UPTEC X 07 029
Examensarbete 20 p
Maj 2007

Predicting refractive index increments for small molecules from molecular descriptors

Emma Haraldsson

UPPSALA UNIVERSITET

Bioinformatic Programme

Uppsala University School of Engineering

UPTEC X 07 029	**Date of issue 2007-05**

Author

Emma Haraldsson

Title (English)

Predicting refractive index increments for small molecules from molecular descriptors

Title (Swedish)

Abstract

The refractive index increment of a number of low molecular weight compounds was measured and the relationship between the refractive index increment and a number of molecular descriptors was investigated. A prediction model of the refractive index increment was calculated and to be used in signal corrections of screening data from Biacore instruments.

Keywords

Biosensors, SPR, Refractive index increment, chemometrics, molecular descriptors

Supervisors

Markku Hämäläinen
Biacore AB

Scientific reviewer

Anders Karlén
Uppsala Universitet

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **32**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Predicting refractive index increments for small molecules from molecular descriptors

Emma Haraldsson

Sammanfattning

När ett läkemedel binder till en annan molekyl hämmas eller ökas molekylens funktion. Biosensorer används inom läkemedelsforskningen för att mäta om, hur starkt och hur fort ett potentiellt läkemedel binder till en biomolekyl. Dessutom vill man veta hur specifik bindningen är. Dessa egenskaper går att undersöka med hjälp av Biacore instrument. Man har den ena molekyl immobiliserad på en yta och den andra molekylen finns i en flödeskanal. När molekylen i flödeskanalen binder till molekylen på ytan ändras brytningsindexet vid ytan och denna förändring registreras av instrumentet. Brytningsindex är ett mått på hur mycket en ljusstråles bana förändras när den passerar mellan två media. Brytningsindexinkrementet är ett mått på hur mycket brytningsindex ökar med ökad koncentration. Förändringen i brytningsindex är proportionell mot mängden molekyl som har bundit in. Om en stor molekyl binder in kommer även brytningsindexet att öka mer än om samma mängd av en mindre molekyl binder in. Signalen som man får från instrumentet brukar därför justeras med hjälp av molekylvikten för de interagerande molekylerna. En ännu bättre justering borde kunna uppnås med hjälp av brytningsindexinkrementvärdet för molekylerna men i de flesta fall vet man tyvärr inte brytningsindexinkrementvärdet för de molekyler man jobbar med.

I det här examensarbetet har möjligheten att prediktera ett ämnes brytningsindexinkrement från molekylstrukturen undersökt. Målet var att skapa en matematisk modell där brytningsindexinkrementet predikteras från beräknade kemiska egenskaper hos molekylen så kallade molekylära deskriptorer.

Examensarbete 20 p i Bioinformatikprogrammet

Uppsala Universitet Mars 2007

Table of contents

1. Introduction:

Many biosensors are using a phenomenon called Surface Plasmon Resonance (SPR) to monitor interactions between molecules. SPR is an optical method that detects changes in refractive index (RI) at the chip surface caused by analyte molecules interacting with the target molecule immobilized on the surface. The SPR response given by the instrument will be proportional to the amount of analyte bound to the available binding sites of the target molecule on the surface. Since the instrument is measuring the change of RI at the chip surface the response will also be dependent on the refractive index increment (RII) of the interacting molecules. The refractive index increment is a quantity describing how RI of a solution increases with increasing concentration of the compound of interest. The change of RI represented by the SPR response is measured at a surface layer consisting of both bulk solution and the interacting molecules. In applications such as binding kinetics, concentration series of analyte is injected and the number of binding sites and the bulk contribution can be eliminated by mathematical modeling.
In screening applications the SPR response is measured at only one single concentration and the signal can only be used for ranking if the compounds have very similar RIIs. For most proteins RII is approximately constant (0.18-0.19[1]). Small molecules however, have a larger variation in RIIs as shown by Davis et al[1]. Molecular weights of small molecules have previously been used to normalize the screening signal.

In this paper results are presented from measurements of RIIs for 50 low LMWs. The relationship between the molecular structures and variation in RIIs is discussed. An attempt to make a prediction model of RII to be used for signal corrections of screening data is also presented.

2. Theory

2.1 Refractive index (RI)

When a light beam passes from one media to another and the two media have different densities, the course of the light beam will change. The angle with which the light beam travels through the first media is called the angle of incidence (Θ_I), and the corresponding angle in the second media is called the angle of refraction (Θ_R) (Figure 1). The refractive index (RI) is calculated as $n_1 \times \sin \Theta_I = n_2 \sin \Theta_R$, where n_1 and n_2 are the refractive indices of the two media respectively. RI is usually denoted n and it is a unitless measure. It is dependent on both the temperature and the wavelength of the light beam. RI decreases when the temperature increases. RI is usually measured at the Sodium D line, at 583.9 nm. Depending on the molecular structure of the compound the RI will be different, but most organic substances have a RI between 1.3 and 1.7[2]. The RI of water is 1.3329[1].

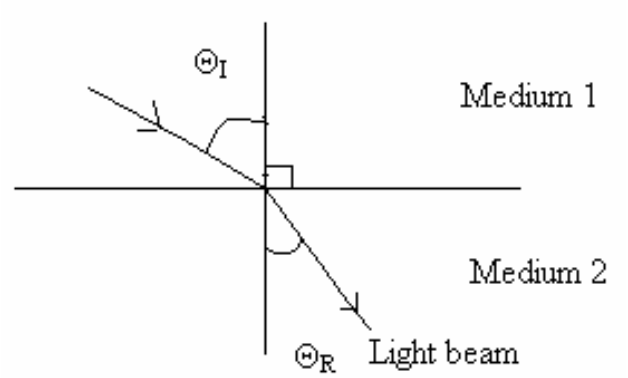


Figure 1 : A light beam passing from one medium to another

The Lorentz-Lorenz equation shows the relation between molar refractivity, density and refractive index[3,4];

$$MR = \frac{n^2 - 1}{n^2 + 1} \times \frac{MW}{\rho}$$ **Equation 1**

where MR is the molar refractivity, MW the molecular weight and ρ is the density of the compound. Molecular weight is measured in g/mol, density in g/cm^3 and n is unitless. MR is an additive measure; it can be calculated from the refractions of all bonds in the molecule[4]. Since n is unitless, the unit of MR is (g/mol)/(g/cm^3)=cm^3/mol, which can be interpreted as the volume taken up by one mole of the substance.

A quantity related to molar refractivity is the molar polarizability[5], which is given by;

$$P = \frac{D - 1}{D + 2} \times \frac{MW}{\rho}$$ **Equation 2**

P is the molar polarizability and D is the dielectric constant of the environment. D is related to n;

$$D = n_{\infty}^2$$ **Equation 3**

The radiation from visible light can only displace electrons, the nuclei is not influenced. If n is measured using visible light we have;

$$P_E = \frac{n^2 - 1}{n^2 + 1} \times \frac{MW}{\rho}$$ **Equation 4**

P_E is the electronic polarization, and it represents the polarizability caused by changes of the molecule's electronic cloud. The right-hand side of the formula ovan is the same as MR, implying that MR is not only related to volume but also to the polarizability of the molecule.

After some rearrangements of the Lorentz-Lorenz equation one obtains:

$$n = \sqrt{\frac{\frac{MW}{\rho} + 2MR}{\frac{MW}{\rho} - MR}}$$ **Equation 5**

Using the formula above it is possible to estimate the refractive index value of a compound.

2.2 Refractive index increment (RII)

The refractive index increment (RII) is a measure of how much the refractive index of a compound increases when the concentration of the compound increases[4]. As seen in Equation 6 the RII can be calculated from RI of the solution and the buffer as long as one knows the concentration of the substances for which the RII is wanted. RII is given by the following formula;

$$dn/dc = \frac{n_{solution} - n_{buffer}}{c_{sample}}$$ **Equation 6**

where dn/dc denotes the refractive index increment of the sample, $n_{Solution}$ and n_{buffer} the refractive index of the solution and the buffer respectively and c_{sample} the concentration of the sample. RII is, as RI dependent on both the temperature and the wavelength[4]. RII is also dependent on the buffer used to dissolve the substance[6]. RII is however independent on the salt concentration in the buffer[7]. The excluded volume of molecule is a measure of the volume of solvent that is displaced by the molecule. Two molecules that have the same RI, but different excluded volumes will have different RIIs.

2.3 Surface Plasmon Resonance (SPR)

In Biacore systems, as well as in many other biosensor systems, RI of a compound is measured using an optical method called Surface Plasmon Resonance (SPR). The method works as follows; a light beam passes from a medium e.g. glass, which has a high RI, to e.g. water, which has a low RI. The light beam hits the glass at a certain angle. Depending on the angle, different amount of the light beam will be reflected. The

part of the light beam that is not reflected is refracted into the glass medium. At a certain angle all light will be reflected, this angle is called the critical angle. An evanescent wave arises and travels a short distance into the medium with the lower refractive index. If a thin layer of a noble metal, e.g. gold, covers the boarder between the two media, metal atoms will absorb energy from the electrical wave and start to oscillate. Electron charge density waves called plasmons are generated. The intensity of the light beam decreases and the loss of energy are registered. If the direction of the wave vector for the plasmons is the same as the direction of the wave vector of the photons the electrons starts to resonate. This phenomenon is called surface plasmon resonance (SPR). The angle at which the energy loss of the incident light is greatest is called the resonance angle (the SPR angle). The strength of the evanescent field increases when passing through the metal surface. The amplitude of the evanescent field decreases with distance. At a distance of ~300 nm the intensity of the wave has decayed to 1/e, which means that about 37% of the intensity is left. The evanescent field interacts with the neighborhood of the metal, which means that optical changes of this region will affect the SPR angle. Changes in the SPR angle reflect changes in the surface concentration[8,9,10].

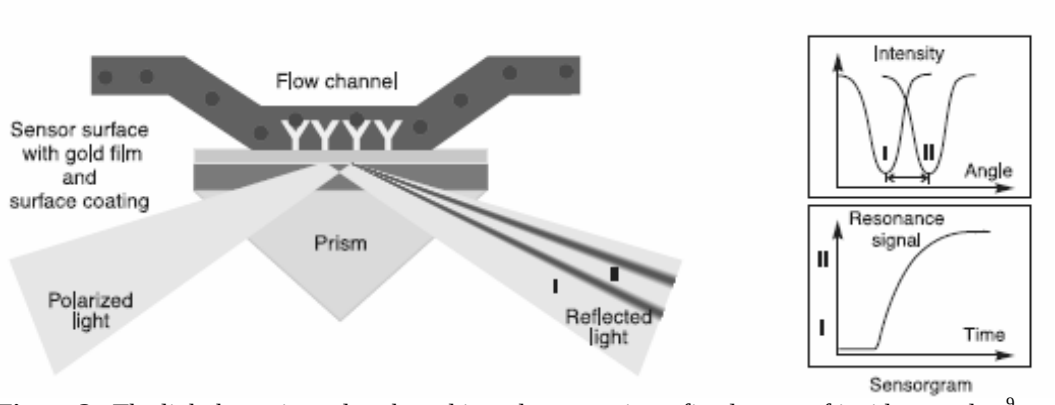


Figure 2 : The light beam is wedge shaped in order to receive a fixed range of incident angles [9] . Illustration used with permission from Biacore AB.

