-body superpositioning problem based on the usual overlap optimization protocol. The two-parameter distance he uses comprises the properties electronic and protonic charge that are assigned to spheres. A first order approximation of the intersection volume is computed. Starting from a few random orientations, local optimizations are performed using analytic first and second derivatives. Proteins as well as small molecules have been aligned via this approach.
- Grant et al. [79] describe rigid-body superpositioning based on van der Waals overlap optimization using Gaussian functions. The optimization is started from four orientations according to the shape quadrupoles with aligned centroids. An application of the method to chirality detection is described as well. Earlier work of the authors on the same subject can be found in [80].
- Cossé-Barbi et al. [81] approach rigid-body superpositioning by *continuous functional overlap*. In a stepwise approach, a pattern in 3D space (e.g., atom locations in a molecular structure) is converted into a continuous function by projecting points onto two planes and interpolating between them with cubic splines. Subsequently, the query pattern is projected onto the planes as well and aligned to the target by local optimization. The structures are prealigned along their first principal component axis. Finally, point pairs of the input patterns that are in close proximity are matched and this matching is used to perform an rms-fit. Besides the usual size-dependence the runtime of

the approach depends on the perturbation of the query pattern with respect to the target pattern.

- McMahon et al. [82] describe a rigid-body superpositioning approach based on the similarity measure of Carbó [9]. The electrostatic potential of the molecules is described by an approximation of the r^{-1}-term with up to three Gaussians. Utilizing a gradient optimization, the authors compare their results with the previously employed simplex optimizer. A single local optimization is carried out, starting from the orientation with aligned centers of mass.
- Nissink et al. [83] tackle rigid-body superpositioning based on electron density overlap optimization using Fourier space methods. The electron density is approximated by sets of Gaussian functions. The application of Fourier space methods, similar to the molecular replacement technique in X-ray crystallography [84], enables separate optimization of rotational and translational parameters. The simplex algorithm is utilized to carry out local optimizations starting from 12 distinct orientations. The overlap integral in Fourier space is approximated assuming the molecules to be placed in an infinite lattice of recurring instances.
- Parretti et al. [85] approach the rigid-body superpositioning problem by Monte Carlo optimization of molecular similarity indices. Molecules are represented with sets of Gaussian functions for steric and electrostatic properties. 3D-QSAR analyses based on the generated alignments are used for evaluation purposes.
- Lemmen et al. [86] describe rigid-body superpositioning based on overlap optimization in Fourier space. The so-called RIGFIT method combines the overlap computation technique first described in [83] and the modeling of different molecular fields by multiple sets of Gaussian functions [87] with an efficient optimization strategy. The preceding approaches are extended by sampling both the translational and rotational space appropriately. Especially the translational optimization was improved substantially, which facilitates the superposition of structures that differ significantly in size (e.g., the placement of molecular fragments). The implementation has been included into the flexible ligand superpositioning system [88] developed earlier

In the following we discuss papers that handle flexible molecules but basically use rigid-body techniques.

- Martin et al. [13] present a semiflexible superposition approach based on a combinatorial matching procedure for a limited number of predefined pharmacophoric points in a set of molecules. A prerequisite is the generation of sets of low-energy conformers for each of the molecules. One of the molecules serves as a reference, its conformations are treated sequentially. Clique detection is used to determine a set of pharmacophoric points that obey distance constraints up to a tolerance δ simultaneously for all molecules. The algorithm iterates, with increasing δ until a given threshold for the pharmacophore size is met. Earlier work of Martin and co-workers on pharmacophore elucidation can be found in [89]. Previous work on clique detection for pharmacophore elucidation can be found in [90].
- Masek et al. [91] approach the rigid-body superpositioning problem following the usual volume overlap optimization protocol. Different physicochemical properties are assigned to two sets of spheres. Local optimizations start from a few random orientations and different conformers are processed sequentially.
- Sanz et al. [92] provide rigid-body superpositioning on the basis of MEP calculations. The so-called MEPSIM program is a collection of tools that allow for integrated computation, analysis, and comparison of MEPs. The central part comprises a previously developed optimizer to overlay different MEPs [93]. The similarity measure used is the *Spearman coefficient* calculated for grid points. Local optimizations are performed using gradient methods and can optionally be restricted to the consideration of a single plane cutting through the molecules, thus reducing the problem to two dimensions. Multiple local optimizations, starting from random orientations, are possible and conformational flexibility considering a single torsional degree of freedom is permitted.
- Jain et al. [94] present rigid-body superpositioning of a set of molecules provided in multiple conformations, starting from a given initial alignment. A neural network is used to discriminate on the basis of an alignment to a reference molecule between active and inactive for each conformer of each of the other molecules. The neural network is trained to best reflect the measured activity by the highest predicted activity. The alignments are optimized on the basis of a gradient that is additionally provided by the neural net. In a few cycles training

of the neural net and optimization are iterated. The molecules are represented by a surface description on the basis of a grid of reference points. In [95] the authors compare the performance of their implementation with two earlier methods, CoMFA and a molecular similarity approach. Related recent work of Jain and co-workers can be found in [96–99].

