ELSEVIER

Integrating molecular design resources within modern drug discovery research: the Roche experience

Martin Stahl, Wolfgang Guba and Manfred Kansy

F. Hoffmann – La Roche Ltd, Pharmaceuticals Division, PRBD-CM, CH-4070 Basel, Switzerland

Various computational disciplines, such as cheminformatics, ADME modeling, virtual screening, chemogenomics search strategies and classic structure-based design, should be seen as one multifaceted discipline contributing to the early drug discovery process. Although significant resources enabling these activities have been established, their true integration into daily research should not be taken for granted. This article reviews value-adding activities from target assessment to lead optimization, and highlights the technical and process-related aspects that can be considered essential for performance and alignment within the research organization.

It has become common sense to use computer-aided techniques for the efficient identification and optimization of novel molecules with a desired biological activity [1]. Judging from the table of contents of medicinal chemistry journals or conference programs, there is a seamless integration of these disciplines into the early drug discovery process. Is this truly the case? General acceptance of a field does not necessarily imply optimal use of the existing resources. With an emphasis on our own experience in cheminformatics and molecular design work at Roche, we examine current practices, highlighting key points to consider for efficient computational project support and to meet future challenges. Throughout the article, we use the acronym CAMM to describe the field of computational approaches to drug discovery. Many other acronyms are also in use, so the choice of CAMM is arbitrary – this acronym can be interpreted as 'computer-assisted molecular modeling' [2] or as 'cheminformatics and molecular modeling'. CAMM includes, in a very broad sense, all molecular modeling and cheminformatics [3] activities aimed at identifying structure-activity relationships and proposing alternative chemical structures with improved affinity, selectivity or physicochemical properties.

Three aspects of today's molecular design research form the basis of our discussion. First, we have to acknowledge that the principles of model-building and prediction-making have hardly changed over the past decade. This is particularly true for many 3D aspects of molecular design: the core routines of some of the most useful software tools are the same as ten years ago. Were it not for the significant increase in computational speed, we might have made little progress in the areas of docking and pharmacophore searching. Our ability to calculate binding free energies has only marginally improved [4], and ranking small-molecule ligands with respect to their relative binding affinities to a protein is still a highly error-prone process. Making proper use of molecular design resources has much to do with the ability to interpret calculation results, while recognizing the limits of the tools and knowing what is realistically achievable. It follows that molecular design is still to a large extent an expert area; the existence of 'high quality, easy to use' modeling tools is still a myth.

Second, the volume and quality of data generated in drug discovery programs is steadily increasing. Data alone is not useful, but needs to be interpreted, distilled into knowledge, and finally used to generate reliable predictions. The broadening basis of high-quality screening results, measured physicochemical properties and crystal structure information provides an enormous chance to learn and empirically improve computational prediction tools where methods based on physical principles fail.

The third aspect concerns the integration of various computational sub-disciplines. Industrial computer-aided design groups were founded in the 1980s with the emergence of the first practically useful computer graphics systems. The initial focus was on the 3D visualization of macromolecules, calculation of ligand conformations and pharmacophore perception. In parallel, QSAR approaches were also refined. During the second half of the 1990s,

Corresponding author: Stahl, M. (martin.stahl@roche.com)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.02.008

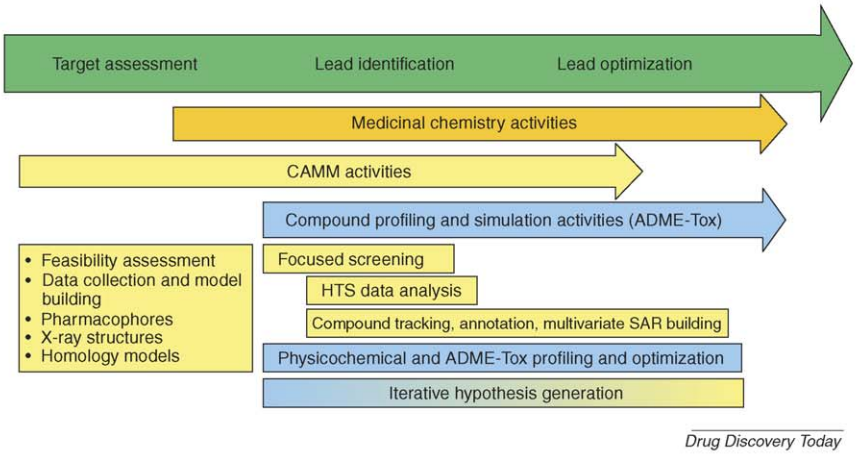

FIGURE 1

Modern drug discovery research process with focus on early CAMM, compound profiling and data interpretation activities. Yellow arrow, box and bars list key CAMM activities. Blue arrow and bars indicate experimental ADME–Tox and compound profiling activities.

when most of the industrial computational groups reached their present size, the field of cheminformatics was introduced, initially with a strong focus on library diversity and design. At that time, bioinformatics emerged as a separate field of growing importance for drug discovery. Today, terms like chemogenomics [5,6] describe the desire to unite all these fields by combining information about ligands, target sequences, target structures and biological activity data [7]. To achieve true synergy, the boundaries between computational subdisciplines have to dissolve to produce a single multifaceted discipline [8].