An investigation of interactions between two molecules is carried out as follows; the ligand is first immobilized onto the sensor surface. The second molecule (the analyte) is then injected in a sample buffer through the flow cell. When and if, the analyte binds to the molecule on the sensor surface (the ligand) RI will increase, causing a change in the SPR angle. The change of the SPR angle will be dependent on the amount of analyte that binds to the ligand, and the change of the SPR angle is measured by the Biacore instrument. The change of the SPR angle is quantified in resonance units (RUs), where 1 RU equals a refractive index change of 10^{-6} which also is approximately an angle shift of 10^{-4} degrees. If 1 pg/mm^2 of protein binds to the sensor surface one will receive a response of ~1 RU[10].

The SPR response received from the biosensor is visualized using a sensorgram; an example of a sensorgram is seen in the right-hand bottom corner of Figure 2. The SPR-signal in RU is plotted against time. The SPR response given by the instrument will be proportional to the amount of analyte bound to the available binding sites of the target molecule on the surface. If the binding affinities between two different molecules is compared and one of the molecules are much larger than the other, the SPR-signal

representing the first molecule will be higher even if the same number of molecules has bound to the ligand molecule on the surface.

$$R_{\max}^{obs} = n * X \qquad \textbf{Equation 7}$$

where $R_{\max}^{obs}$ is the observed instrument response measured in RU, n is RI at the surface and X is a factor used for converting n to $R_{\max}^{obs}$ [1]. The change in n is, as mentioned earlier, detected by the instrument as a shift in the SPR angle, a change of 10^{-6} in RI equals 1 RU. For protein interaction the general RII for proteins can be used as a part of the X factor. Proteins have a RII of around 0.18-0.19, independent on their amino acid composition[1]. Davis et al showed in an earlier study that RIIs for small molecules vary more, a variation corresponding to a factor two.

$$R_{\max}^{obs} = n * \left[\left(dn / dc_{ligand} \right) / \left(dn / dc_{analyte} \right) \right] \qquad \textbf{Equation 8}$$

For protein-protein interactions the ratio will be approximately one, and can therefore be omitted. For small molecule interactions the RII for the interacting molecules need to be known. Since the RII is usually not known, and RIIs are time consuming to measure, the ratio between the molecular weights of the interacting molecules is used instead as an approximation.

The $R_{\max}^{obs}$ value can be compared to the theoretically calculated max response, i.e. all ligand sites are saturated[10];

$$R_{\max} = \frac{MW_{analyte}}{MW_{ligand}} \times ligandresponse \times valence \qquad \textbf{Equation 9}$$

where MW_{ligand} and $MW_{analyte}$ is the molecular weight for the ligand and analyte respectively. Ligand response is the experimental amount of ligand molecules immobilized on the chip surface, and valence is the possible number of analyte molecules that can bind to one ligand molecule. By comparing $R_{\max}^{obs}$ with $R_{\max}$ it is possible to decide the binding affinity of the molecules.

If the same ligand is used and its immobilized level is the same, $R_{\max}^{obs}$ expressed in moles should be the same independent on the molecular weight of the analyte, assuming a linear correlation between RII and MW;

$$R_{norm} = \frac{R_{\max}}{MW} \qquad \textbf{Equation 10}$$

2.4 Quantitative Structure-Activity Relationship (QSAR)

It is important to understand and model the relationship between biological responses and the molecular structure e.g. in drug discovery applications. The molecular structure is described by physical-chemical properties, so-called molecular descriptors. In Quantitative Structure-Activity Relationship (QSAR) analysis the aim is to find a mathematical formula where the biological activity is expressed in terms of molecular

descriptors. Instead of having to measure the biological activity, a QSAR-model makes it possible to predict the biological activity[11].

A method often used to model a variable y as a function of a variable x is linear regression (LR). The relation between x and y is the assumed to be linear;

$$y = a + bx + \varepsilon \qquad \textbf{Equation 11}$$

where ε represents the experimental errors (residuals, noise, model errors etc.), and a and b are constants. If one wishes to model y as a function of more than one variable, LR is extended;

$$y = ax_1 + bx_2 + c + \varepsilon \qquad \textbf{Equation 12}$$

where both x_1 and x_2 are variables, and a, b and c are constants. This extension is called multiple linear regression (MLR)[12]. In order to receive an equation system that is solvable one needs to know the value of y for as many compounds as there are x-variables. In most cases one only knows the y-value for a limited number of compounds, but one has usually a large number of x-variables. In such cases Principal component analysis and Projection to latent structures by means of partial least squares can be used instead.

2.5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a projection method used for pattern recognition. PCA is often used to reduce the dimensionality of a multivariate data matrix and for data classification. PCA reduces the data into a number of principal components (PCs). The first PC describes most of the variation in the data, the second PC second most variation and so on. This means that in most cases the largest part of the variation in the data will be included in the two or three first PCs. How PCs are calculated are described and visualized in Figure 3[11].

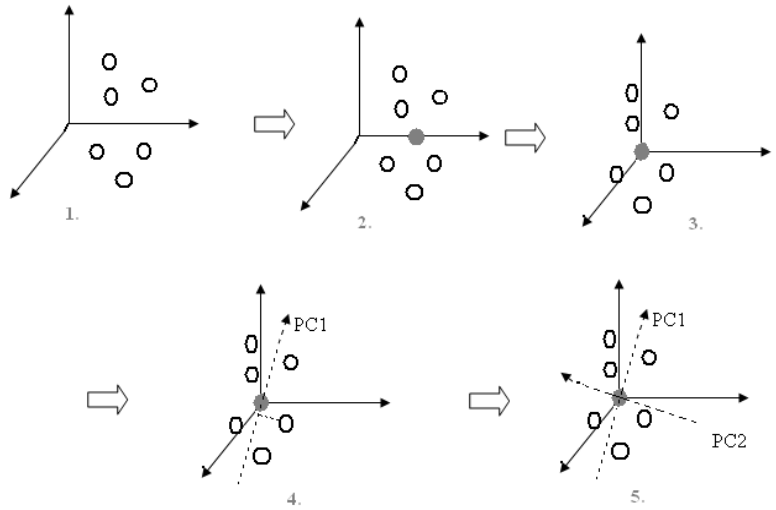


Figure 3. Graphical presentation of principal component analysis.
1. All observations are plotted in the k-dimensional space (in this example k equals three).
2. In the next step the average point of all observations is calculated.
3. The data is mean-centered by the subtraction of the mean value; the origin of the coordinate system is moved so that it overlaps with the average point. The data is also unit variance (UV) scaled; all variables are divided by their standard deviation, i.e. all variables get identical variances.
4. The first PC is calculated. The first PC is the direction vector that goes through the average point, and that best approximates the data in a least square sense. PCs are the eigenvectors of the data matrix. All observations are projected onto this vector. A new coordinate system is obtained, so far only consisting of one dimension. Each observation receives a new coordinate in the new coordinate system; these values are known as scores. From the direction vector one receives the so-called loadings, the loading values for each principal component are the cosine of the angle between the principal component direction vector and each of the axis representing a variable. Loadings are the relative contribution of each variable.
5. A second principal component is calculated, orthogonal to the first one.

The number of PCs needed in order to explain the variation in the data is determent with cross-validation. In cross-validation a certain number of elements are left out when creating the model, the model is then validated with those substances that were left out. This is done iteratively until all substances have been left out once. If the last calculated PC significantly improves the model, the PC is significant and is kept in the model. One more PC is calculated. If the last calculated PC is not improving the model significantly the PC is not included in the model and no more PCs are calculated[13].

The data used in PCA is usually mean-centered and UV scaled (Unit variance scaled), as described in Figure 3. UV scaling means that all values are divided by the standard deviation for the variable they represent. This ensures that all variables will affect the model to the same extent. If the data is not scaled, and it consists of variables with different numerical ranges, variables with large ranges will affect the model to a greater extent than the variables with small ranges. PCA finds the variance in the data and variables with a larger range also have more variance.

2.6 Projection to latent structures by means of partial least squares (PLS)

Projection to latent structures by means of partial least square (PLS), which is a multiple regression extension of PCA, is used to connect two blocks with each other, for example to connect a response to a number of different variables. How the PCs (PLS components) in PLS are calculated is described and visualized in Figure 4[11].

.

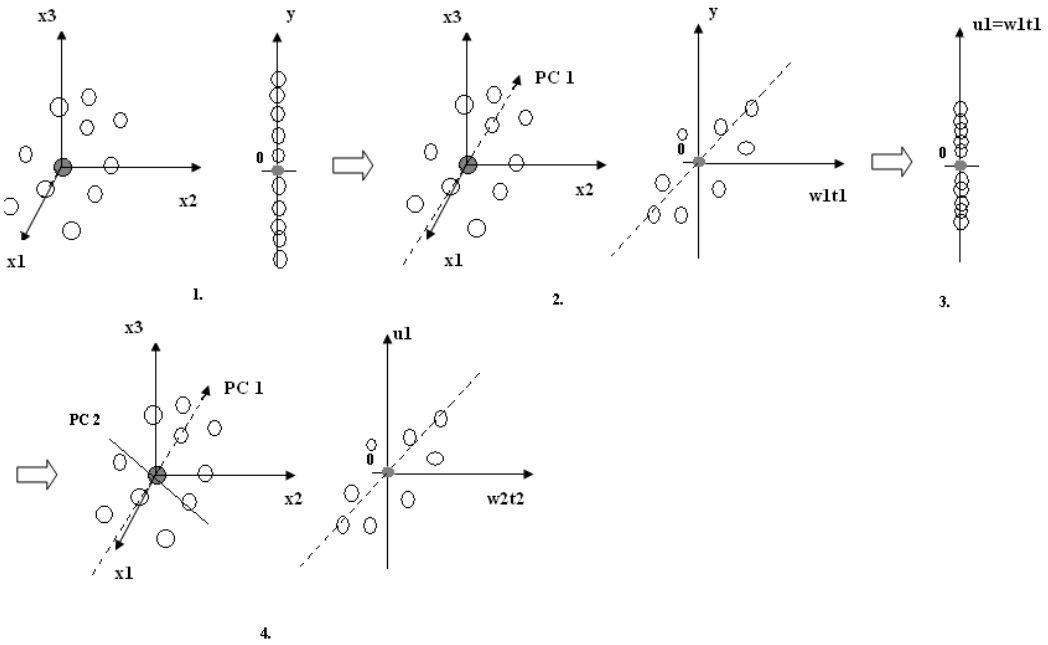


Figure 4.
1. All observations are plotted in the variable space X (in this example three dimensions are used), and also in the response space Y (one dimension is used in this example). The average point in both coordinate systems is calculated, in the plots marked with grey circles.
2. The first PC is calculated. t1 corresponds to the score vector for X, and it corresponds to a new variable containing most of the information in the original x-vector. w1 is the loading vector for y. t1w1 is an estimate of y.
3. t1w1 is subtracted from the original y value.
4. A new PC is calculated.

