ELSEVIER

Multi-dimensional QSAR in drug discovery

Markus A. Lill

Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN 47907, USA

Quantitative structure–activity relationships (QSAR) is an area of computational research that builds virtual models to predict quantities such as the binding affinity or the toxic potential of existing or hypothetical molecules. Although a wealth of experimental data emphasizes the active role of the target protein in the binding process, QSAR studies are frequently restricted to the properties of the small-molecule ligand. This review aims at discussing recent QSAR concepts exploring higher dimensions (simulation of induced fit, simultaneous exploration of alternative binding modes, and solvation scenarios), and their benefit for the drug-discovery process.

Introduction

Over a century after Fischer's formulation of the lock-and-key analogy [1] and 40 years after the seminal contributions of Hansch, Fujita, Free and Wilson [2,3], quantitative structure–activity relationships (QSAR) have matured into a widely used tool, substantially contributing to the drug-discovery process. Originally based on the idea that compounds with similar physico-chemical properties trigger similar biological effects, QSAR are often employed to establish a correlation between structural and electronic properties of potential drug candidates and their binding affinity towards a common macromolecular target.

In drug discovery, QSAR are widely used to identify ligands with high affinity for a given macromolecular target. More recently, the technology has been extended to predict adsorption, distribution, metabolism, elimination, toxicity (ADMET) properties [4] or the oral bioavailability of compounds [5,6]. In the context of the Registration, Evaluation and Authorization of Chemicals (REACH) legislation of the European Union, the prediction of the toxic potential of a drug or environmental chemical using QSAR has spawned much interest [7].

While early QSAR studies were typically based on a single physico-chemical property, such as the solubility or the pK_a value, to explain the biological effect of a molecule (1D-QSAR, see Table 1), Hansch, Fujita, Free and Wilson implicitly included the connectivity of a compound by considering physico-chemical properties of single atoms and functional groups and their contribution to biological activity (2D-QSAR). Nowadays, Hansch-Fujita like QSAR models can also contain 3D-structural descriptors such as the length or width of a substituent.

With an increasing number of three-dimensional structures of proteins becoming available from X-ray diffraction studies, structure-based design (SBD) – the identification of a small-molecule ligand with high affinity by tailoring its structure to the topology of a macromolecular binding pocket using interactive or nowadays automated docking combined with molecular-mechanics (MM) and dynamics optimizations (MD) – surfaced as a promising tool in the 1980s. With an appropriately parameterized force field or empirical scoring function, it became possible to identify the most probable binding mode of any given existing or hypothetical molecule to a target protein. Unfortunately, the quantification of the resulting protein–ligand interactions towards the estimation of binding affinity turned out to be a demanding task. On the one hand, the wealth of devised scoring functions can yield semi-quantitative values at best; free-energy perturbation (FEP) techniques, on the other, are limited to the comparison of similar, structurally related molecules.

The introduction of comparative molecular field analysis (CoMFA) [8] in 1988 represents another milestone in QSAR as, for the first time, such structure–activity relationships were based on the three-dimensional structure of the ligand molecules (3D-QSAR). In 3D-QSAR the ligands' interaction with chemical probes is mapped onto a surface or grid surrounding a series of com-

Corresponding author: Lill, M.A. (mlill@purdue.edu)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.08.004

TABLE 1

Classification of QSAR approaches based on their dimensionality

Dimension	Method	Protein	Refs
1D-QSAR	Affinity is correlated with global molecular properties of ligands, that is one value per property and ligand (pK_a, log *P*, etc.)	No	[33]
2D-QSAR	Affinity is correlated with structural patterns (connectivity, 2D pharmacophore, etc.) without consideration of an explicit 3D representation of these properties	No	[2,3]
3D-QSAR	Affinity is correlated with the three-dimensional structure of the ligands	Possible	[8]
4D-QSAR	Ligands are represented as an ensemble of configurations	Possible	[9–12]
5D-QSAR	As 4D-QSAR + explicit representation of different induced-fit models	Yes	[18,20,22]
6D-QSAR	As 5D-QSAR + representation of different solvation scenarios	Yes	[19]

pounds (superimposed in 3D space). This surface or grid represents a surrogate of the binding site of the true biological receptor. The quality of the QSAR model depends critically on the correct superposition of the ligands, the identification of which is almost impossible in the absence of structural information for the target protein. While this problem has long been recognized, only recently developed 4D-QSAR technologies would seem to provide decent solutions [9–12].