- Klebe et al. [87] describe several structural alignment methods including rigid-body superpositioning based on an efficient overlap optimization. Different molecular fields (steric, electrostatic, and hydrophobic) are described by sets of Gaussian functions. The method extends the SEAL approach [17]. Local optimizations are carried out starting from a limited set of random orientations. Extensions of the method that include simultaneous consideration of the intramolecular conformational strain energy, as well as a multi-molecule alignment algorithm [100] based on the rms-fit of predefined pharmacophoric points, are provided additionally. Recently, an approach to flexible fitting by sequential processing of up to 150 rigid conformers and subsequent flexible post-optimization has been described [101].
- Waszkowycz et al. [102] present two approaches to pharmacophore elucidation. The first method performs the analysis of given static molecular fields, e.g., provided by a CoMFA study [103]. The second method solves pharmacophore mapping by clique-detection [13]. The latter can be equipped with a set of rigid conformers. Both approaches are implemented in the PRO_LIGAND program [104] which facilitates *de novo* design on the basis of the derived pharmacophore.
- Perkins et al. [105] provide a rigid-body superpositioning approach based on surface overlap optimization. The surface overlap volume is evaluated using grid techniques and optimized using simulated annealing. Hydrogen bonding and electrostatic similarity are accounted for in a postprocessing step in order to improve the ranking of results. Sets of conformers are processed sequentially. In earlier work Dean and co-workers provided various other approaches to ligand superpositioning and molecular similarity in general [106–108].
- Barnum et al. [109] provide a semiflexible superposition approach to elucidate common pharmacophores in sets of molecules. Flexibility is accounted for by consideration of multiple conformers that are generated in a preprocessing step with the so-called *poling method* [110]. Features, like hydrogen bonding partners, are identified for each structure and a modified version of the matching algorithm in [111] is used to determine intermolecular matches of growing size. The scoring the authors invented is based on the *maximum likelihood* rule. It takes into account to which degree a matching is common to all structures and an estimate of the rarity of such a matching in non-bonding molecules. An earlier version of the method was implemented in the pharmacophore hypothesis generator in CATALYST [112].
- Finn et al. [113] describe semiflexible superposition of multiple ligands based on pharmacophore matching considering precomputed conformers of each of the molecules. In a two-step approach pairwise matches are extended to multimolecule matches, using different heuristics. Pairwise matches are generated using a stochastic triangle mapping procedure. Earlier work of the authors on the same subject can be found in [114].
- Mestres et al. [115] provide a rigid-body superpositioning approach based on similarity index optimization. The objective function considers the overlap of steric and electrostatic fields represented by sets of Gaussian functions. Local optimizations are performed starting from systematically sampled orientations. The details of the method are presented in a previous study [116], as well as investigations concerning conformer generation in a preprocessing step and the influence of weighting the different field terms. Graphical enhancements via isosimilarity surface plots and the detection of maximum similarity loci are provided. An extension to multi-molecule alignment is proposed.
- Miller et al. [117] describe a rigid-body superposition method that combines clique detection [118] with the previously developed volume overlap optimization procedure SEAL [17]. Clique detection is used to determine reasonable starting positions on the basis of type and distance compatible feature matches of size k where k is typically set to 4. So-called *essential atoms* allow the user to determine specific features of which a certain number is required to be included in every match. This of course greatly reduces the search space and allows for rapid processing – given such constraints. The scoring function used by the optimization procedure has been enhanced with two terms as compared to its predecessor SEAL. Both of them

constrain the search space. Precomputed multi-conformer databases of nontrivial size have been screened and comparisons with crystallographic data have been carried out for comparison purposes. Previous work of Kearsley and co-workers can be found in a variety of molecular similarity studies, see [118–120] and references therein.

The last list of comments is on contributions that provide actual flexible models for superpositioning.

- Sheridan et al. [121] describe flexible superposition of a set of molecules based on distance geometry. Prerequisite is the definition of a pharmacophore which is to be present in each of the considered molecules. The method then finds a low-energy conformation for each of the molecules (if present) that allows alignment of the molecules via the pharmacophore. Earlier work on distance geometry in a variety of QSAR studies has been presented by Crippen and co-workers [122–124]. Distance geometry also has proven to be useful in conformational sampling [125], docking [126], and pharmacophore matching in 3D structural databases [127]. Later work of Sheridan and co-workers on database searching techniques can be found in [128, 129].
- Itai et al. [130] approach the flexible superposition problem based on combinatorial matching of few predefined pharmacophoric points. All possible matchings are enumerated and evaluated via rms-fitting. Discrete sets of conformers are considered by processing torsional increments of specified bonds sequentially in a brute force search.
- Dammkoehler et al. [131] propose flexible superpositioning based on different systematic conformational search methods. Starting with a given pharmacophore that is assumed to be present in each of the molecules, the method (known as the *Active Analogue Approach* [132]) combinatorially searches the conformational space of each of the considered molecules. Initially the distances of n pharmacophoric points of every conformer of the most rigid molecule are stored in an $n(n-1)/2$-dimensional distance map. Subsequently, the distance map of each of the other molecules is evaluated and compared to the reference. While traversing the tree-like search space, the set of valid distance patterns is successively thinned out. The authors invent novel conformational sampling methods and compare these with the commonly used fixed grid sampling technique. In [133] a variant of the active analogue approach, utilizing a different search technique and trying to overcome the problem of choosing a specific gridsize for the distance map, is presented.
- McMartin et al. [134] provide a method to flexibly superimpose a molecule onto a rigid reference. The method is based on a combined Monte Carlo perturbation and energy minimization procedure. Starting from the placement of an anchor fragment which is performed manually, a combination of intermolecular and intramolecular energies is optimized. A Monte Carlo procedure [135] generates random perturbations to the molecule that is to be fitted. Perturbation and optimization are iterated in a large number of cycles in order to cover conformational space. The authors point out that a convex potential for the intermolecular energy term is the most appropriate.
- Jones et al. [136] perform flexible ligand superpositioning based on a genetic algorithm. This method is able to handle even sets of flexible molecules. The chromosomes code the conformation of each molecule as well as intramolecular feature correspondences. The fitness is calculated by an intermolecular conformational energy term, the volume overlay, and an intermolecular matching energy term. The actual orientation is obtained by rms-fitting. The algorithm originates from a flexible docking approach [137]. Previous work of Willett and co-workers on structural alignment can be found in [138, 139] and references therein. See also the section on database screening.
- Lemmen et al. [88] approach flexible superpositioning on the basis of a combinatorial matching procedure. Pairs of molecules are aligned, one of which is considered rigid and the other one is flexibly fitted. The three-step approach employed originates from the related docking program [62]. In essence, the strategy is to decompose the flexible structure into relatively rigid portions, to start the placement using a manually selected portion, and to add the remaining portions in an iterative incremental procedure. The scoring function used to select appropriate intermediate solutions comprises energy-like matching terms for *paired intermolecular interactions* and overlap terms utilizing Gaussian functions to describe different field properties. Recently, Lemmen et al. provided an extensive evaluation of the approach based on experimental data, as well as an application to virtual database screening [140].

