Computers and drug discovery : construction and data mining of chemical and biological databases

Kazius, J.

Chapter 8

General Conclusions and Perspectives

The bioinformatics research that is described in the first chapters of this thesis broadens our knowledge of DNA variants that adversely affect GPCR functioning. Key GPCR variants from literature were discussed, a large database of GPCR variants was constructed, and differences between harmless and disease-causing variants were analysed to understand biological causes for adverse phenotypes and to further insights into important GPCR domains. The cheminformatics research that is documented in the last chapters of the thesis enhances our understanding of chemical causes for an adverse biological activity called mutagenicity. Methods of data mining for such causes were discussed, a set of mutagenic substructures was identified using a knowledge-based approach, and a new cheminformatics method was developed that could rapidly detect sophisticated substructures that cause mutagenicity.

New opportunities for further bioinformatics research are emerging from the field of human genetics. Chapter 2 discussed several types of polymorphic natural variants. The polymorphic character was recently also established for many large parts of human DNA, so-called copy number polymorphisms (CNPs)[1], which can occur zero to several times at one site in the human genome. We need to better understand their function of such CNPs because they affect more human DNA than all single nucleotide polymorphisms (SNPs) put together, and because these CNPs might have a stronger impact on phenotype than individual SNPs. New data on CNPs can spur studies to identify and characterise them and bioinformatics components will be critical in these processes. Experimental studies of the effects of CNPs *in vitro* and in man can be better supported by improved bioinformatics-aided functional predictions for the genes that are affected by CNPs. Opportunities will also arise for new bioinformatics analyses that focus on understanding the neutral, beneficial or adverse effects of CNPs.

Chapter 4 details a study of biological causes for adverse effects on human GPCR function. We found that larger changes at the amino acid level more often caused disease than relatively mild changes. In addition, we found that variants in transmembrane domains, and especially their buried parts, were more likely to adversely affect GPCR functioning. This analysis may be considered nearly exhaustive for the near future. In part, this assumption is based on the size of the dataset as virtually all reported missense variants were included. Additional relationships can therefore only appear when new data on hundreds of additional GPCR variants are reported, but it will likely take years before that many new variants are discovered, if at all. Until such a time, similar statistical investigations of human GPCR variants are unlikely to reveal other biological descriptors that affect disease than those revealed in Chapter 4. Alternatively, new insights may result from considering species differences and/or polymorphisms in GPCRs from other species, preferably other primates. Due to the speed and low costs of robotised sequencing, small and large genetic variants from other species may become available in the near future.

Another path to follow in this line of research is the study of variants in other drug-related protein families, such as proteases, cytochrome P450s, kinases, nuclear receptors or ion channels. For statistically significant results, however, such an analysis requires the availability of at least hundreds of harmless and harmful variants that are spread over different family members. This type of analysis could explain why some variants cause adverse effects, while others lack an impact on phenotype. An advantage of a family-specific approach, which we adopted in Chapter 4, is the promise of identifying key sites and domains in the protein family under investigation. Prior to such an analysis, a database needs to be constructed of family-specific variants. The prospective value is evident if such a database would integrate variants from a diversity of sources and then allow easy online access to the variants of the family, as was the case for the GPCR NaVa database that was reported in Chapter 3. An alternative method is to analyse natural genetic variants from all families at once. Although new general descriptors may turn out to affect the disease potential of genetic variants, such a general approach is unlikely to reveal structural insights.

A further course to pursue is to analyse the effects of variants on other forms of activity rather than adverse human phenotypes. For instance, random mutagenesis studies[2] may cheaply introduce genetic variants and express the variant proteins in systems *in vitro*, and thus enable tests for a specific kind of biological activity. As a result, statistical methods can be employed to determine variant-specific descriptors that affect this activity. Such an academic approach may, moreover, deepen our understanding of the molecular mechanisms and detailed interactions that underlie protein function. Examples of biological effects that could be studied are differences in expression, constitutive activity, protein breakdown and differences in the affinity for ligands, substrates, allosteric modulators, cofactors or other proteins. Finally, in an ideal situation, *in vitro* results from different proteins of the same

family could be compared, such as closely and distantly related receptors from the GPCR superfamily. These results can then prove, or disprove, the common hypothesis that similar distortions in related proteins have similar effects. This hypothesis could not be tested with the individual point mutation studies that have been published so far[3] because these studies were so dispersed that no proper statistical validation of this concept was possible. Moreover, statistical studies of such data may detail when findings from one protein can be extrapolated to other family members in a reliable fashion.

