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

Kazius, J.

Summary

Bioinformatics and cheminformatics are research fields that are largely separated, despite a degree of overlap. As described in the introductory Chapter 1, both disciplines include the construction of biological and chemical databases. Constructed databases can be analysed via methods of data mining to summarise their content and provide new scientific insights. In both disciplines, mining such databases can objectively reveal new causal relationships between chemical and/or biological phenomena, such as chemical and biological causes for adverse effects in man. Reliable insights can then be used to estimate the biological effects of new compounds, genetic variants, proteins and more, often in a resource-efficient manner. Chapter 1 also details the general scope and outline of this thesis.

Several natural genetic variants, both rare mutations and common polymorphisms, produce pathogenic phenotypes. A particular diversity of harmful and harmless genetic variants occurs in human G protein-coupled receptors (GPCRs). The GPCR superfamily is large and regulates a plethora of important physiological processes by transducing extracellular signals over cell membranes. Moreover, GPCRs serve as prominent drug targets. For these reasons, natural variants in human DNA are likely contributors to inter-individual diversity. Chapter 2 therefore discusses the possibilities and limitations of GPCR variants as causes for disease. One observation was that, in comparison to rare GPCR mutations, few GPCR polymorphisms have been convincingly linked to disease susceptibility. Copy number polymorphisms (CNPs) are large DNA segments that vary in how often they are repeated in the human genome. CNPs can contain complete GPCR genes and may therefore be stronger than single nucleotide polymorphisms (SNPs) as predisposition factors for clinical manifestations.

The subject of Chapter 3 is the design and construction of a database of GPCR variants, which is a prerequisite for discovering why some variants are disease-related while others appear harmless. As mutation data are highly dispersed over numerous sources,

investigators of GPCRs would be aided by a single resource that integrates data on natural variants. We therefore constructed the GPCR NaVa database (http://nava.liacs.nl). Where available, variants in this new database include information on their location in the DNA (and in the protein sequence), the involved nucleotides (and amino acids), the average frequency of each allele in the human population, reported disease associations and references to public databases and the scientific literature.

As a conclusion to the Chapters 2 and 3, Chapter 4 aims to relate computable variant-specific properties to pathogenic effects. The wealth of data on natural genetic variants from the GPCR NaVa database was therefore studied through statistical approaches. As expected, olfactory GPCRs occur more often in copy number variants than non-olfactory GPCRs. Non-olfactory GPCRs contain fewer silent polymorphisms and more missense polymorphisms than olfactory GPCRs. Our analysis suggested that stronger risk factors for disease are either variants that cause more radical amino acid substitutions or variants that occur in transmembrane domains, particularly if they occur in the receptor core according to the three-dimensional structure of these transmembrane domains. Very conserved transmembrane positions also appeared to contain most disease mutations. The analysis of the readily accessible data in the GPCR NaVa database showed that disease predisposition of variants depended on several structural features. At the same time, these findings confirmed the relative importance of particular domains and key amino acids for normal GPCR performance.

Disease-causing features of genetic variants can give insights into receptor functioning, but these insights cannot improve the clinical situation of affected patients. On the other hand, chemical features that correlate with biological activities can be considered in the molecular design of pharmaceuticals and thus have potential to benefit patients. In that sense, data mining for chemical rather than biological descriptors may have additional value. Chapter 5 discusses concepts of commonly used techniques to perform data mining, as well as their advantages and pitfalls. Examples were taken from the cheminformatics field to link chemical structure to biological activity, toxicity in particular, by deriving structure-activity relationships. However, the methods of data mining that are discussed in Chapter 5 are independent of the selected chemical or biological descriptors, and may therefore also be useful for bioinformatics research. In data mining, the available dataset is commonly split into a training set, for learning relationships in the form of a model, and a test set, for validating whether these relationships have predictive potential. When relationships are found, it is important to know if they can predict the investigated biological effect. The statistical results that are thought to reflect the reliability of the extracted relationships can be drastically affected by the method of splitting a dataset and the by distribution of the available data. Guidelines for robust validation consist of fully separating test data from training data,

repeated cross-validation with large random test sets and the definition of an applicability domain.

Chapter 6 discusses the substructures that resulted from a semi-computational effort to extract structure-activity relationships for a type of toxicity called mutagenicity. Compounds are often less attractive as marketable drugs when they are mutagenic, that is, when they can introduce mutations into DNA that may be harmful. It is therefore important to increase the reliability and accuracy of mutagenicity predictions by identifying novel toxicophores, which are toxicity-associated chemical substructures. A large mutagenicity dataset was therefore constructed and then analysed to derive and validate toxicophores. An initial substructure search of this dataset showed that most mutagens were detected by applying only eight general toxicophores. From these eight, more specific toxicophores were derived and approved by employing statistical criteria in combination with chemical and mechanistic knowledge from the scientific literature. The resulting twenty-nine toxicophores for mutagenicity are useful for the risk assessment and the design of libraries of potential pharmaceuticals. Examples of a newly identified toxicophore and detoxifying subtructure are the polycyclic planar system and the trifluoromethyl substituent, respectively.

Substructure mining algorithms are important drug discovery tools since they can extract biologically relevant substructures in a rapid and fully automated way. Methods to date, however, only considered a part of all chemical information that is present within a dataset of compounds. Therefore, the overall aim of Chapter 7 was to enable more exhaustive data mining within cheminformatics by designing methods that detect all substructures of any size, shape and level of chemical detail. A means of chemical representation was developed that uses general and/or highly specific features. As a proof-of-concept, the efficient, versatile graph mining system Gaston learned substructures of any size and shape from a mutagenicity dataset that was represented in this elaborate manner. From these substructures, we automatically extracted a set of only six nonredundant, discriminative substructures that represent relevant biochemical knowledge, as detailed in Chapter 6. These results demonstrate the individual and synergistic importance of elaborate chemical representation and mining for nonlinear substructures. As a result, the combination of elaborate chemical representation and Gaston provides an excellent method for two-dimensional substructure mining as this recipe systematically explores all substructures in different levels of chemical detail.

General conclusions about the research described in this thesis are drawn in Chapter 8. In short, this thesis reports several causes for adverse biological effects. In case of GPCR variants, bioinformatics analyses revealed the importance of their position in the receptor structure: variants in transmembrane domains more often caused disease, especially if they were positioned in the receptor core. In case of chemicals, toxicophores were identified as causes through a promising cheminformatics method. Such conclusions were supplemented by an outlook on promising research areas.