BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 23 2005, pages 4289–4296
doi:10.1093/bioinformatics/bti703

Systems biology

Improved approach for proteochemometrics modeling: application to organic compound—amine G protein-coupled receptor interactions

Maris Lapinsh, Peteris Prusis, Staffan Uhlén and Jarl E. S. Wikberg*

Department of Pharmaceutical Biosciences, Uppsala University, Box 591 BMC, SE-751 24 Uppsala, Sweden

Received on June 23, 2005; revised on September 20, 2005; accepted on October 3, 2005
Advance Access publication October 4, 2005

ABSTRACT

Motivation: Proteochemometrics is a novel technology for the analysis of interactions of series of proteins with series of ligands. We have here customized it for analysis of large datasets and evaluated it for the modeling of the interaction of psychoactive organic amines with all the five known families of amine G protein-coupled receptors (GPCRs).

Results: The model exploited data for the binding of 22 compounds to 31 amine GPCRs, correlating chemical descriptions and cross-descriptions of compounds and receptors to binding affinity using a novel strategy. A highly valid model ($q^2 = 0.76$) was obtained which was further validated by external predictions using data for 10 other entirely independent compounds, yielding the high $q^2\text{ext} = 0.67$. Interpretation of the model reveals molecular interactions that govern psychoactive organic amines overall affinity for amine GPCRs, as well as their selectivity for particular amine GPCRs. The new modeling procedure allows us to obtain fully interpretable proteochemometrics models using essentially unlimited number of ligand and protein descriptors.

Contact: jarl.wikberg@farmbio.uu.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Drug discovery relies essentially on combinatorial chemistry and high throughput screening (HTS). Computations (e.g. docking) using the three-dimensional (3D) structure of the target and quantitative structure–activity relationships (QSAR) are also used. However, neither QSAR nor docking can assure that a drug candidate will interact only with the target and not with other members of the proteome. Deriving high-resolution 3D structures is also often problematic.

Ligands bind often to series of proteins and approaches that focus on the differences in the molecular recognition mechanisms and which are able to predict selective interaction partners are warranted. To this end we recently introduced a bioinformatics approach for drug-design termed proteochemometrics (Prusis *et al*., 2001; Wikberg *et al*., 2003, 2004).

In proteochemometrics one analyses the experimentally determined interaction strength of series of ligands with series of proteins. Proteochemometrics is based on quantitative descriptions derived from structural and physicochemical properties of interacting ligands and proteins, which are correlated to interaction affinity using mathematical modeling. In this way, proteochemometrics models the so-called ligand–receptor interaction space (Wikberg *et al*., 2004).

The first proteochemometric studies modeled peptide interactions with chimeric and wild-type melanocortin G protein-coupled receptors (GPCRs) (Prusis *et al*., 2001, 2002) and organic compound interactions with wild-type and chimeric α_1-adrenergic receptors (Lapinsh *et al*., 2001). More recent studies analyzed the binding of organic compounds to multi-chimeric melanocortin receptors (Lapinsh *et al*., 2005) and the interactions of organic amines to a series of 21 different amine GPCRs (Lapinsh *et al*., 2002b). The latter study represented four out of the five biogenic amine GPCR families, namely, serotonin, dopamine, histamine and adrenergic receptors. However, 10 of the receptors were serotonin receptor subtypes, while only one was a histamine receptor and none was a muscarinic acetylcholine receptor. The dataset was thus unbalanced and it also suffered from a large fraction of missing affinity values. Applying proteochemometrics onto it still yielded a statistically valid model. However, the modeling required a very complex description of the data, which involved more than 12 000 cross-terms and higher order cross-terms, which made it very difficult to comprehend the physical meaning of the model (see Lapinsh *et al*., 2002b for details).

The current study was undertaken to derive a more simple and sturdy proteochemometric modeling approach and apply it to the five families of amine GPCRs. To achieve this the modeling algorithms were altered to make the analysis of large-scale datasets affordable, while improving modeling quality. These improvements made the interpretation of the model straightforward, revealing particular molecular interactions that govern the studied compounds' overall affinity for amine GPCRs, as well as each particular compound's selectivity for each particular amine GPCR.