Keeping these three points in mind, we will now turn our attention to CAMM activities in the preclinical drug discovery process. At Roche, CAMM resources are integrated in the medicinal chemistry organization and there is a focus on lead generation support [9,10]. As projects move from target assessment to lead optimization, the nature of CAMM activities will gradually change (Figure 1). This review will roughly follow this path of project maturation, and conclude with sections on software development and teamwork aspects.

Target assessment: enabling informed decision making

When a new target is proposed, there is a need to assess the potential to generate a small molecule ligand with the desired activity and selectivity profile. Such a tractability assessment is based on diverse information from various sources, such as the nature of known ligands, the size, polarity and shape of binding sites in known target 3D structures, homology to related targets, and knowledge of the key amino acids modulating selective binding or functional activity. This requires a close link to bioinformatics and ready access to (proprietary and public domain) compound databases together with biological data and physicochemical properties. By bringing together ligand information with biological and structural data on the target, CAMM experts can become mediators between medicinal chemistry and biology.

Target assessment is closely linked with focused screening, as discussed in the next section. From homology or pharmacophore models, the first rules can be derived on how to select sets of compounds for initial screening experiments – perhaps not yet with the goal of identifying entry points for synthetic chemistry, but of validating assays and finding reference compounds that could be helpful for further assay development. Even at this early stage it is important to reach team consensus on the goals of these early activities because several rounds of screening might be necessary for adjusting assay parameters and finding compounds with the desired properties [11].

Identifying entry points for chemistry: efficient screening

The next key step in any project is identifying suitable starting points for chemistry, a task typically addressed by *in vitro* screening. At first sight, it can seem reasonable to generally screen as many compounds as are available because of the desire not to miss even a single attractive hit. However, in a larger context the choice of screening strategy should depend on additional factors: on the assay cost and throughput, on the level of knowledge of the target, on the likelihood of finding suitable hits in a screening library, and on portfolio considerations. A medium throughput assay could lead to results with higher accuracy. Compound logistics play an essential role – focused screening requires the ability to cherry pick compounds, whereas larger subsets with favorable physicochemical properties could be directly available for screening as a set of plates. Also, it is an often neglected fact that the total amount of time and resources required for an HTS run extends far beyond setting up and running the screen itself. The time required for adaptation of an assay to HTS format, screening and hit list analysis until the delivery of a list of hits validated in a secondary assay is not negligible. A significant amount of work needs to be invested in the analysis of a set of confirmed hits by experimental methods and by means of literature and patent searching.

For these reasons, the pros and cons of HTS, compared with focused subset screening, need to be carefully weighed at the outset of a project. A focused set of compounds could be screened first to arrive at initial hits quickly and to establish early SAR. It is

important to set clear goals for a focused screening campaign to facilitate a later decision on whether or not a full HTS should be run to complement the initial results. The hit rate or enrichment figures typically used in molecular design to judge the degree of success of a particular method are not useful here. The number of tractable and novel hit classes is a much more relevant, if somewhat subjective, measure.

High throughput screening

If HTS is chosen as a screening strategy, the role of CAMM focuses on hit analysis. At Roche, we have implemented a workflow foreseeing a first filter step at the level of primary hits. These are pruned to eliminate compounds with undesirable properties before hit confirmation. Confirmed screening hits are clustered according to common substructures [12,13] and profiled with respect to calculated physicochemical properties. Wherever possible, this data matrix is augmented with measured physicochemical and ADME data from the corporate database. A record is taken of how many times each compound has been screened previously and what activity levels have been measured. Typical false-positives [14] can thus be eliminated and privileged substructures can be identified. This evaluation can be followed by a 'hit expansion' step, whereby compounds structurally related to the hits are retested to identify false negatives [15]. CAMM members and medicinal chemists jointly identify chemical classes and provide an overview of relevant data for a subsequent assessment. In this manner, corporate knowledge is efficiently used and the selection of hit series is based on a clear rationale.

Focused screening

Many computational techniques are available to compile focused compound sets, with most of them falling under the umbrella term 'virtual screening' [16,17]. The secret to success lies in the choice of an appropriate combination of methods. Unlikely candidates should be eliminated, but at the same time the search should not be overly conservative in terms of known ligand features or binding requirements.

For ligand-based virtual screening, we have found it useful to deploy several 2D and 3D similarity metrics [18] on a web interface, so that the corporate compound repository can quickly be searched to examine different aspects of chemical similarity. Results from different searches can then combined by consensus scoring [19] (i.e. select molecules identified by more than one similarity metric) or by data fusion techniques [20] (i.e. compound ranks from different methods are added). In addition to the corporate compound depository, a virtual library of approximately one million compounds from external vendors has been compiled and is regularly updated, and orders for molecules can be placed automatically without bureaucratic overhead. This efficient access to non-corporate chemistry space opens huge opportunities for early SAR exploration. Similarity searches can be performed by any project chemist, whereas more complex search procedures, such as those requiring the derivation of a 3D pharmacophore, are elaborated by an experienced computational chemist. A recent example is shown in Figure 2 [21].

