UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 40

Development of Proteochemometrics—A New Approach for Analysis of Protein-Ligand Interactions

MARIS LAPINS

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006

ISSN 1651-6192
ISBN 91-554-6695-8
urn:nbn:se:uu:diva-7211

Dissertation presented at Uppsala University to be publicly examined in B21, BMC, Husargatan 3, Uppsala, Friday, November 24, 2006 at 09:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English.

Abstract

Lapins, M. 2006. Development of Proteochemometrics—A New Approach for Analysis of Protein-Ligand Interactions. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy* 40. 64 pp. Uppsala. ISBN 91-554-6695-8.

A new approach to analysis of protein-ligand interactions, termed proteochemometrics, has been developed. Contrary to traditional quantitative structure-activity relationship (QSAR) methods that aim to correlate a description of ligands to their interactions with one particular target protein, proteochemometrics considers many targets simultaneously.

Proteochemometrics thus analyzes the experimentally determined protein-ligand interaction activity data by correlating the data to a complex description of all interaction partners and; in a more general case even to interaction environment and assaying conditions, as well. In this way, a proteochemometric model analyzes an "interaction space," from which only one cross-section would be contemplated by any one QSAR model.

Proteochemometric models reveal the physicochemical and structural properties that are essential for protein-ligand complementarity and determine specificity of molecular interactions. From a drug design perspective, models may find use in the design of drugs with improved selectivity and in the design of drugs for multiple targets, such as mutated proteins (e.g., drug resistant mutations of pathogens).

In this thesis, a general concept for creating of proteochemometric models and approaches for validation and interpretation of models are presented. Different types of physicochemical and structural description of ligands and macromolecules are evaluated; mathematical algorithms for proteochemometric modeling, in particular for analysis of large-scale data sets, are developed. Artificial chimeric proteins constructed according to principles of statistical design are used to derive high-resolution models for small classes of proteins.

The studies of this thesis use data sets comprising ligand interactions with several families of G protein-coupled receptors. The presented approach is, however, general and can be applied to study molecular recognition mechanisms of any class of drug targets.

Keywords: chemometrics, QSAR, G-protein coupled receptors, Melanocortin receptors, protein-ligand interactions

Maris Lapins, Department of Pharmaceutical Biosciences, Box 591, Uppsala University, SE-75124 Uppsala, Sweden

ISSN 1651-6192
ISBN 91-554-6695-8
urn:nbn:se:uu:diva-7211 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7211)

List of Papers

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

I. Lapinsh M, Prusis P, Gutcaits A, Lundstedt T, and Wikberg JE (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim Biophys Acta* **1525:**180–190.

II. Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, and Wikberg JE (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci* **11:**795–805.

III. Lapinsh M, Prusis P, Lundstedt T, and Wikberg JE (2002) Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol Pharmacol* **61:**1465–1475.

IV. Lapinsh M, Prusis P, Mutule I, Mutulis F, and Wikberg JE (2003) QSAR and proteochemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes. *J Med Chem* **46:**2572–2579.

V. Lapinsh M, Veiksina S, Uhlen S, Petrovska R, Mutule I, Mutulis F, Yahorava S, Prusis P, and Wikberg JE (2005) Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes. *Mol Pharmacol* **67:**50–59.

VI. Lapinsh M, Prusis P, Petrovska R, Uhlen S, Mutule I, Veiksina S, and Wikberg JE (2006) Proteochemometric modeling reveals the interaction site for Trp9 modified α-MSH peptides in melanocortin receptors. Submitted.

VII. Lapinsh M, Prusis P, Uhlen S, and Wikberg JE (2005) Improved approach for proteochemometrics modeling: application to organic compound–amine G protein-coupled receptor interactions. *Bioinformatics* **21:**4289–4296.

Contents

Abbreviations

2D	two-dimensional
3D	three-dimensional
ACC	auto- and cross-covariances
CoMFA	Comparative Molecular Field Analysis
GOLPE	Generating Optimal Linear PLS Estimations, a method for variable selection
GPCR	G-protein coupled receptor
GRID	three-dimensional molecular descriptors
GRIND	three-dimensional alignment-independent molecular descriptors
K_i	dissociation constant
pK_i	negative logarithm of dissociation constant
MACC	maximum auto- and cross-covariance
MC	melanocortin
MC_1R-MC_5R	melanocortin receptor subtypes 1 to 5
MIF	molecular interaction field
MLR	multiple linear regression
MM	molecular mechanics
MSH	melanocyte simulating hormone
NIPALS	non-linear iterative partial least squares
NMR	nuclear magnetic resonance
OSC	orthogonal signal correction
PCA	principal component analysis
PCM	proteochemometrics
PCR	principal component regression
PLS	partial least-squares projections to latent structures
Q^2	predictive ability of a model
QM	quantum mechanics
QSAR	quantitative structure-activity relationships
R^2	goodness of fit of a model
SDEP	standard deviation of errors of prediction
SED	statistical experimental design
TM	transmembrane region
VIP	variable importance in projection
X	matrix of independent variables
Y	matrix of dependent variables

1 Introduction

Molecular interactions determine all activities performed by living organisms. Detailed analysis of the mechanisms involved provides the key to understanding the function of cells and their response to foreign substances. Even the most modest estimates of the total number of proteins encoded by the human genome suggest around 30,000 proteins (not accounting for multiple gene product variants arising from alternative gene splicing and single nucleotide polymorphism). Such a multitude of entities emphasizes the importance of selectivity for molecular interactions. So-called "drug-like" chemical compounds are estimated to be able to access and influence over 3,000 protein ligand-binding sites in humans (Hopkins and Groom, 2002, Lipinski and Hopkins, 2004). Although there are numerous promiscuous drugs known to interact with multiple proteins, promiscuity of interactions may result in adverse effects and toxicity, which is a main reason for failure in drug development (Kennedy, 1997). Achieving selectivity is a difficult task, however, since major drug target classes, such as cell membrane receptors, proteases, and kinases, comprise numerous members that share similar properties. It is well known that 3D structures of homologous proteins are more conserved than primary sequence and function. Accordingly, proteins that have diverged functionally during evolution may still share the same structural organization and exploit similar molecular interaction mechanisms, thus making the design of selective drugs problematic.

Furthermore, many disease conditions such as central nervous system disorders have a very complicated pathophysiology, and effective medications for these diseases have complex and to a large extent not fully understood pharmacology. Not surprisingly, treatments of diseases such as schizophrenia and depression by drugs selective for single molecular targets are often unsuccessful, and multi-selective drugs (humorously labeled "magic shotguns" in a recent review by Roth et al., 2004, in contrast to "magic bullets") may give better results.

Current drug discovery relies essentially on combinatorial chemistry and high throughput screening. Once a hit compound is found for a particular target it can be optimized by computational methods. Thus, if the 3D structure of a protein has been derived by X-ray or NMR spectroscopy, docking of hit-like structures from virtual compound libraries can be performed. Furthermore, even in the absence of initial hits, *de novo* ligand design can be performed provided that the protein 3D structure is available. However, obtaining high-resolution structures for macromolecules is mostly problematic. Determining the 3D structures of proteins constitutes a bottleneck, in particular for integral proteins. The use of NMR is limited to quite small proteins, and does not work at any practical level for integral proteins. It is also generally acknowledged that membrane-embedded proteins are not readily amenable to existing crystallization methods for determining X-ray structures (Renfrey and Featherstone, 2002). Thus, for example, only one X-ray structure of a G-protein coupled receptor, namely that for bovine rhodopsin, has been determined to atomic resolution (Palczewski et al., 2000). Although the creation of homology models of related proteins could then be attempted, the accuracy of modeling and quality of current docking algorithms and scoring functions are most often insufficient to make such attempts very useful.

Another widely applied computational method for ligand optimization is QSAR. In QSAR, chemical descriptors of series of ligands are correlated to some type of biological activity (e.g., interaction strength with a particular target) by mathematical modeling. A radically new type of QSAR was recently introduced to analyze interactions of multiple ligands with multiple targets, termed proteochemometrics (Wikberg et al., 2004). In proteochemometrics the interaction strength is correlated to a complex description of several targets interacting with several ligands. In a general case, the interaction environment and assaying conditions are also included in the modeling. In this way, a proteochemometrics model analyzes "interaction space," from which only one cross-section would be contemplated by any QSAR model.

A pioneering proteochemometrics study (Prusis et al., 2001) brought the hope that if the approach were extended and used properly it might vastly increase the resolution in the modeling of molecular interactions. Having in hand such models might in turn facilitate ligand design and protein engineering. From a drug design perspective, the thought was that proteochemometrics might find use in the design of drugs showing selectivity between subtly different proteins, and in the design of drugs for a wider variety of targets, such as mutated proteins (e.g., drug resistant mutations of pathogens), as well as in the design of multifunctional drugs with fine-tuned interaction profiles.

2 Aims

The aim of this thesis was to develop and evaluate proteochemometrics modeling of protein-ligand interactions. For this purpose, several data sets for amine GPCRs (studies I, III, and VII) and subtypes of melanocortin GPCRs (studies IV, V, and VI) were collected. One of the studies (II) was more general and was devoted to the development of alignment-independent description of proteins. This was used for the classification of all rhodopsin-like (Class A) GPCRs into 12 families according to the type of their ligand.

Specifically, the goals of the studies were to:

- Evaluate molecular descriptions of ligands and proteins for PCM modeling
- Develop the representation of a PCM object (ligand-protein cross-description)
- Evaluate mathematical algorithms, in particular for modeling of large-scale data sets
- Evaluate methods for proper validation of models in PCM
- Compare PCM and QSAR models
- Develop methods for interpretation of molecular descriptors and their cross-terms
- Evaluate of the use of artificially modified chimeric proteins to increase the resolution of PCM

3 Background

3.1 Mathematical modeling of molecular interactions

3.1.1 Basic assumptions

The basic assumption in mathematical modeling of molecular interactions is that the biological properties of any molecule – from small organic compounds to biomacromolecules – are completely determined by its chemical structure. Differences in biological properties ultimately arise from structural differences (although the opposite statement would not always be true – a structural change does not necessarily cause a change in biological function). A further assumption is that these differences can be expressed in quantitative terms. A modeling study may then aim to find an empirical equation that would relate the structures of compounds to their biological properties, for example the strength of a molecular interaction. This task of finding an empirical equation is, however, as a rule not easy, and careful validation of the obtained quantitative structure-activity relations is always required. Among the risks that may endanger a modeling study are possible differences in the mechanisms of action of molecules (e.g., different binding sites, orientations, and conformations). To avoid this risk, models are often limited to small structural domains of similar molecules. Moreover, even if structures are active for the same reason, they can become inactive for many different reasons. As a result, a large number of descriptors of various molecular properties might be required to allow one to obtain an accurate relationship (which in turn may require assessment of a large number of structures to ensure a statistical significance of the models obtained). In fact, several research fields have evolved under the QSAR umbrella to facilitate modeling of molecular interactions, such as development of relevant chemical descriptors and the development of methods for multivariate data analysis, as well as methods for the statistical design of data sets.

3.1.2 History of structure-activity relationship studies

The first efforts to find structure-activity relationships were undertaken in the latter part of the nineteenth century, when chemists started to investigate the biological effects of chemical compounds (Dunn, 1989). As early as

1868, Crum-Brown and Fraser proposed that the basis of the physiological action of compounds was their chemical structure (Crum-Brown and Fraser, 1868). In particular, they showed in their study that, when quaternized, strychnine, morphine, and other alkaloids lost their characteristic properties and acquired curare-like muscle-relaxing properties. At the turn of the twentieth century, Meyer and Overton, working independently, found a relationship between the anesthetic action of many compounds and their lipophilicity represented by olive oil–water partitioning coefficients (Meyer, 1899, Overton, 1901).

After decades of little additional development, Hammett quantified the influence of the electronic and steric properties of organic acids and bases on their reactivity and on the rates and equilibrium constants of chemical reactions (Hammett, 1940). This provided the basis for works that established QSAR as a new scientific discipline. Pioneering studies conducted by Corwin Hansch et al. in the 1960s used regression analysis to correlate the biological activity of a molecule to both electronic and steric parameters (represented by substituent constants) and lipophilicity (hydrophobic properties) (Hansch et al., 1962, Hansch and Fujita, 1964, Hansch, 1969). At the same time, Free and Wilson (1964) proposed an empirical approach that used indicator variables to relate biological activity to the presence or absence of certain substituents in a molecule. An important milestone in the development of QSAR as a standard technology in drug design was the invention of 3D descriptors, such as GRID developed by Goodford (1985) and CoMFA by Cramer et al. (1988). Protein QSAR studies have been facilitated by the development of multivariate quantitative description of amino acids, first by Hellberg et al., 1987. An example of protein QSAR is a study by Sjöstrom et al, 1995, which exploited the description of amino acid sequences of bacterial proteins to predict their location to different cellular compartments.

Even though the works by Hansch in the 1960s exploited a few molecular property descriptors, it was acknowledged that QSAR is in principle a multivariate problem which is often non-linear (e.g., the use of a square term of logP was required in Hansch models). By increasing the complexity of description, new sophisticated correlation methods were required. Therefore, the development of partial-least square (PLS) algorithms that facilitate analysis of multivariate data must be mentioned among the milestones in QSAR history (PLS was originally introduced by Herman Wold in 1966 for PCA and in 1973 for path modeling, and was finalized to its present form in 1983 by Svante Wold, Harald Martens, and Herman Wold with the PLS regression with orthogonal scores (Wold, 1973, Wold et al., 1983)). The application of multivariate mathematical methods to extract chemically relevant information from data produced in chemical experiments has subsequently become an independent scientific discipline, termed “chemometrics.”

3.1.3 The chemometrics approach to multivariate soft modeling

Chemometrics has developed gradually since the early 1970s when the term was coined by analogy with biometrics, econometrics, etc. (Wold, 1995). An exact definition of chemometrics has been the subject of debate, which can be attributed to its broad applicability in different fields of chemistry. These include optimization of synthetic procedures, multivariate calibration of spectroscopic instruments, design of experiments and libraries of chemical compounds, pattern recognition and classification, QSAR studies, and monitoring and control of complex chemical systems and processes. However, in the absence of a strict definition a reasonable summary of the tasks of chemometrics is given by one of its founders, Svante Wold (1995):

> "How to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into data."

In contrast to the rules of classical physics, which are thought to be universal, chemometrics deals with empirical or so-called "soft modeling." This is due to the nature of chemical measurements, which always contain uncertainty. Thus, an exact quantum mechanical solution has only been found for systems composed of one or two particles, e.g., the hydrogen atom. For small molecules it is possible to arrive at approximate solutions, which may be achieved by applying various theoretical approximations and considering empirical observations. The description of larger chemical systems becomes very computation intensive and requires more rough approximations and more empirical data. Since the modeling of such systems is based on a limited domain of data, it cannot be extrapolated for predictions far beyond the investigated domain.

Moreover, the chemical data are typically multivariate. For example, series of chemical structures can be characterized by numerous descriptors, and numerous parameters can be measured at different time points of a chemical process. Efficient discovery of relationships within the data therefore requires multivariate approaches rather than analysis of one or two variables at a time.

The success of chemometrics due, for the most part, to the tools of multivariate data analysis which can handle data sets with more variables than objects, where variables are multicollinear and contain missing and erroneous data. Two of the most important techniques ("the basic tools") are PCA and PLS (discussed in section 3.3 of this thesis).

3.2 Molecular descriptors

The first step in the establishment of a QSAR relationship is the creation of a mathematical representation of structures by a set of molecular descriptors. Extensive development of QSAR methodologies has given rise to a vide variety of descriptors encoding molecular structure and particular properties (a comprehensive overview of definitions and methodologies for the calculation of more than two thousand molecular descriptors is given by Todeschini and Consonni (2000)).