The statistical analysis of biological causes in Chapter 4 seems relatively straightforward in comparison to the statistical analysis of chemical causes in Chapter 7. It is nevertheless unlikely that further biological insights could result from more complex statistical analyses. In Chapter 4, all descriptors could be analysed individually because of the relatively low number of descriptors that were analysed for association with disease predisposition. Moreover, those biological descriptors that did correlate with adverse effects did not correlate with each other. As a result, no descriptor selection was needed in this study. In contrast, in Chapter 7 the amount of substructure descriptors was high and many descriptors correlated with each other. Here, a selection method was warranted to enable interpretation of an acceptable set of descriptors. Another difference was that even a combination of biological descriptors in Chapter 4 did not reduce the overall error in classification and prediction, while this was the case in Chapter 7 when the set of chemical descriptors was grown. For that reason, the use of different data mining methods in Chapter 4 did not provide added value over individual statistical tests.

Chapter 5 details guidelines for the validation of automated methods that provide predictive models and these guiding principles can be useful for data mining techniques in bioinformatics, cheminformatics and other disciplines. One important criterion that should be taken into account during the validation of future predictive models is the use of repeated cross-validations with unstratified splits into training sets and large test sets. The second important feature of a predictive model is the definition of an applicability domain (AD). In cheminformatics studies, the AD of a model estimates whether reliable predictions are possible for new compounds according to the 'chemical space' of the training set. For instance, if a predictive model is trained and built using only rigid compounds with few hydrophilic groups, like steroids, then it is highly unlikely that this model can reliably predict the activity of flexible compounds with hydrophilic sugar moieties. However, the mere definition of an AD is not enough because the size of the 'chemical space' that it covers needs to be quantified as well. Researchers should therefore report the fraction of public databases such as ChemBank[4], PubChem[5] or DUD[6] that are covered by the 'chemical space' of an AD. If a model can only make accurate predictions for steroids, then the quantification of its AD will be very small because only small fractions of these databases consist of steroids. The

future challenge for cheminformaticians is to provide an intuitive means of defining an AD that can be, and will be, used for any predictive method and any form of biological activity. Even in the field of bioinformatics, the definition and use of ADs are relatively untouched issues and this provides a knowledge gap that must be explored. In bioinformatics, ADs could prove valuable in estimating whether or not properties or activities can be predicted with any reliability for a new protein.

The cheminformatics part of the thesis aimed at linking biological activity to chemical structure in the form of structure-activity relationships. Chapters 6 and 7 focus on chemical causes for the adverse effect called mutagenicity, which can result from a tight binding to DNA by a compound or one of its metabolic derivatives. One application that is specific to the identified set of toxicophores for mutagenicity is the combination of these toxicophores with methods for metabolism prediction. Although few lead compounds will possess mutagenic toxicophores, this does not mean that their major metabolites are also free of toxicophores. Metabolites are often not directly evident from the molecular structure of a lead compound, but computational methods are available to predict major metabolites[7, 8]. These methods have, however, not directly considered the toxicity potential of these metabolites. As the metabolites of such leads are more likely than the leads themselves to possess toxicophores, mutagenicity prediction on the predicted major metabolites of leads may aid in prioritising chemical modifications or even compound classes for follow-up. It would be of interest when such a combined method would prove more accurate than only a toxicophore-based approach to mutagenicity prediction.

Importantly, the results from Chapter 7 show that computerization is possible for a huge part of the knowledge discovery process of Chapter 6, that is, the substructure mining for molecular substructures that cause mutagenicity. Given the recent public availability of large databases of chemicals with experimentally determined activities, numerous scientific research opportunities consist of performing similar analyses to establish whether substructures are equally useful in explaining other biological activities, adverse or beneficial. In addition, the same method can be tested for its ability to identify structural determinants for physico-chemical properties.