The determination of binding energies in QSAR studies is by no means simple. Free energies of binding depend on the ligand–protein interactions as well as on the loss of energy associated with stripping solvent molecules off the small-molecule ligand while moving from the aqueous environment of a cell or a body fluid to a protein binding pocket during the binding process. The loss of conformational entropy upon binding also may contribute significantly to the observed affinity. In addition, accommodation of ligand molecules in a macromolecular binding pocket is facilitated by induced fit, that is the adaptation of the protein to the ligand topology. Induced fit may not only alter the topology of the binding pocket, but also its character (hydrophobic or hydrophilic, dielectric properties) of subsites or solvent accessibility (Figure 1) [13]. Multidimensional QSAR (mQSAR) concepts aim to quantify all these additional contributions to the binding energy. However, a careful consideration of these additional parameters is an absolute necessity, as with a sufficient number of parameters any quantity can be explained—but what would the predictive power of such a model be?

Multiple ligand representation (4D-QSAR)

In 3D-QSAR, bioactive conformation and relative orientations of the ligand molecules must be unambiguously identified in order to derive a predictive model. Particularly, in the absence of structural information on the target protein, the identification of both bioactive conformation and orientation of the ligand molecule is all but obvious. If the 3D structure of the macromolecular target is known, it can be used for this very purpose, for example using docking. Even then, an unambiguous identification of conformation and orientation is frequently difficult (Figure 2). Obviously, a 3D-QSAR model based on ligand conformations and orientations deviating from the bioactive ones may hardly be of any use for predicting purposes.

QSAR based on 'alignment-independent descriptors' (AIDs) [14–16] were proposed to elegantly dodge the alignment issue. Three-dimensional properties of compounds, such as the hydrophobic property projected on its molecular surface, are transformed into position-independent characteristics, such as the terms of a moment expansion of the pysico-chemical fields of a compound. The selection of the conformations is, however, likewise critical for