R^2 and Q^2 are two quantities used for deciding how good the obtained PLS-model is. R^2 is a measure of how much of the variation that is explained by the model. Q^2 is a measure of how good the model predicts the variation. R^2 will increase with increasing complexity of the data, Q^2 however, will first increase with increasing complexity of the model, but at a certain point the predicting ability will not increase anymore but will instead start to decrease[11]. The model has been overfitted, is also tries to explain experimental errors within the data.

2.7 D-Optimal Onion Design (DOOD)

Selection of a subset of compounds that gives a good representation of a large dataset is often needed. It is often practically impossible to measure data from all compounds. The compounds included in the subset need to cover as much of the structural variation found in the whole dataset as possible.

A technique often used to select subsets is Statistical molecular design (SMD). In SMD score vectors calculated by PCA or PLS are combined with statistical experimental design schemes[14]. Such design schemes can sometimes be made by hand. In cases where one has a lot of compounds and many variables, one instead uses algorithms that help creating these schemes. Two methods often used are the Space filling (SF) design and the D-Optimal (DO) design[14,15]. According to the least square criteria the best coefficients in a regression model is given by b=(X'X)X'y. D-Optimal design maximizes

the determinant of the variance-covariance matrix (X'X). This means that the selected compounds will span as much as possible of the property space. A draw back of the DO-design is that it tends to select the most extreme points. The inner regions are often poorly represented. SF-designs main goal is to cover the space as evenly as possible. Space filling works in a similar way as grid based approaches. A grid is placed over the descriptor space or score space. Those points that are found closest to each grid point are chosen. Large SF-designs tend to over represent the inner regions of the candidate set, which means that one will receive unwanted redundancy. Areas that are represented only by a few compounds tend to be poorly represented. A method, developed to overcome the problems with DO-designs and SF-designs is the D-Optimal onion design (DOOD)[15,16]. First a center point is defined as the compound closest to a theoretical center of the experimental domain. The dataset is then divided into subsets i.e layers. Splits are made based on their Euclidean distance to the center point. The number of layers needed depends on the dataset one wishes to investigate, how the experiments are distributed in space, the number of compounds in the dataset and the number of compounds one want in the subset. In the next step, a separate DO design is performed on each layer. DOOD is good to use when the model complexity is not well known. It does not only focus on the most extreme points as DO and it does not mainly focus on the inner regions as SF.

3. Material and Methods

3.1 Softwares

- SMILES codes were created using ChemDraw 8.0 (CambridgeSoft)
- Representative subsets were selected using Modde 8.0 (Umetrics)
- PLS and PCA were performed using SIMCA-P 11.0 (Umetrics)
- Statistical analysis was performed using Statistica 7.0 (StatSoft)
- Plots of raw data were created in Microsoft Office Excel 2003 (Microsoft)

3.2 Molecular descriptors

SMILES codes for each substance used in this project, except the ones included in the Maybridge fragment library, were generated using the ChemDraw software. A SMILES code is a linear notation of a molecular structure. Molecular descriptors for all substances have been generated, using above-mentioned SMILES codes, by Johan Gottfries at Astra Zeneca in Mölndal (Sweden). All descriptors are listed in Appendix B.

3.3 Selection of substances

The substances used in this project can be divided into 4 groups; substances from the literature, halogen substituted molecules, substances belonging to Maybridge fragment library and amino acids.

All substances belonging to the halogen dataset were selected specifically to investigate the influence of halogens on RII.

A subset of compounds was selected from a commercially available fragment library consisting of 500 compounds (Maybridge fragment library). These 500 compounds were known to have a good structural diversity. The aim was to find a subset representing the whole library as good as possible, since it was impossible to measure RIIs for all 500 compounds during this project. A subset containing approximately 30 compounds seemed more realistic. A PCA was performed and it resulted in 18 PCs; a variable reduction had to be made. The molecular descriptors matrix was divided into three groups; molecular descriptors generated using Astra Zenecas in-house program Selma, molecular descriptors generated using the Volsurf procedure and molecular descriptors generated using a various number of different sources[17]. A PCA was performed on each of these three groups. Four PCs from each PCA were selected and all twelve PCs were compiled. A new PCA was performed and four PCs, describing 77% of the variation in the data, were imported into Modde where a DOOD analysis was made. A subset containing 32 compounds was selected.

Amino acids and some other substances already in stock at Biacore were selected mainly to further increase the structural diversity.

3.4 Sample preparation and refractive index increment measurement

Approximately 2.5 mg and 5 mg of each substance was dissolved in 1 ml 10mM PBS buffer containing 5% DMSO, the buffer was prepared following the recipe below. The exact mass of the substance added and the exact mass of the solution obtained was carefully noted.

For all substances, RI was measured at three concentrations, 0 mg/ml, ~2.5mg/ml and ~5mg/ml. RI was measured using an ABBEMAT Digital Automatic Refractometer, measuring at 583.9 nm. The instrument uses an internal solid state Peltier thermostat to control the temperature, which was set to 20 ºC. RI was then measured by placing the sample on the surface of a prism. A light beam was projected towards the bottom side of the sample at different angles, depending on RI of the sample the angle of refraction will change and RI is calculated. The accuracy of the obtained RI is 0.00004[18]. All RI measurements were made the same day as the sample was prepared.

All substances were not soluble in the buffer used. For substances that were significantly but not fully soluble the solutions were further diluted.

RIIs were calculated by plotting the concentrations (g/ml) against RI (RIIconc) and molarity (mol/dm^3) against RI (RIImol). The slope of the line connecting the dots is the RII.

3.4.1 PBS Buffer

$Na_2HPO_4 \times 2\ H_2O$	7.07 g
$NaH_2PO_4 \times H_2O$	1.42 g
KCl	0.20 g
NaCl	5.70 g

The buffer used was prepared by adding 100 ml H_2O to the substances listed above. pH was adjusted to 6.8 giving a 50mM PBS Buffer

Next, 5ml DMS0, 10 ml 50mM PBS buffer and 85 ml H_20 were mixed together, and the pH was adjusted to 7.4, giving a 10 mM buffer containing 5 % DMSO.

3.5 Replicates

Replicated RII measurements were made for seven substances. The substances used were selected to represent all groups of substances, and to cover the RII-variation range. For each substance three replicates were made. The sample preparation was performed the same way as described earlier. RII measurements were performed both on the day the substance was diluted and the day after. An additional buffer was prepared and the same procedure was repeated a second time.

The relationship between RI and concentrations is supposed to be linear. The linearity was tested by diluting the highest prepared concentration of all seven substances. Eight 1:3 dilutions were made and the RI was measured.

3.6 Data analysis and model generation

The obtained RIIs were analyzed together with corresponding molecular descriptors using the SIMCA-P software and Statistica. The goal here was to create a QSAR-model for the prediction of RIIs from the molecular descriptors. A better understanding of what properties that influences the RII was wanted, and also how and if the model can be used to compensate for RII-differences for affinity rankings.

First PLS-models were created and analyzed. The dataset was divided into a training set and a test set. The training set was selected from the dataset by performing a DO analysis in Modde. The structural diversity of the training set was compared to the whole dataset's diversity by analyzing the distribution in structural space. The substances that were not selected to belong to the training set were instead used as a test set. PLS-models representing both RIIconc and RIImol were generated.

Based on the results obtained in the PLS analysis, LR-models were calculated using variables that had been seen to be important for the prediction of RII. Models that are supposed to be used for signal correction should preferably be as simple as possible. Results from predictions of RIIconc and RIImol were then compared.

3.7 Validation

The models obtained were validated using a dataset of LMW fragments binding to thrombin.

RIIs predicted by the models were compared to calculated saturated SPR responses obtained from affinity fitting (R_{max}).

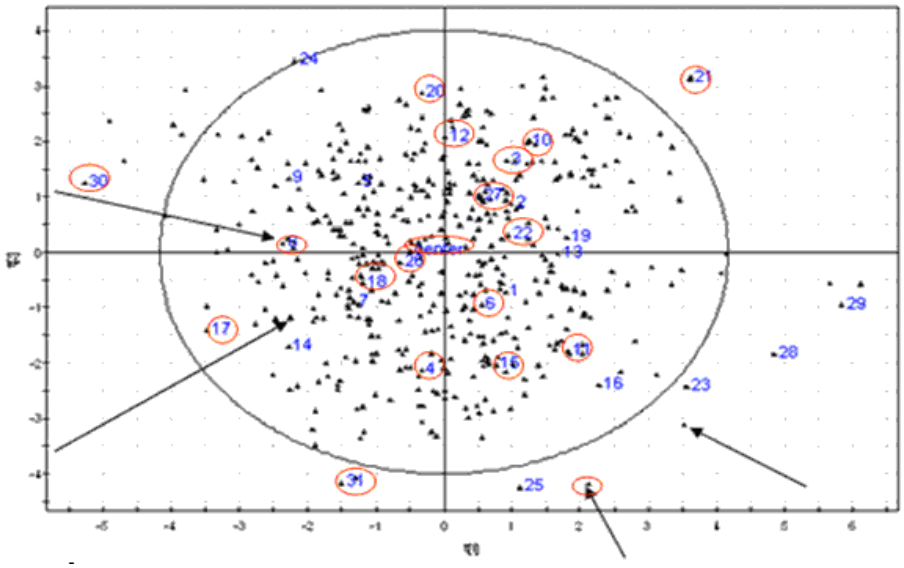
4. Results

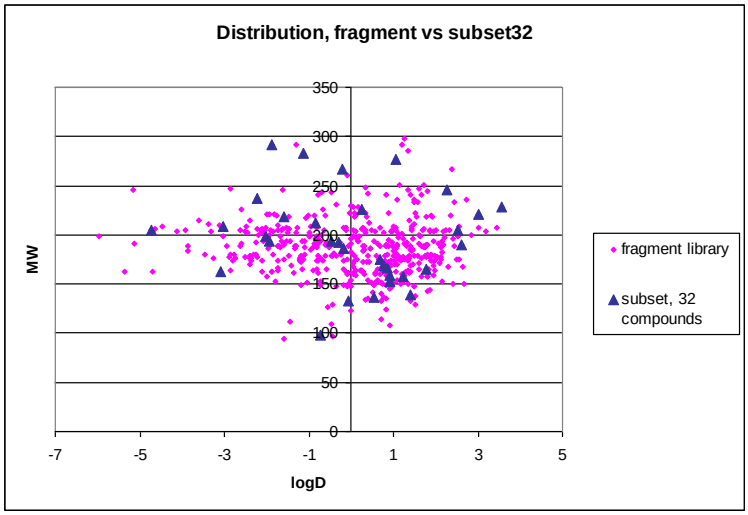
The dataset used in this work has been selected/collected in different ways. The part consisting of halogen substituted molecules had been selected prior to this project, one had earlier observed that halogen substitution resulted in higher RIIs. Substances in the literature part consist of those substances for which RIIs were found. The third part was selected from a fragment library that had been selected for its structural diversity in a prior project. From this fragment library a selection of a subset was made. A number of additional substances were later added, since they were already in stock at Biacore.