Molecular descriptors can be broadly grouped into three major types: measured physicochemical properties, descriptors calculated from 2D structure, and descriptors from experimentally determined, or modeled, 3D structures of molecules.

Measured properties include properties of substances, such as solubility in different liquids, logP, chromatographic properties, melting and boiling points, and different kinds of spectral data.

Calculated descriptors include for example molecular weight and the number of different types of atoms, bonds, rings, and structural fragments. A special type of descriptors are binary or so-called Free-Wilson descriptors (also called “structure keys”), which indicate merely the presence or absence of a certain structural feature rather than counting how many times that feature is present in the molecule (Free and Wilson, 1964, Xue et al., 2003, McGregor and Pallai, 1997). A large collection of structure keys can be hashed into a so-called “molecular fingerprint” of predefined length (Xue et al., 2003, 2003b).

By using algorithms of graph theory, 2D structures can also be represented by a huge variety of topological descriptors and connectivity indices (Todeschini and Consonni, 2000).

Molecular interactions, however, take place in three-dimensional space, advocating the use of descriptors that would directly represent 3D structures of molecules. Where the 3D structures can be superimposed within the studied series of compounds, widely used approaches are CoMFA (Comparative Molecular Field Analysis) (Cramer et al., 1988) and GRID (Goodford, 1985, http://www.moldiscovery.com). In CoMFA, steric and electrostatic interactions of a compound are calculated with a probe atom (sp3-hybridized carbon with +1 charge) at the intersections of a 3D lattice. Compared to CoMFA, GRID offers many different probe groups, thus enabling more thorough description of the capacity of compounds for non-covalent interactions. Moreover, recent improvements to GRID allow the movement of flexible side chains of structure in response to interactions with a probe group to be accounted for (the flexible algorithm was originally designed to handle amino acid side chains) (Goodford, 1998, Afzelius et al., 2004). GRid INdependent descriptors (GRINDs), a successor method to GRID that involves calculation of GRIDs but does not require the superimposition of structures, is presented in the next section.

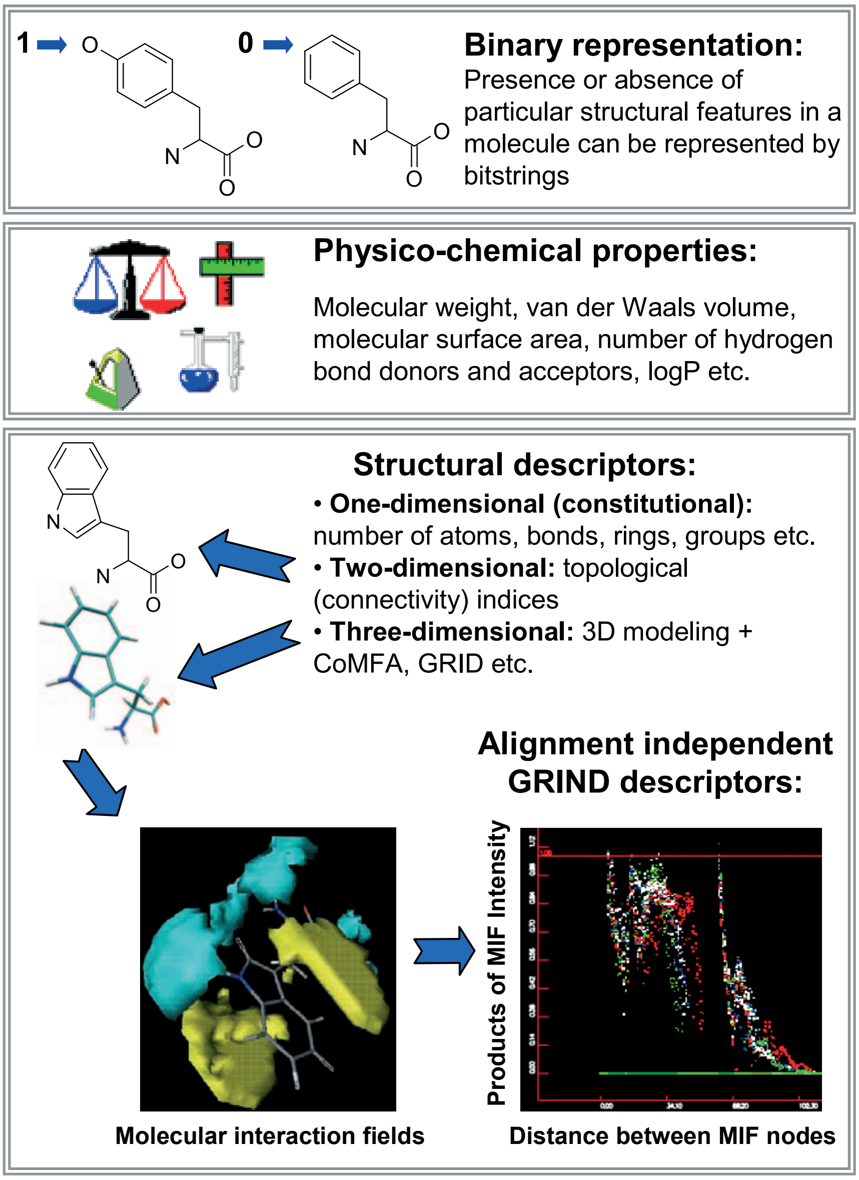

1
O
N
O
O
0
N
O
O
Binary representation:
Presence or absence of particular structural features in a molecule can be represented by bitstrings
Physico-chemical properties:
Molecular weight, van der Waals volume, molecular surface area, number of hydrogen bond donors and acceptors, logP etc.
Structural descriptors:
• One-dimensional (constitutional): number of atoms, bonds, rings, groups etc.
• Two-dimensional: topological (connectivity) indices
• Three-dimensional: 3D modeling + CoMFA, GRID etc.
N
N
O
O
Alignment independent GRIND descriptors:
Products of MIF Intensity
Molecular interaction fields
Distance between MIF nodes

Figure 1. Different types of molecular descriptors

Several recent studies suggest benefits from the extension of descriptors beyond the 3D representation of a single conformer. Concepts of so-called 4D to 6D-QSAR have been presented, which allow different conformations/orientations of a molecule, protonation states, and in some cases induced fit scenarios and solvation models to be accounted for (Vedani and Dobler, 2002, Vedani et al., 2005).

3.2.1 GRIND descriptors

GRINDs are 3D descriptors that represent the ability of a molecule to form favorable interactions with independent pharmacophoric groups. The generation of GRINDs involves the computation of molecular interaction field maps (using GRID software) but does not require the superimposition of 3D structures and is independent of the orientation of structures to each other (i.e., it is alignment independent). GRINDs can thus be computed for any series of molecules (Pastor et al., 2000).

3.2.1.1 Calculation of molecular interaction fields

Calculation of GRINDs starts with positioning of a molecule in a grid box. The interaction energies of a molecule with a pharmacophoric probe group located on the grid points (i.e., grid lattice intersections) surrounding the molecule are then calculated. ALMOND software (Multivariate Infometric Analysis S.r.l., http://miasrl.com) allows the use of up to 4 of 18 different probes, the three most widely used being DRY (hydrophobic probe), O (carbonyl oxygen, i.e., H-bond acceptor), and N1 (amide nitrogen, i.e., H-bond donor). The interaction energy is calculated considering the influence of Lennard-Jones interactions, electrostatic effects, and hydrogen bonding as follows:

$$E_{xyz} = \sum E_{LJ} + \sum E_{hb} + \sum E_q + S$$

where

Σ indicates pair-wise energy summation between the probe and every atom in the molecule

E_{xyz} – interaction energy of a probe having coordinates xyz,

E_{LJ} – energy of the Lennard-Jones potential function,

E_{hb} – energy of hydrogen bonding,

E_q – energy of the electrostatic interactions,

S – entropic term, associated with the hydrophobic (DRY) probe.

The Lennard-Jones potential is an empirical function that explains the attractive and repulsive forces between interacting atoms and is dependent on the Van der Waals radius, polarizability, and effective number of electrons of the atoms. The electrostatic interactions depend on the charges of interacting atoms and the distances between them.

Unlike the former two types of interactions, the calculation of hydrogen bonding accounts for the geometrical arrangement of atoms, i.e., the strength of a bond decreases as the geometry becomes less optimal. (For equations and a full description of the diverse contributions accounted for while calculating molecular interactions, see http://www.moldiscovery.com/docs/).

The whole set of E_{xyz} terms obtained by a certain probe throughout the whole space surrounding the molecule is termed the molecular interaction field (MIF). Hence, using *n* various probe groups result in *n* MIF maps for a molecule.

3.2.1.2 Obtaining auto- and cross-correlograms

The next step in obtaining GRINDs is filtering each MIF map. Only those grid nodes that show the energetically most favorable interactions with the molecule and are concomitantly situated as far as possible from each other are extracted from the MIF map. ALMOND software allows one to choose the number of nodes to extract and the relative importance of each of the two criteria for the selection of nodes. By default, extraction of 100 nodes is suggested for small molecules (however, a larger number might be required for charged structures and structures with more than a few H-bond donors/acceptors).

Once the nodes are extracted from the MIF maps, the distances between all pairs of nodes within the same MIF and from different MIFs are calculated. Moreover, the products of the energy values for each pair of nodes are calculated. These products can now be plotted against the distance between nodes, giving an auto-correlogram for each MIF and a cross-correlogram for each pair of MIFs. Finally, the maxima of products falling within specified distance ranges (smoothing windows) in each correlogram are used as GRIND descriptors for the molecules.

3.2.1.3 Representing molecular shape

It was recently recognized that the GRIND descriptions had deficiencies in some cases since they did not include explicit description of the shape of the molecule (Fontaine et al., 2004). For this reason, the original MIF-based descriptors were complemented by a “molecular shape field” (termed TIP in ALMOND software), based on the local curvature of the molecular surface. In the new settings, using one of the probe groups at a certain repulsion energy outlines the surface of the molecule. The local curvature of the surface is then calculated at each node. Convex regions are considered to be more important than concave because the former may form complementary interactions or cause steric hindrances. Filtering of the thus obtained molecular shape field map is performed by selecting the most convex nodes. The curvature values of the molecular shape field nodes are then employed similarly as energy values of the MIF nodes to create auto- and cross-correlograms.

3.2.2 Modeling of 3D structures

The calculation of GRINDs requires 3D representations of the molecules. Aside from the experimental determination of 3D structure, two groups of modeling methods can be distinguished, theoretical and empirical. The theoretical methods differ in their level of sophistication and can be further divided into the computationally most consuming quantum mechanical (QM) approaches and the less consuming molecular mechanics (MM) methods. QM takes into account the electronic structure of the molecule to calculate the energy of the system, while in MM electrons are ignored and the molecule is treated as a set of balls (heavy atoms) connected by springs (bonds). In the MM approach, the energy of the molecule is calculated from the positions of atoms using a force field. A force field contains a set of potential energy functions (such as bond stretching and bending energy, torsion angle energy, Lennard-Jones interactions and electrostatic interactions of atoms). Many customized packages for QM and MM calculations are available in standard molecular modeling software – the user's choice is mainly dependent on available time and computational resources as well as on the size and type of molecules considered. Once the energy of a molecule and energy gradient, with respect to conformational changes, have been calculated, geometry optimization can be performed in an iterative manner until the molecule reaches some (local) energy minimum. However, although different minimization algorithms have been developed to overcome local energy barriers, none of them has proven to be capable of finding global energy minima from an arbitrary starting geometry. Therefore, the energy minimizations must be performed starting from a large number of initial conformations, which could be selected in a random (Monte Carlo search) manner or by systematic exhaustive modification of angles of rotatable bonds. In this way, ensembles of low energy conformations of molecules are obtained. For flexible structures, several conformations may appear to be energetically similar, and the global minimum conformation can be influenced by the selected force field parameters. A further drawback of theoretical methods is that the molecular environment can only be partially accounted for in MM, while QM simulations are performed in vacuum and thus the molecular environment is completely neglected.

Empirical methods are used for the creation of 3D structures by automatic structure generators, such as Corina and Cobra. The 3D structures are there built based on sets of rules derived from experimental data (e.g., X-ray crystallographic), from force field calculations for pre-defined molecular fragments, as well as from geometric considerations. The primary application of these programs is obtaining 3D models of structures collected in large chemical databases.

3.2.3 Z-scale description of amino acids

The description of large biopolymers might be performed in the same way as for any other chemical molecule. However, to avoid extreme complexity of description one can take the advantage of their polymeric nature. Thus, to encode a protein one can characterize it from the properties of its sequence monomers. In fact, all proteins are constructed from the same common set of twenty amino acids, and all diversity of protein structures and functions results from the differences of these twenty building blocks and their order in the protein primary sequence.

Descriptors for amino acids can be developed based on measured and computed physicochemical properties of the amino acid monomers. One early study used 29 properties, including properties such as liquid chromatographic mobilities in HPLC using nine different solvents, TLC mobilities, and ^{13}C-NMR α-proton shift (Hellberg et al., 1987). PCA was applied to reduce the number of descriptors, giving three principal property scales (z1, z2, and z3) that could be tentatively interpreted as reflecting hydrophobicity, steric, and electronic properties.

In more recent studies z-scales were extended to characterize properties of both natural and synthetic amino acids. Thus, Sandberg et al. (1998) obtained five z-scales from 26 computed and measured properties for 87 coded and non-coded amino acids. It is notable that the three first of the latter z-scales describe about 70% of the variation in the original data, whereas all five describe more than 95% of the variation. Principal property descriptors for the 20 natural amino acids have also been developed using GRID description of amino acids (Cocchi and Johansson, 1993, Crucianni et al., 2004).

3.2.4 Description of biopolymer sequences

3.2.4.1 Alignment-based approaches

With descriptors of amino acids in hand, the amino acid sequence of a whole protein can be translated into a vector of numbers. In cases where the sequences of several proteins are alignable, comparisons of resulting vectors would directly represent the chemical variation between these proteins. One such example is rhodopsin-like GPCRs, where sequence portions of TM regions can be aligned unambiguously. In this case, z-scales of each aligned sequence residue can be used to form a uniform **X** matrix, which is a prerequisite for any subsequent modeling on the data.

With minor alignment gaps (and, accordingly, subtle differences in protein 3D structures), the description of aligned sequences could still be used in forming a uniform matrix. For the modeling, the gaps could be treated as missing data or be represented by some indicator variable showing the presence or absence of amino acid in a particular position.

3.2.4.2 Alignment-independent approaches

Unfortunately, despite the fact that proteins may show substantial conservation in their structural and functional organization, their primary sequences are seldom conserved to the extent that alignments can be made unambiguously.

If protein sequences were aligned wrongly, it would destroy any attempts to compare the difference in their chemical space. Accordingly, methods are preferred that avoid the alignment step and transform the physicochemical descriptions directly into uniform matrices. One such method, called auto- and cross-covariance (ACC) transformations, was developed, describing the changes in physicochemical property or property combinations over sequence stretches of different lengths, over the whole protein sequence (Wold et al, 1993). This is done according to equations:

$$\mathrm{AC}_{z,lag} = \sum_{i}^{n-lag} \frac{V_{z,i} * V_{z,i+lag}}{(n-lag)^{p}}$$

$$\mathrm{CC}_{z_a \neq z_b,lag} = \sum_{i}^{n-lag} \frac{V_{z_a,i} * V_{z_b,i+lag}}{(n-lag)^{p}}$$

where AC represents auto-covariances of the same z-scale and CC the cross-covariances of different z-scales, and where $z = 1, 2, \ldots, Z$ (Z is the number of z-scales), $i = 1, 2, \ldots, n$ (i is the amino acid position in the sequence and n the total number of amino acids), $lag = 1, 2, \ldots, L$ (L is the maximum lag, i.e., the longest sequence stretch used), V is the descriptor value, and p the degree of normalization of the ACC term. The total number of ACC terms depends on the chosen L (maximum lag can be up to the length of the shortest sequence in the dataset) and number of z-scales and is $L*Z^2$.