Although substructure mining has already proven its use, perhaps the method can consider more detailed chemical properties. Potentially interesting properties that have not yet been considered during substructure mining are the partial electronic charge, chirality, atom size, hydrogen-bonding/accepting character and electron-withdrawing/donating character of atoms. The consideration of such extra properties requires that the representation of compounds is significantly extended prior to substructure mining. Another potentially important extension to substructure mining is to enable a search for three-dimensional substructures. One obvious benefit over two-dimensional approaches would be the ability to distinguish between enantiomers (compounds that only differ in chirality). Another

fascinating path would be to explore multiple conformational states of the compounds in the mining process and select the state(s) most relevant for binding. Although such exploration of additional chemical information will come at a computational cost[9], the potential benefits are large.

In light of the above possibilities of the substructure mining method of Chapter 7 and the GPCR-focussed studies of Chapters 2, 3 and 4, it appears especially interesting and promising to automatically discern substructures that are required for binding and/or activation of GPCRs. Such an analysis may result in the detection of substructures that bind selectively to a particular receptor subtype. In addition, substructures may be detected that bind to many different GPCRs in a non-selective way. Of course, substructures can also be uncovered that diminish binding. Next to extracting and cataloguing substructure-specific knowledge cheminformatics methods may moreover predict whether or not a new compound will bind a particular GPCR, given that sufficient binding data are around to construct a reliable model and define its AD.

Recent cheminformatics databases store results from high throughput screening (HTS) campaigns that quantify the binding to a particular protein target for a large set of compounds[4, 5]. With the aid of cheminformatics analyses, these results may reveal chemical characteristics that promote binding to the studied target and these characteristics may be useful in understanding and predicting the binding of untested compounds. Moreover, these screening campaigns may also suggest compounds for targets that are not yet screened. Such extrapolations may be possible thanks to bioinformatics that exploit the available protein sequences of the screened targets and almost all other genes in the human genome[10]. Crystal structures[11, 12] or homology models thereof can further support the extrapolation to other protein targets. The comparison of sequences or structures using bioinformatics methods has long been possible, but the wealth of chemical binding data is new. The binding data of different compounds to various GPCRs, for instance, can be used to predict the binding of new compounds to the screened GPCRs (on the basis of chemical descriptors). These binding data may also be used to estimate the binding of these compounds to untested GPCRs (on the basis of

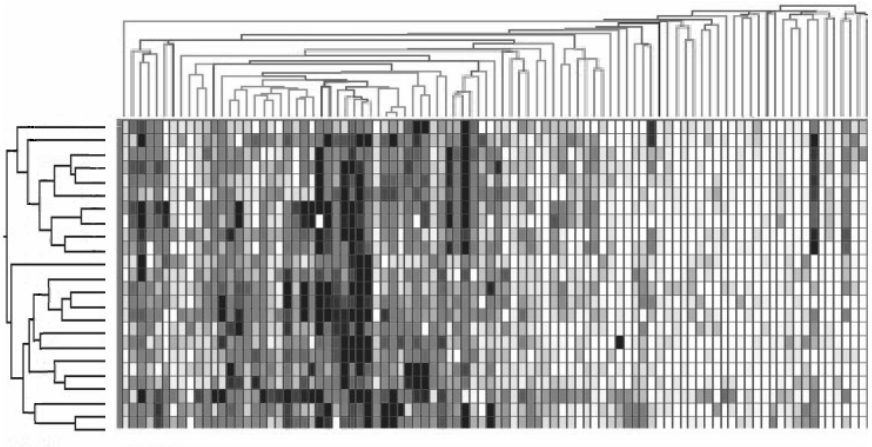

Figure 8.1. Matrix of compounds (rows, clustered by their chemical descriptors) and assays (columns, clustered by their binding profiles), where each point in the matrix quantifies a compound's inhibitory capacity in an assay.