Structure-based virtual screening is widely and successfully applied [22] despite many methodological weaknesses and high error rates [23]. We have found it useful to employ as many search

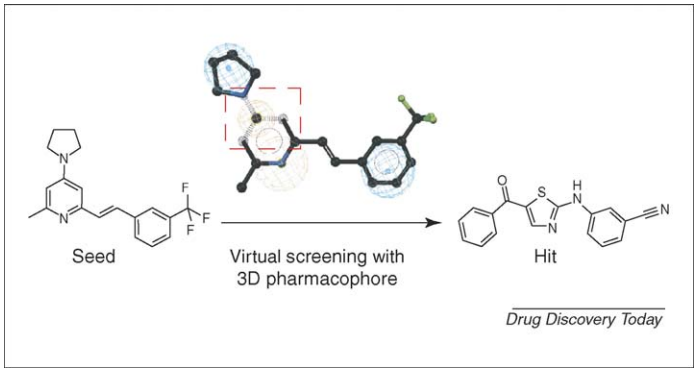

FIGURE 2

Scaffold hopping with 3D pharmacophore. In the quest for neuropeptide Y subtype 5 (NPY5) receptor antagonists, the hitherto unknown 4-aminothiazole hit class was discovered by virtual screening. A reference compound ('seed') was translated into a pharmacophore query consisting of 3D pharmacophoric features, topological elements and spatial constraints. The validity of this query was checked with a calibration set of active and non-active NPY5 receptor antagonists. The trigonal substitution pattern around the aromatic carbon atom (highlighted by the dashed red rectangle) was allowed to rotate and thus enabled the retrieval of a novel, *in vivo* active hit class without intellectual property issues. Conventional success metrics, such as hit rate (7 out of 632 compounds with IC_{50} in the micromolar range), do not reflect the value of this virtual screening campaign.

constraints as possible before and during a large-scale docking run. It does not make sense to dock libraries of a million compounds if many of their members can be excluded beforehand because of undesirable properties. Sequential virtual screening with filters of increasing complexity is a key to success. For most target structures, reasonable assumptions about ligand binding requirements can be made that go beyond the mere location of the binding pocket. These can be included in the docking run through pharmacophores [24] or, if applicable, by fixing a constant ligand fragment in space.

Chemogenomics search techniques

Chemogenomics search techniques [25–30] are a more recent addition to the set of CAMM tools for hit identification. Our experience is that they have proven to be effective for target classes without structure information, especially for G-protein-coupled receptors (GPCRs) [31]. A multidimensional similarity paradigm is at the heart of chemogenomics: ligand structure similarity, target sequence similarity and similarity of biological effects are combined. Biological similarity is determined in terms of affinity fingerprints of compounds against a set of targets [5,32–35]. These fingerprints do not necessarily correlate with molecular structure and derived similarity metrics because structurally different compounds can display the same activity profile. Consequently, complementary search scenarios are possible (Figure 3). Given a novel target 'X', one can search for novel ligands by testing compounds binding to other targets related to X in sequence space. Given a single known ligand, one can identify structurally unrelated hits by selecting compounds with a similar affinity fingerprint.

Significant resources are required to profile a master screening library against a reference panel of receptors, to collect high quality biological data from the public domain and to establish robust sequence-comparison methods. Only the combination of

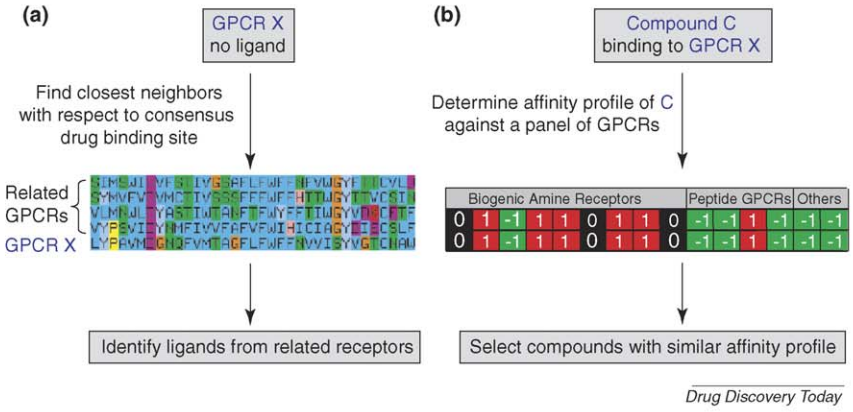

FIGURE 3

Chemogenomics search strategies to identify chemical entry points for G-protein-coupled receptor (GPCR) targets. (a) In the first scenario, the amino acids surrounding the putative binding pocket are identified, and the respective ligands of related receptors are selected for testing. **(b)** In the second scenario, a ligand (compound C) of target receptor X is profiled against a representative panel of GPCRs [affinity ranges are color-coded from red (active) over black (weakly active) to green (inactive)]. This affinity fingerprint is compared with previously profiled compounds and the nearest neighbors with respect to the affinity profile of C are identified. Because biological similarity is not necessarily correlated with chemical similarity, this strategy is a promising route to detect novel chemotypes.

all components allows for an efficient use of the search strategies mentioned above. Such an investment is only warranted for key target families where selectivity is a central issue and might conversely also be exploited in tailored polypharmacology profiles. However, once a chemogenomics data collection exists it fosters the integration of all drug discovery sub-disciplines and allows for faster decision making from target assessment onwards.

