UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 608

Chemogenomics: Models of Protein-Ligand Interaction Space

HELENA STRÖMBERGSSON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2009

ISSN 1651-6214
ISBN 978-91-554-7430-0
urn:nbn:se:uu:diva-89299

Dissertation presented at Uppsala University to be publicly examined in C8:305, Biomedical Centre, Uppsala, Friday, March 27, 2009 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Strömbergsson, H. 2009. Chemogenomics: Models of Protein-Ligand Interaction Space. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 608. 54 pp. Uppsala. ISBN 978-91-554-7430-0.

The large majority of the currently used drugs are small molecules that interact with proteins. Understanding protein-ligand recognition is thus central to drug discovery and design. Improved experimental techniques have resulted in an immense growth of drug target information. This has stimulated the development of chemogenomics and proteochemometrics (PCM) that take target information as well as ligand information into account to study the genomic effect of potential drugs.

This thesis is concerned with modeling protein-ligand recognition, and the aim is to develop models that generalize to the entire protein-ligand space. To this end, protein-ligand interaction data has been extracted and manually curated from public databases, protein and ligand descriptors have been computed, and predictive models have been induced with machine-learning methods.

An introduction to chemogenomics, machine learning, and PCM modeling is given in the thesis summary, which is followed by five research papers. Paper I shows that it is possible to induce interpretable models with a non-linear rule-based method, and paper II demonstrates that local descriptors of protein structure may be used to induce PCM models that cover proteins differing in sequence and fold. In paper III, such local descriptors are used to induce a PCM model on a large dataset that includes all major enzyme classes. This demonstrates that the local descriptors may be used to induce generalized models that span the entire known structural enzyme-ligand space. Paper IV describes a step towards proteome-wide PCM models, and shows that it is possible to predict high- and low-affinity complexes using a set of protein and ligand descriptors that do not require knowledge of 3D structure. Finally, paper V presents a method to visualize and compare protein-ligand chemogenomic subspaces, which may be used to predict unwanted cross-interactions of drugs with other proteins in the proteome.

Helena Strömbergsson, Department of Cell and Molecular Biology, The Linnaeus Centre for Bioinformatics, Box 598, Uppsala University, SE-75124 Uppsala, Sweden

ISSN 1651-6214
ISBN 978-91-554-7430-0
urn:nbn:se:uu:diva-89299 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-89299)

To my family

List of Papers

This PhD thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Strömbergsson H, Prusis P, Midelfart H, Lapinsh M, Wikberg JES and Komorowski J: **Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions**. *Proteins: Structure, Function, and Bioinformatics,* 2006, **63**(1):24-34.

II Strömbergsson H, Kryshtafovych A, Prusis P, Fidelis K, Wikberg JES, Komorowski J and Hvidsten TR: **Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures**. *Proteins: Structure, Function, and Bioinformatics*, 2006, **65**(3):568-579.

III Strömbergsson H, Daniluk P, Kryshtafovych A , Fidelis K, Wikberg JES, Kleywegt GJ and Hvidsten TR. **An interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space**. *The Journal of Chemical Information and Modeling* 2008*,* **48**(11): 2278-2288.

IV Strömbergsson H, Lapinsh M, Wikberg JES and Kleywegt GJ. **Towards proteome-wide interaction models using the proteochemometrics approach**. *In manuscript.*

V Strömbergsson H and Kleywegt GJ. **A chemogenomics view on protein-ligand subspaces**. *Accepted for publication in BMC Bioinformatics.*

Reprints were made with permission from the publishers.

Contents

Abbreviations

DNA	Deoxyribonucleic acid
EC	Enzyme Classification
FN	False Negatives
FP	False Positives
GPCR	G-protein-coupled receptor
IC_{50}	Inhibitory Concentration
K_d	Dissociation constant
K_i	Inhibition constant
MC	Melanocortin
PCM	Proteochemometrics
PDB	Protein Data Bank
PLS	Partial Least Squares
QSAR	Quantitative structure-activity relationship
RMSEP	Root-Mean-Square Error of Prediction
ROC	Receiver-Operating Characteristic
SVM	Support-Vector Machines
TN	True Negatives
TP	True Positives

1. Introduction

This PhD thesis is concerned with modeling protein-ligand interaction, and the aim is to develop models that generalize to the entire protein-ligand space. Molecular recognition is central to all processes in living cells. Antigens binding to antibodies, proteins binding to DNA, and drugs binding to their targets in the body are all examples of molecular recognition between two biomolecules. This work concerns protein-ligand interactions. *Proteins* are large biomolecules expressed from genes through RNA, and composed of amino acids arranged in a linear chain. Most proteins fold into a unique 3D structure which is essential for their function. The chief characteristic of proteins that allows their diverse set of functions is their ability to bind other molecules. The region of the protein responsible for binding another molecule is known as the *binding site* and is often a depression or "pocket" on the molecular surface. A *ligand* is usually a low-molecular-weight chemical compound that is able to form a complex with a biomolecule to serve a biological purpose. A ligand that serves a therapeutical purpose is called a *drug*, and the biomolecule the drug interacts with is called its target. This drug-target recognition is essential for drug function and specificity and an example of a drug-target interaction is shown in Figure1. Traditional drug discovery has been focused on drug design and optimization. However, the wealth of biological information available in the post-genomics era has stimulated the development of new fields such as chemogenomics and proteochemometrics. These disciplines are closely related and intend to obtain a proteome-wide understanding of molecular recognition.

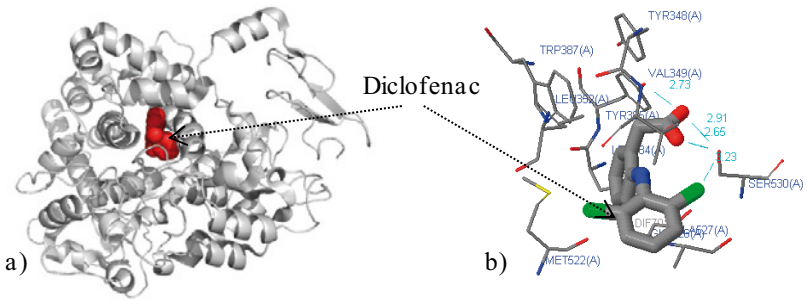


Figure 1. (a) The 3D structure of the enzyme cyclooxygenase in complex with the pain killer Diclofenac® (PDB code 1PXX [1]). (b) Details of Diclofenac®-cyclooxyenase interactions.

1.1 Chemogenomics

Chemogenomics is an emerging approach to drug discovery that lies in the interface of biology, chemistry and informatics [2] (Figure 2). The Human Genome Project [3] was enabled by improved experimental methodologies and it has resulted in a wealth of information on putative targets. Chemogenomics involves measuring the effect of chemical libraries on whole biological systems or screening putative drugs against a selected group of targets. This shifts the focus from traditional drug optimization to drug-target recognition on a proteome-wide basis.

Targets are generally divided into seven categories according to a classic survey of the pharmacopeia by Drews & Ryser [4, 5]. The large majority of the drugs interact with proteins and most of those are receptors or enzymes. The ultimate goal in chemogenomics is to measure the response of all proteins encoded by the human genome [6] to all possible chemical compounds. However, the size of the protein and ligand spaces renders any systematic *in vitro* or *in silico* approach impossible. The *protein space* consists of all expressed gene products, and due to events such as splicing and post-translational modification the number of proteins is in general larger than the number of genes. In humans, the number of genes is approximately 26 000 [7] while the number of expressed proteins is estimated to be more than one million [8]. The *ligand space* is composed of all possible reasonably sized molecules, up to about 600 Da in molecular weight, that contain atoms commonly found in drugs. This space is very large and a commonly quoted estimate is 10^{62} compounds [9]. Since the protein-ligand space is immense, interaction data can only be generated for a minute fraction of all possible protein-ligand complexes, which makes the chemogenomics data grid extremely sparse.

Chemogenomics-based drug-discovery approaches can be divided into three categories that focus on ligands, targets, or the combination of ligands and targets [10]. In *ligand-based* chemogenomics, the basic underlying paradigm is that similar ligands share similar biological activity. Efforts are therefore made to annotate chemical libraries with biological information such as targets, *in vitro* affinity, and toxicology. Ligands commonly bind to more than one target, and cross-interactions between a drug and other proteins in the proteome may cause unwanted side effects. Therefore, it is of interest to design chemical libraries that are specific for a certain target or target family and this has been attempted using machine-learning methods [11]. *Target-based* chemogenomics seeks to obtain selective ligands by comparison of targets from the same family. In sequence-based comparisons, multiple alignments play a central role. Interestingly, Surgand *et al*. [12] have shown that residues in the binding site of most G-protein-coupled re-

ceptors (GPCRs) can be extracted and concatenated into a short ungapped sequence and that the resulting phylogenetic tree is almost identical to the full sequence-based tree. Comparison of binding sites at the sequence level has been successfully used to obtain selective ligands against subfamilies of GPCRs [13, 14]. Structure-based comparisons of binding sites require a protein 3D structure of good quality and are thus limited to certain target families. An *et al.* [15] have recently proposed a rapid method to compare potential ligand-binding envelopes. Their study involved clustering by envelope similarity and a first draft of the human 'pocketome' was proposed. As opposed to the sequence-based study reported by Surgand *et al.*, the ligand-envelope-based phylogenetic tree only partially matched the corresponding sequence-based tree. Chemogenomic approaches focused on *protein-ligand* interactions use 2D and 3D search methods. In 2D protein-ligand chemogenomics, a matrix is set up where each row contains molecular descriptors of a protein-ligand complex linked to some experimentally measured biological activity [12]. Models are induced by statistical or machine-learning methods and are applied for interpretation and prediction of biological activity. This approach was named *Proteochemometrics* by Wikberg *et al.* [16] and it will be described in the following chapters. 3D search methods mainly involve high-throughput docking, where a compound library is docked into the active site of a target. Much effort is spent on developing improved scoring functions that are able to correctly rank target-ligand poses. For instance, Marcou & Rognan [17] have proposed the use of molecular interaction fingerprints for prioritizing the most relevant poses, and Nervall *et al*. [18] have applied molecular-dynamics simulations in combination with a linear interaction method to predict ligand-binding free energies.

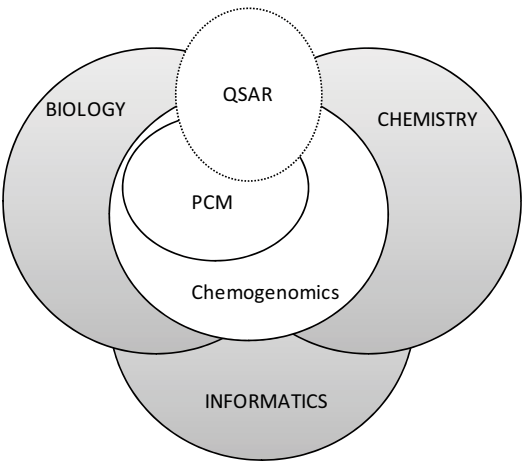


Figure 2. Chemogenomics is an inter-disciplinary field at the interface of biology, chemistry and informatics. Quantitative structure-activity relationship (QSAR) modeling and proteochemometrics (PCM) are disciplines within chemogenomics.