sequence descriptors). One method for such purposes was called proteochemometrics[13] and it makes use of three-dimensional descriptors to represent both the compounds and the protein binding sites. The starting point for such predictions are matrices where each point in the matrix represents the binding capacity, or inhibitory capacity, of a given compound in a given assay, see also Figure 8.1[14]. Retrospective proteochemometrics studies have shown promising results for closely related GPCRs, both in terms of predicting affinities for new ligands to known targets and for known ligands to new targets[15]. Amongst others, an approach such as this might be of particular value in predicting new compounds for orphan GPCRs, as no natural ligands are known yet for these receptors. A drawback of this approach is that three-dimensional conformations of ligands and binding sites are assumed. Another limitation is that proteochemometrics is only applicable to proteins binding sites that are very similar in shape. A different method, called biological spectra analysis[14, 16], linked chemical structure directly to numerous biological activities. For a series of analogous compounds with similar biological activity profiles, very similar new analogues were tested and they also showed the same profile of biological activities[14]. In a follow-up study[16], compounds with known activities in man were then described according to their chemical descriptors and according their *in vitro* activity profiles. Although principles of chemical similarity proved to perform well earlier[14], effects in man correlated better with *in vitro* activity profiles then with similarity on the basis of chemical descriptors. As *in vitro* activity profiles require experimental work, it would be advantageous if biologically relevant chemical descriptors were found as they can rapidly even be computed for non-existing compounds. A drawback of the biological spectra analysis studies is that neither the software nor the examined data are freely available to other researchers. It therefore still remains to be confirmed whether similar analyses on publicly available datasets will give equally promising statistical results. Moreover, none of the above models have specified their applicability domain and they therefore fail to quantify when accurate predictions are possible, either for new compounds or for new proteins.

To test whether such limitations can be overcome, a surge of public data is a good start. For starters, the newly available data can be used to test current bioinformatics and cheminformatics methods and thereby distinguish truly predictive from overfitted methods. This could help to decide which of the current approachs performs best for a new target or a new set of compounds. Only a surge of new data is not enough. As discussed in this thesis, important limitations of current predictive methods also include the inability to provide explanations that are interpretable to experimentalists, and the absence of intuitive reliability estimates for new predictions. New bioinformatics and cheminformatics approaches are thus needed to determine when we can *reliably* extrapolate from these data. New methods should broaden our understanding of which functionalities of compounds and proteins are important for the molecules they interact with. Awareness of the possibilities and impossibilities (see the

above paragraph) of current and future methods is vital. Furthermore, more accurate and more insightful computational approaches can increase the resource-efficiency of hypothesis-driven research. Methods from bioinformatics and cheminformatics are already valuable - and may become priceless - to many scientific disciplines, especially as automated methods for knowledge discovery.

REFERENCES

1. Redon, R.; Ishikawa, S.; Fitch, K. R.; Feuk, L.; Perry, G. H.; Andrews, T. D., et al. Global variation in copy number in the human genome. *Nature* **2006,** 444, 444-54.
2. Beukers, M. W.; van Oppenraaij, J.; van der Hoorn, P. P.; Blad, C. C.; den Dulk, H.; Brouwer, J., et al. Random mutagenesis of the human adenosine A2B receptor followed by growth selection in yeast. Identification of constitutively active and gain of function mutations. *Mol Pharmacol* **2004,** 65, 702-10.
3. Beukers, M. W.; Kristiansen, I.; AP, I. J.; Edvardsen, I. TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol Sci* **1999,** 20, 475-7.
4. Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S., et al. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* **2007,** 18, 18.
5. Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V., et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2006,** 34, D173-80.
6. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J Med Chem* **2006,** 49, 6789-801.
7. Button, W. G.; Judson, P. N.; Long, A.; Vessey, J. D. Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *J Chem Inf Comput Sci* **2003,** 43, 1371-7.
8. Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T., et al. MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J Med Chem* **2005,** 48, 6970-9.
9. Kuramochi, M.; Karypis, G. Discovering frequent geometric subgraphs. *Information Systems* **2007,** 32, 1101-1120
10. Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G., et al. The sequence of the human genome. *Science* **2001,** 291, 1304-51.
11. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A., et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000,** 289, 739-45.
12. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S., et al. High-Resolution Crystal Structure of an Engineered Human {beta}2-Adrenergic G Protein Coupled Receptor. *Science* **2007,** 25, 25.
13. Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J. E. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol Pharmacol* **2002,** 61, 1465-75.
14. Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc Natl Acad Sci U S A* **2005,** 102, 261-6.
15. Lapinsh, M.; Prusis, P.; Uhlen, S.; Wikberg, J. E. Improved approach for proteochemometrics modeling: application to organic compound--amine G protein-coupled receptor interactions. *Bioinformatics* **2005,** 21, 4289-96.
16. Fliri, A. F.; Loging, W. T.; Volkmann, R. A. Analysis of System Structure-Function Relationships. *ChemMedChem* **2007,** 19.