J Comput Aided Mol Des (2009) 23:831–836
DOI 10.1007/s10822-009-9310-3

WARR'S PIECE

Herman Skolnik award symposium honoring Yvonne Martin

Wendy A. Warr

Received: 27 October 2009 / Accepted: 17 November 2009 / Published online: 10 December 2009

Yvonne chose 11 speakers, including many of her colleagues from Abbott. The first talks were devoted to those who had the vision to collect and curate structural data. ACS has mounted these and some of the other presentations online as PowerPoint slides synched with audio [1]. First off was Frank Allen of the Cambridge Crystallographic Data Centre (CCDC) who discussed how small-molecule crystal structure information aids drug discovery and development. He noted that Stahl and co-workers had recently presented a thorough review [2] of the applications of small-molecule crystal conformations in drug discovery. Despite the fact that higher-energy conformations can occur, e.g., the presence of planar biphenyl conformers that are about 6 kJ/mol above the minimum energy twisted form [3], or small conformational variations due to favorable intermolecular interactions, the review showed that crystal conformations are very valuable experimental information in a modeling environment.

Frank's team had also investigated whether higher-energy conformers were a common occurrence in crystal structures [4]. They showed that torsion angles associated with higher strain energy (>5 kJ/mol) appear to be very unusual and, in general, high-energy conformers may well be under-represented in crystal structures when compared with a gas-phase, room-temperature Boltzmann distribution. Since crystallographically determined bond lengths, angles and torsions from the Cambridge Structural Database (CSD) are so useful, CCDC has produced a knowledge base, Mogul, to make them more readily accessible [5]. Validation experiments have shown that, with rare exceptions, such as those noted above, search results from Mogul give precise experimental information on molecular geometrical preferences.

Crystal structures are also the principal source of information on intermolecular interactions such as hydrogen bonding, dipole–dipole, and halogen–halogen interactions. CCDC's knowledge base IsoStar [6] incorporating crystal structure data from the CSD and the Protein Data Bank (PDB), and interaction energies obtained using ab initio intermolecular perturbation theory (IMPT), enables users to visualize and study a wide range of interactions that are important in protein–ligand docking. Recent work at CCDC has included a thorough analysis of the energetic effects of varying the angle at the donor hydrogen in hydrogen bonds [7], together with combined CSD/IMPT studies of some less common interactions [8–11]. Frank cited a valuable review of interactions that are not mediated by hydrogen [12].

Finally Frank discussed the use of crystal structure information in the development, formulation and delivery of active pharmaceutical ingredients (APIs). An unexpected polymorph can have severe financial implications as exemplified in 1998, when supplies of Norvir capsules were suspended, when a new, much less soluble crystal form of the API ritonavir [13] suddenly appeared in production. CCDC has developed a method, based on a statistical analysis of hydrogen bonds in the CSD, to calculate the propensity for formation of different types of hydrogen bonds, given just the chemical diagram of a potential API. Applied now to ritonavir, the logit hydrogen-bonding propensity (LHP) model [14–16] showed that the original polymorphic form I of ritonavir was probably a metastable form, and that a more stable form was highly likely, as proved to be the case. In conclusion, Frank noted that "crystals are windows on the world of atoms" [17], and

W. A. Warr (✉)
Wendy Warr & Associates, 6, Berwick Court, Holmes Chapel, Cheshire CW4 7HZ, UK
e-mail: wendy@warr.com

that the factual content of a database such as the CSD can be converted into information-rich, structural knowledge through diverse software applications.

The PDB also has lessons for us; Helen Berman discussed some of them, and briefly introduced the Protein Structure Initiative (PSI) Structural Genomics Knowledgebase (SGKB). She has recounted the history of PDB in a recent article [18]. The first crystal structures of proteins started to appear in the 1960s; PDB began with just seven structures in 1971 at Brookhaven National Laboratory. In the early days enzymes predominated; the first RNA-containing structures started to appear in 1973. In the 1980s technology started to take off: cloning of genes, expression of proteins, synchrotrons, visualization on computers, and programs for electron density fitting. DNA-containing structures, and protein–nucleic complexes, and then viruses were registered. There was still a problem in persuading people to put data into PDB: despite all the lobbying and determined committee efforts, it took 7 years before data deposition became mandatory for NIH funding.

