CHAPTER 2

Chemoinformatics—Applications in Food Chemistry

Karina Martinez-Mayorga and **Jose L. Medina-Franco**

Contents

Abstract

The aim of the present chapter is to present the current research and potential applications of chemoinformatics tools in food chemistry. First, the importance and variety of molecular descriptors and physicochemical properties is delineated, and then a survey and chemical space analysis of representative databases with emphasis on food-related ones is presented. A brief description of methods commonly used in molecular design, followed by examples in food chemistry are presented, such methods include similarity searching, pharmacophore modeling, quantitative

Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway, Port St. Lucie, Florida 34987

Advances in Food and Nutrition Research, Volume 58
ISSN 1043-4526, DOI: 10.1016/S1043-4526(09)58002-3

structure–activity relationships (QSAR), and molecular docking. The relatedness to virtual screening is emphasized and the perspectives from this field are presented at the end.

I. INTRODUCTION

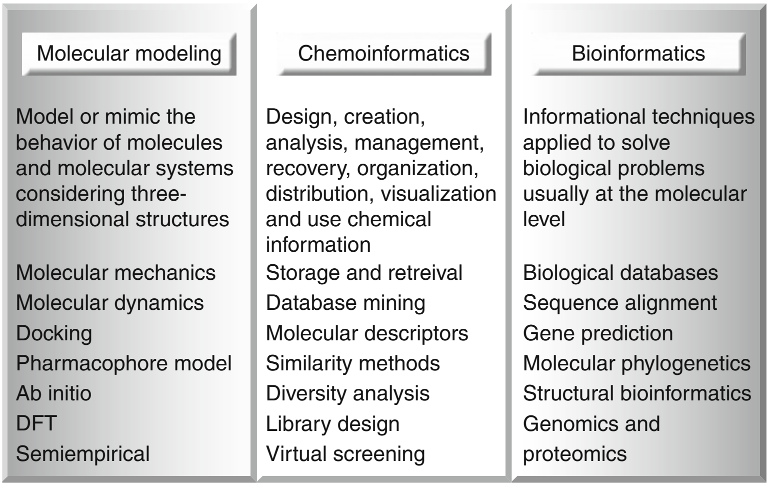
Chemoinformatics is a relatively new research field which has received several definitions. Among the more often cited are:

> The mixing of information resources to transform data into information, and information into knowledge, for the intended propose of making better decisions faster in the arena of drug lead identification and optimization. (Brown, 1998)
>
> Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information. (Paris, 2000)

Accordingly, managing chemical information and computational tools requires knowledge of multiple fields, and in most cases familiarity with a biological system is required in order to generate meaningful models. Chemoinformatics often merits a department of its own in universities and industrial organizations. Very recently the first Master in Science and chemoinformatics courses have been implemented at the University of Sheffield (http://www.shef.ac.uk/is/research/groups/chem/courses.html), the University of Manchester (http://www.informatics.manchester.ac.uk/teaching/), and Indiana University (http://cheminfo.informatics.indiana.edu/). Wild and Wiggings (2006) have described these courses and also the challenges of chemoinformatics education including current trends involving distance education and intensive short courses.

Accordingly, specific journals that are related to or include chemoinformatics in their scope are*: Journal of Chemical information and Modeling*, *Journal of Computer Aided Molecular Design*, *Journal of Chemical Graphics and Modeling*, *Journal of Molecular Modeling*, *Current Computer Aided Drug Design*, and the most recent *Journal of Chemoinformatics* which is *Open Access and Molecular Informatics*. Importantly, while chemoinformatics has been developed with special attention to drug design, the scope is not limited to here and it can be use in other areas like food science, materials science, or polymers.

Figure 2.1 shows a short description and methodologies classified into molecular modeling, chemoinformatics, and bioinformatics. The boundaries are soft; however, it is not uncommon to use a combination of methods. For instance, a pharmacophore or docking model is often used for virtual screening. In general, the outcome is structure–property relationships and in the case of virtual screening the prediction ultimately of new molecules having a certain desired property is the final goal.



Molecular modeling	Chemoinformatics	Bioinformatics
Model or mimic the behavior of molecules and molecular systems considering three-dimensional structures	Design, creation, analysis, management, recovery, organization, distribution, visualization and use chemical information	Informational techniques applied to solve biological problems usually at the molecular level
Molecular mechanics	Storage and retreival	Biological databases
Molecular dynamics	Database mining	Sequence alignment
Docking	Molecular descriptors	Gene prediction
Pharmacophore model	Similarity methods	Molecular phylogenetics
Ab initio	Diversity analysis	Structural bioinformatics
DFT	Library design	Genomics and proteomics
Semiempirical	Virtual screening	

FIGURE 2.1 General classification and definitions of computational methods employed in molecular design in chemistry and biology.

Analysis of chemical information can be performed at different levels depending on the goal and the type and amount of information available. In general, large databases are used for studying the chemical space or as a data source for virtual screening, whereas smaller datasets (<50 compounds) can be used to generate, for instance, pharmacophore or quantitative structure–activity relationships (QSARs) models. In many cases, the system under study involves interactions of small molecules (ligands) with a particular target(s). When the 3D structure of the target is known from X-ray crystallography, NMR, or homology models, the technique of choice is molecular docking. If the 3D structure of the target is unknown, the models are built based on the information of one or more ligands used as reference, such methods being referred to as ligand-based methods, among which, similarity analysis, pharmacophore modeling, and QSAR are the most popular.

Figure 2.2 shows some of these methods and how they can be used in combination; it also shows the outline of the following sections in this chapter. First, the concept of chemical space, including a sketch of molecular descriptors, physicochemical properties, and databases is presented. Then, a brief description of molecular similarity, pharmacophore modeling, docking, and QSAR models with the incorporation of examples of food-related components is described. These methods are mainly used to develop and analyze SAR and the resultant models can be used to perform virtual screening. Comprehensive reviews of each of these methods are described elsewhere (Alvarez and Shoichet, 2005; Leach and Willet, 2003; Varnek and Tropsha, 2008).

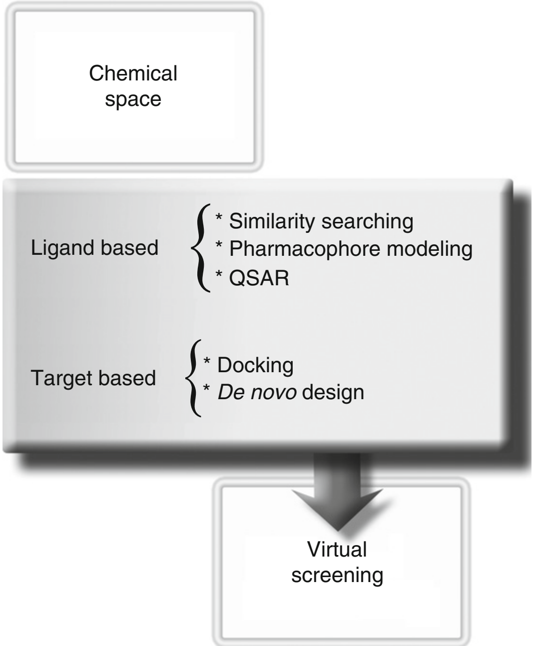


FIGURE 2.2 Exploration of the chemical space includes the classification and analysis of databases. To generate structural models several methodologies have been developed, they can be classified into ligand-based and target-based methods. Models derived from these methods are frequently used as a source for virtual screening.