Table 2. Overview of approaches described in the literature

1st author	[Ref.]	Hardware	N_{sys}	N_{bsp}	N_{atom}	$N_{conf.}$	Best rmsd (Å)	N_{good} (%)
Sheridan	[121]	VAX 11/785	1	1	15–25	∞	–	–
Good	[75]	VAX 3520	1	1	∼40	1	–	–
Martin	[13]	VAX 9000	2	2	20–50	<10	–	–
Feuilleaubois	[77]	DEC DS5000	1	1	9	1	0.21	–
Itai	[130]	SGI (R4000)	3	3	∼50	∞	–	–
Masek	[91]	SGI (R4000)	1	1	∼60	∼10	–	–
Sanz	[92]	VAX	1	1	15–25	1	–	–
Jain	[94]	SGI (R4000)	1	102	20–40	70	–	–
Klebe	[87]	SGI (R4000)	7	24	20–160	1/150	0.38	12 (50%)
Waszkowycz	[102]	SGI (R3000)	2	2	20–50	1/50	–	–
Dammkoehler	[131]	–	1	30	10–60	∞	–	–
Jones	[136]	SGI (R4000)	8	8	10–60	∞	–	–
McMartin	[134]	VAX 8820	3	10	10–60	∞	<1.0	–
Perkins	[105]	SGI (R4000)	4	6	30/500	1–12	0.64	–
Petitjean	[78]	DEC AS2100	2	2	40/>500	1	–	–
Barnum	[109]	MIPS 4400	3	15	50–150	20–300	–	–
Grant	[79]	SGI (R4000)	3	31	30–50	1	0.17	28 (90%)
Cossé-Barbi	[81]	DEC AS2100	1	1	40–50	1	–	–
Finn	[113]	SGI (R4000)	1	1	60–70	<50	–	–
Lemmen	[88]	SUN UltraSparc	9	133	10–160	∞	0.45	88 (67%)
McMahon	[82]	SGI (R4000)	3	27	10–30	1	–	–
Mestres	[115]	–	1	10	∼100	<10	–	–
Nissink	[83]	SGI (IP21)	3	5	40–60	1	0.58	4 (80%)
Parretti	[85]	SGI (R4000)	2	31	50–60	1	–	–
Handschuh	[141]	SGI (R10000)	3	6	40–150	∞	–	–
Lemmen	[86]	SUN UltraSparcII	9	161	10–160	1	<0.6	109 (68%)
Miller	[117]	SGI (R4400)	6	6	20–120	1	0.08	6 (100%)

Column 1 provides the first author and a reference to the article discussed in the text. The second column states which hardware has been used in the different studies. N_{sys} gives the number of systems, i.e., the number of different proteins to which the ligands bind and that may be used to group the ligands into activity classes. N_{bsp} displays the number of superpositions computed. N_{atom} provides a range of the number of atoms considered as an indication of the size of the different examples. $N_{conf.}$ gives a range of the number of conformers in order to indicate the degree of flexibility considered.

Table 2. (continued)