Fragment-based screening

Fragment screening [36,37] should be mentioned here as an additional 'focused screening' technique. Many pharmaceutical companies have established small libraries of several hundred to several thousand low molecular weight substances that are screened by direct-binding methods, such as surface plasmon resonance. In combination with X-ray crystallography, the weak binders identified in this way can provide information for chemistry programs. The challenge for CAMM is to guide the structure-based optimization of these low-affinity hits to potent leads.

Optimization of chemical series: driving quality

The optimization of chemical series starts with the selection of promising compound classes suitable for further optimization. Clearly, the optimization of binding affinity and selectivity is still a central task in this phase. But even at this early stage, all activities should focus on multiple absorption, distribution, metabolism, excretion and toxicity (ADME–Tox)-related parameters in parallel to activity and selectivity. This ensures a high overall quality of the maturing lead series.

The standard procedure for ADME–Tox ranking in the lead identification and early lead optimization phases is typically based on prediction tools and filters focusing on parameters that are easy to calculate [38–41]. Commercially available tools for calculating physicochemical properties and ADME–Tox-related parameters are usually used for this purpose [42–44]. However, results obtained from such tools have to be interpreted with great care [45]. The use of generic models can only be recommended if they have been validated for a particular project. If new compound classes outside of the training sets are evaluated, the result can be misleading. This is of particular importance for ionization constants, lipophilicity and solubility, where the authors have often observed differences between calculated and measured parameters in the range of 1.5–2.0 log units, which is too great for proper ranking and decision-making support. Even more care is necessary if these fundamental parameters are used in combination to predict more complex parameters like distribution coefficients or fraction absorbed. To build reliable models, only validated measurement results with defined quality should be used as input parameters.

These observations have led to a shift in optimization strategy affecting both the generation and the use of experimental data [46–49]. The recommended use of measured values [50] calls for high quality, fast and standardized assays. Nowadays, such assays have a throughput of ~100–500 compounds per week. Measurement cycles need to be synchronized with the chemical design and synthesis cycles to ensure fast feedback loops. In this manner, results can be generated with sufficient speed and quality for interesting lead series and properties to identify potential liabilities early.

The early availability of measurement results enables the interpretation of complex *in vivo–in vitro* relationships and the development of *in silico* models for specific parameters to be optimized for individual series of compounds (Figure 4). Structural effects on distribution, metabolic stability, hERG affinity or phospholipidosis liability can thus be addressed early and with higher predictive power. Generally, the aim of a local model is to rank compounds and not to predict the absolute magnitude of an *in vivo* or *in vitro* effect. Preferably, models should be simple and easy to interpret and thereby directly lead to testable hypotheses. Complex models are difficult to validate and difficult to challenge, and do not inspire researchers to plan further experiments. Models should be continuously refined according to new data becoming available.