In this way, an ACC provides a uniform matrix that captures sets of characteristic physicochemical patterns of the protein. Applying ACC transformations on z-scale encoded sequences has been successfully used for descriptions of polypeptides and proteins (Sjöström et al., 1995, Edman et al., 1999]. One limitation of ACCs is that the local sequence patterns may be hidden by the overall properties of the given sequence. Use of maximum ACCs (in a similar manner as for calculating GRIND descriptions) has recently been proposed for describing peptides (Cruciani et al., 2004). These descriptors would capture extreme property combinations in the sequence; however, their success in the encoding of proteins is not yet proven.

Yet another description method is based on amino acid and/or amino acid pair frequencies in the protein, introduced by van Heel (1991). A drawback of this approach is that the similarities and differences in physicochemical properties of amino acids are ignored.

3.3 Multivariate data analysis methods

Mathematical modeling of molecular interactions aims at finding a relationship between descriptors characterizing the molecules interacting with each other and the experimentally measured strength or activity of these interactions. Having *k* measured or calculated molecular descriptors (*x*-variables) and the measured interaction activity *y*, we are looking for the function $f(x_1, x_2, ..., x_k)$ that would allow us to calculate the value y with as small as possible error *e*:

$$y = f(x_1,x_2,...,x_k) + e$$

Several methods exist that may be useful in deriving this function empirically. These methods can be divided into two major groups: linear and non-linear. Examples of non-linear methods potentially suited for the purpose are neural networks (Zupan and Gasteiger, 1999, Almeida, 2002), support vector regression (Wapnik, 1995, Scholkopf et al., 1998), and non-linear partial least squares (Wold et al., 1989, Baffi et al., 1999). Although non-linear models may seem more adaptable and flexible compared to linear models, they may show disadvantages such a tendency to fit outliers, clustered data, and pure noise (Hawkins et al., 2001, Ngo et al., 2003, Wold et al., 1989, Geladi and Kowalski, 1986). Moreover, even if the fit and predictions within the modeled experimental domain (e.g., series of similar compounds) are good, the extrapolations for compounds outside this domain are very unsafe. In contrast, linear relations are considered more robust and easy to interpret. Examples of linear modeling methods are multiple linear regression (MLR), principal component regression (PCR), and partial least squares projections to latent structures (PLS). PLS is considered “the basic tool” of chemometrics and is the one used in all the studies in this thesis. However, in order to understand PLS it is also necessary to give some details on the two former methods.

3.3.1 Multiple Linear Regression

In the linear case, the measured activity y_o for each object *o* can be expressed by multiple linear regression:

$$y_o = b_1x_{o1} + b_2x_{o2} + \ldots + b_kx_{ok} + e_o$$

or in matrix notation for all objects:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

The solution of this equation is confined to finding the vector of the regression coefficients **b**. However, an infinite number of exact solutions exists in the case that the number of variables exceeds the number of objects (*k*>*o*). Thus, in this case it is impossible to establish a unique relationship between **X** and **y**. If *k*<*o*, an approximate solution which minimizes **e** can be found. The most common method for finding **b** is is to use the least-square solution:

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

However, a problem may still arise in using this approach, since $\mathbf{(X'X)}^{-1}$ cannot be calculated if any of the **X** variables are collinear. Moreover, in cases where some of the **X** variables are nearly collinear the solution is very unstable.

Unfortunately, these requirements of MLR are usually not met in chemical data. The chemical descriptors generally tend to be correlated, e.g., numerous topological descriptors are related to the size of the molecule. Moreover, the explanation of complex molecular interactions may require the use of numerous descriptors. This in turn would necessitate the collection of a large number of chemical objects to perform correlation by MLR.

One option to overcome these problems is to approximate **X** by a new, smaller matrix of uncorrelated descriptors, which would satisfy the requirements of MLR. This can be done by applying PCA.

3.3.2 Principal Component Analysis

PCA is a multivariate projection method, which provides a compression of data sets containing large numbers of variables (Wold et al., 1987). Contrary to the original variables, which are multicollinear, the so-called principal components (PCs) are orthogonal to each other. The first component represents the largest variance in the data set, the second component the largest of the remaining variance, and so on. The major patterns within the original data can therefore often be captured by a small number of components. PCA can thus be used to visualize relationships among objects (e.g., chemical compounds), to reveal correlations among variables (e.g., compound descriptors), and to separate systematic data structure from noise (e.g., true values of compound properties from experimental error). PCA is useful for the identification of outliers and clusters of similar objects, which in certain cases may benefit from being analyzed separately. Moreover, PCA can be used to predict if a novel object is similar to objects in the modeled dataset or if it is different.

PCA starts from the **X**-matrix, consisting of *o* rows (objects) and *k* columns (variables).

From a geometric point of view, each object can be considered as a point in a k dimensional hyperspace. The first step of PCA is the centering of variables to ensure that PCs are drawn through the origin of the hyperspace. The first PC can then be imagined as a line in the hyperspace, drawn so that the sum of squared distances from each point to this line is as small as possible. The coordinates of the objects on the line are termed scores t, while the direction coefficients (angle cosines) between the original variables and the principal component and are termed loadings p. The residuals, $e_{ok} = x_{ok} - t_o p_k$, form a new matrix **E** comprising the unexplained part of the object coordinates. The second and further PCs may then be iteratively calculated from the residual matrix of the previous component (a graphical representation of PCA is outlined in Figure 2). In this way, after calculating A components, the **X** matrix with the size of o rows and k columns is decomposed into two smaller matrices, the score matrix **T** of size o by A and the loading matrix **P** of size k by A, according to the following equation:

$$\mathbf{X} = \mathbf{TP'} + \mathbf{E}$$

where **P’** is the transposed **P** matrix.

The value for descriptor k for observation o, here named x_{ok}, can be recalculated from scores ($t_{o1}, t_{o2}, ..., t_{oA}$) of observation o and loadings ($p_{k1}, p_{k2}, ..., p_{kA}$) of variable k with some residual e_{ok}, according to:

$$x_{ok} = t_{o1} p_{k1} + t_{o2} p_{k2} + ... + t_{oA} p_{kA} + e_{ok}$$

The number of principal components, which would capture all systematic variation of the original data, leaving the non-systematic part (e.g., measurement errors) in the residual matrix, can be determined by cross-validation as described in Eriksson et al., 1997, or by comparing the eigenvalues of principal components.

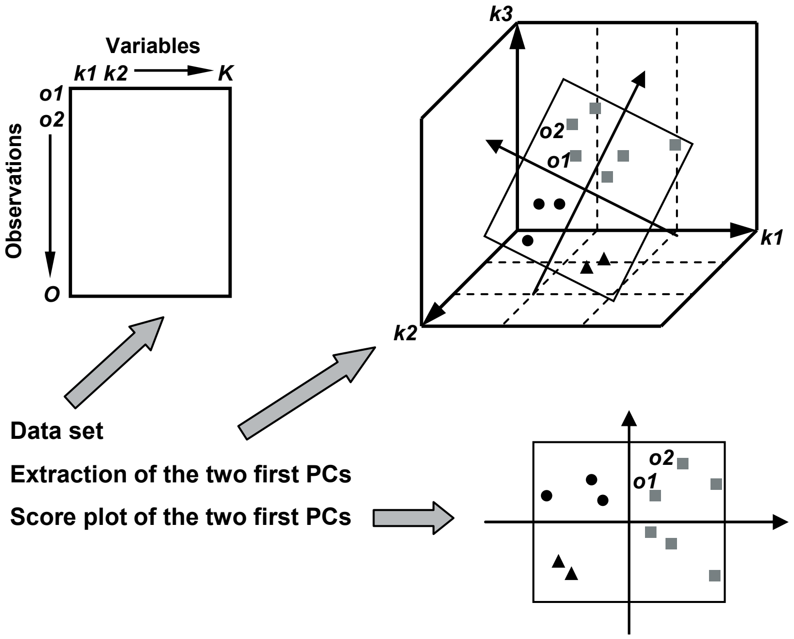

Variables
k1 k2
K
o1
o2
Observations
O
k3
o2
o1
k1
k2
Data set
Extraction of the two first PCs
Score plot of the two first PCs
o2
o1

Figure 2. Outline of Principal Component Analysis. The figure shows the dataset, the directions of the two first principal components and the corresponding score plot. Score plots provide a graphical representation of how the objects are related to each other. Objects with extreme score values are strong outliers that have probably influenced the direction of a PC (weaker outliers, i.e., the ones that have not influenced the PCA model, can be detected by looking at their residues). Similarly, loading plots reveal the importance of variables for the generated PCA model and relationships between variables (correlated variables are close together or opposite one another on the diagonal in the loading plots).

3.3.3 Principal Component Regression

This approach of a combined PCA and MLR is called principal component regression (PCR). PCA satisfies both requirements of MLR, namely it yields uncorrelated components and the number of components cannot exceed the number of objects.

However, the two-step process of PCR may cause some obstacles. The problem is related to the selection of the optimal number of principal components for MLR. On the one hand, we can elect to use the few first components and thus reduce noise (irrelevant effects, random error, etc.), which would end up in the residual matrix. However, we thereby take a risk that useful information will be lost in discarded components while some noise will still remain in the components used for regression. On the other hand, we may elect to use a large number of PCs. However, this would increase the risk for chance correlations with **y**. A solution to this problem is given through the one-step process of PLS.

3.3.4 Partial Least-Squares Projections to Latent Structures

PLS can be considered as an extension of PCA that along with the **X** matrix of predictor variables deals with the dependent **Y** matrix or vector. PLS aims to find the relationship between the two matrices and to develop a predictive model. This is achieved by simultaneously projecting **X** and **Y** to latent variables (components), with an additional constraint to correlate them. (Thus, compared to PCs, PLS components are tilted to maximize covariance between projections of **X** and **Y.)**

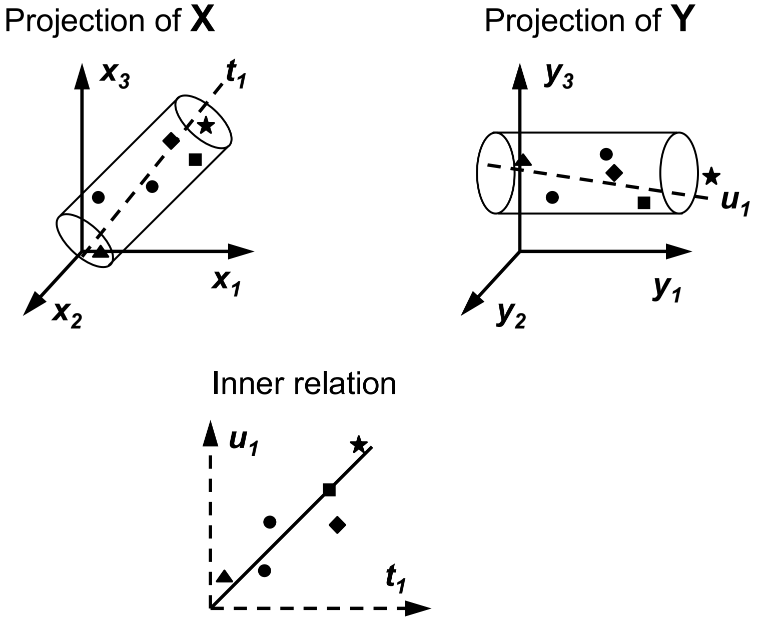


Figure 3. Outline of PLS. The upper left panel shows the projection of **X** to the first PLS component t_1. The upper right panel shows the projection of **Y** to the first PLS components u_1. Cylinders indicate the directions of principal components in PCA (as can be seen, PLS components are tilted compared to those obtained by PCA). The lower panel shows the inner relation, i.e., the correlation between projections of **X** and **Y**.

Mathematically, the approach can be represented by the following equations:

$$\mathbf{X} = \mathbf{TP'}+\mathbf{X}_{\mathbf{res}} \text{ (the approximation of } \mathbf{X}\text{)}$$

$$\mathbf{Y} = \mathbf{UC'}+\mathbf{Y}_{\mathbf{res}} \text{ (the approximation of } \mathbf{Y}\text{)}$$

$$\mathbf{U} = \mathbf{T} + \mathbf{U}_{\mathbf{res}} \qquad \text{(the inner relation)}$$

Here $\mathbf{X}_{\mathbf{res}}$, $\mathbf{Y}_{\mathbf{res}}$, and $\mathbf{U}_{\mathbf{res}}$ are residual matrices. **T** and **U** are score matrices of **X** and **Y**, respectively, while **P’** and **C’** are inverses of loading matrices of **X** and **Y**. Moreover, the PLS algorithm uses additional loadings, **W**, called weights, which express the relationship between **U** and **X** and are used to obtain **X** block scores

$$\mathbf{T} = \mathbf{XW}$$

We can now re-write the above equations to look like a multiple regression model:

$$\mathbf{Y} = \mathbf{XWC'}+\mathbf{Y}_{\mathbf{res}}$$

The first and most widely used PLS algorithm is NIPALS, developed and finalized to its present form in 1983 by Svante Wold, Harald Martens, and Herman Wold (Wold et al., 1983). In a similar fashion as in PCA, the NIPALS algorithm extracts one component at a time.

Since PLS is a projection method it allows us to handle data sets with an essentially unlimited number of variables. Variables can be noisy and collinear. Moreover, the NIPALS algorithm handles missing data well, as long as they are distributed randomly in the data set (Nelsson et al., 1996). PLS allows one to analyze several response variables simultaneously, although performing separate modeling is suggested if responses are not correlated. Separate modeling then gives models with fewer components, which are easier to interpret (Wold et al., 2001).

An important decision in PLS is the choice of the number of components. Each extracted component increases the explained variation of both **X** and **Y**. However, while the first components normally find real correlations between the two blocks, increased model complexity may give rise to chance correlations. Therefore, model validation is required to avoid overfitting, i.e., obtaining a better-fitted model but with reduced predictive ability.

3.3.5 Validation of models

Validation of a model is of crucial importance in any empirical modeling. Validation is used to find optimal model parameters (such as complexity of a PLS model) and to assess the reliability of interpretations and usability of the model for future predictions.

Two statistical parameters are frequently used to characterize a QSAR model. The first is goodness of fit, R^2, which represents the fraction of the explained variance of the dependent variable *y* and is calculated as follows:

$$R^2 = 1 - \text{RSS/SS}$$

where RSS is the sum of squares of the residuals of **y** and SS is the initial sum of squares of mean-centered *y*.

The second is Q^2, which represents the predictive ability of the model assessed by cross-validation. In cross-validation the objects are divided into a number of groups. Models are then developed from the dataset reduced by one of the groups and predictions for the excluded objects calculated. The process is then iteratively repeated until all groups have been omitted once. The Q^2 is then calculated as follows:

$$Q^2 = 1 - \text{PRESS/SS}$$

where PRESS is the sum of squares of the differences between observed and predicted *y* values.

The R^2 values vary between 0 and 1, where a value closer to 1 means a better fit. The Q^2 values normally vary between 0 and R^2; however, negative values may be encountered indicating unpredictive models. In PLS, the R^2 term increases with each extracted component, while the Q^2 value usually reaches a plateau or declines as the model becomes overfitted.

A $Q^2 > 0.4$ is generally considered acceptable, and a $Q^2 > 0.8$ as excellent (Lundstedt et al., 1998). A difference between R^2 and Q^2 of more than 0.2-0.3 is a sign of the presence of chance correlations with irrelevant descriptors, or of outliers in the data, and indicates a risk for erroneous interpretations (Eriksson et al., 1996).