4.1 Selected subset from the Maybridge fragment library

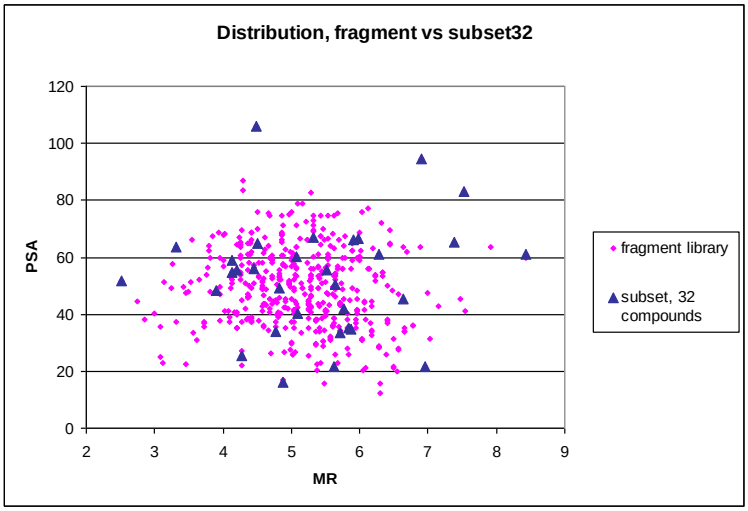
The selection of a subset from Maybridge fragment library was carried out as follows; The DOOD analysis resulted in a subset containing 32 compounds, out of the 500 in the fragment library. Due to delivery problems four compounds were replaced by substances that were close in structural space (figure 5a). Twelve compounds out of the 32 were not soluble in the buffer used, and RII measurements could be made on only 20 compounds. The subset containing 32 compounds selected by DOOD covered the structural space well. The subset containing 20 substances for which RIIs could be measured covered the structural space reasonably well (figure 5c). The smaller subset does not represent molecules with high MR, but higher MR (up to 10) is represented in the halogen dataset. In the smaller subset the representation of molecules with high molecular weight and at the same time high lipophilicity is not covered. Compounds located to the right in figure 5a are all substances that have high values on steric descriptors.

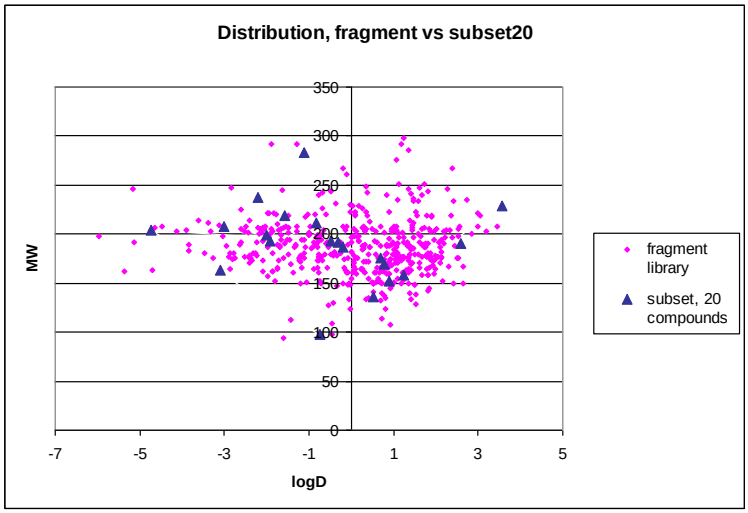
a)

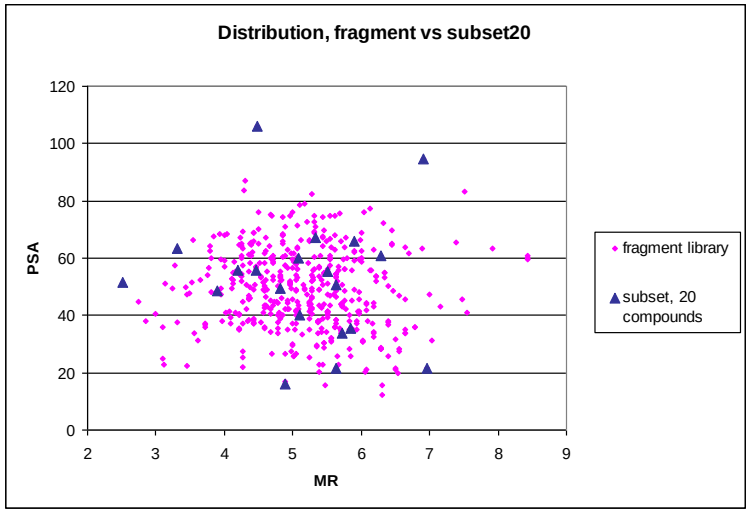



Distribution, fragment vs subset32
MW
logD
fragment library
subset, 32 compounds

b)


Distribution, fragment vs subset32
PSA
MR
fragment library
subset, 32 compounds





c)

Figure 5.

a) Score plot from PCA showing t1-t2 of the Maybridge fragment library. Substances marked with blue numbers represent substances belonging to the subset selected by DOOD analysis. 4 substances had to be replaced and their replacements are marked with arrows. Substances for which RII measurements could be performed are encircled.

b-c) Scatter plots visualizing the distribution of the two subsets (b-32 compounds subset, c- 20 compound subset) in comparison to the fragment library when lgD vs. MW (left-hand side plot) and MR vs. polar surface area(right-hand side plot) are plotted.

4.2 Substances and their RIIs

The results from RII measurements and data found in the literature are compiled in Table 1 and Table 2.

	Molecule name	RIIconc	RIImol			Molecule name	RIIconc	RIImol
1	Benzoic acid	0.1624	0.0198		27	my histidine (L-)	0.2022	0.031
2	2-Flurobenzoic acid	0.1255	0.0176		28	my Valine (L-)	0.1774	0.021
3	2-Bromobenzoic acid	0.1374	0.0276		29	my lysine (L-)	0.2096	0.031
4	2-Chlorobenzoic acid	0.145	0.0227		30	my aspartic acid (DL-)	0.1314	0.018
5	2-Iodobenzoic acid	0.1485	0.0368		31	my glutamic acid (L-)	0.1375	0.02
6	2,6-Difluorobenzoic acid	0.101	0.016		32	Folic acid	0.2222	0.098
7	2,6-Dichlorobenzoic acid	0.15	0.0286		33	AC 12605	0.1731	0.046
8	5-Fluorosalisyclic acid	0.1233	0.0192		34	CC 13501	0.1965	0.038
9	5-Chlorosalicyclic acid	0.1499	0.0259		35	CC 41801	0.2013	0.04
10	5-Bromosalicyclic acid	0.1383	0.03		36	CC 01709	0.1689	0.026
11	5-Iodosalicyclic acid	0.1394	0.0368		37	BTB 09284	0.2437	0.052
12	3,5-Diiodo-2-salisylic acid	0.1337	0.0521		38	BTB 15113	0.2294	0.052
13	3,5-Dibromosalicyclic acid	0.1331	0.0394		39	CC 35509	0.1133	0.022
14	L-phenylalanine	0.2152	0.0355		40	CC 24601	0.1855	0.039
15	4-Fluro-L-phenylalanine	0.1832	0.0335		41	MO 00127	0.1413	0.034
16	4-Chloro-L-phenylalanine	0.2167	0.0433		42	KM 06872	0.1986	0.027
17	2,7-Dichlorofluorescein	0.5604	0.2248		43	BTB 14322	0.164	0.027
18	4,5-Dibromofluorescein	0.337	0.1652		44	CD 04786	0.1868	0.03
19	3,4-Difluorophenylacetic acid	0.1013	0.0174		45	CC 45596	0.2006	0.058
20	3-Iodo-L-tyrosine	0.1677	0.0508		46	CC 43113	0.2242	0.043
21	5-Iodouridine	0.1545	0.0572		47	KM 01757	0.1924	0.055
22	Furosemide	0.1808	0.0598		48	CC 26823	0.1196	0.02
23	Chloramphenicol	0.1824	0.0589		49	KM 07844	0.1411	0.026
24	Diclofenac	0.217	0.069		50	SB 00621	0.1492	0.015
25	my Glycin (L-)	0.1781	0.0134		51	SP 01488	0.1408	0.029
26	my trypthophan (L-)	0.2213	0.0452		52	CC 10209	0.274	0.048

Table 1. List of all substances for which RIIs was measured in this work. RIIconc is given in ml/g and RIImol in dm^3/mol.