II. MOLECULAR DESCRIPTORS AND PHYSICOCHEMICAL PROPERTIES

Chemoinformatics is characterized by the use of large amounts of information. Specific ways to represent the molecules, and to organize and analyze the data have been and continue to be developed. There are different ways to represent the molecules, and they can be classified according to the information that they encode. The most basic level corresponds to representations that depend on or are associated with one dimensional (1D) representation, such as molecular weight. The next level corresponds to the 2D representations associated with the connectivity of the molecules without the consideration of the stereochemistry. In 3D methods the incorporation of stereochemistry conveys not only the specification of the chirality of stereogenic centers but also the possible conformation or conformations. In this regard, the research fields dedicated to conformational analysis have an important impact in reactivity prediction, molecular design, and stability. Examples of 1D, 2D, and

TABLE 2.1 Molecular descriptors and related properties

	Descriptor	Physicochemical and structural properties
1D	Simple filter	Molecular weight
	Complex filter	Basicity, acidity, etc.
2D	Fingerprints	Atom types and substructures in a binary representation
	Physicochemical properties	Lipophilicity, pK_a values, ADME, or toxicity parameters
3D	Quantum chemical descriptors	Change distribution in the molecules
	3D pharmacophore patterns	Targets interaction sites (Hbonds, electrostatic, etc.) Different scaffolds can be recognized
	SAR patterns	Pharmacological activity
	Autocorrelation coefficients	Similarity of shape
	Virtual screening	Pharmacological activity

3D descriptors along with the property represented are summarized in Table 2.1.

A number of methods have been developed to store and share information. It is desirable that such information is stored in an ''inexpensive'' way and, depending on the task, retaining as much structural detail as possible. Based on the molecular representation used, the molecules are stored in suitable formatted files. For instance, for 2D representations, the molecules can be represented by fingerprints, SMILES, or SMARTS strings; such files use a small amount of memory. For 3D representations, typical formats include .sdf, .pdb, .mol, and .mol2; in all these cases the files contain the 3D coordinates of the structures, thus the conformation can be defined.

III. MOLECULAR DATABASES AND CHEMICAL SPACE

The advent of experimental techniques such as combinatorial and parallel chemical synthesis, and high-throughput screening has enabled the production of massive amounts of data. These compound databases have played a key role in drug design (Miller, 2002) and other research areas such as Agrochemistry and Food Chemistry. Current computational capabilities permit the development of molecular databases, not only to store the information but also to link and mine the data. In addition to proprietary databases used in industry and other research groups, there are extensive collections of commercially available resources. Furthermore, institutional and collaborative efforts have made publicly available web-based databases for research and educational purposes. A review of large compound databases used in drug discovery has been published by Scior *et al.* (2007). Representative databases with general, food, and pharmaceutical scope are listed in Table 2.2.

The Generally Recognized as Safe (GRAS) compounds (Table 2.2) is a collection of molecules used in food and beverage products. A fraction of this collection is available from Flavor-Base as a list of chemical names (http://www.leffingwell.com/flavbase.htm). The chemical names can be automatically converted to chemical structures using available software. GRAS compounds have been compared to drugs, revealing significant differences in molecular size, flexibility, and atom composition (Sprous and Salemme, 2007).

The Distributed Structure-Searchable Toxicity (DSSTox) Database Network is a public database containing more than 1000 molecules annotated with toxicity data. The database can be searched online. The chemical structures can also be downloaded from the web site for analysis (*vide infra*).

TABLE 2.2 Examples of large molecular databases used in research

Databases	Web sites
GRAS (Generally Recognized as Safe)	http://www.foodsafety.gov/~dms/eafus.html/ http://www.leffingwell.com/flavbase.htm
DSSTox (Distributed Structure-Searchable Toxicity)	http://www.epa.gov/ncct/dsstox/
SuperScent	http://bioinformatics.charite.de/superscent/
BIOPEP	http://www.uwm.edu.pl/biochemia
ZINC	http://zinc.docking.org/
MMsINC	http://mms.dsfarm.unipd.it/MMsINC.html
PubChem	http://pubchem.ncbi.nlm.nih.gov/
Developmental Therapeutics Program, National Cancer Institute	http://dtp.nci.nih.gov/ http://cactus.nci.nih.gov/
DrugBank	http://www.drugbank.ca/

SuperScent is a database comprising approximately 2000 scents (Dunkel *et al.*, 2008). The database can be browsed online and includes web-based similarity searching. The molecules are linked to PubChem (*vide infra*). Additional databases in the ''Super'' collection are SuperToxic, SuperSite, and SuperTarget. The SuperScent web site (Table 2.2) has links to other public databases such as SenseLab, Pherobase, Scentbase, Flavornet, and Leffingwell.

The Protein and Bioactive Peptide Sequences (BIOPEP) is a public database (Dziuba *et al.*, 1999) containing 707 proteins, 2123 bioactive peptides, 65 allergenic proteins with their epitopes, and 224 sensory peptides and amino acids. Databases include information such as sequence, number of amino acid residues, molecular weight, activity, and references. BIOPEP can be browsed online and the user can search for a specific sequence.

ZINC (Irwin and Shoichet, 2005) is a freely accessible database of increasing use in virtual screening and other computational applications. ZINC contains over 8 million purchasable compounds and it is a very attractive source of compounds to perform chemoinformatic comparisons with other collections.

The MMsINC database (Masciocchi *et al.*, 2009) is a public web-based resource with more than 4 million nonredundant compounds for virtual screening and other computational applications. Substructure and similarity searching can be performed. MMsINC is linked to other databases such as FDA, PubChem, Protein Data Bank, and ZINC.

Table 2.2 also lists three publicly available databases commonly used in drug research. PubChem is accessed through the National Library of Medicine (Austin *et al.*, 2004) and contains chemical structure information and corresponding activity across a number of biological assays. The system links the compound information with biomedical literature and it is possible to perform web-based similarity searching. PubChem also enables one to download files with structures to perform chemoinformatic analysis off-line.

The National Cancer Institute (NCI) database is a freely available collection of more than 200,000 compounds and it is frequently used by research groups to perform virtual screening followed by experimental testing. Table 2.2 contains two links where it is possible to mine the database online and to download structure files for analysis.