An alternative method for estimating the predictive ability of a model, termed Q^2_{cum}, is implemented in the Simca software (www.umetrics.com). In this case, $\text{PRESS}_a/\text{SS}_{a-1}$ is calculated separately for each individual PLS component *a*=1,...*A*, and Q^2_{cum} is estimated as

$$Q^2_{\text{cum}} = 1 - \prod_{a=1}^{A} (\text{PRESS}_a/\text{SS}_{a-1})$$

where Π indicates the product of $PRESS_a/SS_{a-1}$ for individual components. In calculating Q^2_{cum}, the cross-validation groups are rearranged for each component, which can be an advantage if cross-validation is not performed repeatedly. In most cases, Q^2 and Q^2_{cum} give similar values of the predictive ability, although in a few cases a discrepancy is observed (Freyhult et al., 2005).

To assess the statistical significance of the estimated Q^2 value (of the Q^2_{cum} value, using the Umetrics approach), cross-validation can be complemented by response permutation testing as described (Eriksson and Johansson, 1996). In short, the **y**-data are randomly shuffled a number of times. Models are then fitted to the permuted **y**-data, performing cross-validation and extracting the same number of PLS components as in the model for the real **y**-data. The obtained distributions of the R^2 and Q^2 values for the **y**-permuted models may then be compared with the estimates of the real model. If the models on the randomized data result in substantially lower values of R^2 and Q^2 this indicates statistical validity of the original estimates.

Usually cross-validation is performed using five or more randomly formed groups, depending on the size of the dataset. The special case where the number of cross-validation groups are increased to the number of observations (the so-called leave-one approach) is not recommended because the predictability may then be overestimated (Shao, 1993, Wold and Sjöström, 2001).

However, the nature of proteochemometrics data may require more hash validation modes than those with randomly formed groups. This is because a proteochemometrics object is a protein-ligand combination. Hence, even when the object is removed from the data set, its partial description (i.e., description of a protein and a ligand, but not the cross-description) would still be present. In other words, the predictions are made for proteins and ligands already present in the data set, but in other combinations. Since one of the purposes of proteochemometrics is to make predictions for completely novel ligands, we have suggested a cross-validation mode where all objects including the same ligand are assigned to one and the same cross-validation group. Similarly, to assess the predictive ability for new proteins, cross-validation can be performed assigning all objects including the same protein to one and the same cross-validation group.

Several studies prompt a warning that there may be a lack of a direct relationship between the Q^2 of QSAR models and the quality of their external predictions (Peterson et al., 2006, Kubinyi et al., 1998, Doweiko, 2004). This finding (sometimes called the "Kubinyi paradox") may be partially associated with possible redundancies in the data sets, and to cases where high Q^2 values are found by trying different modeling settings and applying feature selection. Nevertheless, external validation should be recommended in all cases when the size of a data set allows setting aside a sufficiently large test set. The whole model development process (i.e., data centering, scaling, variable selection, determination of optimal model complexity etc. – see below) is then carried out using the training set objects. If the predictions for the test set objects come close to the Q^2 according to cross-validation, this would secure the validity of the model.

At last, the interpretation of the results of the modeling can be considered an additional though not exactly defined validation method. Thus, the inspection of plots of the model and the underlying data may facilitate the detection of clusters, outliers, and other possible abnormalities in the data. If the interpretations of scores, loadings, and regression coefficients were found to be reasonable and consistent with existing chemical and biological knowledge, this would serve as an additional proof of model validity.

3.3.6 Elaboration of models

PLS allows one to use a multitude of variables, which may be an advantage if particular variables are only partially relevant. The larger the number of variables the more clear the "systematic" part in data is expected to be (Eriksson and Johansson, 1996, Wold and Sjöström, 1998). A typical data set exploiting 3D descriptions of organic structures may include tens of thousands of variables (Wold et al., 2001). However, not all of these variables would likely encode features related to the interaction activity **y**. PLS modeling of such data might explore the **X**-space instead of maximizing the correlation between **X** and **y**. This in turn would result in models with higher complexity than necessary, leading to complicated interpretations and sometimes also a large increase in prediction errors. Moreover, a high proportion of uninformative variables risks explaining **y** by chance correlations, which in turn would risk ending up with a deteriorated model.

Methods for the elimination of non-relevant variables have therefore been sought. There are several variable selection methods in common use. One category of these exploits the predictive ability as the objective function. In this approach, variables that influence the Q^2 value positively according to cross-validation are retained, while the others are excluded from the model. The methods can use genetic algorithms (Yasri and Hartsough, 2001) or fractional factorial design to create multiple models, each of which is subjected to cross-validation (e.g., the GOLPE method, Baroni et al., 1993). These methods, as well as the other methods mentioned below, may be applied singly or iteratively to improve a model.

Other methods include eliminating variables with negligible regression coefficients, or small variable importance in projection (VIP) values in the PLS model (Wold, 1995), and methods that use cross-validation to estimate the variance of regression coefficients among the cross-validation rounds. The latter type of assessment of the significance of variables is termed jackknifing or uncertainty testing (Westad and Martens, 2000, Martens and Martens, 2000).

An alternative to variable selection methods is to apply signal corrections. One such approach is so-called orthogonal signal correction, or OSC, which has the purpose of pre-processing the **X** matrix so that strong systematic but irrelevant variation is removed (i.e., variation in the **X** matrix being uncorrelated to **y**). Since its original publication (Wold et al., 1998), a number of algorithms have been developed for different OSC filters (Sjöblom et al., 1998, Svensson et al., 2002). In general, a principal component of the **X** matrix that is orthogonal or nearly orthogonal to **y** is calculated (filters, however, differ in the procedures for searching **y**-orthogonal variation and do not produce the same results). The extracted component is then subtracted from **X**, yielding a new corrected **X** matrix. If desired, further OSC of this matrix can then be performed. PLS modeling of OSC-filtered data yields lower complexity models with improved interpretability and has been successfully applied in for example NIR spectroscopy calibration to remove base-line variation and scatter effects. However, it has also been reported to result in overfitted or even degraded calibration models. Moreover, in QSAR studies it is found that despite better cross-validation statistics the external predictions of OSC-filtered models may actually become worse (Bohac et al., 2002). These findings indicate that the benefits of OSC filtering are warranted only if strong structured noise is expected to be present. An improved algorithm, known as orthogonal PLS, has been developed that detects and removes the structured **y**-orthogonal variation in **X** only when it disturbs the interpretation of the model (Trygg and Wold, 2002).

Finally, it should be stressed that rigorous external validation using new data is required to assess whether or not the variable selection or signal correction procedure have improved a model.

3.4 G-protein coupled receptors (GPCRs)

GPCRs comprise one of the most abundant families of cell surface receptors, with more than 1,000 members known in humans. Activated by extracellular stimuli, GPCRs transduce signals to intracellular heterotrimeric G-proteins initiating important signaling pathways in the cell. The extracellular ligands are diverse and numerous, and include biogenic amines, peptides, nucleotides, lipids, and proteins. GPCRs can be broadly divided into odorant/sensory receptors and transmitter receptors, the latter being among the most important targets for drug discovery. Estimates suggest that more than half of currently marketed drugs act as either agonist or antagonist on GPCRs, or interfere with signaling pathways activated by GPCRs (Drews, 2000, Müller, 2000, Wise et al., 2002). However, current drugs target only about 30 of around 500 transmitter receptors (in fact, most of the top-selling therapeutics are directed to the relatively few monoamine GPCRs). The natural ligand has been identified for most receptors, leaving around 160 so-called "orphan" receptors with unknown ligand and function (Wise et al., 2002, Bock and Gough, 2005).

GPCRs share a common structural organization with seven α-helical transmembrane spanning segments, connected by three extracellular and three intracellular loops. The amino terminus of GPCRs is located on the extracellular side and the carboxyl terminus on the intracellular side of the cell membrane.

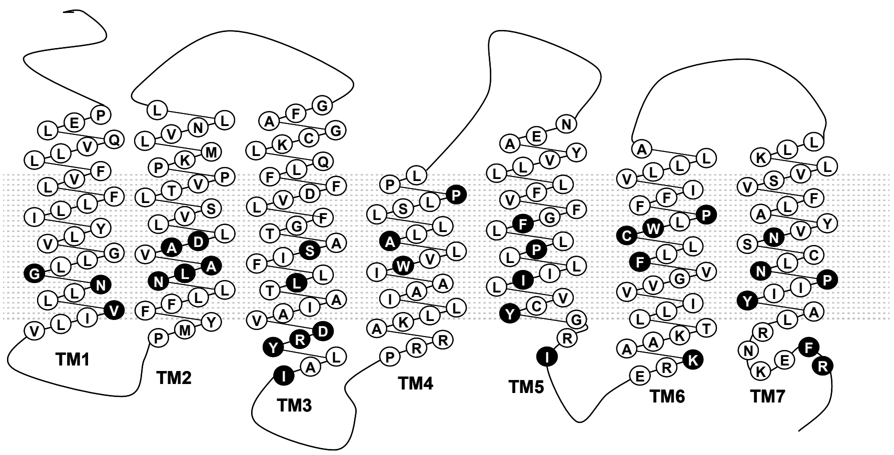


Figure 4. Schematic representation of rhodopsin-like (class A) GPCRs. The most conserved sequence positions according to Baldwin et al., 1997 are colored in black.

GPCRs can be divided into several families, which share little or no sequence similarity. The largest family is rhodopsin-like receptors, followed by secretin-like receptors and metabotropic glutamate/pheromone-like receptors, also termed classes A, B, and C. Each of the seven transmembrane regions of class A GPCRs share at least three amino acid residues, which are highly conserved throughout the family (Baldwin et al., 1997). Thanks to these conserved residues, the sequences of TM regions of class A receptors can be unambiguously aligned, with few exceptions (II, Horn et al., 2003).

Tentative 3D models of TM regions of class A GPCRs can subsequently be derived based on the single resolved X-ray structure of a GPCR, namely bovine rhodopsin (Palczewski et al., 2000). Docking of ligand structures to homology-based models has been undertaken (Evers and Klabunde, 2005, Evers et al., 2005, Bissantz et al., 2005) in virtual screening of chemical libraries. Several studies suggest that docking can be a useful approach, e.g., for lead finding when either little or no information about the active ligands is available (Evers et al., 2005b, Bissantz et al., 2003).

The ligand interaction sites for large proteins and peptides involve mainly the extracellular loops and N-terminus of GPCRs. However, many of these large ligands are also known to have additional points of interactions within the transmembrane cavity. By contrast, the binding pocket for small amines is deeply buried between the transmembrane helices, where the positively charged amide group is thought to interact with the conserved aspartic acid residue in the third TM region of receptors (position 3.31 in Van Rhee code) (Gether, 2000, Jakoby et al., 1999, Van Rhee and Jacobson, 1996). Small peptides (e.g., the melanocortin peptides examined in this thesis) presumably enter the transmembrane cavity, although to a different degree. In particular, site-directed mutagenesis and 3D modeling studies have indicated that binding of the active core of the melanocortin peptides lies within the transmembrane cavity (Prusis et al., 1995, 1997, 2001, Nickolls et al., 2003, Yang et al., 2000), while similar studies for the MC receptor antagonist Agouti-related protein have revealed interactions with the extracellular domain.

In the last decade, with high-throughput screening several small organic compounds have been found that act as antagonist or agonist of some peptide GPCRs but have no structural similarity with the endogenous peptide agonist (Gether, 2000, Dragic et al., 2000, Selnic et al., 2003). Interestingly, mutational studies suggest that these compounds do not always share the binding site with the endogenous peptide, but instead bind in the TM cavity of GPCRs in another way. For example, the interaction sites of tachykinin peptides (substance P, neurokinin A, and neurokinin B) are identified at the N-terminus and extracellular loops of tachykinin receptors TACR1, TACR2, and TACR3. In contrast, the binding site of some small non-peptide antagonists of tachykinin receptors is found to be located in a transmembrane crevice lined by TM 3, 5, and 6 of TACR1 (Gether, 2000). Similarly, the binding of organic antagonists of angiotensin receptor AGTR1 is found to be severely affected by mutations in TM 3 and 7, but unaffected by mutation of residues in the extracellular domains known to affect the binding of the peptide agonist angiotensin (Perlman et al., 1995).

3.4.1 Melanocortin receptors

Melanocortin receptors (MCRs) are members of the Class A (rhodopsin-like) GPCR superfamily. Five MCR subtypes, MC_{1-5}, are known in humans, each having distinct important physiological functions. Thus, MC_{1s}Rs play regulatory roles in the immune system and control melanin pigment formation in the skin; MC_2Rs regulate corticosteroid production; MC_3 and MC_4Rs are involved in controlling sexual and feeding behaviors; and MC_5Rs control sebum secretion from sebaceous glands (Wikberg et al., 2003).

The melanocortin system is regulated by a group of endogenous agonist peptides, the melanocyte-stimulating hormones (MSH) α-MSH, β-MSH, γ-MSH, and adrenocorticotropin (ACTH), and by the endogenous antagonists Agouti and Agouti-related protein. The agonist peptides α-MSH, β-MSH, and γ-MSH show similar patterns of binding affinity and agonistic potencies

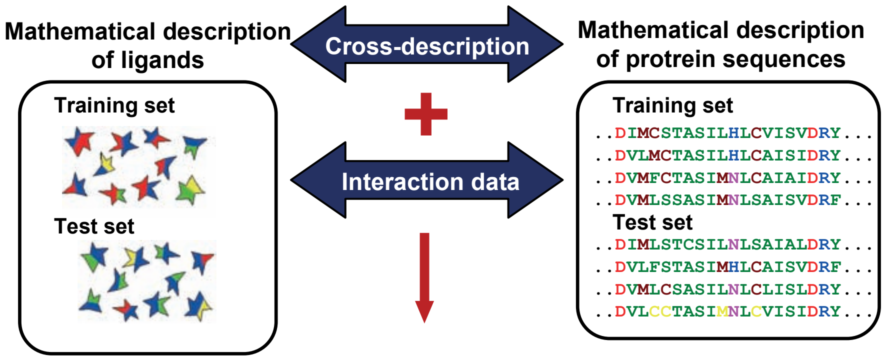
MC_1>MC_3>MC_4>MC_5. The MC_2R is somewhat different since it binds only the ACTH (Schioth et al., 1996).

4 Proteochemometrics

4.1 Outline of proteochemometrics

As its name suggests, proteochemometrics can be considered as an extension of chemometrics. In PCM, chemometric methods are applied to data characterizing proteins and other molecules interacting with proteins. In paper II we used proteochemometrics to perform functional classification of receptor proteins (Class A GPCRs) according to the type of their natural ligand, while in papers I and III-VII we performed quantitative analysis of interactions of series of proteins with series of organic compounds.

An outline of PCM in the studies of protein-ligand interactions is presented in Figure 5. Here the modeling uses the measured interaction activity of series of ligands with varying chemical properties for series of proteins showing sequence variation. The interaction data for training set ligands and training set proteins is correlated to chemical descriptions of ligand-protein pairs by a suitable mathematical method. The data for test set ligands and proteins is used to validate the resulting mathematical model, i.e., to assess its predictive ability. Once the validity has been affirmed, the model can be used for interpretation of factors that determine protein-ligand interaction activity and for prediction of the activity of novel ligands and novel proteins.


Mathematical description of ligands
Cross-description
Mathematical description of protrein sequences
Training set
Test set
Interaction data
Training set
..DIMCSTASILHLCVISVDRY...
..DVLMCTASILHLCAISIDRY...
..DVMFCTASIMNLCAIAIDRY...
..DVMLSSASIMNLSAISVDRF...
Test set
..DIMLSTCSILNLSAIALDRY...
..DVLFSTASIMHLCAISVDRF...
..DVMLCSASILNLCLISLDRY...
..DVLCCTASIMNLCVISIDRY...