	Molecule name	RII
1	Sodium diatrizoate	0.2099
2	Procaine hydrochloride	0.246
3	citric acid	0.1375
4	Tetracaine hydrochloride	0.2292
5	Dextran	0.1512
6	D-glucose	0.1474
7	D-fructose	0.1469
8	glycerol	0.1256
9	Sucrose	0.1486
10	D-mannitol	0.1466
11	Poly(acrylamide)	0.1723
12	Poly(methacrylic acid)	0.1653
13	Ply(vinylpyrrolidone)	0.1403
14	netropsin	0.2023
15	Quinacrine	0.2323
16	Berenil	0.3303
17	Pentamidine	0.2103
18	creatinine	0.1803
19	Davis' threonine	0.2403
20	Neomycin	0.153
21	DB404	0.272
22	Glycin (L-)	0.2153
23	Alanine (B-)	0.2323
24	Valine (L-)	0.2603
25	Leucin (L-)	0.2693
26	Serine (DL-)	0.2103
27	Cystein (L-)	0.2283
28	Phnylalanine (L-)	0.2773
29	tyrosine (L-(-).)	0.2623
30	trypthophan (L-)	0.2873
31	histidine (L-)	0.2433
32	lysine (L-)	0.2563
33	aspartic acid (DL-)	0.2063
34	glutamic acid (L-)	0.2273
35	Isoleucin	0.2723
36	Threonine	0.2263
37	Hydroxyproline	0.2193
38	Methionine	0.2353
39	Glutamine	0.2273
40	alpha-aminobutyric acid	0.2483
41	alpha-aminovaleric acid	0.2623
42	alpha-aminocaproic acid	0.2683
43	Amantadine hydrochloride	0.1814

Table 2. List of all substances for which RIIs were found in the literature. RII is given in ml/g.
1-10 Handbook of Chemistry and physics[19]. λ=589.3 nm, T=20ºC, buffer=water
11-13 Polymer Handbook[4], λ=546 nm, T=25ºC, buffer=water
14-22 Davis[1], average λ=590.5 nm, λ=633 nm and λ=679.5 nm , T=25ºC, buffer=water
23-43 McMeekin[3], λ=589. nm, buffer=water

By comparing RIIs gained from measurements and RIIs obtained from the literature, Table 1 and Table 2, one can see that some differences do exist. An example of this can be seen in the case of amino acids. My measurements are always lower than the ones found in the literature, and the ranking among amino acids is not the same. Further analysis was therefore performed without the literature values. Experimental errors are then hopefully only of the systematical kind.

When examining the obtained RIIs different patterns were seen. As visualized in Figure 6, fluorine substituted structures gives lower RIIs than the corresponding parent molecule. Chlorine, Bromine and Iodine substitution increases the RII, at least when the molar scale is used. The pattern is changed when the weight scale based RIIs are used.

Fluorine substituted molecules still receives the lowest RIIs. For the benzoic acid family for example, benzoic acid receives the highest RII value (Table 1). The three strongly colored substances have the highest RIIs. For those, Folic acid, 4,5-dibromofluorescein and 2,7-dichlorofluorescein, absorbance peaks were measured in a Spectrophotometer. The lowest concentration from each (~2.5mg/ml) was diluted 100 times before the absorbance was measured (Table 3).

Substance	**Wavelength (nm)**	**ABS**
Folic acid	280	1.242
4,5-dibromofluorescein	507	2.128
2,7-dichlorofluorescein	503	1.820

Table 3. Results from absorbance measures in the Spectrophotometer for three strongly colored substances.

4.3 Replicates

A number of replicates were made to estimate the experimental error in RII measurements. Standard deviation values ranged from 0.004 to 0.04 (Table 3). Two substances, 4,5-dibromofluorescein and BTB 14322 showed a variation of 8 and 10 percent respectively. For the other the variation was about 2-4 percent. The variation seen for each substance is visualized in figure 7. The relationship between concentration and RI was also examined, and was proven to be linear (figure 7). No difference was seen between measurement days.

Weight based RII's				**Molarity based RII's**			
Substance	**mean value**	**Standard deviation**	**variance coefficient**	**Substance**	**mean value**	**Standard deviation**	**variance coefficient**
4,5-dibromofluorescein	0.398	0.033	8.364	4,5-dibromofluorescein	0.195	0.016	8.360
Folic acid	0.222	0.008	3.748	Folic acid	0.098	0.004	3.733
2,6-difluorobenzoic acid	0.105	0.004	3.540	2,6-difluorobenzoic acid	0.017	0.001	3.435
Lysine	0.211	0.006	2.875	Lysine	0.031	0.001	2.826
BTB 14322	0.188	0.020	10.392	BTB 14322	0.029	0.001	5.108
MO 0127	0.152	0.006	3.734	MO 0127	0.036	0.001	3.749
Aspartic acid	0.134	0.003	2.416	Aspartic acid	0.018	0.000	2.188
Average		0.011	3.079	***Average***		0.006	2.044

Table 4. Table showing the results obtained from RIIconc measurements performed on a number of replicates (n=7).

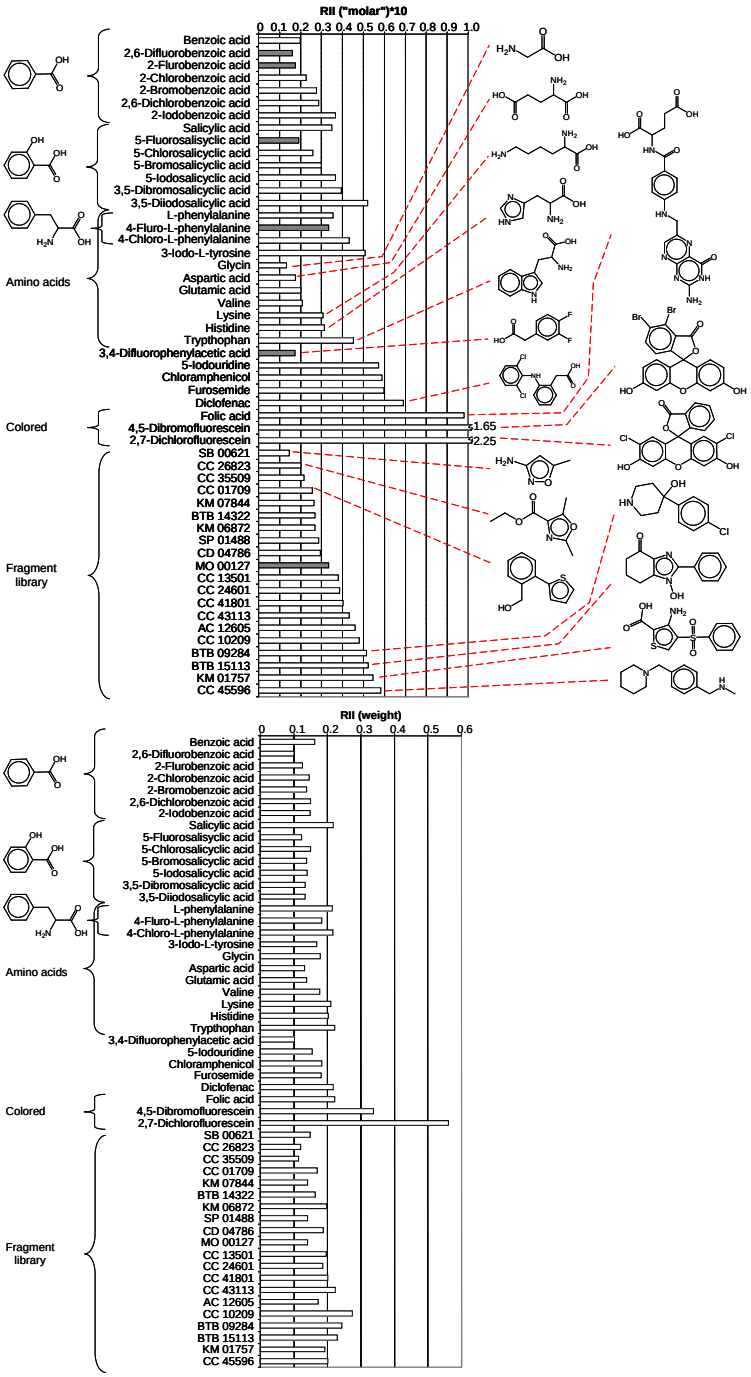
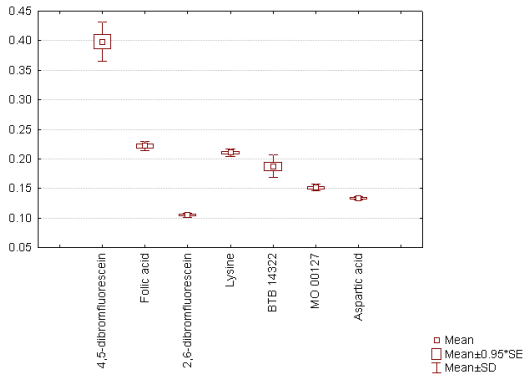
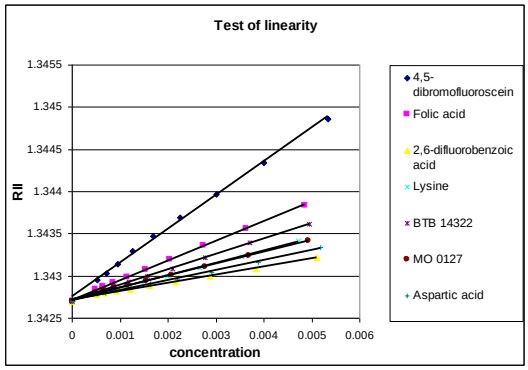


Figure 6. A representation of the structures of a selected number of compounds together with RIImol values for all substances. Bars marked with grey are RIIs belonging to structures containing fluorine. B) RII measured in weight units.



a)



b)

Figure 7.
a) A box plot visualizing the variation within the replicates.
b) Scatter plot of RI versus concentration shows that the relationship between concentration and RI is linear.

4.4 PCA and PLS

The data used from now on contains 52 substances for which RIIs have been measured in this work. Initial PCA indicated that RIIconc correlated poorly with the molecular descriptors. RIImol showed a much better correlation. The same pattern was seen when a PLS analysis was made. The prediction ability for RIImol was significantly better then it was for RIIconc. The model was generated using 38 compounds from the dataset. This training set had been selected by performing a DO analysis in Modde. The rest of the substances in the dataset, 14 compounds, built up the test set. The PLS-model for RIImol predictions explains 95.3% of the variation in y, and the prediction ability of the model is 85.6%. The descriptors that influenced the model most strongly were molar refractivity, molecular volume, molecular weight, molecular surface area, and lgD. They were kept in the model. The two substances that have clearly the highest RIIs are also the two substances that have the most number of rings (5 rings) in their molecular structure. Since extreme values tend to have the strongest influence on the model it is not unexpected that the number of rings was important for the model. In the dataset there are a number of substances that has 1, 2 and 3 rings in their structure, and there are two substances that have 5 rings. If the two fluoroscein substances are excluded the Number of ring descriptor do not influence the model to the same degree. The number of rings variable might be misleading and it was not included in the model. A new model was generated using the above-mentioned variables. The model generated explained 86.3% of the variation in y, and had a prediction ability of 84.7%. The prediction ability was almost as high as when all descriptors were used. When the model was applied to the test set the model showed prediction ability of 66%, but by excluding the worst predicted substance the prediction ability became 92.6% (Figure 8).