1.2 Proteochemometrics Modeling

Proteochemometrics (PCM) [16] is an extension of traditional quantitative structure-activity relationship (QSAR) modeling (Figure 2). In *QSAR*, one series of ligands is characterized by various descriptors. Each ligand is linked to an experimentally measured biological activity and models are induced by statistical or machine-learning methods. The model is used for predictions of the biological activity (typically binding affinity) of new ligands. An interpretation of the model may provide information on which ligand properties are important for prediction of biological activity and may provide guidance for drug design and optimization. Since QSAR is limited to the study of the molecular recognition of one series of ligands interacting with one target, new biological activity measurements have to be made for each model. In *PCM*, the ligand-descriptor matrix (a matrix is a rectangular array of numbers arranged in rows and columns) is extended by protein descriptors. Each row thus contains a unique combination of protein and ligand descriptors linked to some biological activity and this allows for models that span over several series of ligands and targets. Figure 3 illustrates this fundamental difference between QSAR and PCM. A clear advantage of the PCM approach is that activity data stored in databases and in the literature can be reused and that PCM models could in principle generalize to allow predictions of cross-interactions of a given ligand to other proteins in the proteome.

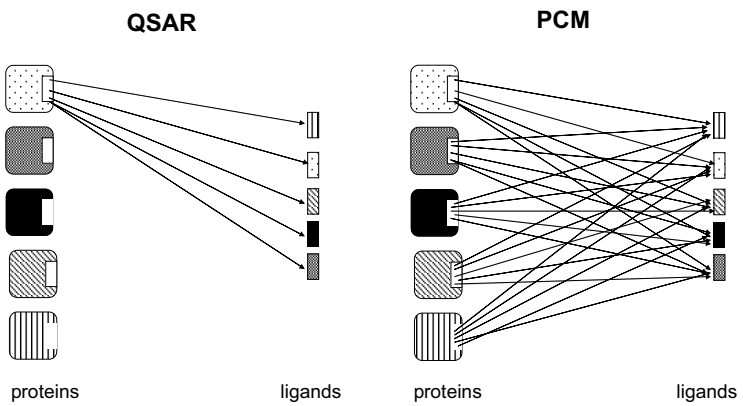


Figure 3. A QSAR model is based on one series of ligands interacting with one target, whereas in PCM several series of ligands and proteins are combined into a single model.

The first PCM models were reported in 2001 by Prusis *et al.* [19] and Lapinsh *et al.* [20], covering melanocortin and adrenergic GPCRs and their ligands, respectively. The receptors and ligands were described by the presence or absence of certain variable protein and ligand parts, and predictive models were computed by Partial Least Squares (PLS) [21]. The PCM approach has since been applied to other drug targets such as amine receptors [22], HIV proteases [23] and cytochrome P450 enzymes [24]. The binary descriptors used in the first studies have been extended to include others such as local descriptors of protein structure [25, 26] and 3D grid-independent ligand descriptors (GRIND) [27, 28]. PCM models have mainly been induced with PLS, but other methods such as rough sets (applied in paper I) and support-vector regression (applied in paper III) have been successfully used as well. An overview of descriptors and methods used in PCM will be provided in the following chapters.

The PCM modeling process is illustrated as a work-flow diagram in Figure 4. The first and perhaps the most important step is to define the purpose of the study. Questions should be answered regarding, for instance, the target families to be included in the model, data availability, and the type of protein and ligand descriptors to compute. This is followed by data collection from public or commercial databases, the literature, or in-house experiments. This data typically consists of information on two interacting biomolecules, such as a receptor and a ligand, linked to some experimentally measured biological activity. Protein and ligand descriptors are then computed and merged into a PCM dataset that consists of protein and ligand descriptors, and a biological activity that is to be predicted. Finally, a model is induced by statistical or machine-learning methods. The model is validated, either *in silico* using, for instance, cross-validation and an external test set, or *in vitro* by binding-affinity assays. The latter is necessary to confirm any biological activity predictions of new biomolecule pairs. Interpretation of the model may lead to an increased understanding of molecular recognition and to new hypotheses that start new loops in the PCM modeling process.

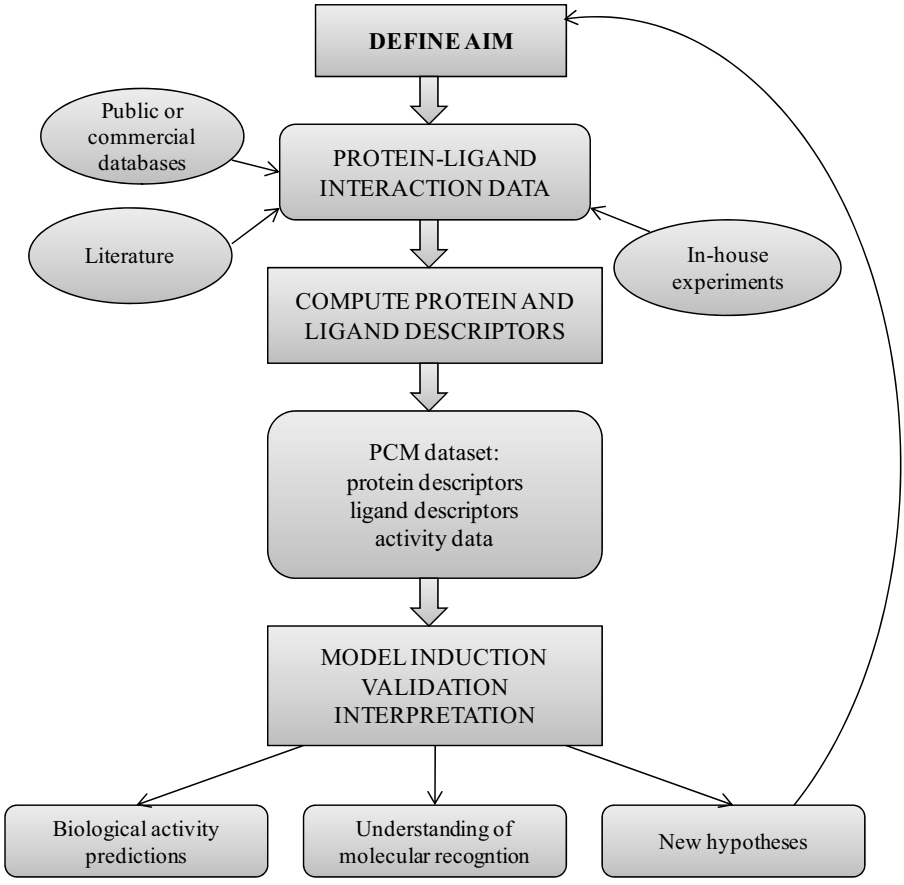


Figure 4. Protochemometrics modeling work-flow.

1.3 Outline of the Thesis

Chapter 2 describes the data sources used in the studies that are included in this thesis, including an overview of protein and ligand descriptors and binding-affinity data. Chapter 3 describes the methodology that has been used and includes a very brief overview of machine learning and validation. Chapter 4 summarizes the results of paper I to V and chapter 5 contains some concluding remarks on PCM and chemogenomics.

2. Proteochemometrics data

A PCM model can be induced from any dataset pertaining to pairs of biomolecules for which descriptors can be computed and biological activity can be measured. PCM was developed to enhance and accelerate the drug-discovery process and to date the large majority of all reported PCM models has been applied to the prediction of protein-ligand binding affinity. This chapter will therefore be concerned with the descriptors and biological activity data used in PCM modeling of protein-ligand interaction.

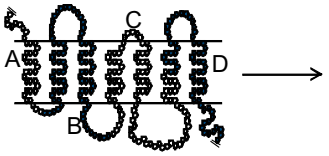
Each row in a PCM dataset consists of descriptors of one ligand that interacts with one protein, and some experimentally measured biological activity. Table 1 shows a toy example of a PCM dataset. In this dataset, there are two protein targets “A” and “B” that interact with ligand “1”, “2” and “3”. The ligands are described by two descriptors: molecular weight and number of rotatable bonds. The proteins are described in a Boolean fashion by the presence (“1”) or absence (“0”) of subunit alpha and beta. The biological activity in this example is the experimentally measured binding affinity of each protein-ligand pair.

Table 1. A toy example of a PCM dataset.

	Ligand descriptors		Protein descriptors		Biological activity
Name	**Molecular weight (Da)**	**Number of rotatable bonds**	**Subunit alpha**	**Subunit beta**	**Binding Affinity**
Protein A- ligand 1	500	9	1	0	5.9
Protein A- ligand 2	750	15	1	0	6.5
Protein A - ligand 3	150	1	1	0	2.6
Protein B - ligand 1	500	9	0	1	8.4
Protein B - ligand 2	750	15	0	1	9.9
Protein B - ligand 3	150	1	0	1	3.0

2.1 Protein Descriptors

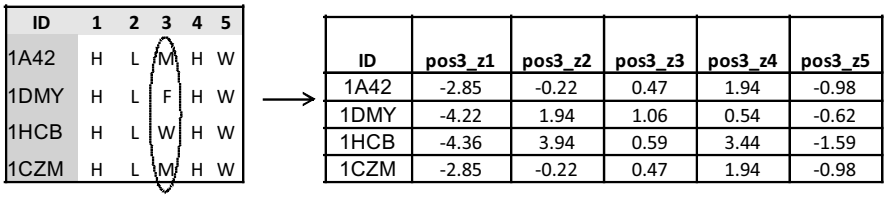
In the first reported PCM models [19, 20, 29], the proteins were described by *binary* attributes, where each column represented the presence or absence of a protein domain, or of amino acids at certain positions in a multiple sequence alignment. In PCM, binary attributes have been used to describe domain and amino-acid composition in melanocortin (MC) [19, 29, 30] and adrenergic [20, 30] GPCRs. Figure 5 shows a schematic representation of a 7-transmembrane helix MC GPCR. The chimeric receptor is composed of four domains (A-D) each originating either from the wild type MC1 or MC3 receptors. The composition of the chimeric receptor is described by four columns (A-D), where "0" means that the domain originates from MC1 and "1" that the domain originates from MC3.



ID	A	B	C	D
Protein 1	0	1	0	1

Figure 5. A chimeric receptor that consists of four domains (A-D). The origin (MC1 or MC3) of each domain is described by four descriptors. In this example, the A domain originates from MC1, the B domain from MC3 and so on.

In later PCM work, *z-scale* descriptors [31] were used which are principal components computed from a set of physicochemical amino-acid properties. To describe a protein with these descriptors, a multiple sequence alignment is performed, and amino-acid variations at certain positions are described by five z-scales. Those z-sales represent essentially hydrophobicity/hydrophilicity (z1), steric/bulk properties and polarizability (z2), polarity (z3), and electronic effects (z4 and z5). The z-scale descriptors have been applied to describe amino-acid variations at certain positions in amine GPCRs [22, 27], melanocortin GPCRs [32, 33], HIV proteases [23, 34, 35], antibodies [36], and cytochrome P450 enzymes [24]. Figure 6 shows a small part of a multiple alignment of four proteins. The amino acids at the variable position 3 are described by five z-scales.