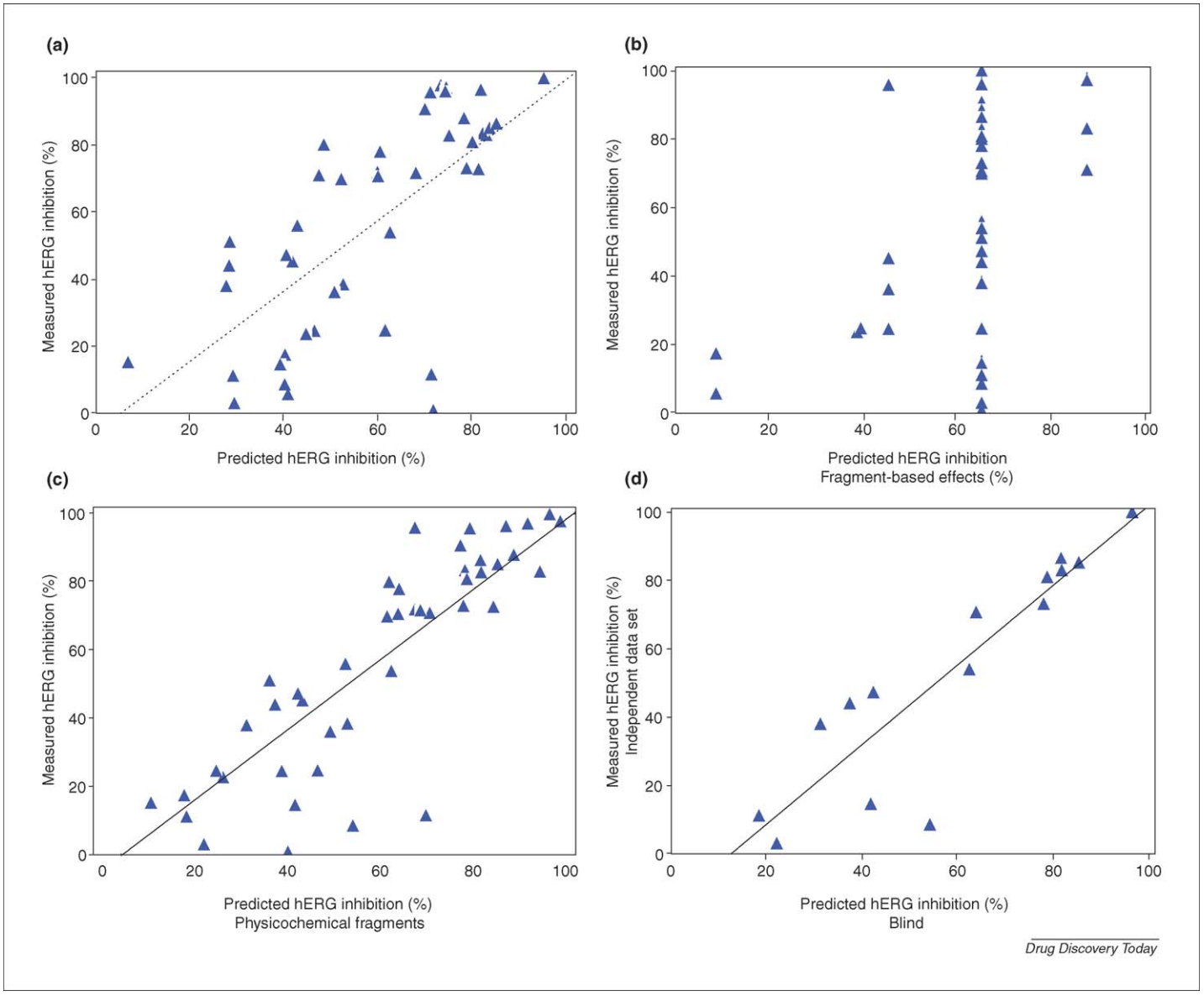

FIGURE 4

Influence of multiple parameters on a local, project specific, *in silico* model for inhibition of the hERG potassium channel. (a) *In silico* model based on physicochemical parameters. **(b)** Model based on structural fragments. **(c)** Combined parameter and fragment model (a+b = final model). **(d)** Result of a blind validation of model (c) with an independent dataset.

If implemented consistently, this process allows project teams to abandon the classic paradigm of sequential filtering in more complex and expensive models. Instead of using *in silico* tools followed by *in vitro* and *in vivo* profiling, model building should continue throughout the process and *in vivo* spot checks should be used early to validate models. In a best-case scenario, a predictive (computational or *in vitro*) model can be found that allows the optimization of *in vivo* parameters. In this manner, assay capacity can be used intelligently to determine only those parameters that will eventually be used for decision making.

Overfitting of models [51] has long been a concern in the CAMM area. Even for an expert it can be easy to over-interpret SAR and to focus too little on true predictive power. Therefore, the criteria mentioned here for model building – an emphasis on experimental data, model simplicity and continuous refinement – should apply not only to ADME–Tox models, but also to all models built by CAMM experts. The principle of frequent cross-checking with experimental data is beneficial in all CAMM areas. For example, current molecular mechanics force fields [52] still have error-prone and incomplete parameter sets. Calculation results based on these force fields should not only be compared with high-level *ab initio* quantum mechanical results but also – more quickly and reliably – with search results in the Cambridge Structural Database (Figure 5). Likewise, for non-bonded interactions one can refer to interfaces to the Protein Databank (PDB), such as Relibase [53–55].

Method development and deployment

Medicinal chemists are increasingly equipped with a very detailed knowledge of many CAMM areas, and many cheminformatics applications are nowadays routinely used by medicinal chemists. The distinction between generalist and expert tools must be made