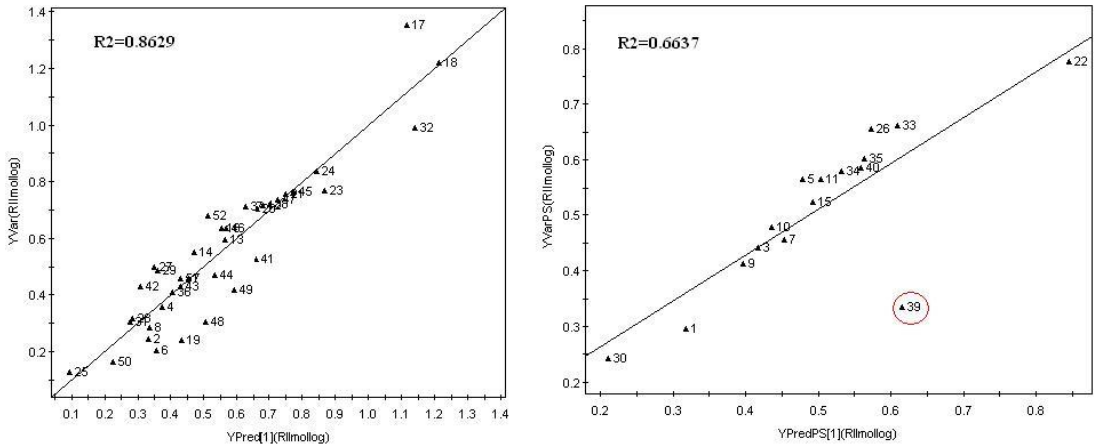


Figure 8.
Predicted vs. observed for a model based on RIImol as y using a selected number of molecular descriptors (MR, MW, MSA, MV, lgD).
a) Applied to the training set, R^2 is 0.86
b) Applied to the test set, R^2 is 0.66. By removing the worst predicted substance, CC 35509 (encircled in the plot), R^2 becomes 0.93

4.5 Linear regression (LR) analysis

The PLS model obtained was a simple one factor model indicating that a simple relation exists between descriptors and RII. The correlation between a number of single descriptors and RIIs was therefore examined. The correlation between the molecular descriptors and RIImol is clearly better than the correlation between the molecular descriptors and RIIconc Figure 9). The correlation between MR and RIImol was the strongest, R=0.83. The correlation between MW and RIImol was lower, R= 0.77. There is a strong correlation between MV and MR (Figure 9). One can also see that there are two substances, located in the right-hand side of the plot, for which the correlation is clearly worse than it is for the other substances.

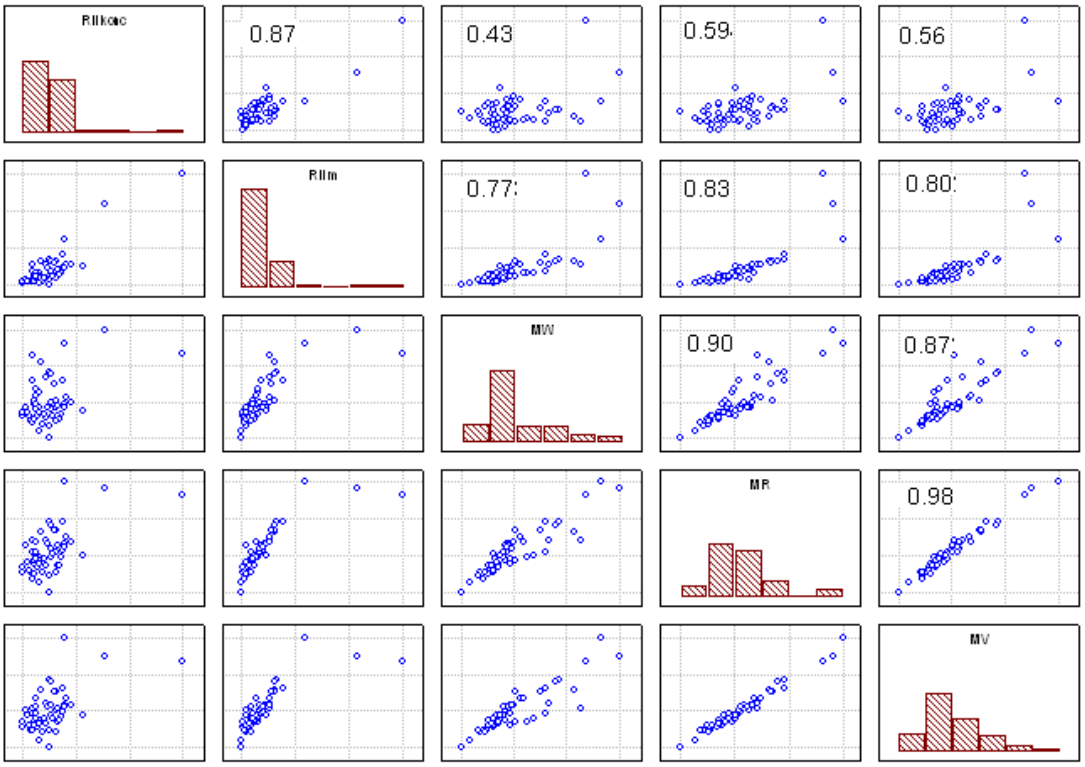
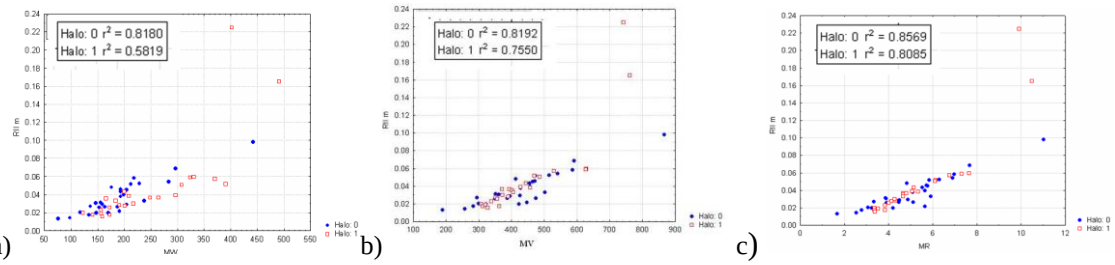


Figure 9. Correlations between a selected number of descriptors and RIIs. One can see that there is an almost linear correlation between MR and MV. The correlation between RIIconc and the molecular descriptors is not as good as the correlation between RIImol and the same descriptors. The three substance for which the correlation is not as good as for the others are the three strongly colored substances.

Since the strongest correlation was seen between RII and MR, MR is easily obtained from many computer software's, models for predicting both RIIconc and RIImol was generated based on only MR. The LR-model generated for predicting RIIconc had a R^2 value of 0.33, while the LR-model for predicting RIImol had R^2 value of 0.86.

All substances were divided into two classes, halogen substituted molecules and non-halogen substituted molecules. The correlation between MW/MV/MR and RIImol was examined. When MW vs. RIImol is examined the correlation is better for non-halogen substituted molecules. For MV vs RIImol halogen substituted molecules correlated better. For MR vs. RIImol the correlation is even better. 2,7-dichlorofluorescein and 4,5-dibromofluorescein do not correlate as well as the other substances.



a) b) c)

Figure 10. Categorized scatter plots.
a) Correlation between MW and RIImol
b) Correlation between MV and RIImol
c) Correlation between MR and RIImol

As already seen the correlation between RIImol and MR is linear except for 2,7-dichlorofluorescein and 4,5-dibromofluorescein (Figure 11). Without these two the correlation between RIImol and MR becomes much better. The R^2 value obtained is 0.92. The model is highly significant and has a Std. error of estimate of 0.005. Folic acid which is colored (even though not as strongly colored as fluoresceins) fits well into the model. Three substances were identifies as outliers in the regression model (SD of residuals were larger than +/-2). They were however not excluded. CC 35509 was also identified as an outlier in the PLS analysis.

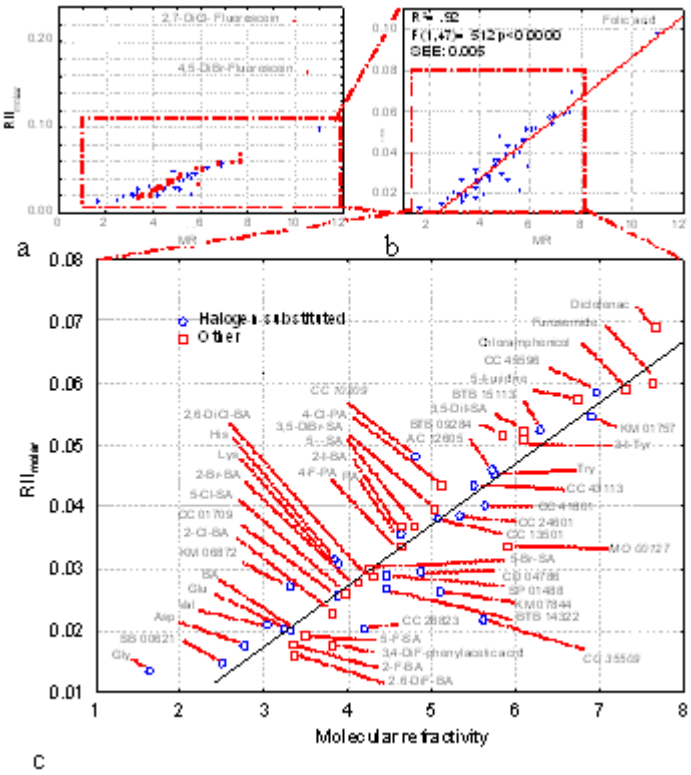
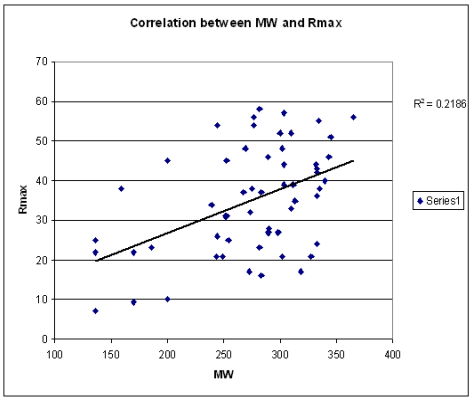


Figure 11.
a) A scatter plot showing RIImol vs. MR. The correlation is linear except for 2,7-dichlorofluorescein and 4,5-dibromofluorescein.
b) LR-analysis gives a model with a high significance level, and a R^2-value of 0.92. The error of estimate is only 0.005.
c) A close up of b. Benzoic acids are abbreviated as BA, Salicylic acids as SA, Phenylalanines as PA and Amino acids are abbreviated with their three letter code. Three substances are marked in italic (CC 35509, MO 00127and CC 10209). Their standard deviation values of residuals are larger than two.