ID	1	2	3	4	5
1A42	H	L	M	H	W
1DMY	H	L	F	H	W
1HCB	H	L	W	H	W
1CZM	H	L	M	H	W

ID	pos3_z1	pos3_z2	pos3_z3	pos3_z4	pos3_z5
1A42	-2.85	-0.22	0.47	1.94	-0.98
1DMY	-4.22	1.94	1.06	0.54	-0.62
1HCB	-4.36	3.94	0.59	3.44	-1.59
1CZM	-2.85	-0.22	0.47	1.94	-0.98

Figure 6. An example of z-scale descriptors derived from a multiple sequence alignment.

Both binary descriptors and z-scale descriptors are very useful in PCM modeling. They are easy to interpret and allow for interpretations such as: "protein domain A" or "amino acid B" are very important for high protein-ligand binding affinity. However, since both descriptor types require a multiple sequence alignment, PCM models based on these descriptors are limited to proteins that are rather similar in sequence and/or structure.

Alignment-independent descriptors can be computed from the amino-acid sequence or the 3D structure of a protein. The PROFEAT [37] server calculates more than 1400 descriptors from the primary structure. The algorithmic implementation of these descriptors is described in detail in the PROFEAT manual [38]. These descriptors have mainly been used to classify proteins by function, but can also be applied for modeling protein-ligand interaction. Descriptors based on amino-acid composition, distribution and transition have been used to classify protein-ligand binding-affinity values in paper IV and to compare protein-ligand spaces in paper V. The selected descriptors were amino-acid frequencies and a set of 187 descriptors proposed by Dubchak *et al.* [39]. These descriptors are computed from seven amino-acid properties whose possible values are divided into three classes each [38]. The properties are hydrophobicity, van der Waals volume, polarity, polarizability, charge, secondary structure and solvent accessibility. Figure 7 shows a few examples of global descriptors computed from chain A of human insulin (DrugBank DB00046). The descriptors are easy to interpret and can be computed from any protein whose sequence is known. However, since the descriptors treat each protein as a single unit, with only a rough measure of sequence order, features such as 3D structure or active-site location are not taken into account. The descriptors are thus mainly useful in large general models of protein-ligand interaction.

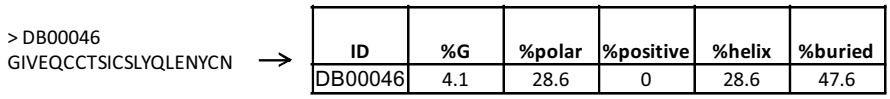


Figure 7. Examples of global protein descriptors computed from the primary structure of human insulin (A chain). The descriptors are percentage glycines (%G), percentage polar amino acids (%polar), percentage positively charged amino acids (%positive), percentage amino acids common in helices (%helix), and percentage buried amino acids (%buried).

Development of descriptors computed from the 3D structure of proteins is an active research area. Examples of descriptors focused on the active site are the SCREEN [40] descriptors that are computed by rolling a spherical probe over the protein surface, and the PocketPicker [41] descriptors that specify the shape of a potential binding-site with regard to its "buriedness" through a grid-based method. General descriptors suitable for representing any part of a protein 3D structure are, for instance, the Gauss-integral-based descriptors developed by Rogen & Fain [42] and the local descriptors of protein structure developed by Hvidsten *et al.* [26, 43]. The latter have been used to induce generalized PCM models from sequentially and structurally diverse sets of enzymes in paper II and III. A local descriptor of protein structure is a set of short continuous backbone fragments, centered on a particular amino acid, that are close in 3D space but not necessarily in the sequence. A descriptor is generated by (a) identifying all amino acids within a radius of 6.5Å from a given amino acid (an amino acid is represented by the point on the line $[C_\alpha,C_\beta]$ that lies 2.5Å away from C_α), (b) for each such amino acid, adding four sequence neighbors, two on each side, to obtain continuous backbone fragments of five amino acids, and (c) merging any overlapping fragments. Local descriptors of protein structure are computed from all amino acids in all domains in ASTRAL [44] which is a representative set of protein domains derived from the Protein Data Bank (PDB) [45]. A library of commonly reoccurring local descriptors has been constructed by (a) for each local descriptor, identifying a group of similar local descriptors and (b) selecting a set of representative, partially overlapping descriptor groups. It has been shown that a library of around 4000 such descriptors can effectively be used to assemble large parts of most proteins in PDB [43]. An example of a local descriptor of protein structure is shown in Figure 8. As the figure shows, the occurrence of local descriptors in a protein structure is encoded by a Boolean array.

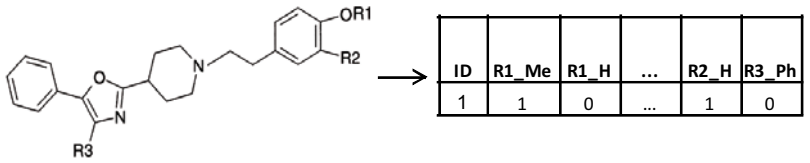
ID	Local descriptor 100	Local descriptor 115	...	Local descriptor 416
Prot 1	1	0	...	0

Figure 8. An example of a local descriptor of protein structure, colored according to secondary structure (yellow strands, red helices), situated close to a ligand (blue). The occurrence of local descriptors is listed in the table, where "1" means that the descriptor is present in the protein structure.

2.2 Ligand Descriptors

Ligand descriptors have traditionally been used in QSAR modeling to establish relationships between molecular structure and function. A molecular descriptor is a numerical value that characterizes some property of a compound. The information content of a descriptor depends on the molecular representation of the compound and the algorithm that is used for the computations. Numerous descriptors have been proposed and several programs that compute such descriptors have been developed, including Dragon [46] and Adriana [47] which compute 3224 and 1224 molecular descriptors, respectively. A list of all published descriptors is beyond the scope of this thesis (for a comprehensive overview, see Todeschini & Consonni [48] or Karelson [49]). Instead, a brief overview of ligand descriptors that have been used in PCM modeling will be given. Molecular descriptors can in general be distinguished by their data type and the dimensionality of the molecular representation from which they are calculated.

The data type of a descriptor is simply whether a descriptor is, for instance, a Boolean, integer, or real number. Boolean descriptors were used in the early PCM models [19, 20, 29] and they contain information on whether an amino acid or chemical group is part of a ligand. Figure 9 shows a ligand with three variable positions (R1-R3), and a selection of Boolean ligand descriptors. For example, the presence of a methyl (Me) group at the variable position R1 is represented by the value "1" in the "R1_Me" column.



R1=Me, R2=H, R3=OCF_3

Figure 9. Example of the use of Boolean ligand descriptors. The generic ligand has three variable positions (R1-R3). In this example, R1 is a methyl group (Me), R2 is a hydrogen atom (H), and R3 is a trifluoromethoxy (OCF_3) group.

Ligand descriptors can also be distinguished by the dimensionality of the molecular representation from which they are calculated [50]. A common division is zero, one, two and three dimensional (0D, 1D, 2D, and 3D) descriptors. In PCM modeling, descriptors of all dimensionalities have successfully been applied. Lists of 0D-3D descriptors are shown in Figure 10 together with a few examples of each descriptor category for the compound Diclofenac®. *0D descriptors* are derived from the chemical formula, and *1D-descriptors* are computed from a ligand represented as a substructure list. Examples include the atom count descriptor "number of carbons" and the molecular weight descriptor "molecular weight". 0D/1D descriptors are generally easy to interpret, but are not likely to provide sufficient discriminating power if used in isolation and they are therefore often used in combination with other descriptors [51]. *2D descriptors* are computed from the graphical representation of a chemical structure. There are many types of 2D descriptors and they are often difficult to interpret. Examples include the physicochemical descriptor "MLOGP" (octanol-water partition coefficient) which is a measure of how lipophilic a molecule is and the Boolean fingerprint "2D fingerprint C-C" which contains information on whether a carbon-carbon (C-C) fragment is present in a compound within a given topological distance. 2D descriptors are frequently used in traditional QSAR and are often sufficient to establish a relationship between structure and function. Finally, *3D descriptors* are generated from 3D conformations and include geometrical, surface-property, and grid property descriptors. Since a 3D structure should reflect the bioactive conformation of a ligand, 3D descriptors may provide important information on properties needed for a ligand to interact with its targets. However, the active conformation of most ligands is unknown. Moreover, 3D descriptors are often difficult to interpret, and their calculation can be time-consuming when dealing with large datasets [51].

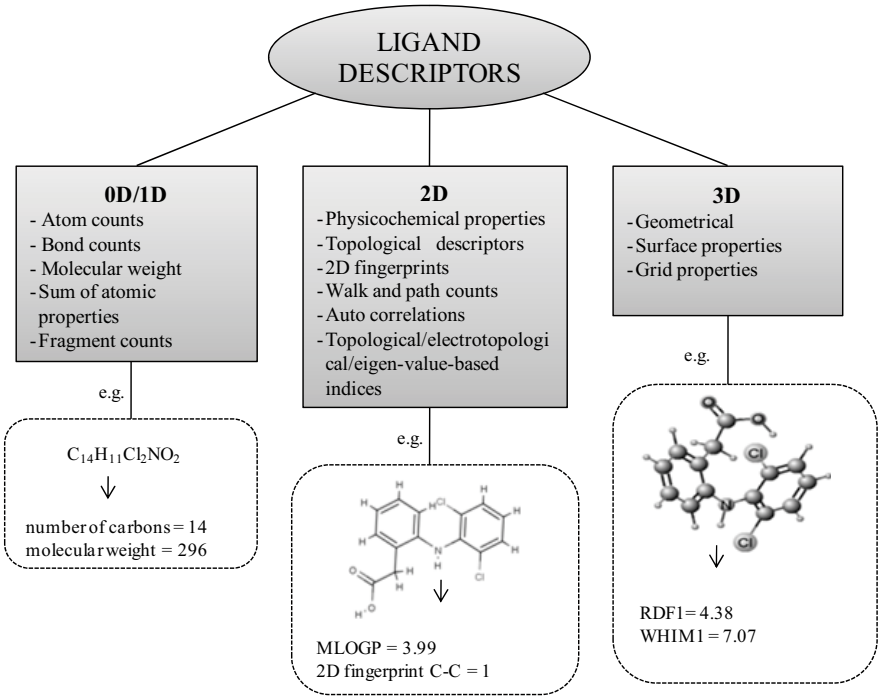


Figure 10. Ligand descriptors classified by the dimensionality of the molecular representation from which they are derived. The ligand used as an example is the pain-killer Diclofenac®.