DrugBank (Wishart *et al.*, 2008) is a public resource containing a collection of nearly 4800 entries including FDA-approved small molecule and protein/peptide drugs, nutraceuticals, and experimental agents. The database has extensive information related to the chemical, pharmacological, and pharmaceutical information of the chemical agent as well as to the drug target. DrugBank can be searched online and the structure files can also be downloaded (*vide infra*).

The continued growth in the number of compounds stored in corporate, commercial, and public databases raises questions regarding the relationships of molecules within compound collections and across different databases. Common questions are related to the assessment of structural diversity of molecules, degree of overlap between compound datasets, and analysis and visualization of the *chemical space*. In fact, the chemical space or chemical universe coverage of compound databases has become a central concept not only in drug discovery but also in other research areas. Analysis and visualization of the chemical space of in-house, commercial, public, and virtual compound databases has a number of applications including diversity analysis, *in silico* property profiling, data mining, virtual screening, library design, prioritization in screening campaigns, and database acquisition (Medina-Franco *et al.*, 2008). Chemical space has been defined as ''the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created.'' (Dobson, 2004) A number of methods have been developed to analyze and visualize such *descriptor space*. Such methods have been applied to compare the chemical space covered by compound collections from different sources including natural products, drugs, metabolites, toxic and nontoxic substances, among other numerous datasets. These comparisons have been reviewed recently by Medina-Franco *et al.* (2008). It is important to mention here that visualization of the chemical space of a molecular database will depend on the molecular representation of the compounds to define the multidimensional descriptor space and the visualization technique to reduce the multidimensional space into a lower, 2D or 3D plot. In other words, the chemical space of a compound database will not be unique (Medina-Franco *et al.*, 2008). However, using the same descriptors to compare different databases enables the assessment of such collection relative to the descriptors used.

IV. CHEMOINFORMATICS IN FOOD CHEMISTRY

An example of a nonpharmaceutical use of chemoinformatics is the work described by Sprous and Salemme (2007) in the flavor portion of the GRAS compounds. Similar analysis could potentially be performed with other databases of, for example, nutraceuticals, flavors, colors, etc. In principle, nutraceuticals exemplify food components with drug-like properties. One of the major challenges in pharmaceutical campaigns is the toxicity of the potential drugs. It should be noted that the toxicity of a given compound is dose-dependent. The National Academy of Sciences Report 2007 (see http://www.nap.edu/catalog.php?record_id=11970#toc) identified the need of a major shift on the assessment of chemical's toxicity, and computational methods are recognized as key components in such transformation. Comparison of the chemical space of food component databases with those of other databases is expected to inform in both directions, that is, from the food chemistry and the pharmaceutical point of view. How do the molecular properties of the nutraceuticals compare with those of drugs and toxic compounds?

Figure 2.3 shows a comparison of the chemical space of the following three databases: DSSTox (1216 compounds), 92 nutraceuticals obtained from ZINC (*vide supra*) and 1490 drugs obtained from DrugBank. The chemical space was obtained by representing the molecules using six physicochemical properties namely molecular weight (MW), number of rotatable bonds (RB), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), topological polar surface area (TPSA), and the octanol/water partition coefficient (SlogP). To obtain the visual representation of

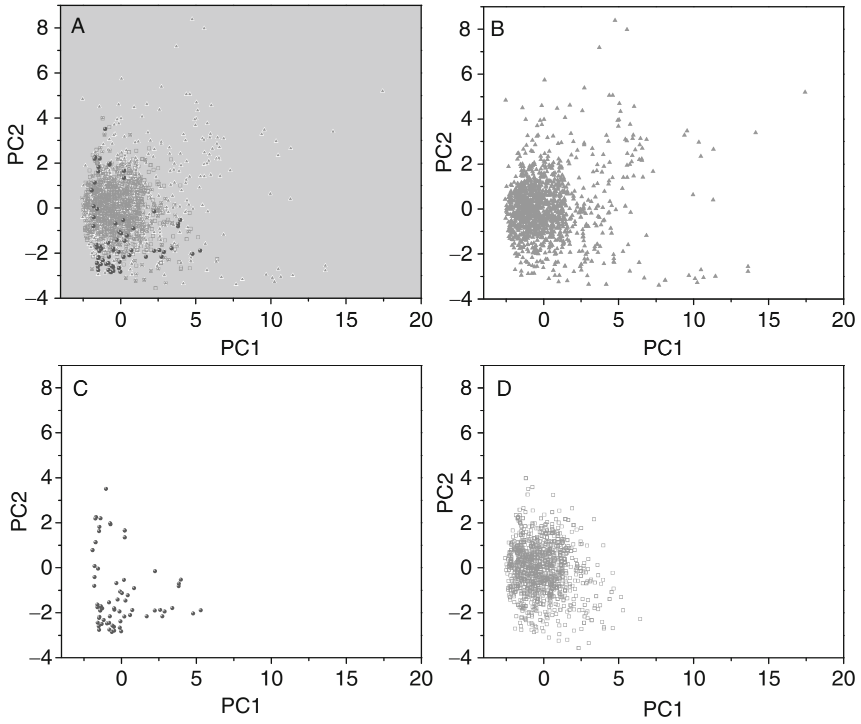


FIGURE 2.3 Chemical space of three databases obtained by principal component analysis of six (scaled) physicochemical properties. The first two principal components (PC1 and PC2) plotted here account for 77.2% of the variance: (A) three databases superimposed; (B) 1216 compounds from the Distributed Structure-Searchable Toxicity (DSSTox ▲); (C) 92 nutraceuticals (○); and (D) 1490 drugs from DrugBank (□).

the chemical space, a principal component analysis was performed considering the six above-mentioned properties (after scaling). A similar procedure has been reported before by our group to compare a collection of drugs and natural products among other compound collections using the same properties (Singh *et al.*, 2009).