•**Modeling –** PLS or other suitable method creates regression equation correlating descriptors to interaction data
•**Validation** is used to assess the optimal complexity of the PLS model and its predictive ability
•**Interpretation** is based on regression coefficients of descriptors and their-formed cross-terms
•**Predictions** can be performed for new ligand-protein combinations, new ligands and new proteins

Figure 5. Outline of one approach for PCM.

The basic correlation method used in our studies is PLS, although other methods such as ridge regression, support vector regression, neural networks, and rough set rules have been attempted by others or might be explored in the future (Strömbergsson et al., 2006, Bock and Gough, 2002, 2003, 2005). The chemical description usually includes three blocks: ligand descriptors, protein descriptors, and ligand-protein cross-terms typically obtained by multiplication of mean-centered descriptors of ligand and receptor. This basic set of descriptors can be extended by other blocks, e.g., cross-terms formed inside the ligand descriptor block or inside the protein descriptor block, or higher order terms, for example squares of any of the above descriptors (the necessity and meaning of particular descriptor blocks will be discussed further below).

The ligand descriptors in our studies included Free-Wilson (binary) description (paper I), GRINDs (papers III-IV, VII), geometrical and topological descriptors (paper IV), and functional group and physicochemical property descriptors (paper V).

Proteins were represented by z-scale descriptions of aligned (I-III, V-VII) primary amino acid sequences or ACC transformed z-scale descriptions of unaligned sequences (II). Sequence similarity descriptors were used when the number of proteins was low (paper IV) and binary descriptors were applied to encode chimeric proteins (papers V, VI).

Interaction activity for use in PCM can be measured by any standard assay, such as radioligand binding (used in all our studies) or enzymatic assay, and can be expressed as a uniform measure such as the logarithmically transformed ligand-protein dissociation constant (pK_i). Often the assay conditions are not identical for all ligand-protein pairs in the dataset, which can influence measurements (for example, the buffer compositions of assay systems are known to affect the interaction strength). In such cases, the differences could be accounted for by complementing the ligand-protein description with assay environment descriptors, which would help in the modeling and further contribute to the understanding of molecular interaction processes.

4.2 Collection of proteochemometric data sets

The advent of genome and proteome research has resulted in well-organized protein databases, allowing the extension of PCM to large-scale modeling and the performing of proteome-wide predictions. The only limitation to the size of a PCM data set is the necessity for ligands to share the same type of binding site in the proteins (although this requirement may not seem stringent, its violation would likely jeopardize the straightforwardness of modeling and interpretations of the model). Until now, unified models have been created for large data sets of GPCR-organic compound and protease-substrate interactions (Kontijevskis et al., unpublished data, Strombergsson et al., 2006); studies of other sets of comparable interactions, in a broad sense, are underway.

Although in the past the interaction data have mainly been gathered in a random fashion, interaction databases already exist both in the commercial and public domains populated by literature data or providing original, systematically collected data (http://www.cerep.fr, www.gpcr.org, http://bind.ca, http://www-mitchell.ch.cam.ac.uk/pld, http://pdsp.cwru.edu).

The number of ligands that can be used in PCM is essentially unlimited. This implies that only a negligible proportion of all possible interaction data can be measured. Hence, approaches for statistical experimental design should be used to collect the most informative dataset with the minimal experimental work (Morgan, 1991, Eriksson et al., 2001). SED methods have already been used to avoid brute force testing of huge chemical libraries (Jamois, 2003, Jamois et al. 2003, Schneider, 2002). By applying SED the informational content of chemical libraries can be retained using a small number of compounds. For example, in one study the size of an informative peptoid library was decreased from more than 100,000 to 90 compounds (Linusson et al., 1998-1999). Several reviews on the application of SED in chemometrics have been published (Wold et al., 2004, Eriksson et al., 2001).

Although most hopes have been placed on large-scale modeling, PCM can also be applied at a high-resolution to model ligand interactions with small families of subtly different proteins. In models for subtypes of MC receptors [papers V-VI], the series of four native proteins were complemented by chimeric three-part constructs, where each part was taken from one of the proteins. Also in this case we applied an experimental design to reduce experimental work without compromising the informational content of the data set.

4.3 The role of cross-terms in ligand-protein description

Using PLS, the correlation of *L* ligand descriptors, *P* protein descriptors, and their cross-terms to interaction activity can be expressed by the regression equation:

$$y = \bar{y} + \sum_{l=1}^{L}(coeff_l * x_l) + \sum_{p=1}^{P}(coeff_p * x_p) + \sum_{1=1,p=1}^{L*P}(coeff_{l,p} * x_l * x_p)$$

where *coeff* are the regression coefficients, which by their sign and magnitude represent how the descriptors are correlated to interaction activity *y*.

Analysis of this equation reveals the role of each block of descriptors in explaining differences in ligand-receptor activity. If we imagine a hypothetical situation where the activity can be explained solely by ligand descriptors (i.e., ligand descriptors were the only ones obtaining non-zero coefficient values), this would only be possible if any one ligand has identical activity for all proteins. If the activity were not affected by properties of proteins, a reasonable interpretation would then be that the ligands interacted with invariant parts of proteins (e.g., the totally conserved amino acids of amine GPCRs in study VII).

In reality, however, the activity of any ligand for different proteins varies, which can then partially be explained in the regression model by the protein descriptor block.

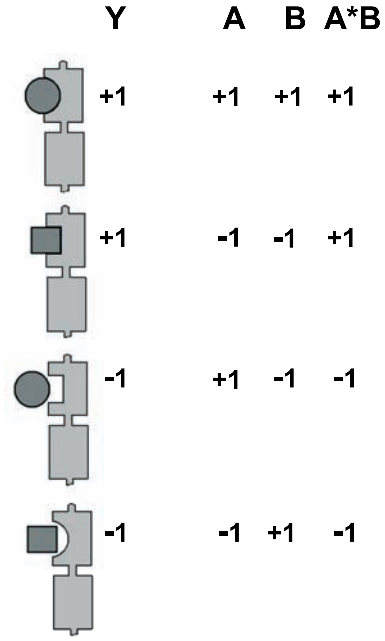
However, a model including only ligand and protein descriptor blocks would neither be able to explain differences in the ligands' selectivity profiles nor proteins' selectivity profiles, which are always apparent in any real data set. Moreover, such a model would predict that any change in a ligand would result in an equal increase or decrease of the activity for all proteins, placing us in the same unlikely situation that ligands interact only with the invariant parts of proteins (a similar interpretation from the protein point of view would suggest that the ligands also interacted with proteins by virtue of some invariant ligand feature). Thus, even if the model could partially explain activity differences, any interpretations would lead to logical flaws and no predictions how to alter the selectivity would be possible.

The situation changes by introducing cross-terms, which represent the complementarity of ligand and protein properties. Used along with the ordinary descriptors, cross-terms allow modeling of the non-linear part of a ligand-protein “activity surface” arising from positive and negative cooperation of ligand and protein properties in the complex process of binding.

The role of cross-terms in explaining ligand-protein binding is further illustrated in Figure 6, where the interactions of two ligands, L1 and L2, with two proteins, P1 and P2, are presented. L1 and L2 differ in some structural property, represented by the descriptor A. This descriptor is assigned the value 1 for L1 and −1 for L2. Other properties of both the ligands are invariant and are therefore not assigned any descriptor. In a similar way, imagine that the proteins differ by some property that is represented by descriptor B attaining values 1 for P1 and −1 for P2. Moreover, the receptors also contain invariant parts that cannot be assigned any descriptors.

Let us further assume that the ligands have the same average activity for the two proteins but opposite selectivity. Thus, L1 binds with high activity to P1 and with low activity to P2, whereas L2 binds with high activity to P2 and with low activity to P1. Accordingly, the activity can be represented by the numbers 1 and −1. Inspection of the figure shows that there is no correlation between A and the activity (since both ligands have the same average activity for the two proteins). Nor is there a correlation between B and the activity, since the proteins do not differ in the ligands’ average activity. By contrast, a cross-term between A and B shows a one-to-one correlation with the interaction activity. Thus, the regression equation for this simulated example would be

$$y = 0 * \mathrm{A} + 0 * \mathrm{B} + 1 * \mathrm{A} * \mathrm{B}$$



Y	A	B	A*B
+1	+1	+1	+1
+1	-1	-1	+1
-1	+1	-1	-1
-1	-1	+1	-1

Figure 6. The role of cross-terms in PCM. Neither ligand nor protein descriptors are correlated with interaction activity (*y*), while the cross-term shows a one-to-one correlation with *y*.

The conclusion can thus be drawn that cross-terms are related to the complementarity between varied parts of ligands and varied parts of proteins. Thus, cross-terms allow one to explain the selectivity of molecular interactions. This contrasts with original descriptors, which explain differences in average affinity. Using real numbers instead of binary descriptors does not change the above reasoning in principle, except that the studied effects are described on a continuous scale rather than in an all or nothing fashion. Using richer descriptions will, however, allow a more detailed interpretation of the nature of the ligand-protein interactions. Thus, for a real data set with numerous structural or physicochemical descriptors, analysis of regression coefficients for ligand-protein cross-terms will allow us to localize the place and type of the molecular interactions.

A further analysis reveals that cross-terms inside the ligand descriptor block are required in situations when two different ligand properties cooperate causing effects on the ligands' ability to interact with the proteins. These cross-terms relate to more complex changes within ligands, such as intramolecular interactions and conformational changes. In an analogous way, protein cross-terms relate to interaction effects within the protein.

4.4 Pre-processing of variables

A standard procedure in chemometrics applied prior to PCA and PLS is centering of variables. Subsequently, scaling of variables to unit variance can be performed. The latter, however, is not compulsory if only one type of variable is employed in the modeling (e.g., GRINDs or z-scales). Yet another kind of scaling is block scaling, where the initial variance of each variable of some type is multiplied by a constant (termed block-scaling weight). Block scaling is required in PCM since the number of cross-terms typically greatly surpasses the number of descriptors of ligands and proteins. Block-scaling weights generally depend on the nature of the data, and the optimal scaling (i.e., scaling producing predictive models) can be found empirically. An effective optimization strategy is Simplex, described by Lundstedt et al., 1998.

The scaling weights of cross-terms determine the degree of non-linearity of a PCM model. Hence, if the scaling weight of ligand-protein cross-terms is set too low, the model is not capable of explaining differences in the selectivity of ligand-protein interactions. Accordingly, the goodness of fit is typically low and the predictive ability may be less than optimal. In the opposite case, when a very high scaling weight is given to cross-terms, the model becomes non-linear. As a result, perfect fits can often be achieved; however, very large mispredictions (in particular for observations on the boundary of the modelled experimental domain) can occur.

There is a possibility that the modeler observes a plateau of Q^2 values over a large range of ratios of the scaling weight for cross-terms versus scaling weights for ligand and/or protein descriptors. In such cases the choice of scaling is a tradeoff between many small or a few large mispredictions, and depends on the goals (and to a large extent on the expertise) of the modeler.

4.5 Interpretation of complex models

Once we have gained an understanding of the meaning of descriptors and cross-terms, a simple way to interpret a proteochemometric model is to inspect the sizes and directions of the regression coefficients of the regression equation (i.e., when modeled using linear methods, such as PLS). A large absolute value of a coefficient for a ligand or protein descriptor would indicate that the property represented by this descriptor has a large influence of the interaction activity. The sign of the coefficient would show the direction of the influence. Similarly, a large absolute value of a coefficient for a cross-term would indicate high importance of the represented combination of features for selective interactions. Having this information, we could gain a detailed picture of the nature of the molecular interactions. Moreover, the information would guide us how to modify a ligand and/or a protein in order to achieve a desired property for their interaction.

The use of PLS coefficients for interpretation is straightforward when we have a clear understanding of the physical meaning of descriptors. However, for large data sets it may be required to transform descriptors into PCA scores (see VII) whose meaning may not always be obvious. However, as PCA scores are derived linearly it is possible to compute the regression coefficients for initial descriptors. If we have computed cross-terms between PCA scores, we can also compute regression coefficients of initial descriptors; in this case the back-derived regression equation will also contain cross-terms of initial descriptors.

A problem in using regression coefficients for model interpretation arises when the number of descriptors becomes large. Since the human brain can only consider a few factors simultaneously, it then becomes difficult to grasp all the particular descriptors. In this case one would prefer to derive some type of overview or more generalized picture, such as having measures for groups of descriptors. For example, we might want to know the importance of one amino acid, a sequence stretch, or some part of a ligand, all of which would be described with several descriptors. One approach to obtaining such an overview could be to sum the absolute values of PLS coefficients for a group of descriptors and to use this sum as a measure for the importance of this group (see I). However, a problem of using such overviews may occure since any ordinary descriptor has indirect non-linear correlations to the interaction activity through the cross-terms. In such cases, the impact on the interaction activity by ordinary descriptors could be amplified or counterweighted by the impact from cross-terms. Such an effect of non-linearity imposes difficulties in summing the real impacts of the descriptors.

A common practical situation is that one also wishes to find out how important a particular descriptor is for a particular ligand or biopolymer. For example, let us assume that the absolute value of the PLS coefficient of some descriptor is large and therefore the descriptor is deemed to be important. However, for some compound the value of this descriptor might be close to zero and accordingly the impact of this descriptor for this particular compounds' activity would be small. Such issues have prompted the development of new interpretation methods.

One approach to reveal the contribution of descriptors to the interaction activity of a particular observation is to use the regression equation for predicting the activity of ligands and biopolymers modified *in silico*. Here the values for each descriptor for this particular observation are replaced, one at a time, with the mean value of the descriptor in the whole dataset, and the activity for the modification predicted using the regression equation. The magnitude and direction of the difference between the activity value of the unmodified and modified observations indicates the contribution of the descriptors for this particular observation. (In more exact terms it indicates the fraction of the total interaction activity being explained by every particular descriptor for this particular observation.) If we want to find the contribution of a group of descriptors (e.g., amino acid, sequence stretch, part of the ligand, etc.) to the activity of a particular observation, we would accordingly simply need to compute the sum of these values for all the descriptors of interest.

Important information can be gained by comparing the contribution of a descriptor (or a group of descriptors) to the activity for all observations containing some particular ligand. This would allow us to separate the contribution to the average affinity of the ligand from the contribution to its selectivity for particular proteins (see III). In a similar fashion, we can estimate the contribution of the properties for the interaction activity and selectivity of some protein.

A practical approach to find the selectivity-determining features of the ligands is to derive a separate regression equation for each one of proteins (see VII for details). As exemplified in Figures 7 and 8, comparisons of equations for pairs of proteins may give clear indications how to modify ligands to alter their selectivity.

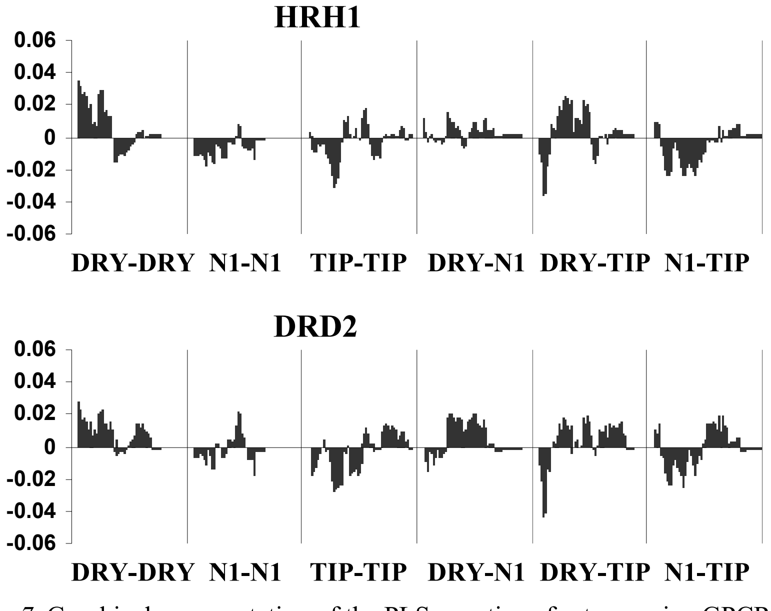


Figure 7. Graphical representation of the PLS equations for two amine GPCRs (HRH1 and DRD2) derived from the PCM model (study VII). Shown are regression coefficients for GRIND descriptors representing auto-correlograms and cross-correlograms of DRY, N1, and TIP molecular interaction fields.