2.3 Protein-Ligand Interaction Data

The affinity between a protein and a ligand is commonly measured with radioligand-binding experiments [52]. A *radioligand* is a radioactively labeled compound that can associate with any biomolecule of interest. The principles of a radioligand protein-ligand binding-affinity assay are quite simple [52]. (1) A preparation containing the protein of interest is incubated with a suitable radioligand for an appropriate period of time. (2) To determine the binding affinity of an unknown ligand (*e.g.*, a putative drug) to the protein, a range of ligand concentrations is added to the equilibrated protein-radioligand preparation. (3) Bound ligands are separated from free ligands and radioactivity measures are used to determine each respective concentration. (4) Non-linear regression analysis is finally used to derive a dissociation equilibrium constant (K_d), which is an absolute measure of protein-ligand binding affinity. If the ligand is an inhibitor, the term inhibition constant (K_i) is often used, but both terms are used interchangeably in the litera-

ture. A radioligand experiment also produces a so-called half-maximal inhibitory concentration (IC_{50}) value, which is the ligand concentration needed to inhibit half of the biological response of a given protein. Although K_i can be derived from IC_{50} through the Cheng-Prusoff correction [53] under certain conditions, the IC_{50} value depends on the radioligand concentration and may therefore vary between experiments. Since K_i is an absolute value, the large majority of PCM studies has been designed to predict inhibition or dissociation equilibrium constants (K_i and/or K_d). The search for PCM interaction data has thus been focused on the retrieval of dissociation constants from the literature and public databases.

The binding of ligands to biomolecules has been studied since the early 19th century [54]. As a result, the scientific literature contains large amounts of protein-ligand interaction data. However, reading scientific publications is time-consuming and attempts have been made to apply text-mining techniques to retrieve protein-ligand binding-affinity data automatically. One example is BUDA (Binding Unstructured Data Analysis) [55], which employs natural-language processing to identify key sentences and phrases in papers and uses a weighted scoring algorithm to rank the probability that a paper contains binding data. Another example is the database GLIDA (GPCR-LIgand DAtabase) [56] which plans to use a text-mining tool to update ligand information from the literature and patents. Although text mining may be useful in the early stages of data retrieval, all data need to be manually curated to ensure a correct linkage between biomolecules and biological activity. This is a very tedious and time-consuming process and in fact one of the major bottle-necks in PCM data modeling.

In recent years, a number of protein-ligand interaction databases have been made publicly available. Table 2 lists public databases that store protein-ligand interaction data. It should be noted that this is not a complete list of repositories since a large amount of interaction data is not available in the public domain. The BindingDB (Binding DataBase) database [57] collects information on all possible protein targets. The database contains more than 48000 affinity values linked to ligand structure, protein sequence and occasionally protein structure. The PDSP (Psychoactive Drug Screening Program) database [58] is focused on receptors. Its affinity values are linked to ligand and protein information in the form a receptor name and possibly a gene identification number. The Brenda database [59] is one of the most comprehensive sources of enzyme information. The affinity values stored in Brenda are linked to ligand structure and enzyme classification (EC) number. Both PDSP and Brenda data need to be manually curated to obtain a mapping between affinity values and protein sequence/structure. The PDL (Protein Ligand Database) [60], AffinDB (Affinity DataBase) [61], BindingMOAD (Mother Of All Databases) [55] and PDBBind (Protein Data

Bank Binding) [62] databases contain relatively few entries, which is due to the fact that these four databases contain a mapping between affinity data, ligand and 3D protein structure.

Table 2. Protein-ligand binding-affinity databases.

Name	**Target focus**	**Affinity Measures**	**Number of entries**
BindingDB	All proteins	K_i, K_d, IC_{50}, EC_{50}, $\Delta G°$, $\Delta H°$, $-T\Delta S°$	~48000
PDSP	Receptors	K_i	~47000
BRENDA	Enzymes	K_i, IC_{50}	~19000
BindingMOAD	All	K_d, K_i, IC_{50}	~3500
PDBBind	All	K_d, K_i, IC_{50}	~3500
AffinDB	All	K_d, K_i, IC_{50}	~700
PLD	All	K_d, K_i	~500

3. Methodology

Computer science is the study of the theoretical foundations of information and computation. Machine learning, databases, and the implementation of computer languages are all areas of computer science that play key roles in chemogenomics and proteochemometrics modeling. Computer programming is needed to perform tasks such as conversion between data formats, descriptor computation and similarity analyses. Databases may be required for organization and storage of large amounts of data, as shown in section 2.3. Finally, machine learning is needed to induce models of protein-ligand interactions.

Central to PCM is model induction and the main steps in this process are illustrated in Figure 11. Prior to model induction, raw data needs to be pre-processed to remove redundant entries and extreme outliers. Models are typically induced by unsupervised or supervised machine-learning methods. All machine-learning methods use data in a tabular form. A toy example of a PCM dataset is shown in Table 1 (Chapter 2). In PCM, the input variables are protein and ligand descriptors, and the outcome variable is binding affinity. There are no absolute boundaries between pre-processing, unsupervised, and supervised learning. This is due to the fact that unsupervised learning methods may be used in data pre-processing, and may as well be part of a supervised-learning modeling process.

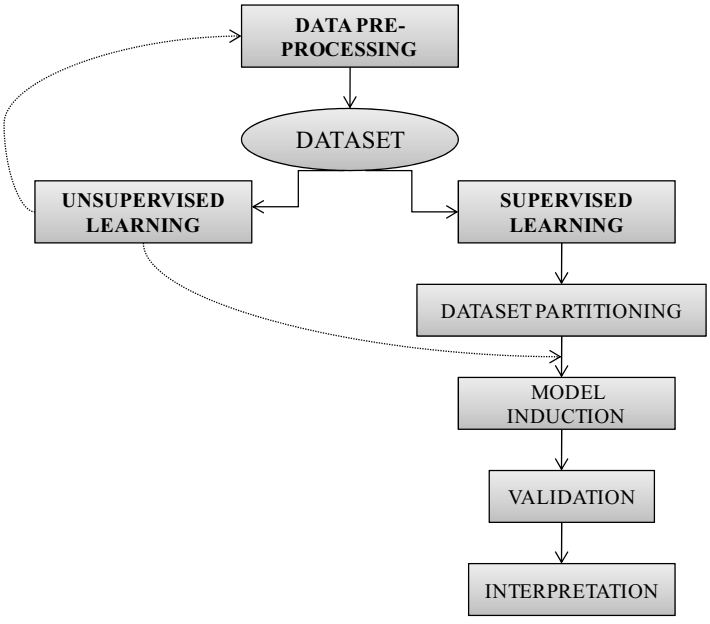


Figure 11. The main steps in the machine learning modeling process.

3.1 Data Pre-processing

Since PCM data usually originates from several different sources, data pre-processing is perhaps the most important step in the modeling process. This step is needed to remove redundant protein-ligand complexes, to transform raw data, and to remove outliers. In this section, a brief overview will be given of methods used in PCM for removal of redundant complexes and transformation of raw data. The unsupervised principal component analysis (PCA) method is generally used for outlier detection in PCM, and it will be discussed in section 3.2.

Detection of redundant complexes

To identify redundant complexes in a PCM dataset, protein and ligand similarity measures are needed. *Protein similarity* is commonly assessed with alignments of two or more sequences or structures. Pairwise sequence-alignment algorithms can be global or local (for an overview, see Durbin *et al.* [63]). A global alignment spans the entire length of the sequences and an optimal alignment can be obtained using the Needleman-Wunsch [64] algorithm. A local alignment identifies (usually) shorter regions of high similarity between the two sequences and such alignments can be obtained with the

Smith-Waterman algorithm [65]. After alignment, similarity measures such as alignment scores and percentage sequence identity can be calculated. Global alignment was used to compare large and diverse PCM datasets in paper III. Although local alignment is generally recommended to compare two dissimilar sequences, it may produce misleadingly high similarity scores if the sequences are of different length but contain one or more regions of high similarity (*e.g.,* shared domains). The global alignment scores are computed over the entire length of the two sequences and are therefore fairer measures of protein similarity if a large number of protein-sequence pairs are to be compared and the alignments cannot be inspected manually for practical reasons. A standard measure of *ligand similarity* is the Tanimoto coefficient [66]:

$$t = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{\sum_{i=1}^{N} (x_{iA})^2 + \sum_{i=1}^{N} (x_{iB})^2 - \sum_{i=1}^{N} x_{iA} x_{iB}} \qquad [1]$$

in which the molecules A and B are represented by vectors $\mathbf{x}_A$ and $\mathbf{x}_B$ of length N with the i^{th} descriptor having the value x_i. The Tanimoto coefficient is commonly computed from ligand 2D fingerprints which are binary descriptors where each bit indicates the presence or absence of a particular substructure. In that case, the Tanimoto coefficient gives a measure the extent to which two molecules share common fragments. A similarity score s was proposed in paper III that can be calculated between two protein-ligand complexes by combining the sequence identity and ligand similarity score:

$$s = \frac{1}{1+\alpha}\left(\alpha t + \frac{p}{100}\right) \qquad [2]$$

where p is the percentage sequence identity and α is a weight factor ($\alpha > 0$; most commonly $\alpha = 1$). A pair-wise comparison of all complexes in a PCM dataset results in the identification of redundant protein-ligand complexes (*i.e.*, s close to 1) that need to be removed prior to model induction. It also illustrates the similarity structure in the data set and can be of help when interpreting the results in later stages of the analysis.

Transformation of raw data

The selection of transformation technique depends on the data type, the amount of data, and the general characteristics of the machine learning algorithm. The large majority of the PCM studies [19, 20, 22-24, 27, 29, 32-35, 67] including paper II, III, and IV have involved methods that are based on distances between points in *n*-dimensional space. A general transformation technique to scale input variables to a certain range is *normalization*. An example of PCM raw data in need of normalization is shown in Table 1 (Chapter 2), where the ligand descriptor "molecular weight" ranges between 150 and 750 Da, while the Boolean protein descriptors can take the values 0 and 1. There are many data normalization methods. *Decimal scaling* simply moves the decimal point and typically maintains the values in a range of -1 to 1. To obtain a better distribution of values on a whole normalized interval, *min-max normalization* can be applied. Min-max normalization subtracts the minimum value of a variable from each value of the variable and then divides the difference by the observed range of the variable. *Standard deviation normalization* has been used extensively in PCM. This method often works well with distance measures, but transforms the data into a form unrecognizable from the original data. The value of object i is transformed and is defined as:

$$v'(i) = \frac{v(i) - mean(v)}{sd(v)} \qquad [3]$$

where $mean(v)$ is the average, and $sd(v)$ is the sample standard deviation of the input variable v.

3.2 Machine Learning

The main goal of machine learning is to automatically extract knowledge from data [68]. Machine learning is commonly divided into unsupervised and supervised learning. In *supervised* learning the goal is to predict the value of an outcome variable based on a number of input variables. In *unsupervised* learning, there is no outcome and the goal is to describe associations and patterns among a set of input variables. In PCM, the outcome variable is typically the binding-affinity value and the input variables are protein and ligand descriptors.

3.2.1 Unsupervised learning

Unsupervised learning is used to characterize data and common approaches are clustering, data compression and outlier detection. PCA methods have been used to visualize PCM data, to reduce data dimensionality, and to detect outliers.