Figure 2.3A shows all databases in the same space. Extensive overlap between the nutraceuticals, drugs, and compounds from the DSSTox database is readily observed. However, there are also clear differences such as the large number of DSSTox compounds located in a different region of the chemical space as compared to drugs and nutraceuticals. For clarity, Figure 2.3B–D shows each compound collection separately but within the same coordinates as used in Fig. 2.3A. From Fig. 2.3B and C the large diversity of compounds in the DSSTox and nutraceuticals databases can be inferred. One should keep in mind that the chemical space depicted in Fig. 2.3 is valid within the framework of the six physicochemical properties used. Table 2.3 shows a comparison of the median, mean, and standard deviation of the distribution of the six properties for each of the three databases. From the median and mean values it can be deduced that several properties do not follow a normal distribution. Comparing the statistics for the six properties it can be concluded that a major difference is MW. This property is obviously associated with the size of the molecules. In general, molecules in the DSSTox have larger MW than the compounds in the nutraceuticals and drugs databases. Notably, the standard deviation is quite large in the DSSTox set indicating a large variability in MW (actually the MW ranges between 46 and 1626). The nutraceuticals have, in general, a lower MW than the drugs but with a greater standard deviation. Interestingly, all three databases

TABLE 2.3 Summary of the distribution of the six physicochemical properties used to compare DSSTox, nutraceuticals, and drug databases

	Median/mean (S.D.)		
Property	DSSTox ($n = 1216$)	Nutraceuticals ($n = 92$)	Drugs ($n = 1490$)
Molecular weight	312/346 (167)	266/256 (115)	310/310 (91)
SlogP	1.6/1.3 (2.9)	0.8/0.3 (4.2)	1.6/1.4 (2.6)
TPSA	67.6/82.6 (61.8)	69.1/83.9 (54.3)	67.1/71.8 (41.5)
RB	5.0/5.3 (4.0)	5.0/4.9 (3.3)	5.0/5.1 (3.3)
HBA	2.0/3.1 (2.9)	1.0/1.9 (2.0)	2.0/2.5 (1.7)
HBD	1.0/1.4 (1.7)	1.0/1.3 (1.3)	1.0/1.2 (1.2)

have, on average, a similar number of RB, a property associated with molecular flexibility. However, considering the smaller size of the nutraceuticals as deduced from the MW it can be concluded that these molecules are more flexible than the drugs and compounds in the DSSTox database.

These are some examples of the analysis that can be performed for three representative datasets in Table 2.2. Similar analysis using these or other molecular representations can be conducted for other databases. In fact, a comparison of the collection of drugs analyzed here with natural products and compounds obtained from PubChem has been published elsewhere (Singh *et al.*, 2009).

V. EXAMPLES OF MOLECULAR SIMILARITY, PHARMACOPHORE MODELING, MOLECULAR DOCKING, AND QSAR IN FOOD OR FOOD-RELATED COMPONENTS

Going deeper into the comparison and classification of chemical structures, a set of computational techniques are available. This section is organized to present a brief description of currently used methodologies, namely; molecular similarity, pharmacophore modeling, molecular docking, and QSAR followed by application of these methods to food-related components.

Of particular importance is to keep in mind the *in vivo* nature of food-related studies, where it is expected the simultaneous activation of several biochemical pathways. For that reason, the selection and applicability of the different methods need to be understood. In addition, the source of the data may play a major role, in most cases the information is collected from different laboratories and the variability of the data might be due not only to the chemical structure but also to the different conditions under which the experiments were performed. As an example, problems arising from the collection of data in olfaction studies have been reviewed by Chastrette (1998).

A. Molecular similarity

Similarity searches are based on the hypothesis that similar molecules will have similar properties (Johnson and Maggiora, 1990). The similarity then is not a unique value; rather it is always referred to a reference compound(s), and also depends on the representation used to describe the molecules (Johnson and Maggiora, 1990). The dependence on the reference molecule and the similarity measure employed is frequently alleviated by the use of *fusion methods*, which consist of extracting a new column in the hit list based on the *maximum*, or *median* similarity among methods or queries (Willett, 2006a,b; Willett *et al.*, 1998). Recently, we have shown the use of both measures (*maximum* and *median*) in one plot, called multifusion similarity maps (Medina-Franco *et al.*, 2007). The changes in biological response with the changes in the structure have been rationalized as activity landscapes (Peltason and Bajorath, 2007). Even in the case of the interaction of one ligand with one receptor, the activity or property landscape can have different shapes. Changes in the biological response associated with proportional changes in the structure are in a region of smooth landscape. However, there are situations where small changes in the structure produce large changes in the biological response. Such features on the property landscape have been termed ''activity cliffs'' (Johnson and Maggiora, 1990). To alleviate the dependence on the molecular representation we have proposed the identification of activity cliffs by using multiple representations and find ''consensus activity cliffs''(Medina-Franco *et al.*, 2009).

Structure–activity relationships have been largely employed for molecular design; these correlations depend on the molecular representation and the activity landscape. The molecular representation depends only on the small molecule, whereas the activity landscape provides information on the ligand–receptor complex, for example, how permissive the binding pocket is. To exemplify the molecular similarity approach, a set of odorants (compared to benzaldehyde) will be presented.

The discovery and characterization of the olfactory receptors in the early 1990s by Buck and Axel (1991) facilitated the understanding of the mechanism of olfaction and biochemical principles involved. The olfactory receptors belong to the G-protein coupled receptor (GPCR) superfamily. These proteins are involved in a variety of biological processes like vision and pain and are one of the principal targets with therapeutic importance, that is, ~45% of currently marketed drugs target GPCRs. Rhodopsin, responsible for dim light vision in eukaryotes, was the first GPCR for which the X-ray crystal structure was resolved (Palczewski *et al.*, 2000), albeit in the inactive state. In addition to the dark or inactive state of rhodopsin, intermediates to the active state (Nakamichi and Okada, 2006a,b; Ruprecht *et al.*, 2004; Salom *et al.*, 2006; Tikhonova *et al.*, 2008) and the X-ray crystal structure of the β-adrenergic receptor have also been resolved (Bhattacharya *et al.*, 2008; Cherezov *et al.*, 2007; Cong *et al.*, 2001; Rasmussen *et al.*, 2007; Rosenbaum *et al.*, 2007). These advances in the GPCR field have allowed the generation of homology models of related proteins, such as the olfactory receptors. The homology models can then be used to analyze interactions of small molecules in the binding site. It can also be used to dock ligand databases into the receptor and thereby to enrich pharmacophore models. In 2001, Zozulya *et al.* (2001) reported the identification of 347 olfactory receptors in the human genome. A complication with olfactory receptors is that the same ligand can bind to multiple receptors while one receptor is able to host different ligands, this is indicative of the high plasticity of olfactory receptors. Knowledge of the structural requirements for binding to the olfactory receptors will help on the development of structure–odor relationships.

A recent review by Zarzo (2007) provides an overview of the different theories that have been proposed on the basis of olfaction. One of the earliest theories is the profile-functional group theory of olfaction, where the association of the functional groups to a particular odor is highlighted. A good example is the thiol (–SH) group, which imparts a rotten eggs or garlic odor to many of the molecules in which it is present. A later theory proposes that the shape of molecules influences their binding to certain nasal receptors. This theory was first proposed by Moncrieff (1967) in 1949 and was latter referred to as ''stereochemical theory'' by Amoore (1963). As early as 1963, Amoore postulated that the size and shape of the molecule determined its odor. As a means of quantification, Amoore (1963) produced silhouette photographs of a scale model of the benzaldehyde molecule, and compared the silhouette of homologous compounds, taking benzaldehyde as the standard. This early form of molecular similarity leads to a reasonably good odor–structure correlation, and he conducted similar analysis on other molecular sets (Amoore, 1967; Amoore *et al.*, 1967).