In the 1990s the macromolecular Crystallographic Information File (mmCIF) format was created as a standard for the representation of structural data. During the same time period, the PDB user base was dramatically expanding. PDB management was moved over to Helen's team at the Research Collaboratory for Structural Bioinformatics. Nevertheless, Helen emphasized throughout her talk that PDB is a *global* resource. The worldwide PDB collaboration was formalized in the 2000s and PDB Europe and PDB Japan were established. The content of the PDB is getting larger and more complex thanks to advances in technology and development of new techniques such as cryo-electron microscopy. In total there are now about 60,000 structures of ever increasing complexity, yet the budget is the same as it was in 2003 when there were fewer than 24,000 structures.

There are formidable challenges in representing the increasing amounts and complexity of the data, making a searchable database and integrating it with other resources, building a scalable system, and meeting the needs of a diverse community. Since, applications are only as good as the quality of the data, all the data have been examined and re-annotated. Just one example is symmetry and coordinate transformations for virus entries [19].

The pipeline for structure determination (structure determination, and data deposition, processing, archiving, distribution and query) has lots of data associated with it; SGKB provides resources for analyses. A searchable database delivers aggregate reports and inventories. There are links to PSI projects and external resources, and to a central community-nominated targets proposal system. The structural genomics gateway with *Nature* delivers research findings, news and events. New editorial content and recently solved structures are also publicized.

The SGKB enables knowledge by connecting sequence information to 3D structures and homology models, providing access to experimental protocols, materials and technologies, and fostering community collaboration. Data sharing leads to new knowledge. Sequence-structure-function relationships are complex. Cases of low sequence identity and the same structure are exemplified by hemoglobin; TIM barrel proteins have the same structure but different functions; lysozymes have different overall structure but the same function. A structural view of biology is closer than we thought.

The next speaker, Jim Dunbar reported on Community Structure-Activity Resource (CSAR), a data repository being built at the University of Michigan [20], to improve docking and scoring through participation of the scientific community. CSAR will disseminate experimental data sets of crystal structures and binding affinities for diverse protein–ligand complexes. Some data sets will be generated in house at Michigan while others will be deposited by external organizations or collected from the literature. Industrial data are especially needed: "give us your dead projects", pleaded Jim.

The Michigan researchers aim to provide high quality data for a diverse collection of proteins and small molecule ligands. Ideal targets will have many high-quality crystal structures (apo structures for 10–20 targets bound to diverse ligands) and affinity data for 25 compounds or more that range in size, scaffold, and log *P*. It is best if the ligand set has several congeneric series that span a broad range of affinity. K_d data is preferred over K_i data, and K_i data is preferred over IC_{50} data. Data for isomerases, ligases, lyases, and non-enzymes are highly desirable, because these data are not well represented in PDB. Until there are depositions, the CSAR center is working on CDK2, CDK2/cyclin A, and LpxC in-house, and on some other targets in collaborations. In 2010, a benchmark exercise using two high quality data sets will examine performance across diverse proteins, will document which systems are most difficult, and will call for new methodology. Details of the data sets and the benchmark exercise are given on the CSAR website [20].

Next Al Leo recounted some of the history that led to Hansch's team discovering that log P_{oct} could be used to extend the Hammett method to bio-reactions [21] and the subsequent measuring, then prediction of log *P*s, and the building of QSARs. The Pomona College MedChem project was founded in 1970 and the Hansch-Fujita team drew post-docs from the United States, Asia and New Zealand. Yvonne herself did a sabbatical at Pomona. The project continued as BioByte.