*To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018

2 METHODS

2.1 Interaction data

Data for 32 organic amine interaction with 31 amine GPCRs were taken from the Psychoactive Drug Screening Program (PDSP) database (http://pdsp.cwru.edu/pdsp.asp) (see Supplementary data for details). The receptors represented five amine GPCR families and included ten serotonin, seven adrenergic, five dopamine, five muscarinic acetylcholine and four histamine receptor subtypes. Most of the organic amines were tricyclic and/or piperidine/piperazine ring containing compounds; the series included approved and candidate drugs (antipsychotics, antidepressants, antiparkinson agents, antihistamines, etc.) as well as some other psychoactive amines.

The affinity values covered a range of more than five logarithmic units. Particular receptor subtypes (e.g. DRD2, ADA1A, 5HT2A and HRH1) showed high average affinity for the compound series, whereas the compounds preferred none of the receptor families as a whole, when compared to any other family.

The large number of observations allowed us to divide the dataset into a work-set comprising 22 compounds that were used for model creation and a prediction set comprising 10 compounds set aside and used after the completion of the proteochemometric model to assess the model's predictive ability. (For further details see Supplementary data).

2.2 Description of organic compounds

Structures of compounds were drawn using ISIS/Draw and converted to 3D by the Corina unit of the Tsar 3.3 (Accelryc Inc., http://www.accelrys.com) software package. Partial atomic charges were derived using the Charge2 unit of Tsar 3.3 and the geometry was optimized by performing energy minimization using the Cosmic utility of Tsar 3.3.

The thus obtained 3D structures were described by grid independent descriptors (GRINDs) (Pastor *et al*., 2000) calculated by Almond 3.1 (Multivariate Infometric Analysis S.r.l., http://miasrl.com) software. GRINDs are alignment independent descriptors that relate to the ability of a compound to form favorable interactions with independent pharmacophoric groups. The generation of these descriptors involves several steps. First, molecular interaction fields (MIFs) are calculated by placing probe groups on grid points surrounding the molecule. Grid nodes that show the energetically most favorable interactions with the molecule and concomitantly are situated as far as possible from each other are then extracted from the MIFs. The distances between each of any two extracted nodes and the products of their energy values are then calculated. Finally, the maxima of products falling within specified distance ranges (smoothing windows) for node pairs of the same MIF (auto-correlograms) and different MIFs (cross-correlograms) are used as descriptors for the molecules.

The Almond software allows the use of up to four MIFs, i.e. it provides four auto-correlograms and six cross-correlograms. The MIFs used herein were obtained using the following probes: DRY (hydrophobic probe), O (sp2 carbonyl oxygen), N1 (neutral flat NH) and Cl (chlorine). Default parameters were selected for the distance between grid points (0.5 Å) and the number of extracted nodes (100 for each MIF). Moreover, the default width was used for the smoothing window (0.8 grid units, i.e. 0.4 Å), resulting in 67 GRINDs in each of 10 correlograms.

We also created four additional sets of descriptors, using in each set three of the four abovementioned MIFs and substituting the fourth by a newly developed molecular shape field (Fontaine *et al*., 2004). Molecular shape was described by using N1 field nodes at a repulsion energy of 1 kcal/mol to outline the surface of the molecule. The local curvature of the surface was then calculated at each node, as described by Fontaine *et al*. (2004). Convex regions were considered to be more important for the shape than concave. This is because the former may form complementary interactions in the receptors' ligand-binding pockets or cause steric hindrances. Here we selected 100 of the most convex nodes (these actually outline the most protruded regions of the molecule and are referred to as TIPs). The TIP–TIP auto-correlogram was generated using curvature–curvature products and cross-correlograms with MIFs were generated using curvature-energy products.

2.3 Description of receptors

Previous 3D modeling and mutagenesis studies indicate that the GPCR ligand-binding pockets for endogenous amines and low molecular weight organic compounds are located in a cavity formed between the receptors' transmembrane regions (Bikker *et al*., 1998; Jacoby *et al*., 1999). We accordingly derived the receptor descriptions from the differences in the physicochemical properties of the seven cell membrane-spanning alpha-helical regions in the receptor series. Receptor amino acid sequences were retrieved from the Swiss-Prot database (http://www.ebi.ac.uk/swissprot) and aligned according to the conserved amino acid positions (Baldwin *et al*., 1997). The amino acids of TM1–TM7 were as follows (using the numbering of the ADA1A human): 25–49, 62–86, 98–122, 145–166, 185–206, 273–292 and 309–328. Of the 159 sequence residues selected 16 were conserved in all receptors. The non-conserved residues were subsequently coded using the three *z*-scale descriptors, *z*1–*z*3, derived by Sandberg *et al*. (1998). Thus, the physicochemical differences in the ligand-binding region of the amine GPCRs were accordingly encoded by a total of $143 * 3 = 429$ descriptors.