Clustering

Cluster analysis is a set of methods for automatic classification of samples into a number of groups [68]. Clustering requires a set of objects and a measure of the similarity of objects pairs. The output is a number of groups that form a partitioning of the data set. All clustering methods comprise the following steps: (1) Define a similarity measure between the objects. For PCM, the protein-ligand similarity measure defined by equation [2] (section 3.1) has been used. (2) Perform the clustering analysis. Clustering analysis can be carried out by hierarchical and non-hierarchical methods. *Hierarchical* clustering methods require calculation of a similarity measure between all pairs of objects and between clusters of objects. According to the bottom-up strategy, data objects (here, protein-ligand complexes) are organized into clusters of increasing size. The most similar objects are first grouped, and these initial groups are merged by their similarity. As the similarity decreases, all subgroups are eventually fused into a single cluster. The relationships between the clusters can be visualized with a dendrogram. *Non-hierarchical* clustering techniques are designed to group objects into K clusters. The number of clusters may either be specified in advance or can be determined by the clustering procedure. Since not all pairs of objects need to be compared, non-hierarchical methods are faster than hierarchical ones and can therefore be applied to very large datasets.

Principal Component Analysis

Many PCM datasets are ill-defined in the sense that they often consist of a very large number of protein and ligand descriptors compared to the number of protein-ligand complexes. Principal Component Analysis (PCA) is a technique that can be used to reduce the number of input variables when there are correlations present between (some of) them. The idea behind PCA is to find principal components $Z_1,Z_2,\ldots,Z_n$ which are linear combinations of the original variables $X_1,X_2,\ldots,X_n$ that describe each object, *i.e.*

$$Z_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + ... + a_{1n}X_n$$
$$Z_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + ... + a_{2n}X_n \quad [4]$$
etc.

The coefficients a_{11}, a_{12} *etc.* are chosen so that the new variables are not correlated with each other. Moreover, the principal components are selected so that the first principal component, Z_1 , accounts for most of the variation in the data, Z_2 accounts for the next largest variation and so on. In PCM, this method has been used to reduce the number of descriptors, and the principal components have been used as input variables to supervised learning methods. PCA has also been used to visualize data in order to compare datasets and to detect outliers. Evidently, this requires that the first two or three components capture a large part of the variation in the original descriptor set.

3.2.2 Supervised learning

The goal of supervised learning is to create a system that can associate a set of input variables with one or more specified outcome variables (Table 1). This section contains an introduction to the supervised machine-learning process as well as the methods that have been used in PCM modeling. The supervised learning process usually begins with dataset partitioning followed by model induction, validation and interpretation (Figure 11).

Dataset partitioning

In machine learning there is always a danger of that a system becomes expert in predicting the data on which it was trained. However, to be useful in practice a model's predictive capability should generalize to cases on which it has not been trained. This capability can be assessed by validation which involves dividing the data into different sets. If there is no shortage of data, the best approach is to divide the data into three parts: a training set, a validation set, and a test set. The training set is used for model induction, the validation set can be used to select the best model, and the test set is used for assessment of the generalization error of the final model. It is important to select a representative training set that covers a large part of the data space as Figure 12 illustrates. For instance, a training set that only consists of objects from square A in Figure 12 would result in a model that cannot predict objects from the remaining squares.

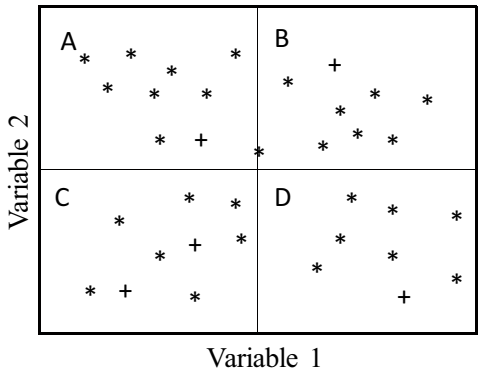


Figure 12. A generic dataset described by two variables. The data space in divided into four squares (A-D). The objects are divided into training set (*) and test set (+).

Model induction

Supervised learning methods can be trained for classification or regression tasks. In *classification*, the algorithm is designed to classify objects into one or several predefined discrete classes. Panel A in Figure 13 shows a training set with objects that belong to two classes (Δ and o). The classification function, shown in panel B, provides the best separation of the objects, and can be used to classify new objects for which the class is unknown.

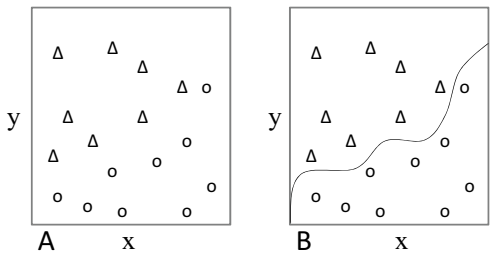


Figure 13. Graphical representation of a classifier that maps objects in a training set (A) described by two variable (x, y) into two classes by a classification function (B).

A *regression* model is a function that maps objects to a real-valued outcome variable. For example, panel A in Figure 14 shows a training set to which a regression function can be fitted. This is shown in panel B, along with a prediction of the real-valued outcome for a new unseen object.

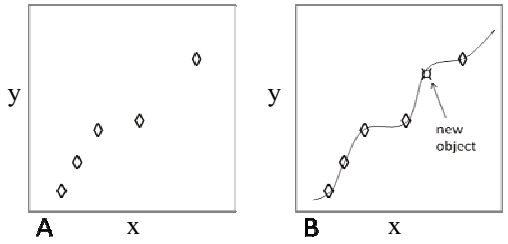


Figure 14. Graphical interpretation of a regression function (y=f(x)) induced from training set objects (A) and used to predict the outcome of a new unseen object (B).

A number of supervised machine learning methods have been developed, and selection of the most appropriate method for model induction depends on the characteristics of the problem and the composition of the available dataset. Table 3 gives an overview of available methods (for a complete overview, see or Kandardzic [68] or Hastie *et al.*[69]). All method categories listed in the table may be used for both classification and regression purposes, although individual methods are normally optimized for one of the two learning tasks. In PCM, partial least squares, decision trees, rough sets, and support vector machines methods have been applied. *Partial least squares* (PLS) [21] combines features from principal component analysis and multiple linear regression and is suitable when the input variables are numerous and highly correlated. PLS identifies components from the input variable matrix **X** that explain as much as possible of the covariance between **X** and the outcome variable matrix **Y**. A *decision tree* [70] consists of nodes where the input variable values are tested. The outgoing branches of a node correspond to all the possible outcomes of the test at the node. The *rough sets* method [71-73] is particularly suitable when dealing with incomplete and uncertain data. The model consists of a number of decision rules that are easy to interpret. *Support vector machines* (SVM) [74] is a general technique in which the data is considered as two sets of vectors in *n*-dimensional space. A SVM will construct a hyperplane in that space that separates and maximizes the margin between the two datasets. In PCM, rough sets and decision trees have been used for classification, PLS has been used for regression, while SVMs have been used for both learning tasks.

Table 3. Supervised learning techniques

Category	**Methods**
Statistical	Bayesian inference Logistic regression ANOVA (analysis of variance) PLS
Decision trees/rules	CLS ID3 C4.5 Rough sets
Neural networks	Feed forward Radial basis function Kohonen
Support vector machines	Linear Non-linear: polynomial, RDF (radial distribution function), Gaussian

Validation

A model induced by a machine-learning algorithm should be validated to verify the predictive ability of the model and to select the most appropriate method for further tuning and refinement. Common methods for validation are: application of the model to a validation set, *k*-fold cross-validation, and randomized shuffling of the outcomes. In a *k*-fold cross-validation the data-set is row-wise randomly divided into *k* blocks. Each block is left out once and a model is induced from the remaining *k-1* blocks. The data in the block that is left out is used to assess the predictive performance of the model. For each block that is left out, a measure of the prediction quality is reported and the mean of the predictive performance is often reported. This approach becomes computationally demanding for large amounts of data and is therefore most suited for small or moderately sized datasets. To investigate whether the predictive performance of a model could have been obtained by chance, it is common to perform a random shuffling of the outcome values. The order of the outcome values is repeatedly randomized and models are induced and evaluated. A significant decrease in predictive performance of models induced from randomized data confirms the validity of the original model. An external test set should be used for model assessment only after model refinement. The prediction results for an external test set are a measure of whether a model is able to generalize to unseen cases. Poor results from test set validation are a sign of model overfitting. This means that the model is too adjusted to the data in the training set, and does not generalize to new unseen cases.

The predictive quality measure depends on whether the modeling task is classification or regression. In classification, the performance of a model is often visualized by a so-called confusion matrix. Figure 15 shows a confusion matrix for a binary classifier with the two outcomes, high and low binding affinity. The complexes that are correctly classified are denoted as true positives (TP) and true negatives (TN), and the complexes that are misclassified are denoted as false positives (FP) and false negatives (FN).

		Predicted	
		High	Low
Actual	High	TP	FP
	Low	FN	TN

Figure 15. A confusion matrix for a binary classifier.

Common measures are the *error rate* which is simply the percentage of a test set that is misclassified (FP+FN) and *accuracy* which is the percentage of a test set that is correctly classified (TP+TN). In applications where certain errors are to be considered more serious than others, a weight on those errors can be applied when calculating the final quality measures. Another classification quality measure is the *Receiver Operating Characteristic* (ROC) curve [75]. The area under the ROC curve is a measure of the discriminatory power of a classifier, and is insensitive to changes in class distribution and the costs of making certain errors. The values in a confusion matrix can be used to compute *sensitivity* and *specificity* which are the true positive and true negative rate respectively, and are computed as:

$$sensitivity = \frac{TP}{TP + FN} \quad [5]$$

$$specificity = \frac{TN}{TN + FP} \quad [6]$$

A ROC curve is plotted as sensitivity *vs.* 1-specificity, as some discrimination threshold is varied. An area under the ROC curve close to 1.0 means that the classifier is able to perfectly map objects into classes, while an area close to 0.5 means that the classifier does not perform better than random guessing. A ROC curve of a moderately good classifier is shown in Figure 16.

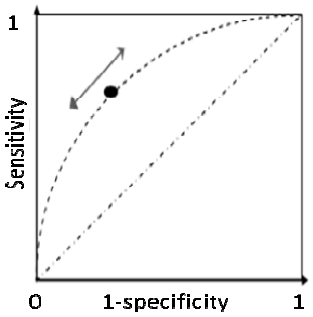


Figure 16. The ROC curve of a moderately good classifier (dashed line). Sensitivity and specificity are calculated as some discrimination threshold (●) is varied. The diagonal curve shows the behavior expected for a random classifier

In regression analysis, there are several ways to quantify the extent to which the predicted outcomes differ from the actual outcome values. One common quality measure used in both QSAR and PLS analysis is the *root-mean square-error of prediction* (RMSEP). RMSEP has the same units as the quantity that is being predicted. It is computed as:

$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{pred_i})^2} \qquad [7]$$

where n is the number of objects in the validation or test set, and y and y_{pred} are the actual and predicted outcome values, respectively. Another common measure is the *coefficient of determination*. Unfortunately, there is no consensus definition of R^2 so its interpretation depends on the context in which it is used. In linear regression, R^2 is commonly a measure of how well a regression line fits actual data points. An R^2 of 1.0 indicates that the predicted values correspond perfectly to the known outcome values. R^2 can be computed as the square of the linear correlation coefficient, but also as:

$$R^2 = 1 - \frac{\sum_i (y_i - y_{pred_i})^2}{\sum_i (y_i - \bar{y}_i)^2} \qquad [8]$$

where $\bar{y}$ is the average of the actual outcome values.