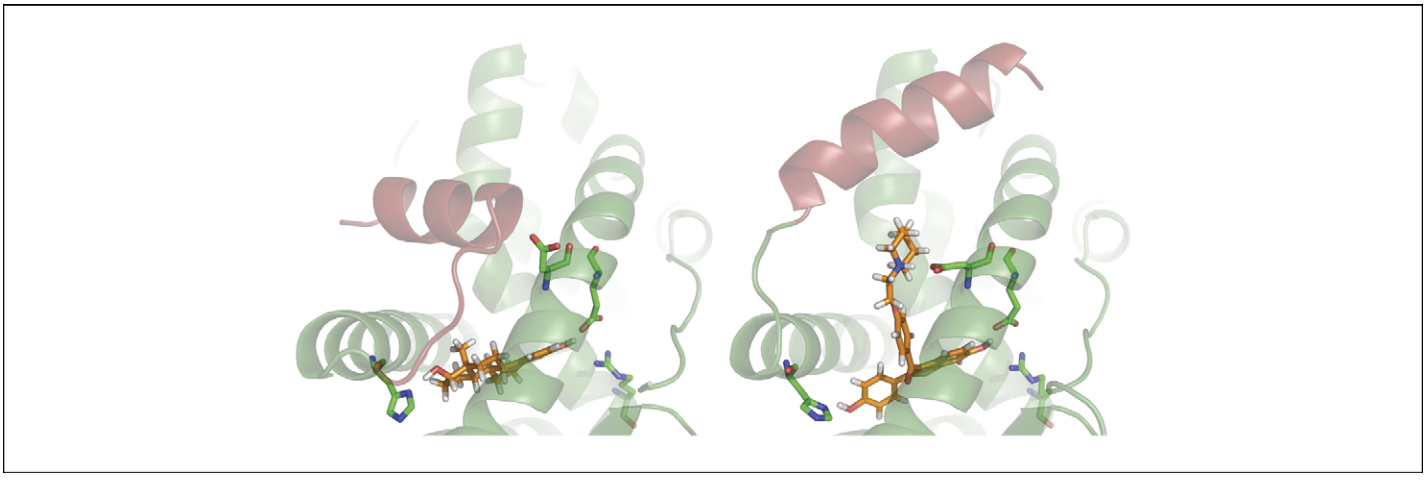

FIGURE 1

Role of induced fit for ligand binding to proteins. Structures of agonist 17β-estradiol (left) and selective-estrogen modulator (SERM) Raloxifene (right) bound to human estrogen receptor were solved by X-ray diffraction. In addition to significant topological changes of helix 12 (red), accommodating either agonist or SERM, Asp 351 is rotated towards the protonated piperidyl N-atom of Raloxifene, forming a salt bridge. Consequently, both hydrophobic field and hydrogen bond propensity spawned by the binding site are altered. Figure was created with PyMol [34].