2.4 Principal component analysis of compound and receptor descriptors

Prior to further computations descriptors were preprocessed. First, the number of descriptors was reduced by applying principal component analysis (PCA).

PCA is a multivariate projection method that can be used to compress datasets containing large numbers of variables. Contrary to the original variables, the so-termed principal components (PCs) are orthogonal to each other (Wold *et al*., 1987; Eriksson and Johansson, 1996). After calculating *A* PCs, the **X** matrix with size *N* rows (objects) and *K* columns (variables) is decomposed into two smaller matrices, the score matrix **T** of size *N* by *A* and loading matrix **P** of size *K* by *A* according to the following equation:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \cdots + \mathbf{t}_A\mathbf{p}_A' + \mathbf{E}, \qquad (1)$$

where $\mathbf{P}'$ is the transpose of the loading matrix, **t** the score vector, $\mathbf{p}'$ the transpose of the loading vector and **E** the matrix of residuals (unexplained part of the data). The majority of the variation within the original data can often be represented by a small number of components. Extracting *N* – 1 components explains all the variation of the original data. PCA was performed using SIMCA-P 9.0 software (Umetrics AB, http://www.umetrics.com).

Since the descriptors of the ligands are not correlated to the descriptors of the receptors we performed the PCA separately on the *z*-scales of the 31 receptors and on the GRINDs of the 22 work-set compounds (the 10 test-set compounds were not included in the PCA; the PC scores for them were calculated by summing the products of value of each GRIND descriptor for the compound with the loading of the respective descriptor). Prior to PCA all descriptors were mean centered and scaled to unit variance. Moreover, in order to fully preserve interpretability of models all components were extracted. Thus, having at hand 22 organic compounds the variance of GRIND descriptors was compressed into 21 components (GRIND-PCs), while the *z*-scale descriptors of 31 receptors were compressed into 30 components (ZSCALE-PCs).

2.5 Calculation of ligand–receptor cross-terms

Ligand–receptor recognition can evidently only partially be explained by linear combinations of ligand and receptor descriptors. For example if the ligands by virtue of some feature (property) interact with non-varied receptor residues, a simple assumption would be that the binding affinity relates linearly with the intensity of this given property. In reality, however, binding is governed by complex processes that depend on the complementarity of the properties of the interacting entities. In proteochemometrics this may

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018

be accounted for by computation of ligand–receptor cross-terms (see e.g. Lapinsh *et al*., 2005 and references therein). Cross-terms were here formed by multiplying the principal components of descriptors of compounds (GRIND-PCs) and receptors (ZSCALE-PCs). In this way one additional descriptor block was obtained comprising $21 * 30 = 630$ descriptors. The total number of descriptors obtained thus became $21 + 30 + 630 = 681$. In fact, for any proteochemometrics dataset this number would be equal to the number of possible ligand–receptor combinations minus one with the use of the present approach for data preprocessing.

2.6 Block scaling of descriptors

In the PCA step the components became scaled relatively to each other. The first component of each block, which encapsulates the major differences between ligands and receptors, obtained the largest variance. The second component obtained the second largest variance, etc. Any further scaling (e.g. to unit variance) would hide the major patterns in the initial data and exaggerate minor non-systematic variations. Furthermore, when cross-terms are computed from the PCA space each cross-term obtains a variance that is proportional to the product of the variances of its originators. Therefore scaling of the cross-term descriptors reflects the significance of the underlying ligand and receptor properties, and accordingly no re-scaling is required.

However, the use of three descriptor blocks for which the descriptors are not directly comparable prompts the need for block scaling. Accordingly, while the mutual scaling of descriptors within each of the three blocks was frozen, each block was scaled to unit variance (i.e. the sum of variances of all GRIND-PCs, all ZSCALE-PCs and all cross-terms being set to unity). Block scaling was then afforded by systematically changing the variance of the blocks in 0.25 variance unit intervals until an optimal model was obtained.