Interpretation

Some machine-learning algorithms produce models that can be interpreted in terms of the input variables. In PCM modeling, it may be interesting to get insight into the ligand and protein descriptors that are important for the prediction of binding affinity. Such information contributes to the understanding of which properties govern molecular recognition which may in turn guide the design of new active compounds.

Rough sets and decision trees are methods that allow for model interpretation. Evidently, this requires that the protein and ligand descriptors used as input are interpretable themselves. This is not the case for a large number of protein and ligand descriptors, even though such descriptors may be very useful for model induction and prediction. A rough sets model consists of a number of decision rules that are easy to interpret, provided of course that the number of rules is not too large. Examples of rule interpretations are shown in paper I. Decision trees are easy to interpret as well since the leaves and branches represent classifications and conjunctions of descriptor features that lead to those classifications. In Paper IV, the decision tree method was used for model induction and interpretation and a pruned tree is presented. The PLS method (*e.g.,* as implemented in the SIMCA program [76]) also allows for some degree of interpretation since it ranks the features that influence the model.

4. Results

In this chapter the results of the papers included in this thesis are summarized. Paper I and II describe PCM models induced from small datasets, and the results show that it is possible to induce interpretable models with a non-linear rule-based method, and that local descriptors of protein structure can be used to induce PCM models that cover proteins differing in sequence and fold. In paper III, those local descriptors of protein structure are used to induce a PCM model on a large dataset that covers all major enzyme classes. Paper IV is a step towards proteome-wide PCM models, and shows that it is possibly to discriminate between low and high binding affinity complexes using a set of simple protein and ligand descriptors that do not require knowledge of the 3D structure of either. Finally, paper V presents a method to visualize and compare protein-ligand chemogenomic subspaces.

4.1 Paper I and II: Proofs of Principle

Paper I shows that it is possible to induce interpretable rule-based models from PCM datasets. Rough sets models were induced from three proteochemometric datasets. The receptors in the datasets were melanocortin and adrenergic GPCRs. The ligands were peptides (native and synthetic) and synthesized chemical compounds. All receptors and ligands in the datasets were alignable which made position-specific information vectors a suitable approach for describing the properties at the variable sites. Rule-based rough sets models were induced from the datasets and the accuracy for the test set was greater than 0.85 for all models. The rule sets provided relevant biological information on receptor-ligand interaction and showed that rough sets can be used successfully to model proteochemometrics datasets.

Paper II introduces a new approach to proteochemometrics modeling. Local descriptors of protein structures based on 3D fragments were used. These descriptors are alignment-independent which in principle allows for models covering a wide variety of proteins. The proteins in the dataset were all hydrolase and lyase enzymes stored in the PLD database [60]. All proteins in this database are co-crystallized with ligands with experimentally measured binding affinity. The ligands were described by 1D-3D traditional QSAR

descriptors. Models were successfully induced by both PLS and rough sets methods and validated with an external test set. A spatial analysis showed that protein descriptors close in space to the enzyme active site are more correlated to binding affinity than those that lie further away from the binding site. The results from this study demonstrate that the fragment-based local protein descriptor approach enables models covering a wide range of proteins that vary in sequence and structure.

4.2 Paper III: Enzyme-Wide Models

In Paper III, the local descriptors of protein structure introduced in paper II were used to induce a model from a large dataset that included members from all major enzyme-classes. Training data was collected from four publicly available protein-ligand interaction databases (PLD [60], Binding-MOAD [55], PDBind [62] and AffinDB [61]). These databases contain binding-affinity values linked to the 3D structure of the protein-ligand complexes. Extraction and annotation required extensive manual curation. Therefore, it is reasonable to assume that the data stored in these databases represent the lion's share of publically available data concerning the binding affinity of protein-ligand complexes of known structure. Each enzyme was characterized by a set of local descriptors of protein structure that describe the binding site of the co-crystallized ligand. The ligands were described by traditional QSAR descriptors. A model was induced by the support vector regression method.

To evaluate the model, a test set consisting of enzyme structures and ligands was manually curated. The data was extracted from Brenda [59], which is one of the most comprehensive enzyme databases. The large majority of the ligands in the test set have not been co-crystallized with their enzymes and their ligand binding sites are therefore generally not known. The test set enzymes were therefore characterized by matching their entire structures to the local descriptor library constructed from the training set.

Both the training and the test set contained enzyme-ligand complexes from all major enzyme classes, and the enzymes spanned a large range of sequences and folds. The experimental binding affinities (pK_i) ranged from 0.5 to 11.9 (0.7-11.0 in the test set). The induced model predicted the binding affinities of the external test set enzyme-ligand complexes with an R^2 of 0.53 and an RMSEP of 1.5. This demonstrates that the use of local descriptors makes it possible to create rough predictive models that can generalize over the entire known structural enzyme-ligand space.

4.3 Paper IV: Towards Proteome-Wide Interaction Models

Paper IV presents an approach to PCM modeling that enables inclusion of all available protein-ligand interaction data. The major obstacle to generalized PCM models lies in the fact that currently applied protein descriptors require sequence or structural alignments or a 3D structure of the proteins. In this study, descriptors that are computed from the amino-acid sequence and that describe general physicochemical properties of the proteins have been evaluated. A PCM model was induced from all protein-ligand interaction data, linked to inhibition constants, stored in the BindingDB database [57]. This amounted to a dataset of 7078 complexes, and included representatives from all major drug-target categories as defined by Drews & Ryser [4, 5]. The dataset was divided into a training, validation and test set. The binding affinity values were discretized into a “low” and a “high” binding affinity class. Protein descriptors were computed with the PROFEAT web server [37] and a set of commonly used QSAR ligand descriptors were computed with the Dragon program [46]. Several machine-learning algorithms were compared and evaluated using the Weka software suite [77], which resulted in the selection of decision trees for further tuning and validation. The model was finally applied to an external test set which resulted in an accuracy estimate close to 80% and an area under the ROC curve close to 0.8. This shows that the sequence-derived protein descriptors are suitable for PCM modeling. Therefore, it may be possible to use this approach to create proteome-wide models.

4.4 Paper V: Chemogenomics and Protein-Ligand Subspaces

Paper V presents an approach to visualize protein-ligand spaces from a chemogenomics perspective. A model was induced from protein and ligand descriptors without any outcome variable. Two chemogenomics protein-ligand interaction datasets were prepared for this study. The first dataset covers the known structural protein-ligand space, and includes all non-redundant protein-ligand interactions found in the PDB [45]. The second dataset contains all approved drugs and drug targets stored in the DrugBank [78] database. To capture biological and physicochemical features of the chemogenomics datasets, sequence-based descriptors were computed for the proteins, and 0, 1 and 2 dimensional descriptors were computed for the ligands. PCA was used to analyze the multidimensional data and to create global models of protein-ligand space. The nearest neighbour method was used to obtain a measure of overlap between the datasets. Despite the gener-

ality of the descriptors used in this study the results show that it is possible to induce a PCA model on the combined set of protein and ligand descriptors. For all ligands in DrugBank that are known to interact with more than one protein, the nearest neighbors in the space defined by the principal components were retrieved. The model captures a large fraction of the known DrugBank cross-interactions. This suggests that for a given complex, this method could be applied to find chemogenomically similar protein-ligand complexes in the proteome in order to define a subset of putative drug targets to study for possible cross-interactions.

5. Concluding Remarks and Future Prospects

Improved experimental methodologies have led to an immense growth in biological data. For instance, during the last 10 years the UniProtKB/Swiss-Prot database [79] has grown from ~50 000 to ~400 000 protein entries [80] and the PDB has grown from ~8500 to ~55 000 3D structures [81]. New techniques such as massive parallel sequencing [82, 83] and high-throughput determination of 3D protein structures [84] will further contribute to the growth of the biological databases. This data "explosion" has led to a wealth of information on drug targets, which in turn has enabled new approaches in chemogenomics and PCM research. The five studies presented in this thesis reflect this trend. The first model (paper I) was based on only 40-60 complexes, whereas the last model (paper V) was based on more than 17 000 complexes. However, even the latter dataset is relatively small compared to the massive amount of information available. This is partly due to the fact that a substantial amount of ligand information is protected by commercial interests. Projects such as PubChem [85], chEMBL[86] and Chebi [87] make ligand information publically available but it will most likely take a long time until the amount of public ligand data matches the amount of protein data available to the public. Another bottle-neck in PCM is the need for experimentally measured biological activities as outcomes in the model. Such measures are continuously published in the scientific literature, but require manual curation to be retrieved.

The key to generalized PCM models is the application of protein descriptors that allow for the inclusion of proteins that vary greatly in sequence, structure and function. Two new descriptor types have been evaluated in this thesis. The first type is local descriptors of protein structure computed from the 3D structure of a protein. The second descriptor type is computed from the amino-acid sequence and covers various physicochemical properties of a protein. Both types allow for the construction of generalized PCM models which might eventually lead to models that cover the entire protein-ligand space. A general conclusion from these studies is that there is a trade-off between generality and precision. That is, the early models were very specific in the sense that they could pinpoint, for instance, which amino acids are important for high binding affinity. The accuracy of these predictions is also very high. The latests and most general PCM model presented in paper IV is able to roughly predict high and low binding affinity but the protein

and ligand descriptors do not allow for any interpretation about which parts of the complex are important for the molecular recognition. A challenge would be to develop models that are *both* specific and general. This would require alignment-independent protein descriptors that can be computed from the amino-acid sequence but still include information specific to certain residues. Another challenge is to develop text-mining tools that can automatically extract protein-ligand interaction data from scientific publications. A major obstacle in this endeavor is that a large part of the scientific literature is not freely accessible and that older publications may only be available in printed form.

Despite the data retrieval and curation problems described in this section, there is a huge potential in the chemogenomics and proteochemometrics approaches. The amount of biological and chemical information will most likely continue to grow. This will require development of computational tools to extract knowledge from data. In recent years there has been a steady growth in the amount of protein-ligand interaction data amassed by manual curation and made available in public databases. As more data becomes readily available, the protein-ligand interaction matrices will be less sparse. This may allow for prediction of unwanted cross-interactions across entire proteomes and thereby facilitate the design of new, specific and efficient drugs.