ϕ rt	Name	Link to previous work, comment
300 min	–	With the embedding step potentially an infinite number $N_{conf.}$ is accessible; however, frequently only a very limited number of conformers will actually be tested.
∼3 min	ASP	[76] Computing times correspond to a single local optimization, carried out on one specific example.
1 min	DISCO	The number of common pharmacophoric points to be searched for is restricted to four or five.
70 min		For the only molecular test case considered, a 9-atom query pattern is searched for in a 22-atom target.
140 min	AUTOFIT	[142] $N_{conf.}$ varies from a few to over a million being inversely related to the number of pharmacophoric points considered.
30 min	MSC	The runtime corresponds to a single rigid-body optimization of which 192 have been carried out on the single application reported, considering different pairs of conformers sequentially.
–	MEPSIM	[93] The application data is taken from the previous study. Considering a single torsional degree of freedom is permitted.
1–2 h	COMPASS	102 molecules are considered simultaneously in a single study.
1 min	SEAL	[17] Iterated rigid-body superposition of conformers, as well as two alternative flexible fitting approaches are provided as extensions.
<5 min	PRO_LIGAND	[104] $N_{conf.}$ corresponds to two alternative approaches suggested.
–	AAA	[132] 30 molecules are considered simultaneously in a single study. Using a dense sampling of torsional angles ensures the consideration of a large portion of conformational space.
15–302 min	GASP	[138] Simultaneous consideration of up to 7 molecules in a single application is reported on.
3.5–33 min	TFIT	Pairs of molecules are considered, one of which is treated as flexible. Manual intervention is required for the initial fitting step.
10–18 min	PLM	[108] In one of the applications a short peptide is placed onto a 10 times bigger template molecule. Only a single conformer is considered in this case.
<1 min/>30 min	–	Alternative figures correspond to two applications performed on vastly different scales, ligand and protein superposition, respectively.
<12 h	CATALYST	[112] Up to six molecules are considered simultaneously in a single study.
–	–	[80] Runtimes are provided for some test cases of varying size; however, they are omitted for the actual applications presented.
2–5 s	–	ϕ rt is given for a single local optimization and depends on the size and perturbation of pattern and target.
0.5–10 h	RAPID	[114] The runtime heavily depends on $N_{conf.}$.
3.5 min	FLEXS	Pairs of molecules are considered, one of which is treated as flexible. The conformational model permits to access millions of conformers for a typical drug-size molecule.
1 s	–	The runtime corresponds to a single local optimization.
–	MIMIC	[116] Successively generated pairwise alignments are used as the starting point for 3- or 4-set multi-molecule alignments.
1–5 min	QUASIMODI	The computing time and the accuracy of results heavily depend on the resolution used to perform the calculations.
1–2 min	–	Results obtained with either the Simplex or the Monte Carlo optimization method are compared.
7 min	–	The runtime heavily depends on the GA parameters used, and these in turn on the size and flexibility of the molecules investigated.
24 s	RIGFIT	[88] The runtime given represents a compromise between efficiency and accuracy of the computation that is competitive with other state-of-the-art alignment tools.
1.2 s	RIGFIT	[17] The predicted superpositions were generated starting from the ligand coordinates derived from the crystal structures.

Best rmsd displays the best achieved rmsd comparing crystallographically determined and computationally predicted superpositions. N_{good} provides the percentage of examples with an rmsd below 1.5 Å. ϕ rt provides the mean of the runtimes for the different superpositions. Name provides the name of the respective software tool and the final column gives an additional link to earlier publications and additional comments.

- Handschuh et al. [141] describe a flexible superposition technique based on a hybrid genetic algorithm/directed tweak method. A GA approach similar to that in [136] with certain substantial differences in the implementation is used in combination with the directed tweak technique first described in [8]. The extensions comprise full atom matching, simultaneous consideration of matches of different size, additional GA operators, and the so-called *directed tournament selection* in order to avoid premature loss of genetic information. The directed tweak technique is used to obtain a high quality matching on the basis of the genetic coding, allowing for conformational adaption. However, intramolecular hindrance is not considered in this step, in order to keep the runtimes low. Simultaneous consideration of several molecules is permitted by the approach.

Comparison of approaches

Comparing different approaches to molecular structural alignment is a difficult task. On the one hand, a test suite of considerable size and diversity is lacking. Such a set of examples, preferably with known ideal solutions, could provide a benchmark for the entire area of molecular similarity research. On the other hand, even the choice of an objective *error function* that would provide a means of measuring *success* is problematic. Obviously, an rms deviation (*rmsd*) of observed versus predicted geometry may be calculated in cases of a known ideal. However, rmsd values may be rather misleading, especially if they adopt larger values. A particular rmsd value can result either from a moderate overall fit or from a convincing fit in some structural portion and an obvious misfit in another part due to conformational and orientational differences. Another problem is the treatment of symmetric or approximately symmetric cases, for which the symmetry-related solutions are far apart from each other in terms of an rmsd. Finally, it is obviously unfair to compare rmsd values resulting from superposition experiments performed on different scales. E.g., 1.5 Å rmsd has a different meaning if two sugar moieties with 10 Å as the largest possible intramolecular distance or two pentapeptides that easily span 30 Å are compared.

Additional difficulties arise from different hardware platforms used, different preconditions assumed, and different degrees of manual intervention required by the respective approach. However, computing times and rmsd values are an important source of information and remain actually the only way of objective comparison unless other precise standards and benchmarks are established. Table 2 collects the available data for the approaches to small molecule superposition described in the previous section. The table also mentions software systems that provide implementation platforms of the discussed methods.

Another point that needs to be taken into account is the applicability of a method. Sometimes only a single highly specific application is provided together with a method (see, e.g., [75, 77, 81, 91, 93, 113, 121]). Frequently, this puts only limited confidence in the general applicability of the respective technique. At the other extreme are application studies providing a variety of substantially different test cases, rigorously evaluated by the aid of quantitative performance criteria such as rmsd, runtime, etc. (see, e.g., [79, 83, 86–88]). Columns 3 and 4 in Table 2 provide the corresponding available data.

Perspectives

In this final section the authors provide their own view on the future of computational methods for the structural alignment of molecules.