Biobyte stores more than 13,000 QSAR equations and a Masterfile which contains about 71,000 structures, with searchable biological activity types, and log P and pK_a values. The QSAR database and Masterfile are now merged into one product called Bio-Loom. The hydrophobic parameter anchors the Hansch databases: measured and calculated physicochemical properties are secondary. The methods complement, but are not replaced by new methods such as combinatorial chemistry, high throughput screening, and docking. Bio-Loom, however, is applicable to the new discipline of fragment-based drug design.

Al devoted the rest of his talk to several examples of the application of Bio-Loom ScaFrag (a scaffold-fragment concept renamed from the parent-substituent concept). Mother nature has designed privileged scaffolds: Al looked at some of these and the different activities found when "secondary", "tertiary" etc., fragments are considered. For example, phenobarbital is a typical GABA agonist but the barbiturate scaffold turns out not to be privileged at all: it is present in 15 matrix metalloproteinase (MMP) inhibitors in the Masterfile (this activity requiring a diphenyl ether and/or a spirolactam secondary ScaFrag) and unexpectedly, also in a high volume industrial chemical. In trying to achieve high MMP-3 selectivity, Hajduk and his colleagues showed that acetylhydroxamic acid has a low but significant activity [22]. Al searched the QSAR database and found that QSAR set 8,746 has 234 equations for nine MMP isozymes; it relates MMP-3 inhibition in 11 acetylhydroxamic acid analogs to ClogP and molecular volume with $r^2 = 0.98$. Looking for MMP-3 activity in Bio-Loom ScaFrag, Al found 278 matches on X–CO–NH–OH, 75 of which are histone deacetylase (HDAC) inhibitors and 130 MMP inhibitors. Sixteen of the 130 are MMP-3 specific: none contains a diphenyl ether fragment, but several have diphenyl cyano.

Al's talk was followed by one about the importance of experimental data curation prior to building QSAR models. Data dependency and data quality are critical issues in data modeling, said Alex Tropsha; cheminformaticians are at the mercy of data providers. Data curation is critical for any cheminformatics data set yet there is no literature on the basics of curation and there is no repository of standard operating procedures. Young et al. [23] have shown that small structural errors in a data set can lead to a significant loss of predictive ability for a QSAR model built on that set.

Erroneous structures lead to errors in the calculation of descriptors. Typical issues are misprints and wrong names, duplicates, mixtures and salts. Using SMILES or carrying out file conversions can lead to errors. Duplicate removal using names or CAS Registry Numbers is not good enough: it is essential to use chemical structures. It is not easy to find duplicates manually in a large data set and sometimes one of the pair is classified as active and one as inactive. When the data set is split in QSAR modeling, the test set may contain a compound identical to one in the training set and the resultant model may seem to give exceptionally good prediction. Duplicates are a common occurrence.

Alex's data curation procedure involves removal of inorganics and mixture components, standardization of aromaticity, normalization of carboxyl and nitro groups, handling of tautomers, removal of duplicates, and finally, manual inspection. Many procedures can be carried out with ChemAxon software and scripts, but without manual inspection it is not easy to know, for example, which component of a mixture is the active one.

Alex concluded with a specific example of what happens if you curate well. His team has collaborated with BioWisdom, a company which has a safety intelligence system based on chemical names extracted from the literature and matched with structures in public databases. Alex's team carries out curation, data analysis and QSAR, using the BioWisdom data. Small clusters were identified where there was high similarity between compounds. It was to be expected that the biological activity of the similar compounds would also be similar but in some cases this was not true, according to MEDLINE searches. It turned out that the initial annotation was inaccurate; re-mining the literature and using curated data produced the expected result.

The theme of the papers now changed to applications. Kent Stewart described the "drug guru" for cheminformatics analysis at Abbott. In the 1990s, a simple nitrogen–carbon switch converted ciprofloxacin into ABT-719 and led to a new series [24]. Kent and his colleagues realized that it would be good to capture such medicinal chemistry lessons so they developed Drug Generation Using Rules (Drug Guru) [25]. Functional group inter-conversions and chemical framework modifications can be encoded as SMIRKS; approximately 350 such rules have been written for Drug Guru. A chemist enters one structure and 50–150 alternatives are typically output.