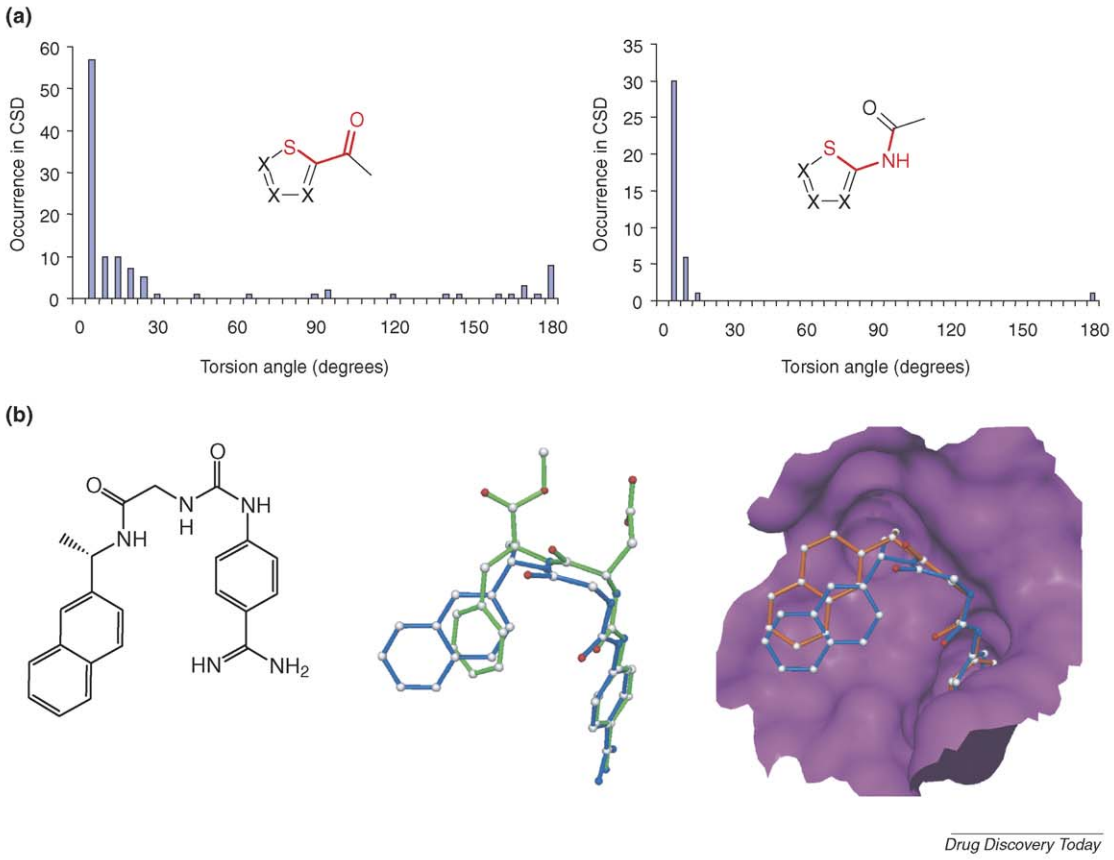

FIGURE 5

The Cambridge Structural Database (CSD) is a valuable source of information on conformational properties of drug-like molecules. (a) Shows just one of many specific conformational effects not captured by many standard force fields. The attractive interaction between divalent sulfur atoms and carbonyl groups leads to pronounced preferred conformations of thiophene-type heterocycles carrying acyl groups (left) or acylamino (right) functions (X matches C, N and O). In **(b)**, a more specific application is shown: the structure of a urea benzamidine inhibitor of the Tissue Factor–Factor VIIa complex claimed by Aventis [60] was predicted (blue structures) before binding modes of related structures were published by Aventis [61] by overlaying a closely related entry (CSD entry code HAFKOO, green structure in the center) with the known binding mode of a benzamidine in the binding site. The structure solved at Roche closely matches the predicted one. Reproduced, with permission, from Ulrike Obst.

wisely because it affects CAMM work in various ways. First, tools accessible to everyone require intuitive interfaces that need to be maintained, and any investment made here will reduce the time available for the development of new methods. Second, the broad availability of a tool has an effect on the way scientists communicate with each other. An expert carrying out a calculation can deliver the result together with an interpretation in a way most suitable for the team, thereby avoiding the danger of over- or mis-interpretation by non-experts. Generally, an investment in interface development is worthwhile if the corresponding calculations are to be performed regularly and leave little room for misinterpretation. Among these are virtual screening techniques requiring few task-specific parameter adjustments like similarity searching, the use of 'logistics and sorting' tools (e.g. for clustering and library design) and query interfaces to chemogenomics or crystal structure databases. At Roche, the only exception is the long-term development of the Roche Molecular Modeling package, MOLOC (www.moloc.ch). The development of this package was started by Paul Gerber in the 1980s, when no commercial software solutions of this type were available. Since then, it has been a key element of CAMM work at Roche.

State-of-the-art software for modeling and prediction is a key asset in drug discovery; however, the central goal of CAMM at Roche is project support. The development of new methods and software (e.g. *in silico* prediction tools and Chemogenomics databases) is therefore actively pursued only when commercial solutions are either not available or prohibitively expensive. Wherever the decision for in-house development is taken, tools should be as simple to extend and maintain as possible.

To be useful in an industrial setting, software tools must be robust and generally applicable. To evaluate tools in this respect, we have developed standardized test sets. Many tools developed in the academic world are based on innovative ideas and algorithms but have not been properly parameterized to yield optimum results. Method development therefore also includes the re-parameterization of software tools. This can be done most efficiently in collaboration with a partner focusing on algorithmic aspects. At Roche, successful collaborations of this kind include the refinement of scoring functions for FlexX [56], the development of the Feature Trees fragment space search module [57] and the development of an enhanced version of the *de novo* design program SkelGen™ [58].

Processes and teams

In a review article written in 1991, James P. Snyder [59] put forward the proposition that 'at the level CADD groups are presently integrated throughout the pharmaceutical industry, there is little chance they will make a fundamental impact on drug discovery in the short term'. Although much has changed over the intervening years, Snyder raised several points still worth considering today. One of his central statements is that without true and active project team participation, contributions by CAMM experts might easily be ignored by a project team, the simple reason being that the 'designer' and the experimentalist are typically not the same person. Molecular design work is conceptual in nature and an element of persuasion is involved in reducing an idea to practice.

This has several direct consequences on resource allocation. Because full integration into a team requires more time than *ad hoc* contributions, it is essential to focus CAMM activities on key projects instead of diluting the efforts over many projects, which in turn requires the courage to say no or to stop activities when little impact on ongoing research is expected. Projects should be selected based on an assessment of whether the participation of CAMM might lead to a high level of inventive collaboration; an assessment that requires experience and an overview of the data available for hypothesis building.