Molecular superposition plays the key role in most methods to analyzing the bonding potential of ligands in the absence of structural data of the target protein under consideration. Basically, there are three established approaches to do this kind of analysis. *Pharmacophore elucidation* extracts the key functional features of binding, *3D-QSAR studies* derive models that allow for estimating binding affinities, and *receptor modeling* provides insight in the specific process of binding to the target protein. The common goal of all these methods is to arrive at computer models which allow to do virtual database screening.

In our opinion, the most promising techniques follow a *filtering protocol* to reduce the number of candidates step by step with increasing sensitivity and accordingly increasing computational costs. However, since the candidate set is reduced in size along the way, the overall effort for each filtering step remains roughly constant.

Many of the methods discussed in the review above can perform single steps of this process. What remains to be done, is to put the pieces together in a reasonable way to arrive at a workbench for virtual database screening. However, this step is not as straightforward as it seems at first sight. E.g., one should consider using alternative filters in parallel rather than applying them sequentially (as usual) since the results stated in the literature clearly indicate that different approaches, which achieve comparable enrichments, pick largely different sets of actives. This workbench is certainly not meant to substitute for the chemist's intuition and laboratory screening procedures. Rather, it should guide the chemist in his or her search for active compounds with certain desired properties and provide ideas for possible and plausible alternatives to the model derived so far. Also, it should help to prioritize experiments to do the most informative ones first. For this purpose, rather than finding the highest number of active compounds, we find it important to learn as much as possible about the target by detecting the boundary between active and inactive compounds in the chemical space under investigation. However, detecting a boundary is only possible if we identify compounds on either side of it. As a point in case, CoMFA samples regions of chemical space in which chemical variation results in a variation of the activity.

Then of course, one should improve the different filters to further enhance the overall performance of the workbench. This is certainly still necessary both in terms of computational speed and in terms of the accuracy of the predictions. The workbench needs to facilitate interactive usage for the different steps of the model building and also the size of the databases possibly to be searched still needs to grow. Also, one should consider that the goal of the first filtering steps is mainly to reject useless candidates with the lowest possible rate of false negatives (i.e., actives that are rejected). In contrast, the last filtering steps mainly aim at selecting active candidates with the lowest possible rate of false positives (i.e., inactives that are selected). Current methods do not really take this change in strategy into account.

In the future, ligand datasets assembled by combinatorial principles will be increasingly important. Of course, computational methods for the structural alignment of molecules should be redesigned for the special purposes of this kind of data. Recurring instances of molecular fragments should be handled as such, which would be more efficient than simple sequential processing.

Usually, there will be no data available on the target protein. In the other case, however, such invaluable data should not be neglected. Even in cases where docking methods do not apply, structural alignment methods may and should benefit from partial information, such as a metal ion or excluded volumes. Finally, methods for affinity prediction have to be improved. The grid imposed by techniques like CoMFA is essentially an artifact of the method and not inherent to the problem. To overcome any grid-spacing problems, the presence or absence of functional groups should be taken into account directly in order to determine which of them are relevant and how to arrive at a model that may predict activity. Also, while most of the available methods aim at lead identification, lead optimization is no less important. Information gained from experiments need to be incorporated into more sophisticated models. Still most of the approaches provide basically pairwise alignments and truly multi-molecule flexible fitting is largely missing. However, our studies exhibited examples for which neither of three pairwise alignments of three molecules A to B, B to C, and C to A reveal the true solution of multiply aligning A, B, and C.

In our opinion, the data-derived discrimination between active and inactive compounds or, more ambitiously, the prediction of activity on the basis of data is basically a learning problem and there is a wealth of theoretically founded machine learning techniques available that go far beyond a simple neural net. These methods need to be explored and incorporated in the software to make the best possible use out of the knowledge bases generated by the experiments.

Acknowledgements

This work has been performed as part of the RELIMO-Project, which was partially funded by the German Federal Ministry for Education, Science, Research, and Technology (BMBF) under grant no. 03 11 620.

References

1. Brint, A.T. and Willett, P., J. Chem. Inf. Comput. Sci., 27 (1987) 152.
2. Martin, Y.C., Bures, M.G. and Willett, P., In Lipkowitz, B. and Boyd, D.B., Reviews in Computational Chemistry, VCH, Weinheim, 1990, pp. 265–294.
3. Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications. ESCOM, Leiden, 1993.
4. Golender, V.E. and Vorpagel., E.R., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications. ESCOM, Leiden, 1993, pp. 137–149.
5. Wermuth, C.-G. and Langer, T., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications. ESCOM, Leiden, 1993, pp. 117–136.