The rules go beyond simply generating isosteres and patent busting: one rule, for example, is done to constrain a conformation. Output structures have a nice even distribution of physicochemical properties such as log P. Drug Guru also provides about 30 rules (e.g., pyridine to pyrimidine) to offset metabolism problems. A high proportion of chemists found that analog structures were "obvious" but 13–21% of structures are in the "got me thinking" category and only 9–16% are improbable.

Of course, Abbott is not the only organization to have developed this sort of software [26–32]. Kent has compared Drug Guru to BROOD from OpenEye, EMIL from CompuDrug, and CBIOSTER from Accelrys. Each program used a different strategy; each generated good and

bad structures, and there was incomplete overlap. Kent gave one example where the Drug Guru "remove chlorine" tool (just one of 50 hERG rules) had been particularly useful in reducing hERG liability.

Anthony Nicholls gave his opinions on hyperparametric modeling. A parameter is anything in a theory that increases your ability to accommodate to the answer. Examples of the types of explicit and implicit parameterization in models are (1) numeric values in some formula, (2) choice of numeric parameters, (3) choice of model or algorithm, (4) "operational" parameters, and (5) choice of question. QSAR parameters are typically type 1 or 2 and practitioners are conscious of them and realize that too many are a bad thing. This is typically because these parameters are "free"', i.e., designed to fit the data. In the QM/MD world there may be just as many parameters, plus a good many more of types 3, 4 or 5 but practitioners would claim they are "fixed" (are not free to accommodate the question under study). The closer you look, however, the more you see that parameters are not as fixed as some would claim.

The consequences of hyper-parameterization in QM and MD ought to give practitioners pause for thought: it becomes very easy to add the occasional free parameter, when there are hundreds of fixed ones. The use of hundreds also implies a lack of understanding of what is giving rise to such hyper-parameterization. The original QSAR formulations of Hansch and colleagues were *not* hyper-parameterized: typically they used only three or four terms. In modern practice, larger and less meaningful sets of descriptors have been adopted in the mistaken belief that this is acceptable because we can validate models. The fig leaves of respectability are tests such as cross-validation and y-scrambling.

It is true that cross-validation is asymptotically able to predict the correct model, but only with vast amounts of data. With finite amounts of data there is a fair chance it will validate an over-parameterized version. In the case of leave-one-out that Nicholls described you can perform an explicit Fischer test to see if the improvement in residual sum of squared differences (RSS) in the model with more parameters is worth it, but instead of using a risk assessment, everyone has "rules of thumb" as to what a cross-validation result means. In y-scrambling a model's over-parameterization is assessed by whether it can also predict random data, but correlation within the sample set really affects y-scrambling. We do not know what r^2 and q^2 numbers *mean* in terms of the risk that the model that better fits the data is the better model, not simply over-parameterized.

There are better ways, ways that weight risk precisely. Nicholls covered just three. His first example was the work of Akaike [33] on "an information content" (AIC). This quantity is like an energy: you lower it either by increasing likelihood or by decreasing the number of parameters. Another method that weights risk precisely is Vapnik's structural risk [34]. The essential form of Vapnik's approach to this problem is the concept of the Vapnik–Chervonenkis dimension, which is like a parameter count. The formula underlies what is called "empirical risk management" theory. The final method is a Bayes approach, called "Bayesian Factor Analysis" in which you are not allowed just to select the parameters for the best fitting model, you have to average over all the potential models. In comparing the different approaches, Nicholls likes the Bayes approach for its generalizability and ease of understanding, although the Akaike method is much easier to use and also allows you to see the consequences of over-parameterization.

Dick Cramer followed Anthony with an application for lead optimization. Optimization is the slowest and most costly stage in discovery. The goal is to find one or more R-groups that confer a combination of multiple required physical and biological properties. *Existing* ideas can be ranked by similarity, docking, or 3D QSAR. Generating *new* ideas should not be limited to synthetic routes and building blocks already on hand yet should consider synthetic feasibility. Splitting an individual molecular structure at any of its acyclic single bonds generates two R-group candidates. A large heterogeneous collection of molecular structures includes more candidate R-groups than complete structures and such R-groups are likely to be synthesizable. The axioms behind the topomer approach [35] are that ligands assembled from shape similar fragment sets tend to share biological activities, and when generating ligand fragment shapes for comparison, rule-based consistency is as effective as physicochemical rigor. Topomer similarity involves steric fields and feature matching; with Topomer CoMFA [36], the 3D alignments are the topomers themselves. Topomer CoMFA enables R-group virtual screening [37].