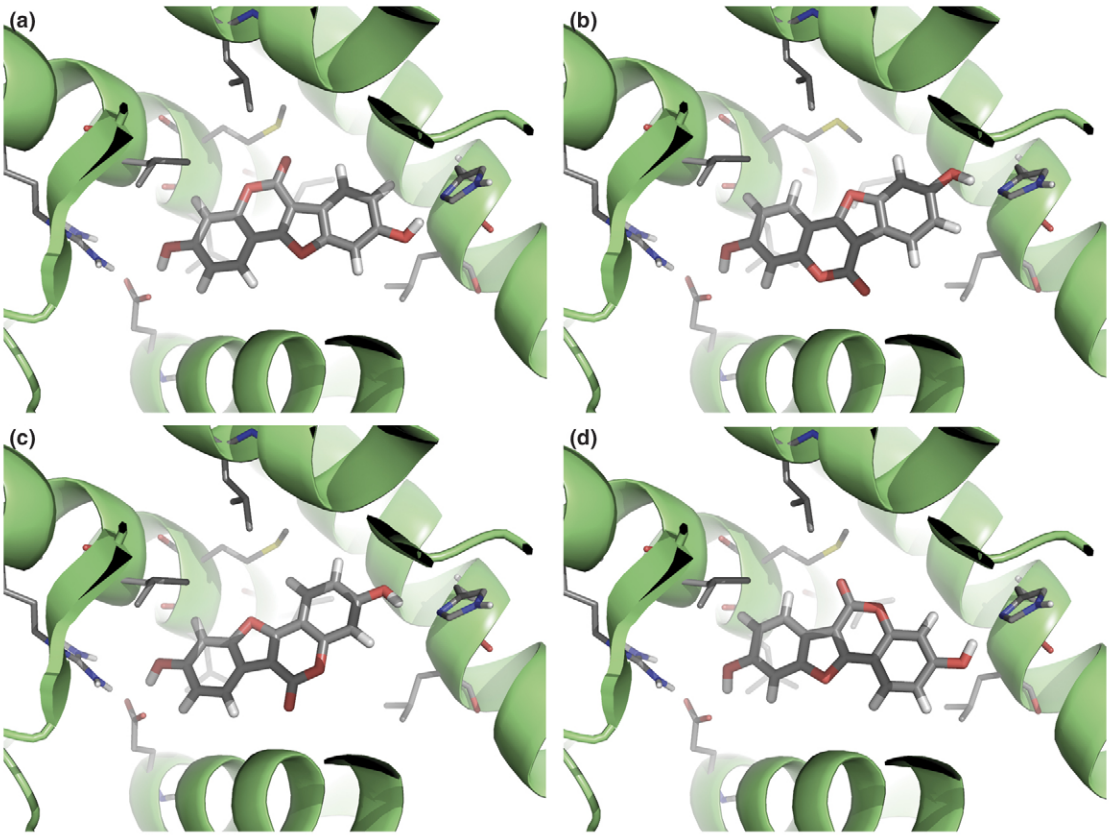

FIGURE 2

Difficulties in identifying the energetically best binding mode. Four different but energetically similar binding modes of Coumestrol binding to human ER_{α} were identified by docking and Monte-Carlo simulations using Yeti [35]. Figure was created with PyMol [34].

the quality of a QSAR model based on such descriptors. Alternatively, 4D-QSAR concepts [9–12] approach the alignment issue by incorporating molecular and spatial variety by representing each molecule in different conformations, orientations, tautomers, stereoisomers or protonation states. The true binding mode (or the bioactive conformation) is then identified by the algorithm underlying the QSAR concept.

In general, two different types of 4D concepts have been developed: One class [9,17] of QSAR makes use of a large ensemble of structurally similar conformations (typically 1000 or more). In 3D-QSAR such as CoMFA even small conformational changes in the ligand can have a profound impact on the ligand-probe interactions and consequently on the results of the QSAR simulation. Sampling the conformational space around the overall binding mode reduces the sharp dependency of the QSAR results on the chosen ligand configuration. This type of 4D-QSAR approach, however, seems not to be capable of dealing with uncertainties in the overall binding mode of a compound. In medicinal-chemistry studies involving non-congeneric series of compounds (e.g. nuclear receptors), in projects where structural information about the protein is not available (e.g. GPCRs), or in toxicity testing of pharmaceutical and environmental chemicals involving promiscuous proteins (e.g. cytochrome P450 enzymes), the identification of the binding mode of a molecule is usually all but unambiguous.

In the second class of 4D-QSAR approaches [12,18–21], a small set of diverse ligand configurations represents independent alternatives for the QSAR modeling. Distinct binding modes with significant root-mean square-deviation and alternative alignment protocols may be explored this way. This 'on-the-fly' generation of conformational alternatives within a single simulation has clear advantages over a serial brute-force examination of alternate binding modes of all ligands in classical 3D-QSAR due to a combinatorial explosion in the number of required QSAR simulations. The ensemble can include different conformations, orientations, tautomers, stereoisomers and protonation states. The underlying algorithm usually selects the configuration with the highest interaction energy to the binding surrogate or a combination thereof (Boltzmann weighted ensemble). As the bioactive entity thereof is identified by the underlying optimization algorithm, the approach can reduce the bias associated with the selection of the bioactive conformer and ligand alignment. In particular, 4D-QSAR can play an important role in identifying the most probable tautomeric form [11], as even X-ray crystallography is often not able to unambiguously determine the protonation state.

4D-QSAR can be interpreted as a feasible extension of 3D-QSAR to address the uncertainties during the alignment process. In addition, it can have fundamental biological relevance, when dealing with multi-mode binding targets. Cytochrome P450 enzymes, for example, are known to accommodate a ligand in

