Drugs, structures, fragments : substructure-based approaches to GPCR drug discovery and design

Horst, E. van der

CHAPTER 7

General Conclusion and Perspectives

7.1 Conclusions

The work presented in this thesis contributes to our knowledge of G protein-coupled receptors (GPCRs) and their ligands. The application of substructure mining on groups of GPCR ligands allowed us to thoroughly analyze the structural characteristics of these ligands, to study the relationships among GPCRs based on their ligands, and to use these findings in a prospective manner, *i.e.* to discover new GPCR ligands.

As reviewed in chapter 2, substructure-based methods are valuable research tools in cheminformatics and medicinal chemistry research. As a complement to structure-based approaches, these methods are deep-rooted in GPCR drug discovery. For instance, the name *privileged structures* was first coined in GPCR research to denote the (perceived) common structural motifs found in GPCR ligands to which a preference for multiple GPCRs could be attributed.

As already discussed in chapter 2, the analysis of the structural parts that constitute molecules, such as rings and ring systems, is biased towards chemists. In contrast, exhaustive substructure-based methods consider all possible substructures that occur in the molecules, and do not suffer from this bias. The application of frequent substructure mining described in chapter 3 thoroughly analyses the structural features that discriminate between ligands that bind to GPCRs and those that do not. This type of large-scale screening of molecular libraries would be virtually impossible without the recent advances in frequent substructure mining algorithms. The algorithm used throughout the first three research chapters, named Gaston, proved very versatile.

Significant differences were found between the background chemical library and the complete set of GPCR ligands. For instance, the substructures found in the screening library used in Chapter 3, predominantly esters and carboxamide moieties, reflected the synthetic reactions used to construct the library. The characteristic substructures found in the merged GPCR sets, alkylamines often linked to aromatic systems, mainly

reflect the presence of aminergic ligands in the set. Hierarchically narrowing down the sets from superfamily, to receptor family, to receptor, and to receptor subtype, revealed well-known motifs as well as new substructural features. These included the imidazole-like substructure common in histamine receptor ligands and the planar ring system consisting of a fused five- and six-membered ring (indole-like substucture) often present in serotonin receptor ligands. Apart from these common recognizable motifs, we also found new, more abstract substructures as typical motifs for a GPCR subset.

Since not only structures recognizable by chemists (ring systems, functional groups) were found but also more abstract structures, we concluded that frequent substructure mining has added advantage for GPCR research. As expected, the use of different chemical representations had the effect of finding different sets of characteristic structural features in a library. Varying the chemical representation of a library can be explored further, however, there will always be a trade-off between specificity and diversity. For instance, the normal chemical structure drawing representation was most specific in identifying compounds of a set but performs worse on libraries that are more dissimilar.

The work of chapter 3 showed that frequent substructure mining is capable of identifying both well-known and new structural motifs that are characteristic for phylogenetic subsets (*e.g.* class, family, subtype) of GPCRs. In chapter 4, this capability was employed to perform the reverse analysis, namely to find a phylogenetic organization for GPCRs based on the common structural features of their ligands. The frequent substructures generated from the ligand sets facilitated a robust analysis of the structural similarity between ligand sets and subsequent hierarchical ordering of GPCRs into a phylogenetic tree. This tree expresses the structural overlap between ligands of different receptors and therefore which receptors are more likely to share a common ligand. Thus, the tree provides insight into possible side-effects and can be used as a ‘deorphanization’ tool to find new ligands for a receptor by starting with ligands from ‘nearby’ receptors. Comparison between the phylogenetic classification

based on the target sequences and the substructure-based tree also provided general insight into the relations among GPCR targets.