2.7 Partial least-squares projections to latent structures

Correlation of descriptions to ligand–receptor affinity was performed by partial least-squares projections to latent structures (PLS) (for an in-depth review of the PLS see Geladi and Kowalski, 1986).

PLS derives a regression equation in which the regression coefficients reveal the direction and magnitude of the influence of *x*-variables on the response. For a proteochemometric model, including *L* ligand descriptors (i.e. GRIND-PCs), *R* receptor descriptors (ZSCALE-PCs) and cross-terms thereof, the equation derived for the response variable (i.e. ligand–protein interaction affinity) is expressed as follows:

$$y = \bar{y} + \sum_{l=1}^{L}(\mathrm{coeff}_l * x_l) + \sum_{r=1}^{R}(\mathrm{coeff}_r * x_r) + \sum_{1=1,\, r=1}^{L*R}(\mathrm{coeff}_{l,r} * x_l * x_r). \qquad (2)$$

The goodness-of-fit of the PLS models was characterized by the fraction of explained variation of **Y** (r^2). The predictive ability was characterized by the fraction of the predicted **Y**-variation (q^2), assessed by cross-validation, as previously described (Wold, 1995; Baroni *et al*., 1993). The q^2 computed using five randomly formed groups was used to adjust the variance of descriptor blocks and to determine the optimal number of PLS components.

Along with estimation of the conventional q^2 parameter we introduced several additional estimates to assess a model's predictive ability. Thus, in order to assess its ability to predict the affinity of novel receptors we repeatedly formed cross-validation groups by excluding one-fifth of the receptors and used the models based on the remaining receptors to compute affinities for the excluded ones, yielding q^2rec. Similarly, to assess the capacity of the model to predict the affinity of new ligands we repeatedly formed cross-validation groups by excluding one-fifth of the ligands, yielding q^2lig. Along with these validations we also performed predictions for the 10 compounds that had not been used in the model creation and thus could not have influenced the scaling and complexity of the PLS model. (The predictive ability for these compounds is here termed q^2ext.) We also performed validation by response permutation as described by Eriksson and Johansson (1996). In short, models were re-calculated 100 times for randomly re-ordered *y*-data and q^2 values were plotted as a function of the correlation coefficient between the original *y* and permuted *y*. The intercept of the regression line (i.e. the correlation coefficient being zero) indicates whether or not the original q^2 value could have been obtained by pure chance.

PLS modeling was performed using SIMCA-P 9.0 and Q2 (Multivariate infometric analysis S.r.l., www.miasrl.com) software. (Q2 was used for repeatedly performed cross-validations using randomly formed groups.)

2.8 Contribution of ligand properties for binding affinity and selectivity

The contributions of the *x*-variables were assessed from the PLS regression coefficients. Since the predictor variables are correlated to the *y*-data by means of the PLS regression equation, the regression coefficients reveal the significance of ligand and receptor properties for the interaction affinity. Thus, the regression coefficient of a compound descriptor represents the direction and magnitude that the underlying property influences the affinity for 'an average' amine GPCR. Furthermore, the coefficients for the cross-terms involving this descriptor summarize the importance of the underlying property for the compound's receptor selectivities.

Since the model included principal components rather than the original GRIND descriptors, regression coefficients were multiplied by the loadings of the original descriptor in each principal component, thereby allowing interpretations of the particular ligand properties represented by each GRIND. In this way, the regression coefficient of a GRIND descriptor could be assessed according to the following equation:

$$\mathrm{coeff}_{\mathrm{GRIND}} = \sum_{a=1}^{21}(\mathrm{coeff}_{\mathrm{GRIND\text{-}PC}_a} * p_{\mathrm{GRIND},\,a}), \qquad (3)$$

where $\mathrm{coeff}_{\mathrm{GRIND\text{-}PC}}$ is the regression coefficients for GRIND-PCs, and $p_{\mathrm{GRIND},a}$ the loading of a given GRIND descriptor in principal component *a*. As $\mathrm{coeff}_{\mathrm{GRIND}}$ represents the change in the calculated average affinity of a compound when the GRIND value increases by 1 SD, it will be further referred to as $\Delta y_{\mathrm{GRIND}}$. Moreover, the contribution of a GRIND descriptor to the affinity for a particular receptor R could be assessed according to the equation:

$$\Delta y_{\mathrm{GRIND,\,R}} = \sum_{a=1}^{21}\Bigg(\Bigg(\mathrm{coeff}_{\mathrm{GRIND\text{-}PC}_a} + \sum_{b=1}^{30}(\mathrm{coeff}_{\mathrm{CROSS}_{a,\,b}} * x_{\mathrm{ZSCALE\text{-}PC}_b,R})\Bigg) * p_{\mathrm{GRIND},\,a}\Bigg), \qquad (4)$$

where $\Delta y_{\mathrm{GRIND},R}$ is the change in calculated affinity of a compound for the particular receptor *R* when the GRIND value increases by 1 SD, and $x_{\mathrm{ZSCALE\text{-}PC}_{b,R}}$ the score for receptor *R* in principal component *b*.

3 RESULTS AND DISCUSSION

3.1 Proteochemometrics modeling

Descriptors of ligands, receptors and their cross-terms were correlated to ligand–receptor interaction affinity using PLS. Several models were created, using descriptors of compounds formed from different combinations of TIP and MIFs (Table 1).

As shown in Table 1, all models based on descriptors derived from different combinations of four TIP/MIFs (i.e. models 1–5) were highly predictive, the q^2s being in the range 0.75–0.77. However, the models differed in their ability to afford predictions for new compounds, as assessed by the q^2lig and q^2ext parameters. Model 2 thus showed the lowest q^2lig value, while model 1 showed the highest (0.51 versus 0.61). Model 2, which lacks the N1 field (i.e. the field resulting from the H-bond donor probe) also showed a lower

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018

Table 1. Results of PLS modeling using GRINDs from different combinations of MIFs

No.	MIFs used	r^2	q^2	q^2lig	q^2ext
1	DRY, N1, Cl, TIP	0.87	0.75	0.61	0.63
2	DRY, O, Cl, TIP	0.90	0.76	0.51	0.57
3	DRY, O, N1, TIP	0.90	0.76	0.55	0.66
4	DRY, O, N1, Cl	0.89	0.76	0.51	0.62
5	O, N1, Cl, TIP	0.90	0.77	0.52	0.65
6	DRY, N1, TIP	0.88	0.76	0.59	0.67
7	DRY, N1	0.86	0.71	0.55	0.61
8	DRY, TIP	0.88	0.73	0.59	0.58
9	N1, TIP	0.88	0.74	0.56	0.40

q^2ext value (0.57) compared with the four other models (0.62–0.67). These results thus indicate that electron-attracting properties (i.e. the location and strength of electronegative atoms and groups in the compounds) are important for the receptor recognition.

By contrast, the q^2lig values were only 0.51–0.55 for the models including the O field (i.e. H-bond acceptor probe field) compared with 0.61 for model 1 which did not use this MIF. These results indicate that the differences in H-bond donating properties of the compounds, such as presence and location of hydroxy groups, yield mainly only chance correlations to the receptor affinity. (In interpreting this, one should keep in mind that all compounds contain an amine group that presumably interact with the conserved aspartic acid residue in TM3 of amine GPCRs. However, the aspartic acid residues and amine groups are invariant in the dataset. Proteochemometrics modeling can therefore not assess their importance.)

Inclusion or omission of the Cl field GRINDs only marginally influenced the predictive ability of the model (c.f. models 1 and 6). By contrast, removing any of the TIP, DRY or N1 fields (models 7–9) significantly reduced the q^2lig and q^2ext values. These latter three fields characterize the shape of the molecule, the strength and location of hydrophobic and H-bond donor moieties and seem to form a minimum set of MIFs required for modeling amine GPCR interactions. The analysis thus showed that model 6 was the best. Accordingly model 6 was used in all subsequent analysis, unless otherwise stated. (In the following sections this model will be referred to as 'the model').

3.2 Assessment of predictive ability of the model

The predictive ability of a proteochemometric model can be evaluated in different ways. The conventional q^2 parameter characterizes the predictions of combinations of ligands and receptors already present in the dataset, but tested in other combinations. However, one of the purposes of our study was to assess the capacity of proteochemometrics to afford predictions for novel yet pharmacologically uncharacterized organic compounds. Therefore, along with the conventional q^2 parameter we estimated q^2lig and q^2ext, which were found to be 0.59 and 0.67 for the model, respectively (Table 1).