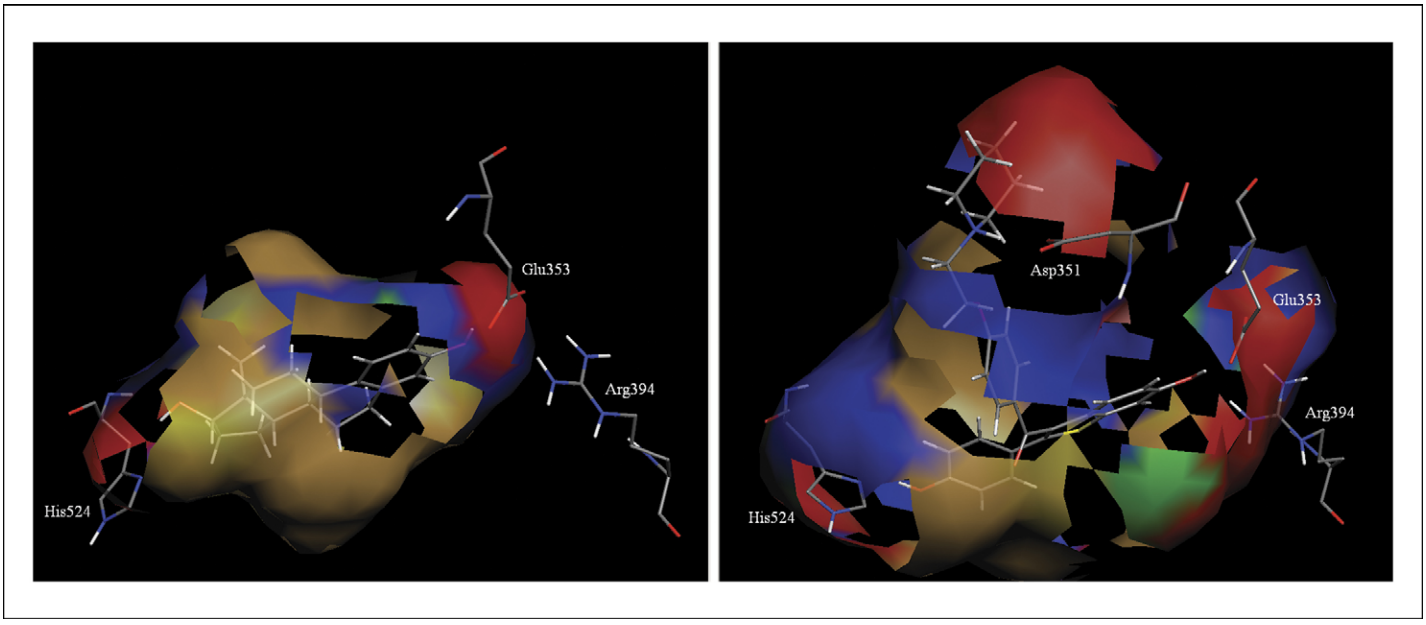

FIGURE 3

Explicit simulation of induced-protein fit in QSAR. Explicit simulation of induced fit by a dual-shell representation of the three-dimensional binding-site model: Different physico-chemical properties are distributed on inner shell (relevant for agonist 17β-estradiol, left) and outer shell (relevant for SERMs like Raloxifene, right). Compare with experimental structure in Figure 1. Coloring: hydrogen-bond accepting character in red, donor in blue and hydrophobic fields in beige. Figure created with Raptor [20].

various binding poses, yielding different metabolic products of a given compound. 4D-QSAR technologies can explicitly account for different ligand configurations in a single simulation. Recently, this has been successfully applied to simulate binding of structurally diverse compounds to cytochrome P450 3A4, representing each small molecule with on average four different binding poses identified by an automated docking procedure [22].

Induced-fit modeling (5D)

The necessity to account for protein flexibility [23] in computer-aided drug design concepts has been recently emphasized in the context of novel structure-based methods [24–28]; flexible-protein docking – that is allowing for the flexibility of the binding pocket, while docking a small-molecule ligand – is nowadays considered state-of-the-art.

The adaptation of this philosophy to the area of QSAR is still in its infancy. To simulate induced fit in an explicit manner, simulation of a topological adaptation of the model of the binding-site surface to the individual ligand molecules has been devised [18]. Herein, the surface of the binding-site model can slightly shrink or expand dependent on the size and topology of the ligand binding to it. As the identification of the correct magnitude and mechanism of induced fit is not possible in absence of the structure of true target protein, different induced-fit protocols (e.g. magnitude dependent on steric, electrostatic, hydrogen-bond or lipophilic potential) are presented as alternative scenarios (5D-QSAR) to the QSAR. An energy penalty is included proportional to the magnitude of induced fit. Furthermore, hydrogen-bonding attributes (donor or acceptor) on the receptor model are allowed to flip, dependent on the hydrogen-bonding properties of the ligand molecule adjacent to it.