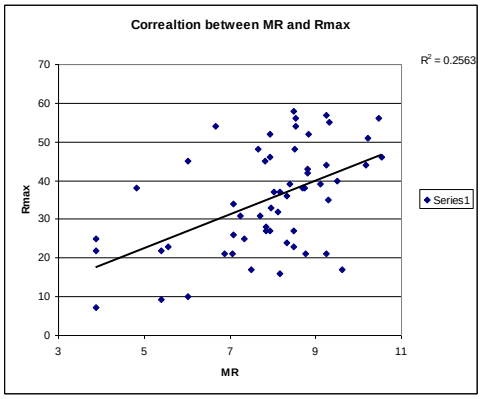
4.6 Validation of models, both PLS-models and LR-models, using saturated SPR responses

An attempt to validate the model on an independent set of LMW fragments was made. The correlation between MW and R_{max} is not good, R^2 is only 0.22 (Figure 12a). The correlation is as bad when the correlation MR vs. R_{max} and RIImol (predicted by LR model) vs. R_{max} is examined, R^2 is only 0.26 and 0.25 respectively. The correlation is independent on if RIIconc or RIImol values have been used. The model based on MR used for predicting logarithmic RIImol values is y=0.1199x -0.0716.

a)



b)



c)

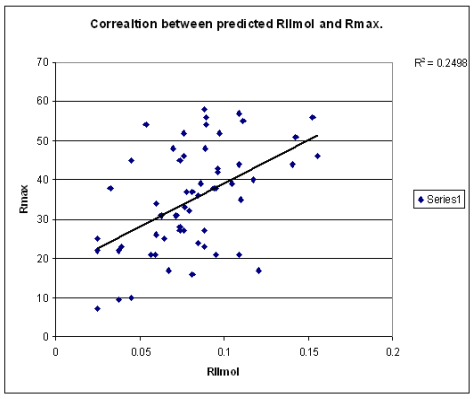


Figure 12 Validation of models on an external dataset
a) Correlation between MW and R_{max}
b) Correlation between MR and R_{max}
c) Correlation between predicted by LR-models based on MR.

5. Discussion

The relationship between the molecular descriptors and molar based RIIs was found to be clearly better than with weights based RIIs. The explanation for this might be that molecular descriptors are molar based. Biacore's biosensors measure the change in mass/area unit at the chip surface i.e. weight as change in refractive index. However, the number of binding sites/area unit is molar dependent.

From the data gained from RII measurements made one could see that fluorinated molecules have a lower RII then corresponding molecule without fluorine. The pattern can be seen for all three homologous families of molecules investigated here. This observation also means that there is not a linear correlation between molecular weight and RII, since adding fluorine to the structure means that the molecule weight of the molecule increases. For the halogen substituted homologous series the parent compound had the highest RII when using weight based RII-scale. However, the molar RII scale agrees much better with previous findings that chlorine/bromine/iodine substituted compounds have high RIIs.

The ability to predict RIIs for the strongly colored substances, Folic acid, 4,5-dibromofluorescein and 2,7-dichlorofluorescein, was shown to be low. RI has a tendency to strongly increase if RI is measured at wavelengths found on the absorbance peak. The results gained from the absorbance measurements show that the absorbance peaks for three strongly colored substances are all found at wavelengths below the one

used for RI measures performed in this work. The high RIIs obtained for these three substances can therefore not be explained with the above mentioned effect.

When comparing the RIIs measured in this project with RIIs found in the literature some differences could be seen. Those differences might be due to the fact that my measurements are made using a buffer which has a higher RI than water, 1.3426 instead of 1.3329. In some cases also the temperature and the wavelength used are different. The use of a different buffer can in this case account for a deviation of ~5%. The deviation seen here is larger than can be explained by buffer differences. The replicates made showed that measurement errors in my data were low; the variance coefficients indicated an average of ~3% error.

The PLS-models that were generated were not better than LR-models based only on MR. For practical reasons an as simple model as possible is wanted. The LR-models based on only one x-variable seem to be appropriate for this limited set of molecules. The average value of the standard deviations obtained from the replicate measurements was 0.011 for weight based RIIs and 0.006 for molar based RIIs. The Standard error of estimate in the LR-model when fluoresceins are not included is 0.005 and 0.013 when fluoresceins are included. This indicates that the LR-model based on MR explain the data to the degree that can be expected considering the measurements errors in the data.

The descriptors that show the strongest correlation with RII is MR. The correlation between MV and RII is almost as good. The correlation is significantly better then molecular weight, especially when molecules containing halogens are used. MR is as MW easily obtained from many computer softwares. MR is computed from the different bonds present in the molecule. Fluorine, bromine, chlorine and iodine are among those atomic groups and structural contributions whose effect on MR is accounted for, those effects are not accounted for in MW. MR gives an approximate measure of the total volume occupied by the molecule[20]. It should therefore be possible to obtain better adjustments of the SPR-signal by using MR as a normalization factor. Since the correlation with RII is better then for MW and both quantities are as easily obtained, it is better to use MR for correction of the SPR response.

The validation of the generated models made on an external data set showed that when it comes to correlation with R_{max} there is no difference between RIIconc and RIImol. The correlation was only slightly better than for MW. Predicted RIIs and R_{max} values correlated badly so did also MR and R_{max}. In order to be able to draw any conclusions from this, one would have to perform highly controlled R_{max} measurements. In the data used there exist some uncertainties regarding changes in surface quality among the experiments.

6. Conclusions

1. Fluorinated molecules have lower RII than the parent molecule.
2. There is no strong linear relationship between increased molecular weight and RII, especially not for halogen substituted molecules.
3. RII of colored molecules cannot be predicted
4. Chlorine, bromine and iodine substituted compounds give higher RII than the parent compounds, but only if RII is expressed in molar scale.
5. Predicted RIIs showed a poor correlation with R_{max}, independent on the scale used.

6. To better validate the use of MR/MV based signal adjustments, focused experiment where R_{max} is controlled need to be done.

7. Acknowledgments

I would especially like to thank my supervisor Markku Hämäläinen for all his help during this project. I would also like to thank Håkan Roos who has acted like a second supervisor during this project and whom have helped me a lot, and Johan Gottfries who have generated all molecular descriptors used. I would also like to thank Åsa Frostell-Karlsson for all her help concerning laboratory work. Finally, I would also like to thank my family and friends for their support through out this project.

8. References

[1] Davis, M Tina, "Determination of the Refractive Index Increments of Small Molecules for Coreection of Surface Plasmon Resonance Data". *Analytical Biochemistry* **284** 348-353 (2000).
[2] Hanson J, Refractometry, *Chemistry Lab Techniques* (2003). http://www2.ups.edu/faculty/hanson/labtechniques/refractometry.html (5 Feb. 2007).
[3] McMekkin T. "Refractive Indices of Amino Acids, Proteins, and Related Substances". *Adv. Chem. Ser.*. **44** 54-65 (1964).
[4] Brandrup J & Immergut E:H. "Polymer Handbook, third edition" *John Wiley and Sons Inc.* (1989).
[5] Carrasco-Velar, Ramon. "Definition of a novel atomic index for QSAR: the refractotopological state", *J Pharm Pharmaceut Sci* **7** 19-26 (2004).
[6] Ball Vincent. "Buffer Dependence of refractive index increments of protein solutions". *Biopolymers I* **46** 489-492 (1998).
[7] Perlmann E. "The specific Refractive Increment of some Purified Proteins". *J. Amer. Chem. Soc*. **70** 2719-2724 (1948).
[8] Biacore Ab's http://www.biacore.com (5 Feb. 2007).
[9] Biacore, "Surface plasmon resonance, technology note 1". (Sept. 2001).
[10] Handbook, version AB, Biacore AB (1998).
[11] Eriksson L. "Instruction to Multi- and Megavariate Data analysis using Projection Methods (PCA & PLS)". *Umetrics* (1999).
[12] Tulsa, "Electronic Statistics Textbook". *StatSoft*. http://www.statsoft.com/textbook/stathome.html(5 Feb. 2007).
[13] Krogsgaard:Larsen P. "Textbook of drugdesign and discovery, third edition". *Taylor and Francis* (2002).
[14] Eriksson L. "Onion design and its application to pharmaceutical QSAR problem". *Journal of Chemometrics* **18** 188-202 (2004).
[15] Olsson I. "D-Optimal onion design in statistical molecular design". *Chemometrics and intelligent laboratory* **73** 37-46 (2004).
[16] Olsson I, "Controlling coverage of D-optimal onion designs and selections". J. Chemometrics **18** 548-557 (2004).
[17] Oprea, Tudor I. "Chemography: The Art of Navigating in Chemical Space". *J. Comb. Chem*. **3** 157-166 (2001).
[18] Kernschen, Dr. W. "Operation Manual ABBEMAT Digital Automatic Region" release 1.0.4 (2001).
[19] Lide, David D. "Handbook of Chemistry and Physics, edition". *CRC Press.* (1995).
[20] Weast, Robert C. "Handbook of Chemistry and Physics, 70th edition". *CRC Press.* (1989).
[21] Molecular Discovery Ltd, "Volsurf manual". Molecular Discovery Ltd (2004).