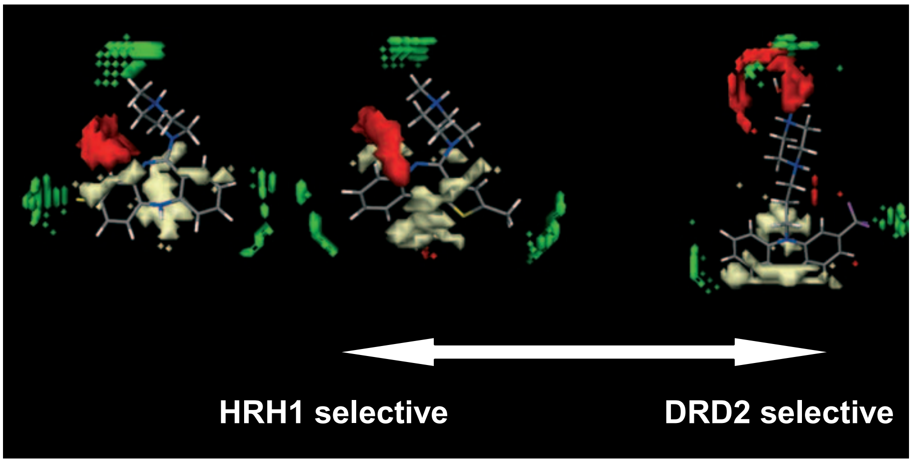


Figure 8. DRY (beige), N1 (red), and TIP (green) molecular interaction fields of two HRH1 and one DRD2 selective compounds.

5 Results and discussion

5.1 Development of proteochemometric technology (study I)

The initial PCM study was directed at analysis of the interaction of a few melanocortin peptides with chimeric MC1/MC3 receptors (Prusis et al., 2001). Although an interpretable and predictive model was obtained, the limited size of the data set comprising ten receptors and four peptides precluded statistical assessment of the predictive ability of the model for new ligands. Therefore, a subsequent study (paper I) was conducted to analyze the interactions of a larger series of twelve structurally related organic compounds (4-pyperidil oxazoles) with series of proteins showing wider structural variation. The protein set comprised ADA1A, ADA1B, and ADA1D adrenergic receptors, eight chimeric receptors constructed by exchanging transmembrane segments of the wild-type receptors, and seven receptors that were further modified by single amino acid changes between wild type receptors. All compounds had antagonistic properties and showed a higher affinity for the ADA1A than for the ADA1B and ADA1D receptors, but affinity ratios for the particular compounds in the series ranged between 3 and 1,000. Thus, some compounds were only slightly selective, whereas others were extremely selective. The sequence positions of the protein sequences varying in the data set were encoded by z-scale descriptors. The quite limited variation of the 4-piperidyl oxazoles, namely changes of substituents at three positions, suggested the use of 24 binary descriptors. The first model was created for interaction of 4-piperydil oxazoles with the eleven wild type and chimeric receptors and was performed stepwise. As expected, neither organic compound nor receptor descriptors alone could provide any valid models. However, in combination, these two descriptor blocks resulted in a reasonable model with a goodness of fit $R^2Y=0.86$ and the predictive ability $Q^2_{cum}=0.79$. Finally, the importance of ligand-receptor and ligand-ligand cross-terms was revealed. By including cross-terms, an excellent model with $R^2Y=0.96$ and $Q^2_{cum}=0.91$ was obtained, showing that cross-terms explain a substantial part of the interaction affinity.

Interpretation of the model was also conducted in two steps. First, we looked at why the compounds generally had a higher affinity for ADA1A than for ADA1B and ADA1D receptors. This analysis was based on scaled sums of absolute values of regression coefficients of protein descriptors (in this study termed “significance of primary variables” or ΣSOP) and showed the importance of transmembrane regions TM2 and TM5. In the second step, we looked for the reasons for differences in the selectivity of the compounds. This analysis was based on the scaled sums of absolute values of regression coefficients of cross-terms (in this study termed “significance of a variable cumulated from cross-terms” or ΣSOC) and revealed that differences in selectivity arose solely due to interactions with TM2. The two measures allowed the separation of the roles of receptor regions in their interactions with the ligands. Thus, ΣSOPs represent the importance of amino acid descriptors for explaining the variation in the affinity of an average ligand for the biopolymers in a data set, while ΣSOCs describe the joint importance of amino acid and ligand descriptors for the creation of ligand selectivity.

The success of modeling allowed us to include the seven point-mutated receptors in the data set. In this case, not only ligand-receptor cross-terms but also receptor-receptor cross-terms appeared to be important, leading to an increase in predictive ability from Q^2_{cum}=0.82 to 0.87. The interpretation of the model obtained gave an even more detailed picture of the ligand-receptor interaction mechanism. Thus, the analysis of the receptor-receptor cross-terms revealed that important interactions occur mainly between amino acids at positions 85 (TM2, residue numbering as in ADA1A), 185 (TM 5), and the non-mutated parts of TM5. The results indicated, in fact, the presence of distant interactions in the receptor. For example, such interactions might affect the receptor conformation. Thus, it is not merely the presence of Val at position 185 of the ADA1A-adrenoceptor that determines the overall high affinities for 4-piperidyl oxazoles, but rather the cooperation of Val185 with other parts of the fifth transmembrane region. Taking the analysis of ΣSOCs of ligand-protein cross-terms further showed that the discrimination of selective from non-selective ligands arose mainly due to interactions with a single sequence residue 86, which was Phe in ADA1A, Leu in ADA1B, and Met in ADA1D receptor. We also found that the ligand-ligand cross-terms were beneficial in all models. These cross-terms may be assumed to reflect changes in molecular shape caused by the various substituents of the 4-piperidyl oxazoles.

5.2 Alignment-independent classification of GPCRs according to the type of interacting ligands (study II)

The structure and function of proteins are determined by the chemical properties of their primary amino acid sequences. Previous computational approaches for functional and structural annotations of proteins have relied mainly on alignment and similarity-based comparisons of protein sequences to proteins with known function. However, relying on the similarities of letter codes may not be an optimal approach; rather a method that explores the similarities and dissimilarities in the physicochemical space of the proteins would seem to be more rational. For this purpose, the protein sequences can be encoded by physicochemical descriptors (e.g., z-scales). Showing similarity in physicochemical properties is, in fact, equivalent to sharing similar structural organization and biological properties. However, direct translation of series of proteins to physicochemical description is only possible if the protein sequences are alignable. Unfortunately, obtaining unambiguous alignments is not always possible. For example, the total length of Class A (rhodopsin-like) GPCR sequences varies from 290 to over 800 amino acids, and only the TM regions are fully alignable. Paper (II) was therefore devoted to the development of alignment-independent descriptions of proteins. The approach was evaluated by classification of 929 cloned rhodopsin-like G-protein coupled receptors (GPCRs) according to the type of ligands they interact with. PCA and PLS-discriminant analysis models were developed for GPCRs of 12 ligand types using alignment-based and alignment-independent approaches. Full sequence alignments were precluded, and the initial alignment-based analysis exploited only the seven transmembrane regions of the GPCRs. The extra and intracellular parts were in this case ignored. Nevertheless, the first components of the PCA model on the data separated the largest GPCR families, i.e., amine, olfactory, glycoprotein hormone, and opsin receptors. When the data were then subjected to PLS-discriminant analysis a highly valid model resulted, which separated all of the receptor families into their ligand binding class.

However, it is known that several GPCR families bind their ligands (e.g., large peptides and the glycoprotein hormones) at the extracellular loops. Hence, utilizing data on the whole sequences would give more adequate representation of the ligand-receptor interaction space. Accordingly, a preprocessing method that can extract the physicochemical properties of full protein amino acid sequences in an alignment-independent way was developed. The approach was based on the auto- and cross-covariance (ACC) functions. To account for the large differences in sequence lengths, modification of the original algorithms was required, namely a normalization of the ACC functions, as well as pre-processing by centering of z-scales according to the distribution of amino acids in membrane proteins. Using this approach of pre-processing and normalization, and after optimizing the maximum lag of the ACC functions, PCA and PLS discriminant models could be obtained with a similar or better resolution and predictability than had been found for the first alignment-based approach. The model was further validated by external prediction of 535 novel GPCRs, with 97% of the predictions being correct. Moreover, 90 orphan GPCRs were tentatively identified to the GPCR ligand binding class. The alignment-independent method could also be used to assess the importance of the principal chemical properties of every single amino acid in the protein sequences for their contributions in explaining the GPCR ligand binding class. A follow-up (alignment-based) study later also provided similar classifications of GPCRs (Gunnarson et al., 2003).

5.3 Modeling of the interaction of amine GPCRs with a diverse set of ligands (study III)

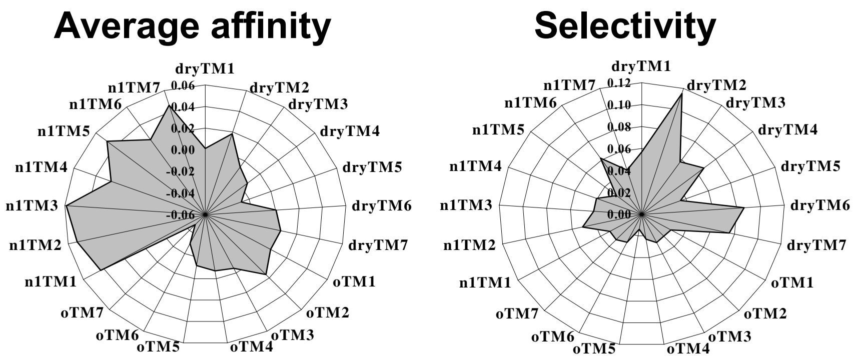
The need to use artificially created proteins seemed to impose restrictions on the application of proteochemometrics on a large scale due to the time needed and cost for creating such altered proteins. The aim of a further study was to apply PCM to wild-type receptors of wider diversity. To this end, we considered the amine subfamily of rhodopsin-like GPCRs. The data set comprised 21 receptors, predominantly serotonin receptor subtypes, but also four dopamine, six alpha- and beta-adrenergic receptors, and one histamine receptor. The ligands tested on these receptors were a relatively diverse series of 23 organic amines, antipsychotics, and alkaloids.

Since it is known that binding of amines and small organic compounds takes place within the TM cavity of GPCRs, only TM segments of receptors were described. Using this strategy also had the advantage that each of the TM parts of amine GPCRs contained several conserved amino acids allowing unambiguous alignments. In total, 159 amino acids were used and described by five z-scales each, resulting in 795 descriptors.

The wide diversity of organic compounds precluded structural superimposition of the molecules. In order to obtain a 3D description of molecular properties relevant to receptor recognition, we used an approach for the calculation of alignment-independent three-dimensional descriptors, so-called GRIND descriptors. Briefly, the interaction energies were calculated with pharmacophoric groups placed on grid points surrounding the molecule. The probe groups used in this study were “DRY,” “O,” and “N1,” representing hydrophobic, H-bond acceptor, and H-bond donor interactions. So-called auto- and cross-correlograms were then used to calculate alignment-independent descriptors from the probe pairs (see section 3.2 for details). In this way six blocks of descriptors, DRY-DRY, O-O, N1-N1, DRY-O, DRY-N1, and O-N1 were obtained giving a total of 230 descriptors.

The large number of protein and ligand descriptors would result in a very large amount of cross-terms. In order to reduce the descriptors to a manageable number, PCA was applied separately to descriptors for each of the seven receptor transmembrane regions and to each of the six blocks of ligand descriptors. Using this strategy, 42 blocks of cross-terms were obtained, each having interpretable, physicochemical meaning and pinpointing the place and type of ligand-protein interactions. It was then shown that using only ordinary descriptors (after PCA) a valid model could not be obtained (R^2Y = 0.41, Q^2cum = 0.31). However, after the addition of ligand-receptor cross-terms and the absolute values of their deviation from the mean values (see III for details) a reasonable model was obtained (R^2 = 0.92 and Q^2cum = 0.75).

The groupings of descriptors and cross-terms sharing a common TM location and property were thereafter used for model interpretations. We created measures for the contribution of TM regions/interaction types to the average affinity and to the selectivity of any one compound. The further use of these measures is illustrated in Figure 9, which shows the results for all 42 combinations of interaction types/transmembrane regions for one of the compounds in the data set, sertindole. As seen from the right panel of the radar plot, the selective discrimination of amine GPCRs of this compound seems to be due to hydrophobic interactions with receptor TM2, 6, and 7.



Average affinity

Selectivity

Figure 9. Example for the interpretation of a proteochemometrics model (study III).

The compounds could subsequently be clustered according to selectivity measures, which indicated the location of compound’s binding pocket and suggested the existence of several distinct (though partially overlapping) interaction sites for different categories of ligands in the amine GPCRs. Finally, the contribution estimates were calculated for every receptor amino acid (and amino acid property) and used to map its influence on the affinity of each amine ligand.

5.4 QSAR and PCM modeling of the interaction of organic compounds with MC receptors (IV)

The aim of the next study was to determine whether the PCM approach would be superior to conventional QSAR modeling in situations where the number of proteins is low. The data set used included interactions of a series of 54 organic compounds with four melanocortin receptor subtypes, MC_1, MC_3, MC_4, and MC_5. Three modeling approaches were attempted: separate QSAR models (i.e., one for each receptor), multiresponse QSAR, and proteochemometrics modeling. The compounds were characterized by topological, geometrical descriptors, and GRIND descriptors. The low number of proteins clearly suggested that the use of extensive descriptions of their sequences would not be sufficient to unambiguously map the ligand interaction site of the receptors. Therefore, only four variables were used for the description of MC receptors in the PCM modeling. These were based on the receptor sequence identities. For comparisons with QSAR models, the proteochemometrics model was validated so that all four observations of a compound were always included in the same cross-validation group. Thus, when predicting the affinity of a compound for each receptor, no information was present in the model about binding of this compound to the other receptors. Statistical analysis showed that all three modeling approaches gave a similar predictive ability, the standard deviation of errors of prediction (SDEP) in the PCM approach ranging from 0.44 to 0.56 for particular receptors (the corresponding Q^2_{cum} being 0.71), while the SDEP of QSAR models ranged from 0.41 to 0.57. It should be noted that such predictive ability is close to the accuracy of biological measurements. An interesting finding of the study was that, although the combined use of different types of ligand descriptors was applied, PLS coefficients indicated a dominating role of GRINDs in the models. All three approaches used revealed determinants of high affinity of particular compounds. However, use of only four proteins did not allow mapping of the ligand binding pocket in the MC receptors.