Compounds may bind to different sub-pockets of the binding site as a consequence of induced fit and, hence, experience different interaction fields of the protein. Recently, a method has been developed that is able to anisotropically simulate induced fit (Figure 3) [20]. Two spatially separated shells allow both the simulation of local protein adaptation as well as large conformation changes such as those involved in agonism and antagonism in nuclear receptors. Variations in the distribution of properties between inner and outer shells represent the different physico-chemical nature of different subpockets resulting from induced-fit protein motions. This enables the simultaneous simulation of agonists and antagonists, a situation in which larger conformational changes can be expected than in studies of the agonist state only. A prerequisite of such studies is however the existence of a significant overlap in the binding modes and a coherent mechanism of antagonism among the ligands.

A wealth of structural data, for example for neuroaminidase [29], has shown that hydrogen-bond attributes of amino acids sculpting the binding pocket can be either involved in intramolecular or intermolecular hydrogen bonds with the ligand. Consequently, both the interacting fields and the topology of the binding site can change for each individual ligand. In a binding-site surrogate this can be achieved by combining a steric adjustment to the topology of the ligand and a component modifying the physico-chemical properties on the surrogate due to attraction or repulsion between ligand and receptor model.

Conclusions

Considering the importance of a correct ligand alignment as input for 3D-QSAR as well as the often significant human bias possibly associated with this step, mQSAR approaches provide a promising alternative to classic 3D-QSAR for drug-discovery purposes. Even if the drug target's three-dimensional structure is known to atomic resolution, a decent selection of binding poses resulting from a docking study reflects a smaller human bias than just using a single

docking pose as input for QSAR. Thus, mQSAR would seem to be best suited for quantifying binding affinities in combination with ligand docking to a macromolecular structure [30]. Furthermore, the quantification of entropy is an important area of ongoing research. While current approaches focus on ligands configurational entropy [31], the protein's contribution to this quantity should be addressed in future approaches [32]. As a wealth of experimental data suggests the relevance of protein flexibility upon ligand binding, induced fit should be employed to identify the binding mode, for example during docking, as well as to quantify protein–ligand interactions using QSAR. The energetic quantification of induced fit, however, remains a current challenge.

In summary, mQSAR methods promise to be valuable extensions of classical 3D-QSAR, both conceptually and often with respect to reliability in predicting affinities of new chemical entities. However, it is an absolute necessity to carefully include these additional dimensions in QSAR, as mQSAR could increase the ratio of degrees-of-freedom to the number of biological data points. The use of Boltzmann weights, or specifying domains (neighboring surface points of the model are constraint to similar properties) are just a few examples for proper treatment of new dimensions. An increase in the number of degrees-of-freedom is however not even a necessary consequence of mQSAR, as moving from a representation of the binding site on a 3D grid (e.g. in CoMFA) to a surface representation (e.g. [19,20]), for example, could even reduce this number. Furthermore, extensive challenging of any established mQSAR model by employing internal and external test set, applying a scramble test and by means of consensus scoring, the use of different scoring functions to estimate the binding affinity, leads to more reliable models.