6. Summary in Swedish

Livet på jorden tros ha uppstått ur en ”soppa” av små kemiska föreningar. Dessa molekyler gick samman till större enheter och bildade basen för de biomolekyler som utgör byggstenarna och förutsättningen för allt jordiskt liv. Biomolekylerna organiserade sig till allt större enheter och system vilket utmynnade i de celler som, till exempel, människan utgörs av. En cell består i huvudsak av biomolekylerna DNA, RNA, proteiner, och fetter. Figur 17 visar en schematisk bild av en cell.

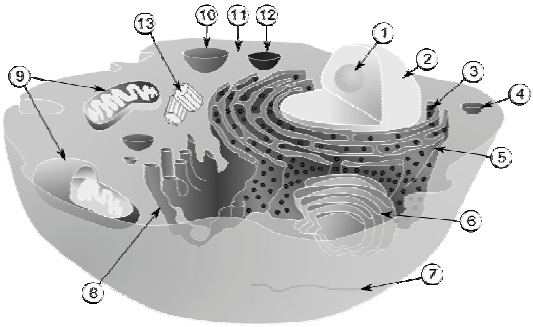


Figure 17. Schematisk bild av cellens olika delar. Arvsmassan i form av DNA förvaras inuti cellkärnan (1 och 2). Ribosomerna (3) är cellens proteinfabrik. Vissa kemiska processer sker i vesiklar (4). Det korniga och glatta endoplasminska nätverket (5 och 8) har flera funktioner och är bland annat viktigt för proteiners veckning och transport. Golgiapparaten (6) sätter ”adresslappar” på proteiner som visar vart i cellen de skall transporteras. Cellskelettet (7) ger cellen stabilitet. Mitokondrierna (9) är cellens ”kraftverk” där energi genereras. Vakuolen (10) är ett hålrum i cellen. Cytoplasma (11) är utrymmet mellan cellens kärna och det omslutande membranet (7). Lysosom (2) är cellens ”soptipp” där nedbrytning sker, och centriolen (13) är slutligen en viktig komponent i celldelning.

Det centrala dogmat handlar om informationsflödet i cellen. Genetisk information förvaras i form av DNA i cellkärnan. DNA avläses till RNA som vandrar ut ur cellkärnan in i cytoplasman där RNA sekvensen används som mall för att bygga proteiner. Man kan något förenklat säga att en *gen* utgör en DNA sekvens som genom RNA beskriver hur motsvarande protein skall byggas upp. Alla processer som sker i en cell är beroende av fysisk kontakt mellan biomolekyler. Denna interaktion leder till informationsutbyte, transport, utökad eller minskad produktion osv. När små kemiska föreningar binder in till en eller flera biomolekyler leder detta ofta till en biologisk respons. Dessa små molekyler kallas *ligander* och om responsen har en medicinsk tillämpning kallas liganden för ett läkemedel. Figur 1 (Kapitel 1) visar hur det smärtstillande läkemedlet Diclofenac binder in till målproteinet cyklooxygenas.

Den här avhandlingen handlar om protein-ligand interaktion. Arbetet ligger inom proteokemometri och kemogenomik, som är tvärvetenskapliga områden i gränslandet mellan biologi, kemi och datavetenskap. Målet inom *kemogenomik* är att få en övergripande bild av hur alla proteiner (det s.k. *proteomet*) reagerar på tillsatsen av en viss ligand. *Proteokemometri* är en del av kemogenomiken och syftet är mer specifikt att skapa modeller som kan förutsäga interaktionsstyrkan mellan serier av ligander och proteiner. Eftersom antalet humana proteiner är uppskattat till mer än en miljon och antalet möjliga ligander kan uppgå till den astronomiskt höga siffran 10^{62} är det omöjligt att prova alla proteiners respons på alla ligander. Studier inom kemogenomik och proteokemometri fokuserar därför oftast på närbesläktade grupper av proteiner och ligander, istället för att försöka omfatta hela den kända protein-ligand rymden.

Samtliga arbeten (I-V) som ingår i avhandlingen har använt data som hämtats från öppna databaser. Den största databasen som innehåller information om proteiner är Uniprot. Vad gäller ligander så är mycket information tyvärr inte offentlig utan i privat ägo. Inititativ som den öppna databasen PubChem har som syfte att öka tillgängligheten av information om ligander. Andra exempel på databaser är Brenda, AffinDB, BindingDB och PDBBind, som innehåller information om protein-ligand interaktioner. Information om proteiner och ligander har använts som bas för beräkning av deskriptorer. En *deskriptor* är ett sätt att numeriskt beskriva en egenskap hos en molekyl. Till exempel kan en deskriptor vara ett mått på hur benägen en ligand är att tycka om vatten, eller ett mått på hur laddat ett protein kan vara. Det finns mängder av protein- och liganddeskriptorer och urvalet av deskriptorer beror dels på vad modellen skall förutsäga och dels på huruvida det är viktigt att modellen skall vara tolkningsbar.

Artikel I-IV beskriver proteokemometrimodeller som har som syfte att förutsäga styrkan av protein-ligandinteraktioner. Artikel I omfattar tre små dataset där både proteiner och ligander är beskrivs med binära deskriptorer. Modellen är regelbaserad med regler såsom ”OM R1 i liganden är en metylgrupp OCH aminosyraposition 15 är en Arginin DÅ hög bindningsstyrka”. För att denna metod skall klara av att generalisera till nya fall krävs en avgränsning av bindingsaffinitetsvärdena i ”hög” och ”låg” bindningsstyrka. Artikel II visar att det går att använda en ny typ av deskriptorer som är baserade på den 3-dimensionella (3D) strukturen av ett protein. Modellen är gjord på ett relativt litet dataset bestående av lyaser och hydrolaser (proteiner som är enzymer) och deras ligander. I artikel III är datasetet från artikel II utbyggt och innehåller en stor mängd enzymer som varierar både i sekvens, struktur och biologisk funktion. Både i artikel II och III används metoden ”partial least squares” som till skillnad från den regelbaserade metoden i artikel I skapar en modell där ett numeriskt värde på bindningsstyrka förutsägs. Ett stort externt kontrollset genererades genom att länka enzym- och ligandinformation med bindingsaffinitetsvärden. Resultaten från detta kontrollset visar att modellen kan användas för att förutsäga bindingsstyrkan för okända enzymer-ligandkomplex. I artikel IV utprovas generella proteindeskriptorer som beräknas från aminosyrasekvensen istället för 3D strukturen. Eftersom sekvenser är mycket lättare att bestämma än 3D strukturer var förhoppningen att dessa deskriptorer skulle möjliggöra mycket större och mer generella PCM modeller. Deskriptorerna beskriver allmänna egenskaper, som procent laddade aminosyror. För liganderna beräknades väl utprovade deskriptorer som ofta används i traditionella QSAR modeller. Interaktionsdata samt information om proteiner och ligander hämtades från den stora databasen BindingDB. Den bästa modellen skapades med hjälp av så kallade beslutsträd. Andelen korrekt klassificerade komplex i det externa testsetet var nära 0.8. Detta visar att enkla proteindeskriptorer kan användas för PCM. I artikel V presenteras en metod att visualisera och jämföra protein-ligandrymder. Interaktionsdata hämtades från de stora databaserna PDB och DrugBank. PDB innehåller strukturell information, medan DrugBank innehåller information om läkemedel och deras målproteiner. Globala deskriptorer av samma typ som i artikel IV beräknades för alla protein-ligandkomplex. Databasernas protein-ligandrymder visualiserades med hjälp av principalkomponent analys, och jämfördes med närmsta-granne metoden. Rymderna är inte helt överlappande, vilket inte är särskilt överraskande eftersom PDB innehåller många proteiner som inte är målproteiner för läkemedel och DrugBank innehåller många proteiner vars 3D struktur är mycket svår att bestämma. En genomgång av de ”närmsta grannarna” för varje läkemedel i DrugBank som är kända att interagera med mer än ett protein visar dock att en stor andel kända korsinteraktioner förutsägs av metoden.

De arbeten som presenteras i den här avhandlingen har gått från små modeller baserade på ett part dussin protein-ligandkomplex till ganska stora studier som innefattar mer än 15 000 komplex. Detta avspeglar en allmän trend, både i biologi och kemi, mot allt större och mer omfattande studier. För att utforska protein-ligandrymden krävs dels ett ofantligt antal mätningar och dels kraftiga beräkningsverktyg för att hantera och analysera datat. Det är dock värt besväret eftersom det kommer att leda till en ökad förståelse av protein-ligandinteraktion, samt till modeller som kan förutsäga möjliga korsinteraktioner av potentiella läkemedel med andra proteiner i proteomet.

7. Acknowledgements

I wish to thank:

Gerard Kleywegt for being a real role-model scientist and an excellent advisor. I could not have wished for a better one. Thank you Gerard for your patience and many stimulating scientific discussions. I still have a lot to learn, but you should know that I am very grateful for your efforts in teaching me the trade!

My supplementary advisors Torgeir R. Hvidsten, Peteris Prusis, and Jarl ES Wikberg for their support and encouragement. Special thanks to you Torgeir for your excellent work on the local descriptors of protein structure.

My co-authors Pavel Daniluk, Jan Komorowski, Krzysztof Fidelis, Andriy Kryshtafovych, Maris Lapinsh, and Herman Midelfart.

Everybody at the Linnaeus Centre for Bioinformatics for making it fun to go to work. Many thanks to Gunilla for all the administrative support (and therapeutical chats) over the years. I am also very grateful to all present and former PhD students whose generosity and good spirits have created a sharing and open research environment at the LCB. Special thanks to Adam for all the help with graphics (all roads seem to lead to R!) and to Jacke for being a very good colleague during the home stretch of this thesis.

The BMC Computing Department. Many thanks to Emil for help with the Brenda database, Gustavo for sorting out my hardware problems, and Nils-Einar for network support.

Åke Nilsson whose enthusiasm for biology inspired me to opt for university studies.

My family and friends for their love and support. Thank you Andreas for always believing in me and Sonja for constantly reminding me of what is important in life.

References

[1] Rowlinson, S. W.; Kiefer, J. R.; Prusakiewicz, J. J.; Pawlitz, J. L.; Kozak, K. R.; Kalgutkar, A. S.; Stallings, W. C.; Kurumbail, R. G.; Marnett, L. J. A novel mechanism of cyclooxygenase-2 inhibition involving interactions with Ser-530 and Tyr-385. *J Biol Chem* **2003,** *278*, 45763-45769.

[2] Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat Rev Genet* **2004,** *5*, 262-275.

[3] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001,** *409*, 860-921.

[4] Drews, J. Drug discovery: a historical perspective. *Science* **2000,** *287*, 1960-1964.

[5] Drews, J.; Ryser, S. Classic drug targets. *Nat Biotechnol* **1997,** *15*, 1297-1350.

[6] Stockwell, B. R. Chemical genetics: ligand-based discovery of gene function. *Nature Rev Genet* **2000,** *1*, 116-125.

[7] Eyre, T. A.; Ducluzeau, F.; Sneddon, T. P.; Povey, S.; Bruford, E. A.; Lush, M. J. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res* **2006,** *34*, D319-321.