It should be self-evident that optimal communication is fostered in one-on-one or small group meetings. At Roche, it has been common practice for more than a decade to arrange regular interactive 'design sessions' in front of a computer screen or in a room equipped with a 3D projection facility to address complex 3D structural aspects. The high level of stimulation and scientific exchange achieved in such sessions is unparalleled. It can hardly be reached if contact with project team members is reduced to exchanging lists of compounds and proposals by e-mail.

Computational chemists are neither isolated creative minds nor anonymous operators of semi-automated software tools, but data miners, facilitators and, ultimately, drug hunters. Each compound can be characterized with dozens of numbers ranging from calculated properties to measured physicochemical and biological data. Multivariate techniques allow the clustering of compounds with respect to common property profiles and the detection of outliers. The challenge for computational chemists is to derive patterns and trends and to convert numbers to more qualitative statements. Results need to be communicated clearly, with as little technical detail as possible but without oversimplification, to eventually convincing project teams to gather missing, but needed, data. Newly designed ligand structures need to be scrutinized and refined in team discussions. Because the task in lead generation is to scan the chemical space widely and to identify promising series for further optimization, computational chemists also need to monitor compound statistics and SAR to ensure a representative coverage of chemical space and to constantly refine working hypotheses. Beyond the horizon of an individual project, there is an increasing necessity to capture the knowledge gained in a way that it is useful for other projects. This is a task that CAMM experts can significantly contribute to.

Conclusion

It is almost a trivial statement that today computational methods pervade all aspects of drug discovery research. This is underlined by new approaches to hit finding and optimization, such as chemogenomics and fragment screening, that rely on computational methods as a core discipline. Higher success rates and better integration of CAMM is not primarily based on the improvement of computational methods, but on a better way of dealing with the shortcomings of existing methods. At the same time, the early drug discovery process has become a much more data-rich field. A key challenge for today's CAMM work is to provide timely and accurate analyses of experimental data from various fields. Finally, making optimum use of CAMM resources requires a research culture based on mutual trust and team-oriented working processes.

Acknowledgements

We are grateful to all the Roche scientists who helped to make the dream of a truly integrated drug discovery process come true: the members of the molecular design/cheminformatics and molecular properties groups, and one decade's worth of drug discovery project teams. Their enthusiasm has been a thoroughly enjoyable experience.

References

1 Jorgensen, W.L. (2004) The many roles of computation in drug discovery. *Science* 303, 1813–1818
2 Gund, P. *et al.* (1980) Three-dimensional molecular modeling and drug design. *Science* 208, 1425–1431
3 Olsson, T. and Oprea, T. (2001) Cheminformatics: A tool for decision-makers in drug discovery. *Curr. Opin. Drug Discov. Devel.* 4, 308–313
4 Stahl, M. and Schulz-Gasch, T. (2004) Scoring functions for protein ligand interactions: a critical perspective. *Drug Discov. Today. Technol.* 1, 231–239
5 Caron, P.R. *et al.* (2001) Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* 5, 464–470
6 Kubinyi, H. and Müller, G. (2004) *Chemogenomics in Drug Discovery*. Wiley–VCH
7 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58
8 Tropsha, A. (2000) Recent trends in computer-aided drug design. *Curr. Opin. Drug Discov. Devel.* 3, 310–313
9 Alanine, A. *et al.* (2003) Lead generation - enhancing the success of drug discovery by investing in the hit to lead process. *Comb. Chem. High Throughput Screen* 6, 51–66
10 Roche, O. and Guba, W. (2005) Computational chemistry as an integral component of lead generation. *Mini Rev. Med. Chem.* 5, 677–683
11 Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894
12 Stahl, M. and Mauser, H. (2005) Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model* 45, 542–548
13 Stahl, M. *et al.* (2005) A robust clustering method for chemical structures. *J. Med. Chem.* 48, 4358–4366
14 Roche, O. *et al.* (2002) Development of a virtual screening method for identification of ''frequent hitters'' in compound libraries. *J. Med. Chem.* 45, 137–142
15 Shanmugasundaram, V. *et al.* (2005) Hit-directed nearest-neighbor searching. *J. Med. Chem.* 48, 240–248
16 Walters, W.P. *et al.* (1998) Virtual screening – an overview. *Drug Discov. Today* 3, 160–178
17 Oprea, T. and Matter, H. (2004) Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* 8, 349–358
18 Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* 7, 903–911