For predicting similarity in a biological property, topomer similarity is at least as effective as any other ligand similarity metric. For predicting magnitudes of biological properties 3D-QSAR has performed best, and Topomer CoMFA is a 3D-QSAR method that can supply such predictions objectively, automatically, and rapidly. Dick presented the results of retrospective validation studies [37, 38]. Topomer alignments yielded satisfactory CoMFA in all but one of 25 data sets. Using a new method, leave-out-one-R-group (LOORG) validation, predicted r^2 was 0.5; finding the best R-groups by ROC gave an SE of pIC_{50} prediction of 0.72. Workers at Bayer Healthcare have successfully used Topomer Search technology as a compound similarity measure in predicting hERG activity [39]. Another example demonstrated prospective ADMET

prediction success in a commercial discovery project: PubChem was mined, Topomer CoMFA was applied, and new structures were verified by synthesis and biological testing. There is no doubt context-ignorant topomeric alignments are very effective in 3D-QSAR: 15 successes in 15 trials is not simple good luck. Topomer alignments are obsessive about aligning like with like.

Next, Derek Debe talked about integrating data and analysis. Yvonne once did a study of what scientists want from software vendors. A chemist faced with 3,000 active compounds from a screen wants to be able to press just one button and get the 3–10 best series, and have requests automatically set up requesting additional experiments, compound purchases etc. To do this, Yvonne reckoned 120 searches had to be run, and 30,000 data items tracked, plus 18,000 more data items from physicochemical property work, using ten different computer programs and lots of databases. The system must be flexible because when the next screen is evaluated all the calculations are different. Additionally there must be tight integration between proprietary and external data. Lead optimization presents yet more challenges. Other essentials are a single, intuitive interface; sharing project information; flexible data pivoting and averaging, decision support, and so on.

Abbott chose Synaptic Science's SEURAT [40] to meet those needs. The system took 1 year and fewer than three FTEs to implement. Results from a database search appear in a spreadsheet with links to forms, analytics and SAR tables. Project management information is automatically updated and details can be emailed to the boss. Reporting is pain-free. Comprehensive compound data is available in one interface and any combination of WOMBAT, MDDR, WDI and Aureus GPCR can be selected. Compounds and assays in a sparse matrix can be reordered to squash the most useful data together at the top. Derek demonstrated mousing over, averaging, and the calculation engine. Calls to ChemAxon LibMCS clustering and to Pipeline Pilot can bring data into the spreadsheet. Data can be exported to Spotfire. Other decision support tools include kinome mapping and ChemAxon's R-group analysis. SEURAT is not yet perfect but it is well on the way to meeting Abbott's requirements.

Another Abbott speaker, Philip Hajduk described a probabilistic framework for interpreting similarity measures that directly relates the similarity value for a pair of compounds to a quantitative expectation that the two will actually be equipotent. This probabilistic approach using belief theory has been applied to compound subset selection for virtual and high-throughput screening. If 100 analogs of a hit are chosen at 10% belief to make new hits for SAR development then about 10% of those analogs should be active.

A screening collection contains multiple analogs of a number series of compounds but the question arises as to how many analogs are necessary to represent each series to ensure that an active series will be identified. Retrospective studies have been published [41, 42]; Philip spoke about application to *new* sets. Molecules are clustered at a certain level of belief (say 20%) so that if one cluster member is active then a certain percentage (here 20%) of other members will also be active. Belief levels of 10% produce very tight clusters. Philip described validation based on Abbott data. The approach outperforms conventional methods of subset selection. The method is enabled in Pipeline Pilot.