References

1 Fischer, E. (1894) Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* 27, 2984–2993
2 Hansch, C. *et al.* (1964) *p*–σ–π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* 86, 1616–1626
3 Free, S.M. *et al.* (1964) A mathematical contribution to structure–activity studies. *J. Med. Chem.* 7, 395–399
4 Norinder, U. (2005) *In silico* modelling of ADMET—a mini review of work from 2000 to 2004. *SAR QSAR Environ. Res.* 16, 1–11
5 Martin, Y.C. (2005) A bioavailability score. *J. Med. Chem.* 48, 3164–3170
6 Yoshida, F. *et al.* (2000) QSAR model for drug human oral bioavailability. *J. Med. Chem.* 43, 2575–2585
7 Worth, A.P. *et al.* (2007) The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods. *SAR QSAR Environ. Res.* 18, 111–125
8 Cramer, R.D. *et al.* (1988) Comparative molecular-field analysis (CoMFA) .1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 110, 5959–5967
9 Hopfinger, A.J. *et al.* (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* 119, 10509–10524
10 Ekins, S. *et al.* (2000) Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metab. Dispos.* 28, 994–1002
11 Vedani, A. *et al.* (2000) Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *J. Med. Chem.* 43, 4416–4427
12 Lukacova, V. *et al.* (2003) Multimode ligand binding in receptor site modeling: implementation in CoMFA. *J. Chem. Inf. Comput. Sci.* 43, 2093–2105
13 Brzozowski, A.M. *et al.* (1997) Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* 389, 753–758
14 Todeschini, R. *et al.* (1994) New molecular descriptors for 2D and 3D structures. *Theory. J. Chemometrics* 8, 263–272
15 Bravi, G. *et al.* (1997) MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. *J. Comput. Aided Mol. Des* 11, 79–92
16 Silverman, B.D. *et al.* (1996) Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* 39, 2129–2140
17 Pan, D. *et al.* (2003) Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* 43, 1591–1607
18 Vedani, A. *et al.* (2002) 5D-QSAR: the key for simulating induced fit? *J. Med. Chem.* 45, 2139–2149
19 Vedani, A. *et al.* (2005) Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* 48, 3700–3703
20 Lill, M.A. *et al.* (2004) Raptor: combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling: application toward the simulation of structurally diverse ligand sets. *J. Med. Chem.* 47, 6174–6186
21 Lill, M.A. *et al.* (2006) Combining 4D pharmacophore generation and multidimensional QSAR: modeling ligand binding to the bradykinin B2 receptor. *J. Chem. Inf. Model.* 46, 2135–2145
22 Lill, M.A. *et al.* (2006) Prediction of small-molecule binding to cytochrome P450 3A4: flexible docking combined with multidimensional QSAR. *Chem. Med. Chem.* 1, 73–81
23 Teague, S.J. (2003) Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* 2, 527–541
24 Carlson, H.A. (2002) Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* 6, 447–452
25 Zavodszky, M.I. *et al.* (2004) Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins* 57, 243–261
26 Sousa, S.F. *et al.* (2006) Protein–ligand docking: current status and future challenges. *Proteins* 65, 15–26
27 Mizutani, M.Y. *et al.* (2006) Effective handling of induced-fit motion in flexible docking. *Proteins* 63, 878–891
28 Sherman, W. *et al.* (2006) Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* 49, 534–553
29 Taylor, N.R. *et al.* (1998) Dihydropyrancarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. *J. Med. Chem.* 41, 798–807
30 Vedani, A. *et al.* (2007) Simulating alpha/beta Selectivity at the Human Thyroid Hormone Receptor: Consensus Scoring Using Multidimensional QSAR. *Chem. Med. Chem.* 2, 78–87
31 Dervarics, M. *et al.* (2006) Development of a chirality-sensitive flexibility descriptor for 3 + 3D-QSAR. *J. Chem. Inf. Model.* 46, 1431–1438
32 Homans, S.W. (2005) Probing the binding entropy of ligand–protein interactions by NMR. *Chem. Bio. Chem.* 6, 1585–1591
33 Overton, C.E. (1901) *Studien über die Narkose.* Gustav Fischer Verlag, Jena, Germany
34 DeLano, W.L. (2002) *The PyMOL Molecular Graphics System.* DeLano Scientific, Palo Alto, CA, USA
35 Vedani, A. *et al.* (1990) A new force field for modeling metalloproteins. *J. Am. Chem. Soc.* 112, 4759–4767