19 Charifson, P.S. *et al.* (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* 42, 5100–5109
20 Salim, N. *et al.* (2003) Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* 43, 435–442
21 Guba, W. *et al.* (2005) Novel and potent NPY5 receptor antagonists derived from virtual screening and iterative parallel chemistry design. *Bioorg. Med. Chem. Lett.* 15, 1599–1603
22 Congreve, M. *et al.* (2005) Structural biology and drug discovery. *Drug Discov. Today* 10, 895–907
23 Shoichet, B.K. (2004) Virtual screening of chemical libraries. *Nature* 432, 862–865
24 Hindle, S.A. *et al.* (2002) Flexible docking under pharmacophore constraints. *J. Comput. Aided Mol. Des.* 16, 129–149
25 Jacoby, E. (2001) A novel chemogenomics knowledge-based ligand design strategy - Application to G protein-coupled receptors. *Quantitative Structure–Activity Relationships* 20, 115–123
26 Schuffenhauer, A. *et al.* (2002) An ontology for pharmaceutical ligands and its application for *in silico* screening and library design. *J. Chem. Inf. Comput. Sci.* 42, 947–955
27 Müller, G. (2003) Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* 8, 681–691
28 Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 5, 262–275
29 Crossley, R. (2004) The design of screening libraries targeted at G-Protein coupled receptors. *Curr. Top. Med. Chem.* 4, 581–588
30 Savchuk, N.P. *et al.* (2004) Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr. Opin. Chem. Biol.* 8, 412–417
31 Kratochwil, N.A. *et al.* (2005) An automated system for the analysis of GPCR transmembrane binding pockets: Alignment, receptor-based pharmacophores and their application. *J. Chem. Inf. Model.* 45, 1324–1336
32 Weinstein, J.N. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349
33 Dixon, S.L. and Villar, H.O. (1998) Bioactive Diversity and Screening Library Selection via Affinity Fingerprinting. *J. Chem. Inf. Comput. Sci.* 38, 1192–1203
34 Beroza, P. *et al.* (2002) Chemoproteomics as a basis for post-genomic drug discovery. *Drug Discov. Today* 7, 807–814
35 Fliri, A.F. *et al.* (2005) Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U. S. A.* 102, 261–266
36 Fattori, D. (2004) Molecular recognition: The fragment approach to lead generation. *Drug Discov. Today* 9, 229–238
37 Hartshorn, M.J. *et al.* (2005) Fragment-based lead discovery using x-ray crystallography. *J. Med. Chem.* 48, 403–413
38 Yoshida, F. and Topliss, J.G. (2000) QSAR model for drug human oral bioavailability. *J. Med. Chem.* 43, 2575–2585
39 van de Waterbeemd, H. *et al.* (2001) Property-based design: Optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* 44, 1313–1333
40 Chin, D.N. *et al.* (2004) Integration of virtual screening into the drug discovery process. *Mini Rev. Med. Chem.* 4, 1053–1065
41 Martin, Y.C. (2005) A bioavailability score. *J. Med. Chem.* 48, 3164–3170
42 Yu, H. and Adedoyin, A. (2003) ADME-Tox in drug discovery: Integration of experimental and computational technologies. *Drug Discov. Today* 8, 852–861
43 Egan, W.J. *et al.* (2004) In silico prediction of drug safety: despite progress there is abundant room for improvement. *Drug Discov. Today. Technol.* 1, 381–387
44 Cirovic, D.A. *et al.* (2005) In silico methods and predictive tools along the drug discovery value chain. *Pharmaceutical Discovery* 5, 32–35
45 Stouch, T.R. *et al.* (2003) In silico ADME/Tox: Why models fail. *J. Comput. Aided Mol. Des.* 17, 83–92
46 Avdeef, A. and Testa, B. (2002) Physicochemical profiling in drug research: a brief survey of the state-of-the-art of experimental techniques. *Cell. Mol. Life Sci.* 59, 1681–1689
47 Kansy, M. *et al.* (2004) Advances in screening for membrane permeability: High-resolution PAMPA for medicinal chemists. *Drug Discov. Today. Technol.* 1, 349–355
48 Saunders, K.C. (2004) Automation and robotics in ADME screening. *Drug Discov. Today: Technol.* 1, 373–380
49 Taylor, D.L. and Giuliano, K.A. (2005) Multiplexed high content screening assays create a systems cell biology approach to drug discovery. *Drug Discov. Today. Technol.* 2, 149–154
50 Krejsa, C.M. *et al.* (2003) Predicting ADME properties and side effects: The BioPrint approach. *Curr. Opin. Drug Discov. Devel.* 6, 470–480
51 Hawkins, D.M. (2004) The problem of overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1–12
52 Mackerell, A.D. (2004) Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* 25, 1584–1604
53 Bergner, A. *et al.* (2001) Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects. *Biopolymers* 61, 99–110
54 Günther, J. *et al.* (2003) Utilising structural knowledge in drug design strategies: Applications using Relibase. *J. Mol. Biol.* 326, 621–636
55 Hendlich, M. *et al.* (2003) Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* 326, 607–620
56 Stahl, M. and Rarey, M. (2001) Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* 44, 1035–1042
57 Rarey, M. and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* 15, 497–520
58 Stahl, M. *et al.* (2002) A validation study on the practical use of automated *de novo* design. *J. Comput. Aided Mol. Des.* 16, 459–478
59 Snyder, J.P. (1991) Computer-assisted drug design. 1. Conditions in the 1980s. *Med. Res. Rev.* 11, 641–662
60 Klingler, O. *et al.* (2001) Factor VIIa Inhibitory (thio)urea derivatives, their preparation and their use. Aventis Pharma GMBH (DE), WO0194301.
61 Klingler, O. *et al.* (2004) Structure-based design of amidinophenylurea-derivatives for factor VIIa inhibition. *Bioorg. Med. Chem. Lett.* 14, 3715–3720