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 SD 0.02. Thus, a whole half of all data points could be omitted without endangering model validity. Finally, the predictive ability of the model for new receptors was also assessed; the q^2rec being 0.62 (thus revealing the potential use of proteochemometrics in finding ligands for yet biologically/pharmacologically uncharacterized GPCRs). Moreover, the model was also validated by response permutations. The negative q^2 intercepts (–0.34 using five cross-validation groups and –0.48 using two groups) obtained from this analysis show that randomized data produce non-predictive models.

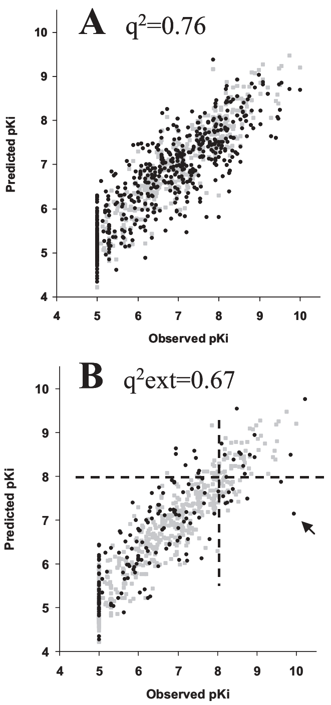

Fig. 1. Relation of predicted versus observed pK_i values derived from the PLS model. (**A**) shows predictions from conventional cross-validation using five random groups (black symbols). (**B**) shows results from external validation, i.e. black symbols are the predictions of the 10 compounds that had not been used during model creation (for details see text). Goodness-of-fit of the model (gray symbols, calculated versus observed pK_i values) is shown in both (A) and (B).

Results of validations are graphically shown in Figure 1. Results for cross-validation with five groups are shown in Figure 1A and Figure 1B shows results for external predictions. As can be seen from Fig. 1A, the predictive ability for compound–receptor combinations is very good, the average prediction error being <0.5 pK_i units. As seen from Fig. 1B, for only one test-set compound–receptor interaction a misprediction by >2 pK_i units occurred, while for the remaining observations the average prediction error was 0.55. The misprediction >2 pK_i units occurred for roxindole for the 5HT1A receptor. In fact this misprediction can be explained by the absence of any compounds with high affinity for the 5HT1A receptor in the work-set. Despite this the model still correctly predicted that roxindole has the highest affinity for the 5HT1A receptor among all test-set compounds, thus showing that an experimentalist would get proper guidance on the direction of the affinity also in this case.

Another way of viewing the predictive ability of the model is to set an arbitrary cutoff limit, such as a $pK_i > 8$, which in a real setting might be a selection criteria for a candidate compound. As is shown

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018

in Figure 1B, on 20 occasions a $pK_i > 8$ is predicted for test-set compound–receptor combinations. Of these, 13 combinations have indeed an experimentally determined $pK_i > 8$, while for the remaining 7 the measured pK_i value was 6.95 or higher. Furthermore, the model predicts high affinity for a number of ligand–receptor interactions for which measured data are not available results. These indicate the potential use of the model in screening of compound databases for high affinity binders to particular amine GPCRs.

3.3 Interpretation of the model

Contribution of ligand properties for binding affinity and selectivity was assessed by estimating the contributions of GRINDs according to the $\Delta y_{\mathrm{GRIND}}$ and $\Delta y_{\mathrm{GRIND},R}$ measures. A $\Delta y_{\mathrm{GRIND}}$ is a regression coefficient of a GRIND descriptor and shows the descriptor's influence on the compounds' overall (average) affinity for amine GPCRs. A $\Delta y_{\mathrm{GRIND},R}$ value can be considered as a regression coefficient of a GRIND for a particular receptor. Comparing the latter for different receptors thus allows one to assess the influence of particular GRINDs for a compound's selectivity for any receptor pair. In Figure 2A are plotted $\Delta y_{\mathrm{GRIND}}$ values for DRY–DRY, N1–N1, TIP–TIP auto-correlograms, and DRY–N1, DRY–TIP and N1–TIP cross-correlograms. The GRINDs are for each correlogram arranged in order of increasing distance between node pairs, the vertical separators representing a distance range from 0 to 26.8 Å. Similarly, in Figure 2B–F are given examples for $\Delta y_{\mathrm{GRIND},R}$ for particular receptors.