To exemplify a molecular similarity method, we employed here a 3D shape-based molecular similarity approach using OpenEye scientific software (OpenEye). A set of 27 molecules (Amoore, 1971) were compared to benzaldehyde (query molecule). The representation used here is based on the volume of each molecule. A conformational ensemble is built for the molecules in the database, whereas the conformation of the query remains fixed (the chemical nature of benzaldehyde does not entail different conformers, though in many cases the conformation of the query molecules might be complex and crucial). After the conformers of each molecule in the data set are built, each one of them is compared with the query and a similarity value is computed. For the particular program employed here (ROCS), the similarity is quantified as a score formed by two terms, one takes into account the chemical nature of the molecules while the other relies on molecular shape, such score is referred to as combo score. The maximum similarity value is 2 which can only be obtained from the comparison of a molecule with itself in the exact same conformation (perfect match). The normalized values (from 0 to 1) for the odor and combo score similarities are compared in the graph shown in Fig. 2.4. As can be observed, as the combo score increases, the odor similarity to benzaldehyde also increases. This correlation shows that part of the odor similarity was captured by the molecular

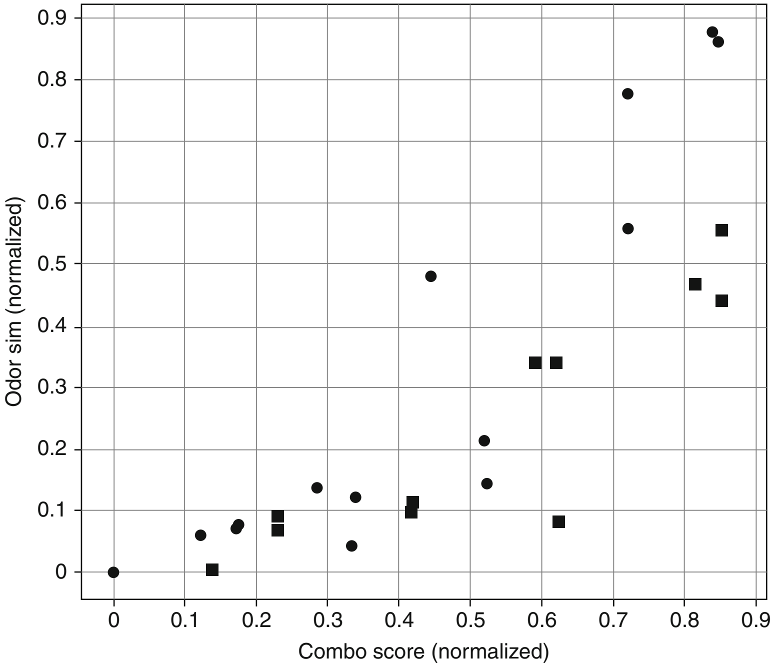


FIGURE 2.4 Odor–structure similarity plot. *X*-axis represents 3D shape similarity; *Y*-axis shows odor similarity to benzaldehyde. Shape by functional group: squares (CHO) and circles (NO).

similarity employed here. Other examples have been reported recently (Martinez-Mayorga *et al.*, 2008; Yongye *et al.*, 2009).

Limitations in the rational design of new odorants have been described (Sell, 2006, 2008). Moreover, minute structural differences resulting on significant odor changes have discouraged attempts at using SAR and rational design. In line with the "*activity cliff*" concept (Johnson and Maggiora, 1990) (*vide supra*), there are several examples of what can be called "*odor cliffs*", a collection of such examples was published by Sell (2006).

It is likely that activity landscapes are heterogeneous (Peltason and Bajorath), in other words, for the same system there are regions where smooth SAR can be found, but there are other regions with many activity cliffs. Even for one receptor, the presence of activity cliffs is not uncommon; however, they are rather important because they give information about regions in the receptor-binding pocket that are very selective. When more than one receptor is involved in the biological process corresponding to an expected property (as in the case of olfactory receptors), a much more complicated situation arises, where simultaneous activity landscapes are possible.

Concerning the selectivity of receptors, when they allow the binding of very different ligands they are regarded as *permissive* (Peltason and Bajorath, 2008). In those cases, ligands of different sizes, chemical nature, and conformation can be accommodated, showing that the binding site is flexible. When a receptor is able to carry only specific ligands, it is called *selective* and is expected to have multiple *activity cliffs*. Thus, smooth landscapes are the best scenario for the development of structure–activity relationships, not only for similarity searches but also for pharmacophore and QSAR models (Maggiora, 2006).

In many cases, multiple biochemical pathways are activated by the intake of therapeutical drugs. The activation of side mechanisms coincides with the consequent side effects. This concept has led to research areas like polypharmacology as well as re-purposing studies. In such cases, the possibility of activating of multiple biochemical pathways is employed to explore alternative therapeutic uses of the same compound. In addition to the ligand–receptor process, it is necessary to take into consideration the other mechanistic features that might also be involved and affect the analysis, such as the metabolism, transport based on carry proteins, co-factors, influence of metals, etc.

B. Pharmacophore model

In pharmacophore modeling, the aim is to obtain a hypothesis that best describes the chemical features and conformation of molecules responsible for biological activity. The features generally evaluated are hydrogen bond acceptors, hydrogen bond donors, aliphatic hydrophobic groups, negatively ionizable groups, positively ionizable groups, and aromatic hydrophobic groups.

In addition to the pharmacophoric features, current programs (Dixon *et al.*, 2006; Kirchmair *et al.*, 2007) allow the generation of excluded volumes that are meant to mimic regions occupied by the receptor. The pharmacophore generated can be then used for virtual screening of a compound collection. A detailed review with practical aspects and successful examples of pharmacophore-based virtual screening in the pharmaceutical field is described by van Drie (2003). An example of a pharmacophore model is presented in Fig. 2.5 which shows several pharmacophoric features for a methanesulfonamide ligand in the binding pocket of a kinase receptor, produced with LigandScout (Kirchmair *et al.*, 2007).