[8] O'Donovan, C.; Apweiler, R.; Bairoch, A. The human proteomics initiative (HPI). *Trends Biotechnol* **2001**, 178-181.

[9] Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods.* **2000,** *44*, 235-249.

[10] Rognan, D. Chemogenomic approaches to rational drug design. *Br J Pharmacol* **2007,** *152*, 38-52.

[11] Savchuk, N. P.; Balakin, K. V.; Tkachenko, S. E. Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr Opin Chem Biol* **2004,** *8*, 412-417.

[12] Surgand, J. S.; Rodrigo, J.; Kellenberger, E.; Rognan, D. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins* **2006,** *62*, 509-538.

[13] Guba, W.; Green, L. G.; Martin, R. E.; Roche, O.; Kratochwil, N.; Mauser, H.; Bissantz, C.; Christ, A.; Stahl, M. From astemizole to a novel hit series of small-molecule somatostatin 5 receptor antagonists via GPCR affinity profiling. *J Med Chem* **2007,** *50*, 6295-6298.

[14] Martin, R. E.; Green, L. G.; Guba, W.; Kratochwil, N.; Christ, A. Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach. *J Med Chem* **2007,** *50*, 6291-6294.

[15] An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* **2005,** *4*, 752-761.

[16] Wikberg, J. E. S.; Lapinsh, M.; Prusis, P., Proteochemometrics: a tool for modeling the molecular interaction space. In *Chemogenomics in drug discovery*, Kubinyi, H.; Müller, G., Eds. Wiley-VCH: Darmstadt, **2004**; pp 289-309.

[17] Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J Chem Inf Model* **2007,** *47*, 195-207.

[18] Nervall, M.; Hanspers, P.; Carlsson, J.; Boukharta, L.; Åqvist, J. Predicting binding modes from free energy calculations. *J Med Chem* **2008,** *51*, 2657-2567.

[19] Prusis, P.; Muceniece, R.; Andersson, P.; Post, C.; Lundstedt, T.; Wikberg, J. E. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. *Biochim Biophys Acta* **2001,** *1544*, 350-357.

[20] Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim Biophys Acta* **2001,** *1525*, 180-190.

[21] Geladi, P.; Kowalski, B. R. Partial least-square regression: a tutorial. *Anal. Chim. Acta* **1986,** *185*, 1-17.

[22] Lapinsh, M.; Prusis, P.; Uhlén, S.; Wikberg, J. E. Improved approach for proteochemometrics modeling: application to organic compound--amine G protein-coupled receptor interactions. *Bioinformatics* **2005,** *21*, 4289-4296.

[23] Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J. E. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinformatics* **2008,** *9*, 181.

[24] Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. S. Generalized proteochemometric model of multiple cytochrome P450 enzymes and their inhibitors. *J Chem Inf Model* **2008,** *48*, 1840-1850.

[25] Hvidsten, T. R.; Kryshtafovych, A.; Fidelis, K. Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. *Proteins* **2008**.

[26] Hvidsten, T. R.; Kryshtafovych, A.; Komorowski, J.; Fidelis, K. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics* **2003,** *19 Suppl 2*, ii81-91.

[27] Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol Pharmacol* **2002,** *61*, 1465-1475.

[28] Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* **2000,** *43*, 3233-3243.

[29] Prusis, P.; Lundstedt, T.; Wikberg, J. E. Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. *Protein Eng* **2002,** *15*, 305-311.

[30] Strombergsson, H.; Prusis, P.; Midelfart, H.; Lapinsh, M.; Wikberg, J. E.; Komorowski, J. Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. *Proteins* **2006,** *63*, 24-34.

[31] Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjoström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem* **1998,** *41*, 2481-2491.

[32] Lapinsh, M.; Prusis, P.; Petrovska, R.; Uhlén, S.; Mutule, I.; Veiksina, S.; Wikberg, J. E. Proteochemometric modeling reveals the interaction site for Trp9 modified alpha-MSH peptides in melanocortin receptors. *Proteins* **2007,** *67*, 653-660.

[33] Kontijevskis, A.; Petrovska, R.; Mutule, I.; Uhlén, S.; Komorowski, J.; Prusis, P.; Wikberg, J. E. Proteochemometric analysis of small cyclic peptides' interaction with wild-type and chimeric melanocortin receptors. *Proteins* **2007,** *69*, 83-96.

[34] Kontijevskis, A.; Prusis, P.; Petrovska, R.; Yahorava, S.; Mutulis, F.; Mutule, I.; Komorowski, J.; Wikberg, J. E. A look inside HIV resistance through retroviral protease interaction maps. *PLoS Comput Biol* **2007,** *3*, e48.

[35] Kontijevskis, A.; Wikberg, J. E.; Komorowski, J. Computational proteomics analysis of HIV-1 protease interactome. *Proteins* **2007,** *68*, 305-312.

[36] Mandrika, I.; Prusis, P.; Yahorava, S.; Shikhagaie, M.; Wikberg, J. E. Proteochemometric modelling of antibody-antigen interactions using SPOT synthesised peptide arrays. *Protein Eng Des Sel* **2007,** *20*, 301-307.

[37] Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* **2006,** *34*, W32-37.

[38] Reference Manual for PROFEAT. http://jing.cz3.nus.edu.sg/prof/prof_manual.pdf

[39] Dubchak, I.; Muchnik, I.; Holbrook, S. R.; Kim, S. H. Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A* **1995,** *92*, 8700-8704.

[40] Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006,** *63*, 892-906.

[41] Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem Cent J* **2007,** *1*.
[42] Rogen, P.; Fain, B. Automatic classification of protein structure by using Gauss integrals. *Proc Natl Acad Sci U S A* **2003,** *100*, 119-124.
[43] Hvidsten, T. R.; Kryshtafovych, A.; Fidelis, K. Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. *Proteins* **2009,** *In press*.
[44] Brenner, S. E.; Koehl, P.; Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* **2000,** *28*, 254-256.
[45] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000,** *28*, 235-242.
[46] Dragon. Talete srl. Via V. Pisani 13, 20124 Milano, Italy. http://www.talete.mi.it/main_exp.htm
[47] Adriana code 2.2. Henkestraße 91, 91052 Erlangen, Germany. www.molecular-networks.com
[48] Todeschinini, R.; Consonni, V., *Handbook of molecular descriptors*. Wiley-VCH: Weinheim, **2000**; Vol. *11*.
[49] Karelson, M., *Molecular descriptors in QSAR/QSPR*. Wiley-Interscience: New York, **2000**.
[50] Terfloth, L., Calculation of structure descriptors. In *Chemoinformatics*, Gasteiger, J.; Engel, T., Eds. Wiley-VCH: Darmstadt, **2003**; pp 401-431.
[51] Leach, A. R.; Gillet, V. J., Molecular descriptors. In *An introduction to chemoinformatics*, Kluwer Academic Publishers: Dordrecht, **2003**; pp 53-75.
[52] Haylett, D. G., Direct measurement of drug binding to receptors. In *Textbook of receptor pharmacology*, Foreman, J. C.; Johansen, T., Eds. CRC Press: Boca Raton, **2003**; pp 153-182.
[53] Cheng, Y.; Prusoff, W. H. Relationship between the inhibition constant (K_I) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem Pharm* **1973,** *22*, 3099-3108.
[54] Jenkinson, D. H., Classical approaches to the study of drug-receptor interactions. In *Textbook of receptor pharmacology*, Foreman, J. C.; Johansen, T., Eds. CRC Press: Boca Raton, **2003**; pp 4-72.
[55] Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res* **2008,** *36*, D674-678.
[56] Okuno, Y.; Yang, J.; Taneishi, K.; Yabuuchi, H.; Tsujimoto, G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res* **2006,** *34*, D673-677.

[57] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **2007,** *35*, D198-201.
[58] Roth, B. J.; Driscol, J. PSDP - NIMH Psychoactive Drug Screening Program. http://pdsp.med.unc.edu/
[59] Barthelmes, J.; Ebeling, C.; Chang, A.; Schomburg, I.; Schomburg, D. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* **2007,** *35*, D511-514.
[60] Puvanendrampillai, D.; Mitchell, J. B. Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics* **2003,** *19*, 1856-1857.
[61] Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res* **2006,** *34*, D522-526.
[62] Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J Med Chem* **2005,** *48*, 4111-4119.
[63] Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G., *Biological sequence analysis*. Cambridge University Press: Cambridge, **2002**.
[64] Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **1970,** *48*, 443-453.
[65] Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **1981,** *147*, 195-197.
[66] Leach, A. R.; Gillet, V. J., *An Introduction to chemoinformatics*. Kluwer academic publisher: Dordrecht, **2003**.
[67] Lapinsh, M.; Prusis, P.; Mutule, I.; Mutulis, F.; Wikberg, J. E. QSAR and proteo-chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes. *J Med Chem* **2003,** *46*, 2572-2579.
[68] Kantardzic, M., *Data mining - concepts, models, methods, and algorithms.* Wiley-Interscience: Piscataway, **2003**.
[69] Hastie, T.; Tibshirani, R.; Friedman, J., *The elements of statistical learning*. Springer-Verlag: New York, **2001**.
[70] Kantardzic, M., Decision trees and decision rules. In *Data mining - concepts, models, methods, and algorithms*, Wiley-Interscience: Piscataway, **2003**; pp 139-164.
[71] Komorowski, J.; Pawlak, Z.; Polkowski, L.; Skowron, A., Rough sets - a tutorial. In *Rough-fuzzy hybridization - A new trend in decision making*, Pal, S. K.; Skowron, A., Eds. Springer Verlag: Singapore, **1999**; pp 3-98.
[72] Pawlak, Z. Rough sets. *Int. J. Comp. Inf. Sci.* **1982,** *11*, 341-356.
[73] Pawlak, Z., *Rough sets - theoretical aspects of reasoning about data.* Kluwer Academic Publishers: Dordrecht, **1991**.
[74] Hastie, T.; Tibshirani, R.; Friedman, J., Support vector machines and flexible discriminants. In *The elements of statistical learning*, Springer-verlag: New York, **2001**; pp 437-504.

[75] Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982,** *143*, 29-36.
[76] SIMCA-P+ 10.5. http://www.umetrics.com
[77] Witten, I. H.; Frank, E., *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufman: San Francisco **2005**.
[78] Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **2008,** *36*, D901-906.
[79] UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res* **2008,** *36*, D190-195.
[80] UniProtKB/Swiss-Prot protein knowledgebase release 56.6 statistics. http://www.expasy.org/sprot/relnotes/relstat.html
[81] Protein Data Bank - growth of released structures per year. http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100
[82] Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **2008,** *24*, 133-141.
[83] Wold, B.; Myers, R. M. Sequence census methods for functional genomics. *Nat Methods* **2008,** *5*, 19-21.
[84] Chandonia, J. M.; Brenner, S. E. The impact of structural genomics: expectations and outcomes. *Science* **2006,** *311*, 347-351.
[85] PubChem. http://pubchem.ncbi.nlm.nih.gov/
[86] ChEMBL. http://www.ebi.ac.uk/chembl/
[87] Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcantara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **2008,** *36*, D344-350.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 608

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-89299

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2009