A prospective application of frequent substructure mining of GPCR ligands is first described in chapter 5. We constrained ourselves to the subfamily of adenosine receptors as templates for the whole superfamily of GPCRs. In this particular case we screened a commercial vendor library for new ligands for the A_{2A} adenosine receptor. New ligands were found and prioritized that were truly different from existing A_{2A} adenosine receptor ligands. This analysis was quite successful, as we learned that many compounds recognized the adenosine receptors when tested in an experimental setting. For reasons of cost-effectiveness we included one vendor only; it is to be expected that the possible inclusion of more libraries would increase the diversity of ligands even further. The findings in chapter 5 also formed the starting point for a more extensive screening effort within the TI Pharma consortium "The GPCR forum", in which we virtually screened a company library for a number of GPCRs. The results of this study will be published elsewhere. In chapter 6, our screening concept was extended from screening existing compounds to exploring chemical space. In contrast to analyzing existing significant substructures present in e.g., a vendor library, chemical structures were modified through small substructural changes to iteratively yield 'better' compounds, a typical example of *de novo* design. For this we used a derivative of our Molecule Evoluator software, called the Molecule Commander. This prototype was used as a first component in a typical drug design workflow (PipelinePilot®) that combined requirements such as affinity for the desired target (the adenosine A_1 receptor in this case) with constraints such as Lipinski's Rule of Five. The hit-rate of this molecular evolutionary optimization procedure was high, *i.e.* at least two out of six scaffolds were active. This offers the perspective of a chemist formulating a research question and/or drawing one or more molecules, after which the computer is instructed to search compound databases, optimize molecules, and present a range of best possible ideas to the chemist.

7.2 Future perspectives

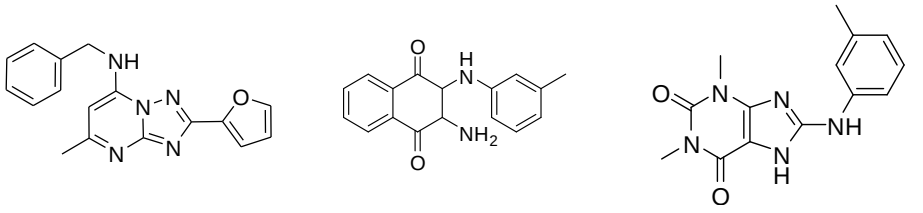
7.2.1 Cheminformatics

From the work described in this thesis we have learned to appreciate the role of cheminformatics in the process of drug design and discovery. In fact, that contribution may even be increased. In particular the use of more abstract substructures as described in Chapter 3 brought significant advantages. In retrospect we have the impression that the vocabulary of the medicinal chemist is largely based on compound features important for chemical reactivity. The explicit use of 'rings', 'bonds' and 'functional groups' as substructures may largely stem from a medicinal chemist's background in synthetic organic chemistry. For the desired biological effect it would be helpful – and Chapter 3 is proof of that – to focus on the pharmacologically relevant substructures. Algorithms such as GASTON can help us with that.

At the same these substructure mining methods can also be improved. In the current set-up substructures can be anywhere in the molecule, and lack information about the precise location, e.g. at the 'end' of a branch or mapped onto the 'core' of the molecule. This also has an impact on the connectivity of an atom, which is essentially lost in this approach. The representation of bonds can also be problematic from the way molecules are represented in databases. For instance, the representation of an aromatic ring system can be quite different in databases, and this may have consequences for the perceived substructure. Likewise, keto/enol tautomerism can cause ambiguity in the process of defining substructures. When the substructures are eventually used for virtual screening purposes – as in Chapter 5 – another question arises: if one substructure is good, would two of the same be better? That is not necessarily the case for a biologically active molecule, but without further precautions a simple summation of fragments may lead to erroneous conclusions, as e.g., more fragments mean higher molecular weights, which are not desirable *per se*.

An important aspect of cheminformatics has been and still is the availability of publicly accessible, open databases. In the world of bioinformatics there has always been the notion that databases holding e.g., genomic or protein information, should have 'open access'. This has long been different in the world of cheminformatics. Many compound databases were either prohibitively expensive for academic use or simply not accessible (e.g., compound collections in big pharma). During the course of this PhD project this situation has drastically improved, most impressively by two initiatives, one in the US, the other in the UK. PubChem at the NIH (US) now holds information on all aspects of available compounds, including screening data on selected compounds, literature data, and vendor availability. At the EBI (Hinxton, UK) the ChEMBL database was made publicly available from a commercial database through the generous support of the Wellcome Trust. It is a database of drug-like small molecules, with bioactivities abstracted and curated from available scientific literature and patents. Currently, data on clinical progress of compounds is also being integrated into ChEMBL. This development is reflected in this thesis: in Chapter 3 we used the then available GPCR compound database GLIDA, in Chapters 4 and 5 we also used (pre)releases of ChEMBL. As with all databases care should be taken as errors tend to propagate; we noticed several errors ourselves, such as wrong bioactivity data, wrong structures, all probably inevitable in such huge data compilations. In fact, this is the consequence of the transformation from information in published documents such as scientific papers towards data storage. This involves the human mind and possible interpretation errors, the conversion of flat text into structured data such as in databases, and potentially many more 'conversion' errors. A better approach would be to remove these intermediate steps of flat text publishing and data extraction and instead make the data directly available in structured format. However, when data is entered in standard databases, most of the context that would normally be provided in an article is lost. Some data providers attempt to offer some context as extra fields in database tables. However, this is done fairly ad-hoc and not in a standardized manner, and these additions therefore lack any real meaning. These shortcomings are increasingly realized nowadays, but aligning and integrating proprietary and public