With cumulative belief assessment of diversity it is possible to see how different a new compound set is relative to an existing one and to find which subset maximizes coverage while minimizing testing. Given two sets, tested and untested, you calculate the cumulative probability that a given untested will be represented by a compound in the tested set. If the probability is sufficiently high, there is no need to include the new compound in the tested set. Using this method 19,838 compounds have been added to the Abbott collection, 13,707 of which have $>75\%$ coverage and ≥ 15 analogs at 10% belief. Philip is still working on subset coverage, using random or diversity subsets.

Steven Muchmore is also working with belief theory but his talk concerned application to similarity data fusion for use in analog searching and lead hopping [43]. One difficulty in effective lead hopping is in meaningfully combining results from different measures of similarity. Steven's approach is based on benchmarking of ten different similarity methods against a database of more than 150,000 compounds with activity data against 23 protein targets. Principles of decision theory can then be applied to combine the evidence from different similarity measures in a way that builds on the strengths of the individual methods and at the same time maintains a quantitative estimate of the likelihood that two molecules will have similar activity. The approach is implemented in an application called Lead Hopper, with a simple interface where a chemist enters a structure, hits a button and gets back an Excel spreadsheet together with alignment information. This application is now integrated in the high throughput screening process at Abbott and usage has really taken off. Steven concluded with a brief exposition of prospective analyses.

The symposium concluded with a presentation by the awardee herself. Yvonne's slides are available on the web [1]. With increasing numbers and types of 3D ligand-macromolecule structures becoming available every year, it is time to ask whether ligand-based methods are obsolete when there is a structure on which to base a design. Yvonne's presentation presented observations that suggest that careful analysis of ligand structure-activity

relationships provides independent information that contributes the discovery of ligands with the desired profile of potency, novelty, selectivity, etc. The second edition of Yvonne's book [44] *Quantitative Drug Design* is in the press and some of the examples she gave were research that she has done for the book using older publications that she has revisited in the light of 21st century advances, plus examples from her contribution to *Burger's Medicinal Chemistry*, also in the press.

She also touched on recent publications on lead hopping [43, 45]. She and Steven Muchmore have deployed a lead hopping application (mentioned above) that uses belief theory to combine the results of ROCS, Daylight, and ECFPlowbar6 similarities [45]. Ligand-based methods are useful in lead hopping; in pharmacophore-based enhancement of affinity and 3D identification of novel cores; and in potency predictions. Challenges to predicting potency include energetics of water, conformers, tautomers [46], and ionization states, and the energetics of subtle changes in the 3D structure of a target in response to ligand binding or changes in the environment.