5.5 Mapping of the interaction site for organic compounds in MC receptors (V)

In study V, we undertook a further analysis to model the recognition site for organic compounds in the MCRs by using both native and chimeric receptors. Twelve chimeric melanocortin receptors were designed based on statistical experimental design; each chimera contained parts from three of the $MC_{1,3\text{-}5}$ receptors (Figure 10, Panel A). We identified four highly conserved sequence stretches, making it possible to divide the receptors into five segments and to create two sets of chimeras. For the first set, the combinations took place at the end of TM3 and at the end of TM5. In the second set, the combination sites were located at the beginning of TM2 and in the middle of TM6. Unfortunately, we failed to obtain full-length constructs for some chimeras, whereas some others showed very low levels of expression, making their use unfeasible. To obtain a working set of receptors, we therefore combined the two sets so that the final set included 4 native and 13 chimeras as depicted schematically in Panel B of Figure 10. As can be seen, each of the receptors can be considered as consisting of five segments, each of which was characterized by four Free-Wilson (binary) descriptors.

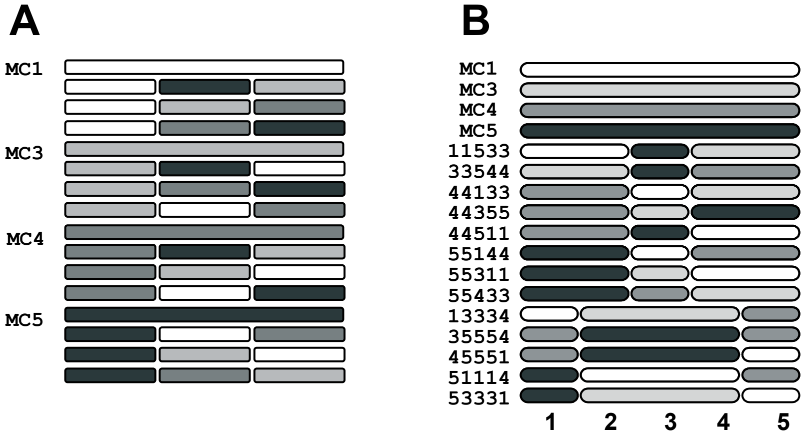


Figure 10. Chimeric receptors. A. Set of three-part chimeras according to SED. B. Obtained "five-part" chimeric receptors (study V).

The interaction affinities of 18 organic compounds were measured for all 17 receptors, data for 14 compounds were used in the modeling, while data for the 4 remaining compounds were used for validation of the model. In this study, the descriptors of organic compounds represented different physicochemical properties (e.g., molecular weight, van der Waals volume, electronegativity, polarizability, molar refractivity, polar surface area, log P, etc.) and the numbers of functional groups and structural fragments in the molecule.

Along with the conventional cross-validation with randomly formed groups (Q^2 being 0.79), the predictive ability was also estimated for new receptors (Q^2rec=0.76) and new organic compounds (Q^2lig=0.68) as well as for the four test set compounds (Q^2ext=0.73; in the latter case the compounds were completely excluded from the modeling and could not influence the centering of descriptors, calculations of cross-terms, optimal scaling, or complexity of the PLS model). It is notable that one of the test set compounds showed the highest affinity in the whole data set (pK_i=7.4) and was correctly predicted by the model.

In the second step of the modeling, we correlated the z-scale descriptors of amino acids located inside the TM cavity of each of the five MC receptor segments to the PLS regression coefficients for the four binary descriptors of the corresponding segment. We also created a 3D model of MC receptors and mapped the molecular interaction site. Interestingly, for the MC_4 receptor selective compounds the interactions occur with amino acids in TM1 and TM3, as well as deeply in the TM cavity located Met203, whereas these amino acids appeared insignificant for MC_1 selective compounds (of crucial importance being the interactions with TM4-TM6).

5.6 Localization of the interaction site for Trp9 modified alpha-MSH peptides in melanocortin receptors (VI)

The study aimed to find the receptor residues that determine the activity and selectivity of alpha-MSH analogues. In the study, we constructed nine alpha-MSH peptide analogues by exchanging the Trp9 residue in the core of the alpha-MSH with another natural or artificial amino acid. We measured the affinities of the thus obtained ten peptides for the four native $MSH_{1,3\text{-}5}R$ and for 15 chimeric MSH receptors (essentially the same series of receptor chimeras as in V; however, we succeeded in obtaining two additional chimeras thus improving the quality of the dataset). All peptides in a series were MC_1R selective. Those peptides containing an aromatic amino acid in the modified sequence position (namely Trp, Nal, d-Trp, and d-Nal) showed, however, up to a hundredfold greater selectivity for MC_1R versus other native MC receptors, while the other peptides generally had fairly low affinities for the receptor series and were less selective for MC_1R.

Several roles can be distinguished for each residue of each peptide and each receptor in the ligand-receptor recognition process. Thus, a particular amino acid in a certain sequence position can stabilize conformation(s) of the peptide or the receptor. A residue can be important at the early stage of ligand-receptor interactions by facilitating the ligand’s passage into its binding cavity. A residue can also stabilize the ligand-receptor complex in its bound stage.

Since such molecular interactions comprise complex dynamic processes, it is problematic to model them using traditional QSAR approaches, and they are generally overlooked in 3D modeling studies. Our study therefore attempted to test whether all such roles of ligand and protein residues can be differentiated by a single multivariate model. A valid, predictive (Q^2=0.74), and easily interpretable proteochemometrics model was obtained. The interpretation of the model showed that the MC_1R/MC_3R selectivity of the peptides arose mainly from the interactions with non-conserved amino acids in receptors' transmembrane regions TM1, TM2, and TM3. Moreover, interactions with TM3 gained additional importance in creating MC_1R/MC_4R and MC_1R/MC_5R selectivity. The interpretation also suggested that the $MC_3R/MC_{4\text{-}5}R$ selectivity of peptides was determined by interactions with those sequence residues in TM2-3 that are identical or have similar properties in MC_1 and MC_3Rs but are different in the MC_4 and MC_5Rs.

We also wanted to identify the 3D location of sequence residues found to be important for creating the high affinity and selectivity of α-MSH for the MC_1R. For this purpose, we used a homology model of the MC_1R based on the crystal structure of bovine rhodopsin. An interesting finding was that, although the important amino acids resided in distinct parts of the receptor primary sequence, they formed clusters in the 3D structure of the receptor transmembrane cavity. Thus, we found that several of the important residues were located close to the extracellular side of the receptors' TM regions. Interactions with these amino acids likely facilitate the entry of the peptides into the TM cavity. Other important residues were found to be located deeply in the receptors' transmembrane cavity. These amino acids are either directly involved in interactions with the peptides in their bound state, or they may have an indirect effect by influencing the overall conformation of the receptor. The importance of representatives of both groups of residues was confirmed using the data from a parallel study involving site-directed mutagenesis (Prusis et al., 2006). This study provided data on the changes in MC_1R/MC_4R selectivity arising from single and cassette mutations in MC_4R to the corresponding amino acid(s) in MC_1R. We used our PCM model to predict these changes in selectivity, and the predictions correlated well with the measured values, SDEP being as low as 0.19 pK_i units. Our study thus proved the hypothesis that the whole complexity and dynamic nature of ligand-protein interactions can be revealed, and that the different roles of ligand and protein residues can be distinguished by a single PCM model.

5.7 Improved large-scale PCM model for amine GPCR-organic compound interactions (study VII)

The data set of study III suffered from the fact that only three of the five biogenic amine GPCR families were well represented, and it included only one histamine receptor and no acetylcholine muscarinic receptors. Moreover, a large proportion of affinity data was absent. These missing data most likely imply no or negligible affinity, which researchers unfortunately are reluctant to report. However, for the sake of PCM such data are equally important as data for high affinity binding. Yet another obstacle in study III was the broad diversity of scaffolds of ligand structures, several compounds appearing to be outliers and leading to deterioration of the model quality. Applying PCM still yielded a statistically valid model; however, the modeling required a very complex description of the data, which made it very difficult to comprehend the physical meaning of the model.

Therefore, in a new modeling attempt we collected data for the interaction of 32 tricyclic and/or piperidine/piperazine ring containing amines with representatives of all amine GPCR families, namely 7 adrenoceptors, 10 serotonin, 5 dopamine, 4 histamine, and 5 muscarinic acetylcholine receptors. The large number of organic compounds allowed us to set aside ten of them as a test set, thus ensuring that proper validation of the obtained models could be carried out. Several improvements were achieved over the preceding study. Firstly, the 3D structures of compounds were obtained using two approaches: 1) by simulated annealing and 2) by use of the rule-based Corina software. In the latter case, the conversion process was very fast, taking only a few seconds per molecule, thus allowing it to be used to screen large compound databases. Moreover, Corina seemed to provide some advantage. In our previous study (III) the compounds predominantly containing several hydrophobic moieties tended to bunch together in the 3D structures for flexible molecules. In the present study, we found that Corina-generated extended structures produced superior models compared with simulated annealing.

The next attempt in improving the modeling was the optimization of settings in calculating the GRIND descriptors. (As shown in a recent study that evaluated the settings of a similar 3D description of molecules (i.e., CoMFA descriptors) for nine different data sets (Peterson et al., 2006), the default settings did not produce the best models obtained from a total of 6120 combinations of settings.) A major finding of our study was the high usefulness of the newly developed molecular shape descriptors.

Finally, we applied PCA separately to all ligand descriptors and protein descriptors, extracting all components, and showed that this pre-processing method produces models identical to those obtained from calculating all cross-terms of original descriptors (though the number of cross-terms in the latter case would exceed 300,000, essentially precluding PCM using available academic and commercial software). Accordingly, the main finding of the study was the possibility to apply PCM to large-scale data sets with essentially unlimited size. The model was validated for its predictive ability for new protein-ligand combinations (Q^2=0.76), new proteins (Q^2=0.62), and new ligands (Q^2=0.59). The model was further subjected to repeatedly performed cross-validation with two random groups (i.e., validation with half of all observations excluded). This very harsh validation mode certified model sturdiness, the Q^2 of 100 repeats being 0.68 with a standard deviation of 0.02. Thus, a whole half of all data points could be omitted without endangering model validity. Finally, the high quality of the model was confirmed by predictions for the test set compounds (Q^2ext=0.67).

In the real settings the selection of a lead compound for further optimization could be based on some pre-selected cutoff of the affinity value. We selected such a predefined limit, higher than 10 nanomolar (pK_i>8). In 20 cases such affinity was predicted for test set compound-protein combinations. Of these, 13 combinations had indeed an experimentally determined pK_i >8, while for the remaining seven the measured pK_i was 6.95 or higher. Such results indicate the potential use of the model in screening of compound databases for high affinity binders to particular amine GPCRs.

A clear advantage over the previous approach (study III) was the full interpretability of the PCM model. We created a PLS regression equation for each of the proteins, based on the PCA scores of GRINDs and PLS regression coefficients of these PCs and their cross-terms. In this way, the ligands' selectivity between each pair of proteins could be revealed. Moreover, we linked the regression coefficients of GRIND descriptors to molecular interaction fields surrounding the ligands, exemplified in the paper for HR1H/DRD2 selective compounds.

6 Future outlook

Proteochemometrics is simple in its idea and intuitively perceivable approach for the analysis of interactions. We have shown here in a series of studies how proteochemometric models can be created and validated, and how PCM can be adapted for the analysis of large-scale datasets yielding models that are straightforward to interpret in a chemical sense.

One may be surprised that the benefits of the PCM approach have been realized as late as the turn of the twenty-first century. Just a few decades ago proteochemometrics would hardly had been conceivable due to the lack of data for the interaction of ligands with multiple macromolecules. A whole research career of a biologist could then be filled with the cloning and characterization of a few single proteins. Nowadays, the momentous revolution of genomics and proteomics has provided computational biologists with enormous amounts of sequential and structural data, and the new Holy Grail is to map the entire molecular interaction networks that govern the functions of cells, organs, and whole organisms.

Obtaining interaction data has also become relatively fast and cheap thanks to the modern techniques of molecular biology. Thus, a few recent studies exploiting protein microarrays have succeeded in simultaneously measuring the interactions of ligands with hundreds of macromolecules, such as kinases (Fabian et al., 2006) and certain structural domains of proteins (Jones et al., 2006). A typical number of entries in interaction databases compiled from literature and patent data already exceeds several hundred thousand biological activity values (www.aureus-pharma.com, www.gvkbio.com, www.accelrys.com). Although several problems persist, such as the neglect of reporting negative interaction data and insufficient description of assay conditions, we may expect that the amount of systematically collected interaction data will increase exponentially in the near future. This opens up new prospects, and calls for approaches for mega-scale data analysis. Proteochemometric modeling might then appear among the most efficient ways of mapping the molecular recognition mechanisms – either by general large-scale models or by more specific high-resolution models. It must be emphasized that proteochemometric models can be created in an iterative and agglomerative manner, and are not limited to any particular type of mathematical descriptions or correlation algorithms. Ultimately, proteochemometrics may develop into an indispensable tool in biology for extracting relevant information from the vast quantities of complex biological data.

7 Acknowledgements

The work presented in this thesis was carried out at the Division of Pharmaceutical Pharmacology, Department of Pharmaceutical Biosciences at Uppsala University, Sweden. I wish to express my deep and sincere gratitude to all who made this thesis possible. In particular, I would like to thank:

Professor Jarl Wikberg, my supervisor, for welcoming me in your laboratory and continuous support and encouragement throughout my stay in Sweden. For sharing your vast knowledge, experience, and never-failing enthusiasm.

My unofficial supervisor Peteris Prusis for teaching me computer chemistry, for fruitful scientific discussions and for assistance in all kind of personal situations.

Professors Torbjörn Lundstedt and Claes Post for introducing us to the exciting fields of multivariate analysis and experimental design and for giving insights into the uneasy process of drug design.

My co-author biologist Ramona Petrovska for taking care of cells and assays, and pleasant working conditions in the whole lab.

Professor Staffan Uhlén for serving an example of enthusiasm and devotion.

Our hard-working chemists Felikss Mutulis, Ilze Mutule, Sviatlana Yahorava, and Oleg Yahorav.

My all-time best co-author Santa Veiksina. Thank you!

I would also like to thank my other Latvian friends and colleagues who work or have worked in Uppsala: Ilona, Kaspars, Ruta, Ilga, Sandra, Andris, Baiba, Jana, Janis, and Maija.

I would like to thank Erica Johansson, Annika Häger, Kjäll Åkerlund, and Magnus Jansson for help with administrational, technical, and computer matters.

All former and present Ph.D students of our department, which I have shared a fantastic time of being Ph.D student: Amanda, Jonas, Chao, Katarina, Tina, Carolina, Erika, Sadia, Lisa, Madeleine, Mathias, Kristina, Martin, Ola, Aleksejs, and all others.

My girlfriend Ginta Elsberga. Thank you for your understanding, loving and accompanying me through all these years. For your patience and insistence that has allowed you to acquire a new language and start a new career.

8 References

Afzelius L, Zamora I, Masimirembwa CM, Karlén A, Andersson TB, Mecucci S, Baroni M, Cruciani G (2004) Conformer- and alignment-independent model for predicting structurally diverse competitive CYP2C9 inhibitors. J Med Chem. 47, 907-914.

Almeida JS (2002) Predictive non-linear modeling of complex data by artificial neural networks. Curr Opin Biotechnol. 13, 72-6.

Baffi G, Martin EB and Morris AJ (1999) Non-linear projection to latent structures revisited (the quadratic PLS algorithm). Comput Chem Eng. 23, 395-411.

Baldwin JM, Schertler GF, Unger VM (1997) An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J Mol Biol. 272,144-64.

Baroni M, Costadino G, Cruciani G, Riganelli D, Valigi R, Clementi S (1993) Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. Quant Struct-Act Relat. 12, 9-20.

Bissantz C, Bernard P, Hibert M, Rognan D (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets? Proteins. 50, 5-25.

Bissantz C, Schalon C, Guba W, Stahl M (2005) Focused library design in GPCR projects on the example of 5-HT(2c) agonists: Comparison of structure-based virtual screening with ligand-based search methods. Proteins. 61, 938-52.