data sources into a single system is a difficult and time consuming task. Hence it does not come as a surprise that duplication and redundancy are common across companies, institutes and academic laboratories. An interesting initiative in this respect is the European concerted action called Open PHACTS. The members involved in this consortium (both from academia and industry) aim to create an open platform, Open Pharmacological Space, which will be freely accessible for knowledge discovery and verification. It will provide a growing body of data on small molecules, their pharmacological profiles, pharmacokinetics, ADMET data, biological targets and pathways in a semantically interoperable format. The latter means for instance that the concepts in the database to describe e.g., biological context (potency, activity, activation, inhibition, etc) are interpreted in an unambiguous, formalized way.



0.77 µM
(Compound 7, Table 5, Chapter 5)

0.8 µM
(Compound 50 of Table 1 Katritch *et al.*[1])

1.2 µM
(Compound 12 of Table 1, Carlsson *et al.*[2])

Figure 1. Most potent hit compound from the A_{2A} substructure-based screening (Chapter 5) and two examples of hits with roughly the same affinity identified in the structure-based screening studies of Katritch *et al.*[1] and Carlsson *et al.*[2]

During the research done in this thesis remarkable developments in the elucidation of GPCR structures occurred. In the early phases of the research (2006, 2007) no other structures than rhodopsin were available. Rhodopsin has low sequence homology to most other GPCRs, including adenosine receptors, and hence homology models based on the rhodopsin structure were considered highly speculative, at least by us. That all

changed with the elucidation of the 3D structure of the β_2-adrenergic receptor, followed by a number of other receptors.[3] Interestingly, some of these receptor structures have been successfully used for virtual screening, by docking commercially available compounds into the ligand binding site and prioritizing them on their energy score.[1,2,4] High 'hit rates' were observed and chemical diversity in these hits was also significant. This might suggest that ligand-based methods are obsolete; however, we showed that the hits from the ligand-based approaches are also viable, and, most importantly, different from the ones found in the structure-based screening (see Figure 1). In other words, both methods are highly complementary. In fact, we also benefitted from the structural knowledge obtained in recent years. In Chapter 4 we constructed a phylogenetic tree of GPCRs based on the approx. 40 amino acid residues that are close to the ligand binding site as defined by Gloriam *et al.* from the then available crystal structures.[5] Another option would be to feed the compounds stemming from a structure-based search into the ligand-based approach we took in Chapter 6, i.e. a further optimization and modification of the chemical structure through evolutionary computation. It would be interesting to see whether the combination of the two approaches would also allow us to further expand on all pharmacological characteristics of new compounds. Currently, affinity is the almost exclusive determinant used, whereas ligand efficacy (i.e. agonism, antagonism, inverse agonism) is largely ignored, let alone concepts such as ligand-biased signaling. Two recent computational papers are seminal in this respect, and they seem to define a new avenue for research. Coworkers of the D.E. Shaw Research Institute embarked on long timescale molecular dynamics calculations by using a supercomputer (Anton) that speeds up these lengthy calculations by orders of magnitude. In the first paper they analyzed the route of entry and mechanism of drug binding to G-protein-coupled receptors, using a number of available GPCR structures.[6] In the second paper the activation mechanism of the β_2-adrenergic receptor was studied, taking advantage of both agonist- and antagonist-bound structures.[7] There appears to be a relatively good match between protein dynamics at the microsecond scale and pharmacological

observations, such as ligand association and dissociation kinetics and the interconversion between an active and an inactive conformation.