6. Kabsch, W., Acta Crystallogr., A32 (1976) 922.
7. Redington, P.K., Comput. Chem., 16 (1992) 217.
8. Hurst, T., J. Chem. Inf. Comput. Sci., 34 (1994) 190.
9. Carbó, R., Leyda, L. and Arnau, M., Int. J. Quant. Chem., 17 (1980) 1185.
10. Hodgkin, E.E. and Richards, G., Int. J. Quant. Chem., Quantum Biol. Symp., 14 (1987) 105.
11. Lamdan, Y. and Wolfson, H.J., IEEE International Conference on Computer Vision, Tampa, FL, 1988, pp. 238–249.
12. Nussinov, R. and Wolfson, H.J., Proc. Natl. Acad. Sci. USA, 88 (1991) 10495.
13. Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1992) 83.
14. Bron, C. and Kerbosch, J., Communications of the ACM, 16 (1973) 575.
15. Crippen, G.M. and Havel, T.F., Distance Geometry and Molecular Conformation. Research Studies Press, Taunton, 1988.
16. Devillers, J. (Ed.) Genetic Algorithms in Molecular Modelling. Academic Press, London, 1996.
17. Kearsley, S.K. and Smith, G.M., Tetrahedron Comput. Methodol., 3 (1990) 615.
18. Johnson, M.A. and Maggiora, G.M. (Eds) Concepts and Applications of Molecular Similarity. John Wiley & Sons, New York, NY, 1990.
19. Dean, P.M., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design. Blackie Academic & Professional, London, 1995, pp. 1–23.
20. Good, A.C., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design. Blackie Academic & Professional, London, 1995, pp. 24–56.
21. Rouvray, D.H., In Sen, K. (Ed.) Molecular Similarity I. Topics in Current Chemistry, Vol. 173, Springer-Verlag, Heidelberg, 1995, pp. 1–30.
22. Klopman, G., J. Am. Chem. Soc., 106 (1984) 7315.
23. Klopman, G., Quant. Struct.-Act. Relat., 11 (1992) 176.
24. Petke, J.D., J. Comput. Chem., 14 (1993) 928.
25. Crippen, G.M., J. Comput. Chem., 16 (1995) 486.
26. Ghose, A.K. and Crippen, C.M., In Ramsden, C. (Ed.) Comprehensive Medicinal Chemistry: the Rational Design, Mechastic Study, and Therapeutic Application of Chemical Compounds. Vol. 4. Pergamon Press, Oxford, 1990, pp. 715–753.
27. Crippen, G.M., J. Comput. Chem., 8 (1987) 943.
28. Gasteiger, J. and Li, X., Angew. Chem. Int. Ed. Engl., 33 (1994) 643.
29. Gasteiger, J. and Zupan, J., Biol. Cybern., 70 (1993) 189.
30. Polański, J., Gasteiger, J., Wagener, M. and Sadowski, J., Quant. Struct.-Act. Relat., 17 (1998) 27.
31. Apaya, R.P., Luchese, B., Price, S.L. and Vinter, J.G., J. Comput.-Aided Mol. Design, 9 (1995) 33.
32. Blanley, F.E., Edge, C. and Phippen, R.W., J. Mol. Graph., 13 (1995) 165.
33. Robinson, D.D., Lyne, P.D. and Richards, W.G., J. Chem. Inf. Comput. Sci., 39 (1999) 594.
34. Mezey, P.G., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design. Blackie Academic & Professional, London, 1995, pp. 241–268.
35. Matter, H. and Rarey, M., In Jung, G. (Ed.) Combinatorial Organic Chemistry. John Wiley & Sons, New York, NY, 1999.
36. Humblet, C. and Dunbar Jr., J.B., In Venuti, M.C. (Ed.) Annual Reports in Medicinal Chemistry. Vol. 28, Chapter VI. Topics in Drug Design and Discovery. Academic Press, London, 1993, pp. 275–284.
37. Willett, P., J. Mol. Recognition, 8 (1995) 290.
38. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
39. Lauri, G. and Bartlett, P.A., J. Comput.-Aided Mol. Design, 8 (1993) 51.
40. Bartlett, P.A., Shea, G.T., Telfer, S.J. and Waterman, S., In Roberts, S.M. (Ed.) Molecular Recognition: Chemical and Biological Problems. Royal Society of Chemistry, Cambridge, 1989, pp. 182–196.
41. Moock, T.E., Henry, D.R., Ozkabak, A.G. and Alamgir, M., J. Chem. Inf. Comput. Sci., 34 (1994) 184.
42. Thorner, D.A., Wild, D.J., Willett, P. and Wright, P.M., J. Chem. Inf. Comput. Sci., 36 (1996) 900.
43. Thorner, D.A., Willett, P., Wright, P.M. and Taylor, R., J. Comput.-Aided Mol. Design, 11 (1997) 163.
44. Wild, D.J. and Willett, P., J. Chem. Inf. Comput. Sci., 36 (1996) 159.
45. Hahn, M., J. Chem. Inf. Comput. Sci., 37 (1997) 80.
46. Hahn, M., J. Med. Chem., 38 (1995) 2080.
47. Wang, T. and Zhou, J., J. Chem. Inf. Comput. Sci., 38 (1998) 71.
48. Xu, J., J. Chem. Inf. Comput. Sci., 36 (1996) 25.
49. Rarey, M. and Dixon, J.S., J. Comput.-Aided Mol. Design, 12 (1998) 471.
50. Orengo, C., Curr. Opin. Struct. Biol., 4 (1994) 429.
51. Holm, L. and Sander, C., Proteins, 19 (1994) 165.
52. Godzik, A., Protein Sci., 5 (1996) 1325.
53. Lattman, E.E. (Ed.) Critical Assessment of Techniques for Protein Structure Prediction (CASP2). Supplement 1 to Proteins: Structure, Function, and Genetics, 1997. See also http://PredictionCenter.llnl.gov/casp3/Casp3.html.
54. Gibrat, J.-F., Madej, T. and Bryant, S.H., Curr. Opin. Struct. Biol., 6 (1996) 377.
55. Holm, L. and Sander, C., J. Mol. Biol., 233 (1994) 123.
56. Taylor, W.R. and Orengo, C.A., J. Mol. Biol., 208 (1989) 1.
57. Fischer, D., Wolfson, H., Lin, S.L. and Nussinov, R., Protein Sci., 3 (1994) 769.
58. Alesker, V., Nussinov, R. and Wolfson, H.J., Protein Eng., 9 (1996) 1103.
59. Norel, R., Lin, S.L., Wolfson, H.L. and Nussinov, R., J. Mol. Biol., 252 (1995) 263.
60. Sandak, B., Nussinov, R. and Wolfson, H.J., Comput. Appl. Biosci., 11 (1995) 87.
61. Rarey, M., Wefing, S. and Lengauer, T., J. Comput.-Aided Mol. Design, 10 (1996) 41.
62. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
63. Diederichs, K., Proteins, 23 (1995) 187.
64. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A., Proc. Natl. Acad. Sci. USA, 89 (1992) 2195.
65. Lattman, E.E., Acta Crystallogr., B28 (1972) 1065.
66. May, A.C.W. and Johnson, M.S., Protein Eng., 8 (1995) 873.
67. May, A.C.W. and Johnson, M.S., Protein Eng., 7 (1994) 475.
68. Fredman, M.L., Bull. Math. Biol., 46 (1984) 553.
69. Poirrette, A.R., Artymiuk, P.J., Rice, D.W. and Willett, P., J. Comput.-Aided Mol. Design, 11 (1997) 557.
70. Connolly, M.L., J. Appl. Crystallogr., 16 (1983) 548.
71. Wallace, A.C., Borkakoti, N. and Thornton, J.M., Protein Sci., 6 (1997) 2308.
72. Escalier, V., Pothier, J., Soldano, H. and Viari, A., J. Comput. Biol., 5 (1998) 41.