In addition to pharmaceutical applications, pharmacophore models have also been generated to analyze structure–odor relationships. In spite of complications arising from the need for more information on the mechanistic basis of olfaction, successful models have been developed

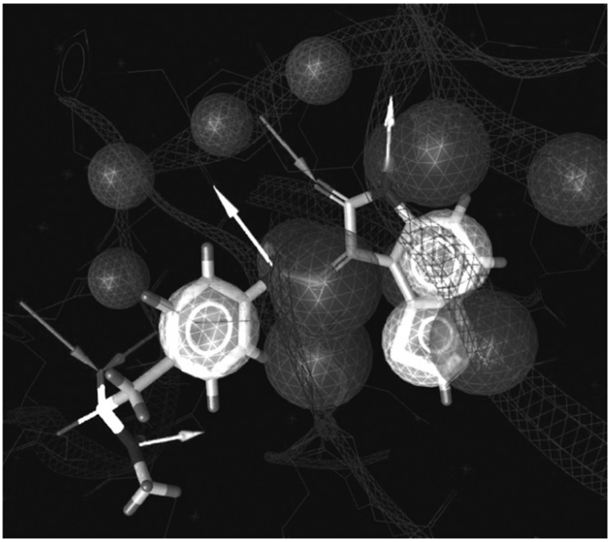


FIGURE 2.5 Pharmacophore model (produced using LigandScout) showing a methanesulfonamide ligand in the binding pocket of a cyclin-dependent kinase (PDB code: 1KE6). Pharmacophoric features: white arrow, hydrogen bond donor; gray arrow, hydrogen bond acceptor; white spheres, aromatic group; gray spheres, excluded volumes; protein is shown as mash ribbons.

(Kraft *et al.*, 2000). This particular type of pharmacophore has been termed an olpfactophore and a review by Kraft includes the development of olfactophores and SAR for the major odor notes of relevance in perfumery: ''fruity,'' ''marine,'' ''green,'' ''floral,'' ''spicy,'' ''woody,'' ''amber,'' and ''musky'' (Kraft *et al.*, 2000). In addition to perfumery, odors are also of great importance in the food industry, for example, in the development of flavors.

C. QSAR and QSPR

Among the computational methods available, QSARs, or more general, quantitative structure–*property* relationships (QSPR) have been widely used not only in drug design and environmental chemistry but also in food-related studies. QSPR studies are grounded in the concept that a property (e.g., biological activity, reactivity, toxicity, volatility, etc.) depends on the molecular structure and that is possible to find a mathematical or *quantitative* relationship between that property and a suitable molecular representation (e.g., some combination of descriptors).

Depending on the molecular representation used, several types of QSAR studies can be carried out, such as 2D- or 3D-QSAR studies, where molecules are represented using 2D or 3D descriptors, respectively (*vide supra*). The first investigations in QSAR aimed to identify linear relationships between the property and molecular descriptors, which were either calculated or measured experimentally. Current QSAR models can be quite complex.

One of the goals of QSAR studies is to help explain retrospectively the response or property of a molecule with a rationale based on molecular structure. A second major goal and challenge of QSAR or QSPR studies is to develop models that are able to predict quantitatively the property of new molecules either real or virtual compounds. Thus, successful predictive QSAR models can have a tremendous impact in the design of new molecules. Furthermore, predictive models are useful to perform *in silico* predictions of the properties of new structures. In virtual screening, those molecules that are predicted to have the desired property according to the QSAR model are selected as best candidates. Reviews, examples, caveats, and modified versions of QSAR are described elsewhere (Kubinyi, 1997a,b; Wermuth, 2008). Some recent examples reported in the food chemistry field are summarized in Table 2.4.

D. Molecular docking

Automated molecular docking aims to predict the best conformation of a molecule to fit into the target binding pocket (Kitchen *et al.*, 2004). The Protein Data Bank (Berman *et al.*, 2000) is a widely used source of target crystal structures as well as 3D coordinates of conformational ensembles obtained by NMR. In the absence of crystallographic structures, however, homology models can be used instead. Predicting the preferred conformation of a molecule with a target frequently comprises two major steps namely docking and scoring. Several docking programs are able to dock hundreds or even thousands of molecules relatively quickly so that virtual screening of large compound collections is feasible. Examples of docking programs are Autodock, (Morris *et al.*, 2009) Gold, Glide (Glide, 2008), and Fred (McGann *et al.*, 2003). For virtual screening proposes, Glide, for instance, has implemented a high-throughput docking algorithm that is designed for fast filtering. Then a more refined docking protocol can be used, to follow up.

One of the main areas of improvement in docking is the refinement of scoring functions to predict more accurately ligand–target interactions. In some cases biologically-relevant ‘‘poses’’ or docking solutions are found although they are not necessarily scored as the top ranked solutions.

Figure 2.6 shows the docking of a small molecule with the enzyme PTP-1B (PDB code 2F71) using the program Autodock.

A food-related example of molecular docking was published by Pripp (2007) on angiotensin converting enzyme inhibitory dipeptides. In an

TABLE 2.4 Recent examples of QSAR models in food chemistry

Computational method	Relevant property	Food relatedness	Title	References
Quantitative structure–activity relationships (QSAR)	Carcinogens	Food contact materials	Structure–activity relationship analysis tools: validation and applicability in predicting carcinogens	*Regul. Toxicol. Phaimacol.* 2008, 50, 50–58
Docking and virtual screening	Angiotensin converting enzyme inhibitors	Food derived peptides—milk	Docking and virtual screening of ACE inhibitory dipeptides	*Eur. Food Res. Technol.* 2007, 225, 589–592
Quantitative structure–property relationships (QSPR)	Aroma release	Yogurt	Effect of thickeners on aroma compound behavior in a model dairy gel	*J. Agric. Food Chem.* 2008, 55, 4835–4841
QSAR	Prediction of chemical residues	Tuna	A new risk framework for predicting chemical residue(s)—preliminary research for PCBs and PCDD/Fs in farmed Australian Southern Bluefin Tuna (*Thunnus maccoyli*)	*Chem. Eng. Process.* 2007, 46, 491–496
Molecular dynamics (MD), docking, QSAR	Mutagenic potency	Cooked meat	Mutagenic potency of food-derived heterocyclic amines	*Mutation. Res.* 2007, 616, 90–94

QSAR	Cognitive and neurological deficiencies	Milk	Quantitative structure–activity relationship of prolyl oligopeptidase inhibitory peptides derived from b-casein using simple amino acid descriptors	*J. Agric. Food Chem.* 2006, 54, 224–228
QSAR	Angiotensin converting enzyme inhibitors	Food derived peptides – Milk	Structural requirements of angiotensin I-converting enzyme inhibitory peptides: QSAR study of di- and tripeptides	*J. Agric. Food Chem.* 2006, 54, 732–738
QSAR	Antimicrobial, ACE-inhibitory and bitter tasting peptides	Several—review	Quantitative structure–activity relationship modeling of peptides and proteins as a tool in food science	*Trends Food Sci. Technol.* 2005, 16, 484–494
3D-QSAR (CoMFA)	Antibacterial	Food safety	3D-QSAR, synthesis, and antimicrobial activity of 1-alkylpyhdinium compounds as potential agents to improve food safety	*Euro J. Med. Chem.* 2005, 40, 840–849
QSAR	Bioaccumulation potential of organic chemicals in aquatic food webs	Food safety	A generic QSAR for assessing the bioaccumulation potential of organic chemicals in aquatic food webs	QSAR *Comb. Sci.* 2003, 22, 337–345