Bock JR, Gough DA (2002) A new method to estimate ligand-receptor energetics. Mol Cell Proteomics. 1, 904-10.

Bock JR, Gough DA (2003) Whole-proteome interaction mining.Bioinformatics. 19, 125-34.

Bock JR, Gough DA (2005) Virtual screen for ligands of orphan G protein-coupled receptors. J Chem Inf Model. 45, 1402-14.

Bock JR, Gough DA (2005) Virtual screen for ligands of orphan G protein-coupled receptors. J Chem Inf Model. 45, 1402-14.

Bohac M, Loeprecht B, Damborsky J, Schuurmann G (2002) Impact of orthogonal signal correction (OSC) on the predictive ability of CoMFA models for the ciliate toxicity of nitrobenzenes. Quant Struct-Act Relat. 21, 3-11.

Cocchi M, Johansson E (1993) Amino acids characterization by GRID and multivariate data analysis. Quant Struct-ActRelat. 12, 1-8.

Cramer RD III, Patterson DE, and Bunce DJ (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc. 110, 5959-67.

Cruciani G, Baroni M, Carosati E, Clementi M, Valigi R, Clementi S (2004). Peptide studies by means of principal properties of amino acids derived from MIF descriptors. Journal of Chemometrics. 18, 146-155.

Crum-Brown A and Fraser TR (1868-9) On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebia, Codeia, Morphia, and Nicotia. Trans Roy Soc Edinburgh, 25, 151-203.

Doweyko AM (2004) 3D-QSAR illusions. J Comp-Aid Mol Des. 18, 587-596.

Dragic T, Trkola A, Thompson DA, Cormier EG, Kajumo FA, Maxwell E, Lin SW, Ying W, Smith SO, Sakmar TP, Moore JP (2000) A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembrane helices of CCR5. Proc Natl Acad Sci U S A. 97, 5639-44.

Drews J (2000) Drug Discovery: A Historical Perspective. Science. 287, 1960-1964.

Dunn WJ III (1989) Quantitative structure-activity relationships (QSAR). Chemom Intell Lab. 6, 181-190.

Edman M, Jarhede T, Sjöström M, Wieslander Å (1999) Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and Escherichia coli: a multivariate data analysis. Proteins, 35, 195-205.

Eriksson L and Johansson E (1996) Multivariate design and modeling in QSAR. Chemom Intell Lab. 34, 1-19.

Eriksson L, Johansson E, Kettaneh-Wold N, Wikström C, Wold S (2001) Design of Experiments: Principles and Application, Umetrics.

Eriksson L, Johansson E, Lindgren F, Sjöström M, Wold S (2002) Megavariate analysis of hierarchical QSAR data. J Comp- Aid Mol Des. 16, 711-726.

Eriksson L, Johansson E, Wold S (1997) Quantitative structure-activity relationship model validation, in: F. Chen, G. Schuurmann (Eds.), Quantitative Structure-Activity Relationships in Environmental Sciences - VII, SETAC Press: Pensacola, FL, 381-397.

Evers A, Hessler G, Matter H, Klabunde T. (2005b) Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. J Med Chem. 48, 5448-65.

Evers A, Klabunde T (2005) Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. J Med Chem. 48, 1088-1097.

Fabian MA, Biggs WH 3rd, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, Ford JM, Galvin M, Gerlach JL, Grotzfeld RM, Herrgard S, Insko DE, Insko MA, Lai AG, Lelias JM, Mehta SA, Milanov ZV, Velasco AM, Wodicka LM, Patel HK, Zarrinkar PP, Lockhart DJ (2005) A small molecule-kinase interaction map for clinical kinase inhibitors. Nat Biotechnol. 23, 329-36.

Fontaine F., Pastor M. and Sanz F (2004) Incorporating Molecular Shape into the Alignment-free GRid-INdependent Descriptors. J. Med Chem. 47, 2805-2825.

Free SM Jr and Wilson JW (1964) A mathematical contribution to structure-activity studies. J Med Chem. 7, 395-399.

Freyhult E, Prusis P, Lapinsh M, Wikberg JE, Moulton V, Gustafsson MG (2005) Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. BMC Bioinformatics, 6:50.

Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 1986, 185, 1-17.

Gether U (2000) Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocr Rev. 21, 90-113.

Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem. 28, 849-857.

Goodford PJ (1998) Atom movement during drug-receptor interactions. Rational Mol Des Drug Res. 42, 215-230.

Gunnarsson I, Per Andersson P, Wikberg J, Lundstedt T (2003) Multivariate analysis of G protein-coupled receptors. J Chemom. 17, 82-92.

Hammett LP (1940) Physical Organic Chemisrry, New York: McGraw-Hill.

Hansch C and Fujita T (1964) A method for the correlation of biological activity and chemical structure. J Am Chem Soc. 86, 1616-1626.

Hansch C, Maloney P, Fujita T (1962) Correlation of biological activity of phenoxyacetic acids with Hammet substituent constants and partition coefficients. Nature(London), 194, 178-180.

Hansch C. (1969) A Quantitative Approach to Biochemical Structure-Activity Relationships. Acct Chem Res. 2, 232-239.

Hawkins DM, Basak SC, Shi X (2001) QSAR with few compounds and many features. J Chem Inf Comput Sci. 41, 663-670.

Hellberg S, Sjöström M, Skagerberg B, Wold S (1987) Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem, 30, 1126-1135.

Hopkins AL, Groom CR (2002) The druggable genome. Nat Rev Drug Disc. 1, 727-730.

Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G. (2003) GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res. 31, 294-7.

Jacoby E, Fauchere JL, Raimbaud E, Ollivier S, Michel A and Spedding M (1999) A three binding site hypothesis for the interaction of ligands with monoamine G protein-coupled receptors: implications for combinatorial ligand design. Quant Struct-Act Relat. 18, 561-572.

Jamois EA (2003) Reagent-based and product-based computational approaches in library design. Curr Opin Chem Biol. 7, 326-30.

Jamois EA, Lin CT, Waldman M (2003) Design of focused and restrained subsets from extremely large virtual libraries. J Molec Graph Modell. 22, 141-9.

Jones RB, Gordus A, Krall JA, MacBeath G (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature. 439, 168-74.

Kennedy T (1997) Managing the drug discovery/development interface Drug Discovery Today, 2, 436-444. Nat Rev Drug Discov. 3, 353-9.

Kubinyi H, Hamprecht FA, Mietzner T (1998) Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem. 41, 2553-2564.

Linusson A, Wold S, Nordén B (1998-1999) Statistical molecular design of peptoid libraries. Mol Div. 4, 103-114.

Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature. 432, 855-861.

Lundstedt T, Seifert E, Abramo L, Thelin B, Nyström Å, Pettersen J, Bergman R (1998) Experimental design and optimization. Chemometr Intell Lab. 42, 3-40.

Martens H, Martens M (2000) Modified Jack-knife estimation of parameter uncertainty in bilinear modelling (PLSR). Food Qual Prefer. 11, 6-15.

McGregor MJ, Pallai PV (1997) Clustering of large databases of compounds: using MDL "Keys" as structural descriptors. J.Chem.Inf.Comput.Sci. 37, 443-448.

Meyer H (1899) Lipoidtheorie der Narkose. Arch Exp Pathol Pharm. 42, 109-118.

Morgan E (1991) Chemometrics: Experimental Design, John Wiley & Sons, Inc.: New York.

Müller G. (2000) Towards 3D structures of G protein-coupled receptors: a multidisciplinary approach. Curr Med Chem. 7, 861-88.

Nelson PRC, Taylor PA, and MacGregor JF (1996) Missing data methods in PCA and PLS: Score calculations with incomplete observations. Chemom Int Lab 35, 45-65.

Ngo SH, Kemény S, Deák A (2003) Performance of the ridge regression method as applied to complex linear and nonlinear models. Chemom Intell Lab. 67, 69-78

Nickolls SA, Cismowski MI, Wang X, Wolff M, Conlon PJ, Maki RA (2003) Molecular determinants of melanocortin 4 receptor ligand binding and MC4/MC3 receptor selectivity. J Pharmacol Exp Ther. 304, 1217-27.

Overton CE (1901) Studien über die Narkose. Jena: Fischer.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. Science. 289, 739-45.

Pastor M, Cruciani G, McLay I, Pickett S and Clementi S (2000) Grid-Independent Descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem. 43, 323-3243.

Perlman S, Schambye HT, Rivero RA, Greenlee WJ, Hjorth SA, Schwartz TW (1995) Non-peptide angiotensin agonist. Functional and molecular interaction with the AT1 receptor. J Biol Chem. 270, 1493-1496.

Peterson SD, Schaal W, Karlen A. (2006) Improved CoMFA Modeling by Optimization of Settings. J Chem Inf Model. 46, 355-64.

Prusis P, Frandberg PA, Muceniece R, Kalvinsh I, Wikberg JE (1995) A three dimensional model for the interaction of MSH with the melanocortin-1 receptor. Biochem Biophys Res Commun. 210, 205-10.

Prusis P, Muceniece R, Andersson P, Post C, Lundstedt T, Wikberg JES (2001) PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. Biochem Biophys Acta. 1544, 350-357.

Prusis P, Muceniece R, Mutule I, Mutulis F, Wikberg JE (2001) Design of new small cyclic melanocortin receptor-binding peptides using molecular modelling: Role of the His residue in the melanocortin peptide core. Eur J Med Chem. 36,137-46.

Prusis P, Schioth HB, Muceniece R, Herzyk P, Afshar M, Hubbard RE, Wikberg JE (1997) Modeling of the three-dimensional structure of the human melanocortin 1 receptor, using an automated method and docking of a rigid cyclic melanocyte-stimulating hormone core peptide. J Mol Graph Model. 15, 307-17.

Prusis P, Uhlen S, Petrovska R, Lapinsh M, Wikberg JE (2006) Prediction of indirect interactions in proteins. BMC Bioinformatics. 22, 167.

Renfrey S, Featherstone J (2002) Structural proteomics. Nat Rev Drug Discov. 1, 175-6.

Roth BL, Sheffler DJ, Kroeze WK (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat Rev Drug Discov, 3, 353-9.

Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem. 41, 2481-2491.

Schneider G (2002) Trends in virtual combinatorial library design. Curr Med Chem. 9, 2095-101.

Scholkopf B, Bartlett P, Smola A, Williamson R (1998) Support vector regression with automatic accuracy control. In Proceedings of the Eighth International Conference on Artificial Neural Networks; Niklasson, L., Boden, M., Ziemke, T., Eds.

Selnick HG, Barrow JC, Nantermet PG, Connolly TM. (2003) Non-peptidic small-molecule antagonists of the human platelet thrombin receptor PAR-1. Curr Med Chem Cardiovasc Hematol Agents. 1, 47-59.

Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc. 88, 486-494.

Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S (1998) An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemom Intell Lab. 44, 229-244.

Sjöström M, Rännar S, Wieslander Å (1995) Polypeptide sequence property relationships in Escherichia coli based on auto cross covariances. Chemom Intell Lab. 29, 295-305.

Strombergsson H, Prusis P, Midelfart H, Lapinsh M, Wikberg JE, Komorowski J (2006) Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. Proteins. 63, 24-34.

Strombergsson H, Kryshtafovych A, Prusis P, Fidelis K, Wikberg JE, Komorowski J, Hvidsten TR (2006) Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. Proteins, 2006 Aug 31; [Epub ahead of print].

Svensson O, Kourti T, MacGregor JF (2002) An investigation of orthogonal signal correction algorithms and their characteristics. J Chemometr. 16, 176-188.

Todeschini R and Consonni V (2000) Handbook of Molecular Descriptors (in the Series of Methods and Principles in Medicinal Chemistry - Volume 11 (Eds.: R. Mannhold, H. Kubinyi, H. Timmerman)) WILEY - VCH.

Trygg J, S. Wold (2002) Orthogonal projections to latent structures (O-PLS). J Chemometr. 16,119-128.

Van Heel M (1991) A New Family Of Powerfull Multivariate Statistical Sequence Analysis Techniques. J Mol Biol. 251, 877-887.

Van Rhee AM and Jacobson KA (1996) Molecular architecture of G protein-coupled receptors, Drug Develop Res. 37, 1-38.

Vapnik, V (1995) The Nature of Statistical Learning Theory. Springer-Verlag: Heidelberg, Germany.

Vedani A, Dobler M (2002) 5D-QSAR: the key for simulating induced fit? J Med Chem, 45, 2139-2149.

Vedani A, Dobler M, Lill MA (2005) Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. J Med Chem, 48, 3700-3703.

Westad F, Martens H (2000) Variable selection in NIR based on significance testing in partial least squares regression (PLSR). J Near Infrared Spec. 8, 117-124.

Wikberg JES, Mutulis F, Mutule I, Veiksina S, Lapinsh M, Petrovska R, Prusis P (2003) Melanocortin receptors: ligands and proteochemometrics modeling. Ann N Y Acad Sci. 994, 21-6.

Wise A, Gearing K, Rees S (2002) Target validation of G-protein coupled receptors. Drug Disc Today. 7, 235-46.

Wold H (1966) in Research Papers in Statistics, Festschrift for J. Neyman (Ed.: F. David), , NY: Willey, 411-444.

Wold H (1973) in Multivariate analysis III (Ed.: P.R. Krishnajah) NY: Academic press, 383-407.

Wold S (1995) Chemometrics; what do we mean with it, and what do we want from it? Chemom Intell Lab. 30, 109-115.

Wold S (1995) in Chemometric Methods in Molecular Design (Ed.: D. van de Waterbeem) VCH Verlagsgesellschaft: Weinheim, vol 2, 195-218.

Wold S and Sjöström M (1998) Chemometrics, present and future success. Chemom Intell Lab. 44, 3-14.

Wold S, Antti H, Lindgren F, Ohman J (1998) Orthogonal signal correction of near-infrared spectra. Chemom Intell Lab. 44, 175-185.

Wold S, Esbensen K and Geladi P (1987) Principal component analysis. Chemom Intell Lab. 2, 37-52.

Wold S, Jonsson J, Sjöström M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures Ana. Chim Acta. 277, 239-253.

Wold S, Josefson M, Gottfries J, Linusson A (2004) The utility of multivariate design in PLS modeling. J Chemom. 18, 156-165.

Wold S, Kettaneh-Wold N and Skagerberg B (1989) Non-linear PLS modelling. Chemom Intell Lab. 7, 53-65.

Wold S, Martens H, Wold H (1983) in Lecture Notes in Mathematics, (Eds.: A. Ruhe, B. Kågström), Heidelberg: Springer Verlag, 286-293.

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab. 58, 109-130.

Wold S, Trygg J, Berglund A, Anti H (2001) Some recent developments in PLS modeling. Chemom Intell Lab. 58, 131-150.

Xue L, Godden JW, Stahura FL, and Bajorath J (2003) Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys. J Chem Inf Comput Sci. 43, 1218 - 1225.

Xue L, Godden JW, Stahura FL, and Bajorath J (2003B) Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. J Chem Inf Comput Sci. 43, 1151 - 1157.

Yang YK, Fong TM, Dickinson CJ, Mao C, Li JY, Tota MR, Mosley R, Van Der Ploeg LH and Gantz I (2000) Molecular determinants of ligand binding to the human melanocortin-4 receptor. Biochemistry. 39, 14900-11.

Yasri A, Hartsough D (2001) Toward an Optimal Procedure for Variable Selection and QSAR Model Building. J Chem Inf Comp Sci. 41, 1218-1227.

Zupan J, Gasteiger J (1999) Neural Networks in Chemistry and Drug Design. Second Edition. Wiley-VCH:Weinheim.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 40

Editor: The Dean of the Faculty of Pharmacy

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7211