73. Klebe, G., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications. ESCOM, Leiden, 1993, pp. 173–199.
74. Bures, M.G., In Charifson, P.S. (Ed.) Practical Application of Computer-Aided Drug Design. Marcel Dekker, New York, NY, 1997, pp. 39–72.
75. Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Chem. Inf. Comput. Sci., 32 (1992) 188.
76. Good, A.C., J. Mol. Graph., 10 (1992) 144.
77. Feuilleaubois, E., Fabart, V. and Douchet, P., SAR QSAR Env. Res., 1 (1993) 97.
78. Petitjean, M., J. Comput. Chem., 16 (1995) 80.
79. Grant, J.A., Gallardo, M.A. and Pickup, B.T., J. Comput. Chem., 17 (1996) 1653.
80. Grant, J.A. and Pickup, B.T., J. Phys. Chem., 99 (1995) 3503.
81. Cossé-Barbi, A. and Raji, M., J. Comput. Chem., 18 (1997) 1875.
82. McMahon, A.J. and King, P.M., J. Comput. Chem., 18 (1997) 151.
83. Nissink, J.W.M., Verdonk, M.L., Kroon, J., Mietzner, T. and Klebe, G., J. Comput. Chem., 18 (1997) 638.
84. Crowther, R.A., In Rossmann, M.G. (Ed.) The Molecular Replacement Method. Gordon & Breach, New York, NY, 1972, pp. 174–178.
85. Parretti, M.F., Kroemer, R.T., Rothman, J.H. and Richards, W.G., J. Comput. Chem., 18 (1997) 1344.
86. Lemmen, C., Hiller, C. and Lengauer, T., J. Comput.-Aided Mol. Design, 12 (1998) 491.
87. Klebe, G., Mietzner, T. and Weber, F., J. Comput.-Aided Mol. Design, 8 (1994) 751.
88. Lemmen, C. and Lengauer, T., J. Comput.-Aided Mol. Design, 11 (1997) 357.
89. Martin, Y.C., Tetrahedron Comput. Methodol., 3 (1990) 15.
90. Takahashi, Y., Maeda, S. and Sasaki, S.-I., Ann. Chimica Acta, 200 (1987) 363.
91. Masek, B.B., Merchant, A. and Matthew, J.B., J. Med. Chem., 36 (1993) 1230.
92. Sanz, F., Manaut, F., Rodríguez, J., Lozoya, E. and López-de-Briñas, E., J. Comput.-Aided Mol. Design, 7 (1993) 337.
93. Sanz, F., Manaut, F., Sanchez, J.A. and Lozoya, E., J. Mol. Struct., 230 (1991) 437.
94. Jain, A.N., Dietterich, T.G., Laterop, R.H., Chapman, D., Critchlow, R.E., Bauer, B.E., Webster, T.A. and Lonzano-Perez, T., J. Comput.-Aided Mol. Design, 8 (1994) 427.
95. Jain, A.N., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.
96. Ghuloum, A.M., Sage, C.R. and Jain, A.J., J. Med. Chem., 42 (1999) 1739.
97. Ruppert, J., Welch, W. and Jain, A.N., Protein Sci., 6 (1997) 524.
98. Welch, W., Ruppert, J. and Jain, A.N., Chem. Biol., 3 (1996) 449.
99. Jain, A.N., J. Comput.-Aided Mol. Design, 10 (1996) 635.
100. Gerber, P.R. and Müller, K., Acta Crystallogr., A43 (1987) 426.
101. Klebe, G., Mietzner, T. and Weber, F., J. Comput.-Aided Mol. Design, 13 (1999), 35.
102. Waszkowycz, B., Clark, D.E., Frenkel, D., Li, J., Murray, C.W., Robson, B. and Westhead, D.R., J. Med. Chem., 37 (1994) 3994.
103. Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
104. Clark, D.E., Frenkel, D., Levy, S.A., Li, J., Murray, C.W., Robson, B., Waszkowycz, B. and Westhead, D.R., J. Comput.-Aided Mol. Design, 9 (1995) 13.
105. Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 479.
106. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 107.
107. Danziger, D.J. and Dean, P.M., J. Theor. Biol., 116 (1985) 215.
108. Perkins, T.D.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 7 (1993) 155.
109. Barnum, D., Greene, J., Smellie, A. and Sprague, P., J. Chem. Inf. Comput. Sci., 36 (1996) 563.
110. Smellie, A., Teig, S.L. and Towbin, P., J. Comput. Chem., 16 (1995) 171.
111. Ferro, D.R. and Hermans, J., Acta Crystallogr., A33 (1977) 345.
112. Sprague, P.W., Perspect. Drug Discov. Design, 3 (1995) 21.
113. Finn, P.W., Kavraki, L., Latombe, J.-C., Motwani, R., Shelton, C., Venkatasubramanian, S. and Yao, A., Proceedings of the 13th Annual Symposium on Computational Geometry, ACM Press, New York, NY, 1997.
114. Finn, P.W., Halperin, D., Kavraki, L., Latombe, J.-C., Motwani, R., Shelton, C. and Venkatasubramanian, S., In Lin, M. and Manocha, D. (Eds) 1996 ACM Workshop on Applied Computational Geometry, LNCS, Springer-Verlag, Heidelberg, 1996.
115. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Mol. Graphics Mod., 15 (1997) 114.
116. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Comput. Chem., 18 (1997) 934.
117. Miller, M.D., Sheridan, R.P. and Kearsley, S.K., J. Med. Chem., 42 (1999) 1505.
118. Sheridan, R.P., Miller, M.D., Underwood, D.J. and Kearsley, S.K., J. Chem. Inf. Comput. Sci., 36 (1996) 128.
119. Kearsley, S.K., Smith, G.M., Sallamack, S., Fluder, E.M., Andose, J.D., Mosley, R.T. and Sheridan, R.P., J. Chem. Inf. Comput. Sci., 36 (1996) 118.
120. Kearsley, S.K., Underwood, D.J., Sheridan, R.P. and Miller, M.D., J. Comput.-Aided Mol. Design, 8 (1994) 565.
121. Sheridan, R.P., Nilakantan, R., Dixon, J.S. and Venkataraghavan, R., J. Med. Chem., 29 (1986) 899.
122. Crippen, G.M., J. Med. Chem., 22 (1979) 988.
123. Crippen, G.M., J. Med. Chem., 23 (1980) 599.
124. Crippen, G.M., J. Med. Chem., 24 (1981) 198.
125. Crippen, G.M., Smellie, A.S. and Richardson, W.W., J. Comput. Chem., 13 (1992) 1262.
126. Billeter, M., Havel, T.F. and Kuntz, I.D., Biopolymers, 26 (1987) 777.
127. Clark, D.E., Willett, P. and Kenny, P.W., J. Mol. Graph., 11 (1993) 146.
128. Sheridan, R.P. and Venkataraghavan, R., J. Comput.-Aided Mol. Design, 1 (1987) 243.
129. Sheridan, R.P., Rusinko III, A., Nilakantan, R. and Venkataraghavan, R., Proc. Natl. Acad. Sci. USA, 86 (1989) 8125.
130. Itai, A., Tomioka, N., Yamada, M., Inoue, A. and Kato, Y., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications. ESCOM, Leiden, 1993, pp. 173–199.
131. Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B. and Marshall, G.R., J. Comput.-Aided Mol. Design, 9 (1995) 491.