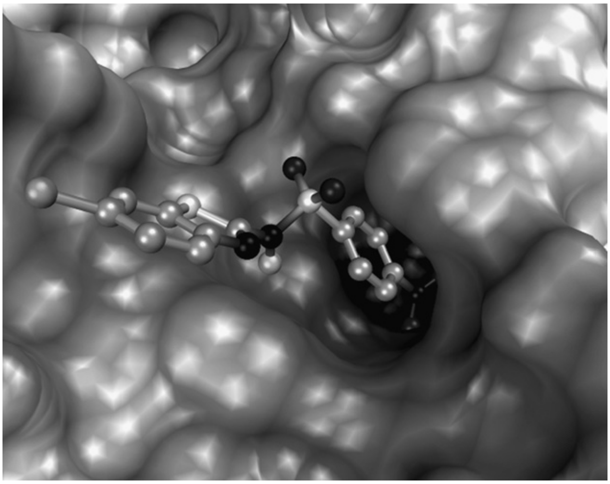


FIGURE 2.6 Docking of a small organic molecule into the catalytic site of PTP-1B (PDB code 2F71).

analysis of bioactive peptides derived from food proteins, he found that docking and virtual screening were feasible techniques to identify promising bioactive peptides. As pointed out in his paper, this approach does not fully replace experimental work, though it can contribute to the molecular understanding of bioactivity, and prioritization in virtual screening (Pripp, 2007).

VI. CONCLUDING REMARKS AND PERSPECTIVES

The ultimate goals of the methods described in Section IV fall in the understanding of structure–activity relationships and the discovery of new molecules. SAR is an inherent part of the derivation of the models, whereas the virtual screening of databases against the models not only serves as a means of validation but also provides candidates with a better chance (compared to random selection) of having the desired property.

The development of new descriptors, modified versions of established methods, or the creation of new approaches is constantly evolving in the scientific community, not only in the molecular modeling field but also in the analysis, mining, and processing of the chemical information.

All of the computational methods presented here have been used in connection with food or food-related components, and they can potentially be applied to other areas in food chemistry.

The better mechanistic pathways for the property response are understood, then the more suited methods will become in analyzing the very rich, diverse, and important body of information that comprises food chemistry.

ACKNOWLEDGMENTS

The authors are thankful to Dr. Terry Peppard (Robertet Flavors, Inc.) for valuable discussions and for proof reading this book chapter. This work was supported by the State of Florida, Executive Officer of the Governor's Office of Tourism Trade and Economic Development. We thank OpenEye Scientific Software for providing OMEGA, ROCS, FRED, and VIDA programs.

REFERENCES

Alvarez, J. and Shoichet, B. (2005). ''Virtual Screening in Drug Discovery''. Taylor & Francis Group, LLC CRC Press, Boca Raton, FL.

Amoore, J. E. (1963). Specific anosmia: A clue to the olfactory code. *Nature* **214**, 1095–1098.

Amoore, J. E. (1967). Specific anosmia: A clue to the olfactory code. *Nature* **214**, 1095–1098.

Amoore, J. E. (1971). Stereocehmical and vibrational theories of odour. *Nature* **233**, 270–271.

Amoore, J. E., Palmieri, G., and Wanke, E. (1967). Molecular shape and odor: Pattern analysis by PAPA. *Nature* **216**, 1084–1087.

Austin, C. P., Brady, L. S., Insel, T. R., and Collins, F. S. (2004). Molecular biology: NIH molecular libraries initiative. *Science* **306**, 1138–1139.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.

Bhattacharya, S., Hall, S. E., and Vaidehi, N. (2008). Agonist-induced conformational changes in bovine rhodopsin: Insight into activation of G-protein-coupled receptors. *J. Mol. Biol.* **382**, 539–555.

Brown, F. K. (1998). Chemoinformatics: What is it and how does it impact drug discovery? *Ann. Rep. Med. Chem.* **33**, 375–384.

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **65**, 175–187.

Chastrette, M. (1998). Data management in olfactation studies. *SAR QSAR Environ. Res.* **8**, 157–181.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G. F., Thian, F. S., Kobilka, T. S., Choi, H.-J., Kuhn, P., Weis, W. I., Kobilka, B. K., and Stevens, R. C. (2007). High-resolution crystal structure of an engineered human 2-adrenergic G protein coupled receptor. *Science* **318**, 1258–1265.

Cong, M., Perry, S. J., Hu, L. A., Hanson, P. I., Claing, A., and Lefkowitz, R. J. (2001). Binding of the beta2 adrenergic receptor to *N*-ethylmaleimide-sensitive factor regulates receptor recycling. *J. Biol. Chem.* **276**, 45145–45152.

Dixon, S. L., Smondyrev, A. M., Knoll, E. H., Rao, S. N., Shaw, D. E., and Friesner, R. A. (2006). PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput. Aid. Mol. Des.* **20**, 647–671.

Dobson, C. M. (2004). Chemical space and biology. *Nature* **432**, 824–828.

Dunkel, M., Schmidt, U., Struck, S., Berger, L., Gruening, B., Hossbach, J., Jaeger, I. S., Effmert, U., Piechulla, B., Eriksson, R., Knudsen, J., and Preissner, R. (2009). SuperScent—A database of flavors and scents. *Nucl. Acids Res.* **37**, pp. D291–D294.

Dziuba, J., Minkiewicz, P., Nalecz, D., and Iwaniak, A. (1999). Database of biologically active peptide sequences. *Nahrung* **43**, 190–195.

Glide (2008). Schrödinger, LLC, New York, NY, version 5.0.

Irwin, J. J. and Shoichet, B. K. (2005). ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182.

Johnson, M. and Maggiora, G. M. (1990). ''Concepts and Applications of Molecular Similarity''. Wiley, New York.

Kirchmair, J., Ristic, S., Eder, K., Markt, P., Wolber, G., Laggner, C., and Langer, T. (2007). Fast and efficient *in silico* 3D screening: Toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **47**, 2182–2196.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949.

Kraft, P., Bajgrowicz, J. A., Denis, C., and Frater, G. (2000). Odds and trends: Recent developments in the chemistry of odorants. *Angew. Chem. Int. Ed.* **39**, 2980–3010.

Kubinyi, H. (1997a). QSAR and 3D QSAR in drug design. 1. Methodology. *Drug Discov. Today* **2**, 457–467.

Kubinyi, H. (1997b). QSAR and 3D QSAR in drug design. 2. Applications and problems. *Drug Discov. Today* **2**, 538–546.

Leach, A. R. and Willet, V. J. (2003). ''An Introduction to Chemoinformatics''. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Maggiora, G. M. (2006). On outliers and activity cliffs—Why QSAR often disappoints. *J. Chem. Inf. Model.* **46**, 1535.