References

1. Oral Presentations from the Fall (2009) ACS National meeting. http://www.softconference.com/ACSchem/slist.asp?C=3109, Accessed 13 Oct, 2009
2. Brameld KA, Kuhn B, Reuter DC, Stahl M (2008) J Chem Inf Model 48(1):1
3. Brock CP, Minton RP (1989) J Am Chem Soc 111(13):4586
4. Allen FH, Harris SE, Taylor R (1996) J Comput Aided Mol Des 10(3):247
5. Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, Purkis LH, Smith BR, Taylor R, Cooper RI, Harris SE, Orpen AG (2004) J Chem Inf Comput Sci 44(6):2133
6. Bruno IJ, Cole JC, Lommerse JPM, Rowland RS, Taylor R, Verdonk ML (1997) J Comput Aided Mol Des 11(6):525
7. Wood PA, Allen FH, Pidcock E (2009) CrystEngComm 11:1563
8. Allen FH, Baalham CA, Lommerse JPM, Raithby PR (1998) Acta Crystallogr Sect B Struct Sci B54(3):320
9. Maccallum PH, Poet R, Milner-White EJ (1995) J Mol Biol 248(2):374
10. Maccallum PH, Poet R, Milner-White EJ (1995) J Mol Biol 248(2):361
11. Deane CM, Allen FH, Taylor R, Blundell TL (1999) Protein Eng 12(12):1025
12. Paulini R, Mueller K, Diederich F (2005) Angew Chem Int Ed 44(12):1788
13. Bauer J, Spanton S, Henry R, Quick J, Dziki W, Porter W, Morris J (2001) Pharm Res 18(6):859
14. Galek Peter TA, Fabian L, Motherwell WDS, Allen Frank H, Feeder N (2007) Acta Crystallogr B 63(Pt 5):768
15. Galek PTA, Allen FH, Fábián L, Feeder N (2009) CrystEngComm 11:2634
16. Chrisholm J, Pidcock E, van de Streek J, Infantes L, Motherwell S, Allen FH (2006) CrystEngComm 8(1):11
17. Raymo C (1991) The virgin and the mousetrap. Collected Science Musing from the Boston Globe, Viking, New York
18. Berman HM (2008) Acta Crystallogr A 64(Pt 1):88
19. Lawson CL, Dutta S, Westbrook JD, Henrick K, Berman HM (2008) Acta Crystallogr Sect D Biol Crystallogr D64(8):874
20. Community Structure-Activity Resource (CSAR) (2009) http://www.csardock.org/, Accessed 13 Oct, 2009
21. Fujita T, Iwasa J, Hansch C (1964) J Am Chem Soc 86(23):5175
22. Hajduk PJ, Shuker SB, Nettesheim DG, Craig R, Augeri DJ, Betebenner D, Albert DH, Guo Y, Meadows RP, Xu L, Michaelides M, Davidsen SK, Fesik SW (2002) J Med Chem 45(26):5628
23. Young D, Martin T, Venkatapathy R, Harten P (2008) QSAR Comb Sci 27(11–12):1337
24. Li Q, Chu DTW, Claiborne A, Cooper CS, Lee CM, Raye K, Berst KB, Donner P, Wang W et al (1996) J Med Chem 39(16):3070
25. Stewart KD, Shiroda M, James CA (2006) Bioorg Med Chem 14(20):7011
26. Wagener M, Lommerse JPM (2006) J Chem Inf Model 46(2):677
27. Sheridan RP (2002) J Chem Inf Comput Sci 42(1):103
28. Ertl P (2003) J Chem Inf Comput Sci 43(2):374
29. Southall NT, Ajay (2006) J Med Chem 49(6):2103
30. Lewell XQ, Jones AC, Bruce CL, Harper G, Jones MM, McLay IM, Bradshaw J (2003) J Med Chem 46(15):3257
31. Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, Wood JM, Colclough N, Law B (2006) J Med Chem 49(23):6672
32. Kennewell EA, Willett P, Ducrot P, Luttmann C (2006) J Comput Aided Mol Des 20(6):385
33. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second international symposium on information theory. Budapest: Akademiai Kiado, 267
34. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
35. Jilek RJ, Cramer RD (2004) J Chem Inf Comput Sci 44(4):1221
36. Cramer RD (2003) J Med Chem 46(3):374
37. Cramer RD, Cruz P, Stahl G, Curtiss WC, Campbell B, Masek BB, Soltanshahi F (2008) J Chem Inf Model 48(11):2180
38. Cramer RD, Wendt B (2007) J Comput Aided Mol Des 21(1–3):23
39. Nisius B, Goeller AH (2009) J Chem Inf Model 49(2):247
40. Synaptic Science SEURAT. http://www.synapticscience.com, Accessed 16 Oct, 2009
41. Lipkin MJ, Stevens AP, Livingstone DJ, Harris CJ (2008) Comb Chem High Throughput Screen 11(6):482
42. Harper G, Pickett SD, Green DVS (2004) Comb Chem High Throughput Screen 7(1):63
43. Muchmore SW, Debe DA, Metz JT, Brown SP, Martin YC, Hajduk PJ (2008) J Chem Inf Model 48(5):941
44. Martin YC (1978) Quantitative drug design. A critical introduction [with reference to structure-activity relationships]. Dekker, New York
45. Martin YC, Muchmore S (2009) QSAR Comb Sci 28(8):797
46. Martin Y (2009) J Comput Aided Mol Des in press