132. Marshall, G.R., Barry, C.D., Bosshard, H.D., Dammkoehler, R.D. and Dunn, D.A., In Olson, E.C. and Christoffersen, R.E. (Eds) Computer-Assisted Drug Design. Vol. 112. American Chemical Society, Washington, DC, 1979, pp. 205–222.

133. Ghose, A.K., Logan, M.E., Treasuywala, A.M., Wang, H., Wahl, R.C., Tomczuk, B.E., Gowravaram, M.R., Jaeger, E.P. and Wendoloski, J.J., J. Am. Chem. Soc., 117 (1995) 4671.

134. McMartin, C. and Bohacek, R.S., J. Comput.-Aided Mol. Design, 9 (1995) 237.

135. Chang, G., Guida, W.C. and Still, W.C., J. Am. Chem. Soc., 111 (1989) 4379.

136. Jones, G., Willett, P. and Glen, R.C., J. Comput.-Aided Mol. Design, 9 (1995) 532.

137. Jones, G., Willett, P. and Glen, R.C., J. Mol. Biol., 245 (1995) 43.

138. Payne, A.W.R. and Glen, R.C., J. Mol. Graph., 11 (1993) 74.

139. Pepperrell, C.A. and Willett, P., J. Comput.-Aided Mol. Design, 5 (1991) 455.

140. Lemmen, C., Lengauer, T. and Klebe, G., J. Med. Chem., 41 (1998) 4502.

141. Handschuh, S., Wagener, M. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 38 (1998) 220.

142. Kato, Y., Inoue, A., Yamada, M., Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 6 (1992) 475.