The findings of the D.E. Shaw Research Institute bring me to the last point of this Chapter, i.e. the collaboration between all players in the world of drug discovery. Not too long ago drug discovery was an endeavor almost exclusive to big pharma. However, recently, government-sponsored initiatives, private charities and academic institutions have emerged that challenge this dominant position. Most notable example for me is TI Pharma, the Dutch public-private partnership that funded the research described in this thesis. Examples of privately-funded foundations that have a significant influence on global drug research and development today are e.g., the Bill & Melinda Gates Foundation, the Cystic Fibrosis Foundation and the Michael J. Fox Foundation.

The Bill & Melinda Gates foundation is a charity that invests heavily in research and development of new vaccines, drugs, and diagnostics. It is mainly focused on developing countries, which is reflected in their funding of drug discovery and development for the treatment of HIV, malaria, tuberculosis, and polio. The foundation collaborates directly with several institutions such as the Scripps Research institute (molecular pathways of vaccines), Oxford university (target discovery for prevention of malaria), and Harvard university (identification of host factors to control HIV infections). The Cystic Fibrosis Foundation supports cystic fibrosis research and clinical trials through grants and specialized treatment centers. Research efforts funded by the foundation resulted in identification of the gene responsible for cystic fibrosis as well as several therapies, both approved and in the pipeline, for the treatment of cystic fibrosis. An important industrial partner for the Foundation is Vertex Pharmaceuticals, which is developing several small molecule drugs for this life-threatening condition. The Michael J. Fox Foundation is another example of an institution with significant influence focused at a specific disease. The foundation is dedicated to finding a cure for Parkinson's disease by funding research efforts and the

development of therapies. Funded research ranges from biomarker discovery to translational research.

In addition to aforementioned charitable foundations, academic institutions are also emerging as important players in the drug discovery and development field. A typical example with significant breadth is the Vanderbilt center for neuroscience drug discovery. This institute aims to bring its findings beyond the proof-of-concept stage and move towards advanced toxicity testing in animals in at least three of their programs before beginning clinical trials.

In conclusion new ways of 'drug hunting' are being explored, and I am happy to have been actively involved in one of such initiatives. The emerging 'open source' approach to drug discovery would be the logical consequence of trends described in this thesis, such as the use of publicly accessible databases, the availability of open source software, and an increased data sharing between public and private partners. It would have the added benefit of a further maturation of the cheminformatics field.

7.3 References

(1) Katritch, V.; Jaakola, V.-P.; Lane, J. R.; Lin, J.; IJzerman, A. P.; Yeager, M.; Kufareva, I.; Stevens, R. C.; Abagyan, R. Structure-Based Discovery of Novel Chemotypes for Adenosine A2A Receptor Antagonists. *J. Med. Chem.* **2010**, *53*, 1799-1809.

(2) Carlsson, J.; Yoo, L.; Gao, Z.-G.; Irwin, J. J.; Shoichet, B. K.; Jacobson, K. A. Structure-Based Discovery of A2A Adenosine Receptor Ligands. *J. Med. Chem.* **2010**, *53*, 3748-3755.

(3) Salon, J. A.; Lodowski, D. T.; Palczewski, K. The Significance of G Protein-Coupled Receptor Crystallography for Drug Discovery. *Pharmacological Reviews* **2011**, *63*, 901 -937.

(4) Kolb, P.; Rosenbaum, D. M.; Irwin, J. J.; Fung, J. J.; Kobilka, B. K.; Shoichet, B. K. Structure-based discovery of beta2-adrenergic receptor ligands. *Proc Natl Acad Sci U S A* **2009**, *106*, 6843-6848.

(5) Gloriam, D. E.; Foord, S. M.; Blaney, F. E.; Garland, S. L. Definition of the G Protein-Coupled Receptor Transmembrane Bundle Binding Pocket and Calculation of Receptor Similarities for Drug Design. *Journal of Medicinal Chemistry* **2009**, *52*, 4429-4442.

(6) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences* **2011**, *108*, 13118 - 13123.

(7) Dror, R. O.; Arlow, D. H.; Maragakis, P.; Mildorf, T. J.; Pan, A. C.; Xu, H.; Borhani, D. W.; Shaw, D. E. Activation mechanism of the β2-adrenergic receptor. *Proceedings of the National Academy of Sciences,* **2011**.