Inspecting Figure 2 reveals that DRY–DRY correlogram descriptors (representing the ability of a molecule to form hydrophobic interactions) are the most important for all receptors. On one hand the positive $\Delta y_{\mathrm{GRIND}}$ and $\Delta y_{\mathrm{GRIND},R}$ values for the DRY–DRY descriptors at node distances from 0 up to 4 Å indicate that the presence of a hydrophobic group is in general of high importance for ligand binding to the amine GPCRs. Moreover, the presence of an additional DRY field at a distance 6–8 Å from the first one gives a further positive contribution to the binding. On the other hand, $\Delta y_{\mathrm{GRIND}}$ values for DRY–DRY descriptors at distances between 8 and 20 Å are close to zero, showing that distantly located hydrophobic groups have only minor impact on the average affinity of the compounds for the amine receptors. However, such interactions may still be important for the selectivity of the compounds for particular receptors. For example, for the DRD2 receptor they yield a positive influence on the affinity (Fig. 2C).

Positive values are also given to $\Delta y_{\mathrm{GRIND}}$s of DRY–N1 cross-correlogram descriptors; the optimal distance between hydrophobic and H-bond acceptor MIFs being 6–10 Å. Moreover, inspection of Figure 2B–F reveals that DRY–N1 descriptors have highly positive $\Delta y_{\mathrm{GRIND},R}$ values for 5HT2A and DRD2 receptors (at distances up to 14 Å), whereas these descriptors contribute only marginally for other receptors. Thus, the mutual location of hydrophobic and H-bond acceptor moieties not only determines the average affinity of compounds for amine GPCRs but also is important for receptor subtype selectivity.

By contrast, most N1–N1 descriptors show slightly negative $\Delta y_{\mathrm{GRIND}}$ values. Thus, the mutual location of several H-bond acceptor groups appears to have low contribution to the ligands' average affinity. However, as revealed in Figure 2B–F N1–N1 descriptors show very negative $\Delta y_{\mathrm{GRIND},R}$ values for particular receptors at several distance ranges, such as for the 5HT2A and ACM1. Negative $\Delta y_{\mathrm{GRIND}}$ and $\Delta y_{\mathrm{GRIND},R}$ values are also given to N1–TIP cross-correlogram descriptors at distances from 4 to 16 Å (an exception is the ACM1 receptor), whereas at larger distances $\Delta y_{\mathrm{GRIND},R}$ values for 5HT2A and DRD2 obtain positive values. Negative values are also assigned to short distances (up to 4 Å) of the DRY–TIP cross-correlogram descriptors, suggesting that very protruded hydrophobic moieties do not contribute favorably to the binding of ligands to the amine GPCRs. Comparisons of all three correlograms including the TIP field reveal the importance of the overall shape of a molecule for receptor selectivity. Thus, interactions over large distance ranges yield positive coefficients for 5HT2A and DRD2 (and somewhat lower for ADA1A) but not for HRH1 and ACM1.

In a further analysis we linked the patterns of Figure 2 to the MIFs of particular compounds showing HRH1 or DRD2 selectivity.