Martinez-Mayorga, K., Medina-Franco, J. L., Giulianotti, M. A., Pinilla, C., Dooley, C. T., Appel, J. R., and Houghten, R. (2008). Conformation-opioid activity relationships of bicyclic guanidines from 3D similarity analysis. *Bioorg. Med. Chem.* **16**, 5932–5938.

Masciocchi, J., Frau, G., Fanton, M., Sturlese, M., Floris, M., Pireddu, L., Palla, P., Cedrati, F., Rodriguez-Tome, P., and Moro, S. (2009). MMsINC: A large-scale chemoinformatics database. *Nucl. Acids Res.* **37**, D284–D290.

McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A., and Brown, F. K. (2003). Gaussian docking functions. *Biopolymers* **68**, 76–90.

Medina-Franco, J. L., Maggiora, G. M., Gulianotti, M. A., Pinilla, C., and Houghten, R. A. (2007). A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug. Des.* **70**, 393–412.

Medina-Franco, J. L., Martínez-Mayorga, K., Bender, A., Marín, R. M., Giulianotti, M. A., Pinilla, C., and Houghten, R. A. (2009). Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. *J. Chem. Inf. Model.* **49**, 477–491.

Medina-Franco, J. L., Martinez-Mayorga, K., Giulianotti, M. A., Houghten, R. A., and Pinilla, C. (2008). Visualization of the chemical space in drug discovery. *Curr. Comput. Aid. Drug Des.* **4**, 323–334.

Miller, M. A. (2002). Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* **1**, 220–227.

Moncrieff, R. W. (1967). ''The Chemical Sence''. CRC Press, Cleveland.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* Doi: 10.1002/jcc.21256.

Nakamichi, H. and Okada, T. (2006a). Chrystallographic analysis of primary visual photochemistry. *Angew. Chem. Int. Ed.* **45**, 4270–4273.

Nakamichi, H. and Okada, T. (2006b). Local peptide movement in the photoreaction intermediate of rhodopsin. *Proc. Natl. Acad. Sci. USA* **103**, 12729–12734.

Omega version 2.2.1 OpenEye Scientific Software: Santa Fe, NM, USA, www.eyesopen.com..

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Trong, I. L., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M., and Miyano, M. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**, 739–745.

Paris, G. (2000). ''Balancing the Needs of the Recruiters and the Aims of the Educators''. http://www.warr.com.

Peltason, L. and Bajorath, J. (2007). Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chem. Biol.* **14**, 489–497.

Peltason, L. and Bajorath, J. (2008). Molecular similarity analysis in virtual screening. *In* ''Chemoinformatics Approaches to Virtual Screening'', (A. Varnek and A. Tropsha, eds), pp. 120–147. RSC Publishing, Cambridge, UK.

Pripp, A. H. (2007). Docking and virtual screening of ACE inhibitory dipeptides. *Eur. Food. Res. Technol.* **225**, 589–592.

Rasmussen, S. G. F., Choi, H.-J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R. P., Sanishvili, R., Fischetti, R. F., Schertler, G. F. X., Weis, W. I., *et al*. (2007). Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383–387.

ROCS version 2.3.1 OpenEye Scientific Software: Santa Fe, NM, USA, www.eyesopen.com.

Rosenbaum, D. M., Cherezov, V., Hanson, M. A., Rasmussen, S. G. F., Thian, F. S., Kobilka, T. S., Choi, H.-J., Yao, X.-J., Weis, W. I., Stevens, R. C., and Kobilka, B. K. (2007). GPCR engineering yields high-resolution structural insights into beta2 adrenergic receptor function. *Science* **318**, 1266–1273.

Ruprecht, J. J., Mielke, T., Vogel, R., Villa, C., and Schertler, G. F. X. (2004). Electron crystallography reveals the structure of metarhodopsin I. *EMBO J.* **23**, 3609–3620.

Salom, D., Lodowski, D. T., Stenkamp, R. E., Trong, I. L., Golczak, M., Jastrzebska, B., Harris, T., Ballesteros, J. A., and Palczewski, K. (2006). Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *Proc. Natl. Acad. Sci. USA* **103**, 16123–16128.

Scior, T., Bernard, P., Medina-Franco, J. L., and Maggiora, G. M. (2007). Large compound databases for structure-activity relationships studies in drug discovery. *Mini-Rev. Med. Chem.* **7**, 851–860.

Sell, C. S. (2006). On the unpredictability of odor. *Angew. Chem. Int. Ed.* **45**, 6254–6261.

Sell, C. S. (2008). Rational odorant design: Fantasy or feasibility? *Perfum. Flavor.* **33**, 48–52.

Singh, N., Guha, R., Giulianotti, M. A., Pinilla, C., Houghten, R. A., and Medina-Franco, J. L. (2009). Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.* **49**, 1010–1024.

Sprous, D. G. and Salemme, F. R. (2007). A comparison of the chemical properties of drugs and FEMA/FDA notified GRAS chemical compounds used in the food industry. *Food Chem. Toxicol.* **45**, 1419–1427.

Tikhonova, I. G., Best, R. B., Engel, S., Gershengorn, M. C., Hummer, G., and Costanzi, S. (2008). Atomistic insights into rhodopsin activation from a dynamic model. *J. Am. Chem. Soc.* **130**, 10141–10149.

van Drie, J. H. (2003). Pharmacophore discovery—Lessons learned. *Curr. Pharm. Desgn.* **9**, 1649.

Varnek, A. and Tropsha, A. (2008). ''Chemoinformatics Approaches to Virtual Screening''. RSC Publishing, Cambridge, UK.

Wermuth, C. G. (2008). The Practice of Medicinal Chemistry. Elsevier, Academic Press, London, UK.

Wild, D. J. and Wiggings, G. D. (2006). Challenges for chemoinformatics education in drug discovery. *Drug Discov. Today* **11**, 436–439.

Willett, P. (2006a). Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* **25**, 1143–1152.

Willett, P. (2006b). Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053.

Willett, P., Barnard, J. M., and Downs, G. M. (1998). Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucl. Acids Res.* **36**, D901–D906.

Yongye, A. B., Appel, J. R., Giulianotti, M., Dooley, C. T., Medina-Franco, J. L., Nefzi, A., Houghten, R., and Martinez-Mayorga, K. (2009). Identification, binding mode prediction and structure-activity relationships of potent triamine and piperazine opioid ligands. *Bioorg. Med. Chem.* **17**, 5583–5597.

Zarzo, M. (2007). The sense of smell: Molecular basis of odorant recognition. *Biol. Rev.* **82**, 455–479.

Zozulya, S., Echeverri, F., and Nguyen, T. (2001). The human olfactory receptor repertoire. *Genome Biol.* **2**, 18.11–18.12.