9. Appendixes

9.1 Appendix A; Abbreviations

Low Molecular Weight Compounds – LMWs
Refractive Index Increment – RII
Molar Refractivity – MR
Molecular Weight –MW
Molecular volume - MV
Surface Plasmon Resonance – SPR
Refractive Index – RI
Resonance angle – SPR angle
Resonance Unit – RU
Quantitative Structure-Activity Relationship – QSAR
Linear Regression – LR
Multiple Linear Regression – MLR
Principal Components Analysis – PCA
Principal Component – PC
Unit variance – UV
Projection to latent structures by partial least squares – PLS
Explained variance – R^2
Prediction ability – Q^2
Statistical Molecular Design –SMD
Space Filling design – SF-design
D-Optimal Design –DO-design
D-Optimal Onion Design – DOOD
Refractive Index Increments values expressed in weight scale - RIIconc
Refractive Index Increments values expressed in molar scale –RIImol

9.2 Appendix B; Molecular descriptors

Descriptors	**Explanation**
Distribution coefficient (lgP)	Distribution coefficient. Calculated as $\log(c_{octanol}/c_{water})$, a measure of the hydrophilicity of the molecule
Calculated distribution coefficient (ClgP)	Calculated logP values
lgD7.4	Distribution coefficient that takes all neutral and charged forms of the molecule into account. Here measured at pH 7.4
lgD6.5	Distribution coefficient that takes all neutral and charged forms of the molecule into account. Here measured at pH 6.5
Molecular weight (MW)	Steric descriptor. Measurement of the size of the molecule
Number of donors	Number of hydrogen donating atoms
Number of acceptors	Number of hydrogen accepting atoms
Molecular refractivity (MR)	Steric descriptor. Measure of the volume taken up by a molecule in a solution
Molecular volume (MV)	Steric descriptor. Measurement of the size of the molecule
Molecular surface area (MSA)	Measurement of the size of the molecule
Polar surface area (PSA)	The amount of polar atoms on the surface of the molecule
Non-polar surface	The amount of non-polar atoms on the surface of the molecule

area (NPSA)	
%PSA	The amount of polar atoms calculated as percentage
%NPSA	The amount of non-polar atoms calculated as percentage
Polar atoms (PAT)	Number of polar atoms in the molecule
Non polar atoms (NPAT)	Number of non-polar atoms in the molecule
Acid	Is the molecule an acid or not
Base	Is the molecule a base or not.
Neutral	Is the molecule neutral
Zwitter ion	Is the molecule a zwitter ion
Number of bonds	Number of rotational bonds
Lipinski score	Number of parameter satisfying Lipinski's rule of five
Amphilic moment	The amphilic moment is used as a measure of the molecules ability to penetrate a membrane. It is calculated as the length of the vector pointing from the center of the hydrophobic domain to the center of the hydrophilic domain.
Critical packing parameter	The ratio between the hydrophobic and the lipophilic parts of the molecule can be used to predict the packing of the molecular packing, such as the formation of micells.
Capacity factors	A measure of the rate of hydrophilic regions in comparison to the total molecular surface. Measured at -0.2, -0.5, -1.0,-1.2,-3.0,-4.0,-5.0,-6.0 kcal/mol
Local minima of interaction energy distances	Given the three best local minima of interaction energy when the probe is interacting with a target molecule.
Hydrophobic regions	A measure of how hydrophobic a molecule is. It is defined as the envelope accessible by solvent water molecules. Measured at -0.2, -0.4, -0.6, -0.8,-1.0,-1.2,-1.4,-1.6 kcal/mol
Local interaction energy minima	Given the three best local interaction energy minima when the probe is interacting with a target molecule.
Molecular globularity	The molecular globularity is a measure of how special a molecule is. This entity is also related to the flexibility of the molecule.
The hydrophilic-lipophilic balance	The hydrophilic-Lipophilic balance tell if the molecule is more hydrophilic or more lipohpilic.. It is calculated as the ratio between hydrophilic regions measured at a certain energy level and hydrophilic regions at a certain energy level. Measued at -0.6, -0.8 kcal/mol
Hydrophobic integy moments	A measure of the unbalance between the position of the center of mass and the position of the hydrophobic regions. Measured at -0.2, -0.5, -1.0,-1.2,-3.0,-4.0,-5.0,-6.0 kcal/mol
Integy moments	A measure of the unbalance between the position of the center of mass and the position of the hydrophilic regions. Measured at -0.2, -0.5, -1.0,-1.2,-3.0,-4.0,-5.0,-6.0 kcal/mol
Volsurf molecular weight	Molecular weight, Steric descriptor. Measurement of the size of the molecule
Polarizability	The relative tendency of the molecule to develop a charge distribution
Volume/surface ratio	The rate between the volume and the surface area of the molecule is a measure of the rugosity (how wrinkled the surface is) of the molecule.
Volsurf molecular surface area	Molecular surface Steric descriptor. Measurements of the size of the molecule
Volsurf molecular volume	Molecular volume. Steric descriptor. Measurements of the size of the molecule
Hydrophilic regions	A measure of how hydrophobic a molecule is. It is defined as the envelope accessible by solvent water molecules. Measured at -0.2, -0.5, -1.0,-1.2,-3.0,-4.0,-5.0,-6.0 kcal/mol
Hydrogen bonding	Represents the capability of the molecule to form hydrogen bonds at different energy levels. Measured at -0.2, -0.5, -1.0,-1.2,-3.0,-4.0,-5.0,-6.0 kcal/mol
Volsurf Best volumes	Represents the three best hydrophilic regions generated when a water molecule interact with the molecule in question. Measured for both H_2O and dry probe, at -1.0, -3.0 kcal/mol
Elongation	Elongation is a measure of how far the molecule can reach when it is stretched.
Fixed Elongation	The fixed elongation is calculated when considering a part of the molecule as rigid. The ratio between the elongation and the fixed elongation represents the portion of

	the extension given the rigid part.
Diffusivity	A representation of how easily a solute transfers in a given fluid when influenced by a concentration gradient.
Volsurf ClogP	Calculated logP values

[13] Krogsgaard, "Text book of drugdesign and discovery"
[21] Molecular Discovery Ltd, "Volsurf manual".

9.3 Appendix C; Maybridge subset substances, CAS names

Product name	CAS name
MO07110	3-(1-Pyrrolidinsylsulfonyl)aniline
CC29209	(4-Methyl-2-phenyl-1,3-thiazol-5-yl)methanol
AC12605	1-(3-Methoxyphenyl)piperazine
CC13501	4-Methyl-3,4-dihydro-2H-1,4-benzoxazine-7-carboxylic acid
CC04409	2-Quindinylmethanol
CC41801	2-Pyrid-3-ylbenzoic acid
CC39222	Methyl 4H-furo[3,2-b]pyrole-5-carboxylate
CC01709	1,3-Benzodioxol-4-ylmethanol
KM00316	3-(tert-Butyl)-1H-pyrazol-5-amine
BTB09284	4-(4-Chlorophenyl)-4-hydroxypiperidine
BTB15113	1-Hydroxy-2-phenyl-1,5,6,7-tetrahydro-4H-benzimidazol-4-one
CC35509	(2-Thien-2-ylphenyl)methanol
BTB01858	2-Morpholino-5-(trifluoromethyl)aniline
GK04786	4-Hydrazinothien[2,3-d]pyrimidine
CC24601	2-Morpholinoicotinic acid
MO00127	1-(4-Fluorobenzyl)-5-oxo-3-pyrolidinecarboxylic acid
KM06872	2,1,3-Benzoxadiazol-5-ol
BTB14322	Indoline-2-carboxylic acid
CD09182	3-Amino-2-phenyl-1H-inden-1-one
CD04786	2-(1H-Pyrrol-1-ylmethyl)piperidine
CC45596	N-Methyl-N-[4-(piperidin-1-ylmethyl)benzyl]amine di hydrochloride
CC43113	(2-Morpholinopyrid-4-yl)methylamine
BTB08846	Methyl 4-(methylthio)-6-oxo-2-phenyl-1,6-dihydropyrimidine-5-carboxylate
SB01761	Indan-2-amine
KM01757	3-Amino-4-(phenylsulfonyl)thiophene-2-carboxylic acid
CC26823	Ethyl 2,5-dimethyl-1,3-oxazole-4-carboxylate
KM07844	Ethyl 1,4-dimethylpiperazine-2-carboxylate
HTS07558	2-(2-Hydroxyethyl)-3-methyl-1-oxo-1,5-dihydropyrido[1,2-a]benzimidazole-4-carbonitri
CC30013	tert-Butyl 4-[4-(aminomethylphenyl] tetrahydo-1-("H)-pyrazinecarboxylate
AC21377	3-Amino-5-methylisoxazole
SP01488	5-(aminosulfonyl)-1-methyl-1H-pyrrole-2-carboxylic acid
CC10209	(5_phenyl-1,3-oxazol-4-yl)methanol

9.4 Appendix C; Non-solvable substances

1. 4-Bromo-L-phenylalanine
2. 4-Iodo-L-phynylalanine
3. Iopanoic acid
4. Iophenoxic acid HPLC
5. 3,5-Diiodo-L-thyronine
6. Amiloride
7. Hydrochlorothiazide
8. Hexachlorophene
9. Prednisone
10. Tetracycline hydrochloride
11. L-thyroxine sodium salt pentahydrate
12. Amantadine hydrochloride
13. Carbamazepine
14. MO 07110
15. CC 29209
16. CC 04409
17. CC 39222
18. KM 00316
19. GK 02837
20. CD 09182
21. BTB 08846
22. SB 01761
23. HTS 07558
24. CC 30013

9.5 Appendix D; A100 experiment

A test was made to extract RII of a subset of substances from the Maybridge fragment library using the Biacore A100 instrument. The hope was that the results would be possible to use for comparisons between values obtained from Biacore A100 with the results obtained from the Refractometer.

The test was performed as follows:

1. 5 mg of all substances was used.
2. All substances were dissolved in 100 % to a concentration of 100 mM
3. A part of the received solutions was then diluted into a concentration of 50 mM
4. The solutions obtained from both 2 and 3 were diluted in 10 mM PBS buffer in order to reduce the concentration of DMSO to 5 %, a 1:19 dilution.
5. 100 % DMSO was also diluted into 5 % DMSO using 10 mM PBS buffer

All samples were transmitted onto a 96 well Microtiter plate. The samples were added to the Microtiter plate in a way ensuring that the same sample was injected into all four flow cells. The sensor chip used was Biacore's CM5 chip. After the run, the information wanted was extracted from spot three, where no protein was immobilized on the surface. After each sample 50% DMSO was injected in order to remove sample that might have bound to the surface.

The result from the first run showed a variation in the obtained values that was not consistent with the pattern one would expect only from changes in RII. A conclusion made was that the variation in concentration of DMSO was the reason for this, a variation due to pipeting that was not as exact as needed. A number of solutions containing only DMSO and PBS buffer, no samples, has also been included in the run. These solutions should have given approximately the same signal, but that was not the case. Their values varied from 300 RU to 900 RU. One could also see that some air spikes and that some substances had bound to the surface, this affects the signal but not that much.

After gathering of advised concerning a good pipeting technique a new A100 run was prepared. The run was made on a 96 well Microtiter plate consisting of only blank probes (DMSO + buffer) .The variation was lower, but instead the signal seemed to decrease with time. This was probably due to the fact that all probes first were added to the Microtiter plate and then the plate was enclosed with cover foil.

A new run with blank probes was prepared. This time the cover foil was cut into smaller peaces, making it possible to fill two rows and then directly enclose those. To every third row, the rows where the split between two peaces of cover foil were, only buffer was added. Data from these rows were later disregarded. The result from this run showed a variation due to differences in DMSO concentration of 50 RU. The result was though to be good enough and a new run with the Maybridge subset was prepared using this stepwise enclosing technique.

The result from this run showed better results then the first, but still the bulk variation from the DMSO in the solutions made it impossible to extract RII from the data. DMSO had to be added since all substances are not dissolvable in only the 10 mM PBS buffer. Unfortunately this was an unsuccessful experiment, and there was no time for further investigations on how to extract RIIs from the Biacore A100.