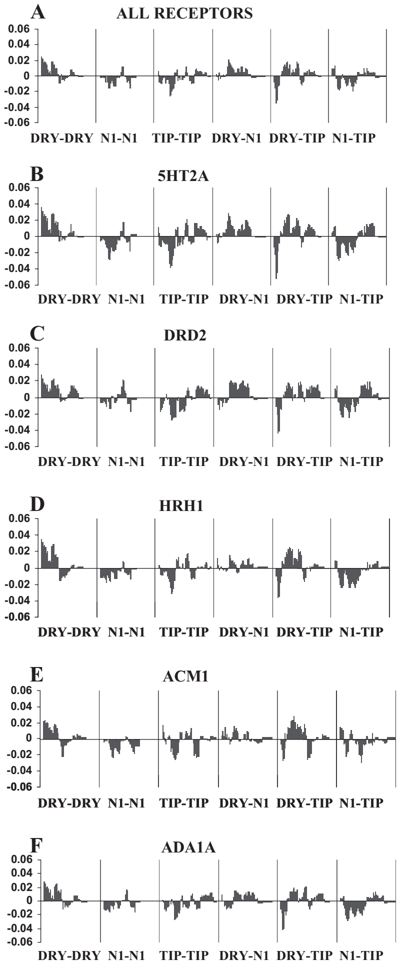

Fig. 2. Contribution of GRIND descriptors for explaining the binding affinity of organic compounds for amine GPCRs. (**A**) shows the PLS regression coefficients of GRINDs computed according to Equation (3). (**B**–**F**) show regression coefficients of GRINDs for 5HT2A, DRD2, HRH1, ACM1 and ADA1A receptors, respectively, computed according to Equation (4). Increments on the *Y*-axes indicate the change of affinity in pK_i units when a GRIND value is increased by 1 SD. The interval between the vertical separators represents the distance range 0–26.8 Å for each particular GRIND (see text for further details).

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018

We selected these two receptors as a demonstration case because their affinity profiles are distinct. In particular, clozapine, chlorpromazine, olanzapine and several other compounds show significantly higher affinity for HRH1, while haloperidol, aripiprazole, risperidone, fluphenazine and some other compounds prefer DRD2. Inspection of the $\Delta y_{\mathrm{GRIND},R}$ values for these two receptors (Fig. 2C and D) reveals several patterns. Firstly, the DRY–DRY correlograms reveal that hydrophobic interactions influence ligand affinity for DRD2 and HH1R differently. For distances between DRY nodes of up to 8 Å the $\Delta y_{\mathrm{GRIND,HH1}R}$s show larger positive values than the $\Delta y_{\mathrm{GRIND,DRD2}}$s, while at distances >8 Å the $\Delta y_{\mathrm{GRIND,\,HRH1}}$s, but not the $\Delta y_{\mathrm{GRIND,\ DRD2}}$s, are negative. The presence of one, or two closely located, strong hydrophobic groups is thus needed to render a compound HRH1 selective, while several distantly located hydrophobic moieties are needed to create a DRD2 selective one. Secondly, the TIP–TIP correlograms reveal that the overall shape of the molecule is important for selectivity. Thirdly, it can be seen from the DRY–N1 and N1–TIP correlograms that a strong N1 field situated at certain distances from the DRY and TIP nodes may improve the affinity for the DRD2 with-out affecting the affinity for the HRH1.

The patterns of the foregoing paragraph are further visualized in Figure 3 by showing the MIFs around some HRH1/DRD2 and DRD2/HRH1 selective compounds. Compounds are there arranged in the order of their relative preference for the two receptors, with the MIFs represented as follows: DRY in beige, N1 in red and TIP in green. Inspections of Figure 3 reveal that distantly located DRY fields are present only for the two most DR2D selective compounds, namely haloperidol and risperidone. For haloperidol these fields, which appertain to the chlorophenyl and fluorophenyl moieties of the compound, are much weaker than the fields for the HRH1 selective compounds (clozapine, olanzapine and chlorpromazine). Moreover, also for the two remaining DRD2 selective compounds (fluphenazine and risperidone), the DRY field descriptors computed at distances from 3 to 6 Å show lower values than for any of the three HRH1 selective compounds.

Comparisons of structures also reveal systematic changes in the overall shape of the molecules, which are altered from rounded for the more HRH1 selective ones to elongated for the DRD2 selective ones.

Finally, inspection of the relative location of the N1 and TIP fields shows that only for the DRD2 selective compounds the N1-TIP nodes exist with locations of >16 Å from each other. Over shorter distances the values for the GRINDs overlap for both the HRH1 and DRD2 selective compounds.

4 DISCUSSION

In proteochemometrics the strength of ligand–protein interactions is correlated to chemical descriptions and cross-description (‘cross-terms’) of the interacting moieties. Cross-terms would not be needed if the ligands interacted with the invariant parts of the proteins only (e.g. with the 16 entirely conserved amino acids of the amine GPCRs transmembrane regions). However, differences in the binding affinity profiles of the ligand series arise since recognition is governed by complementary properties of receptors and ligands. Supplementing the descriptions by cross-terms is then used to reveal how ligand and receptor property combinations affect the interaction strengths.

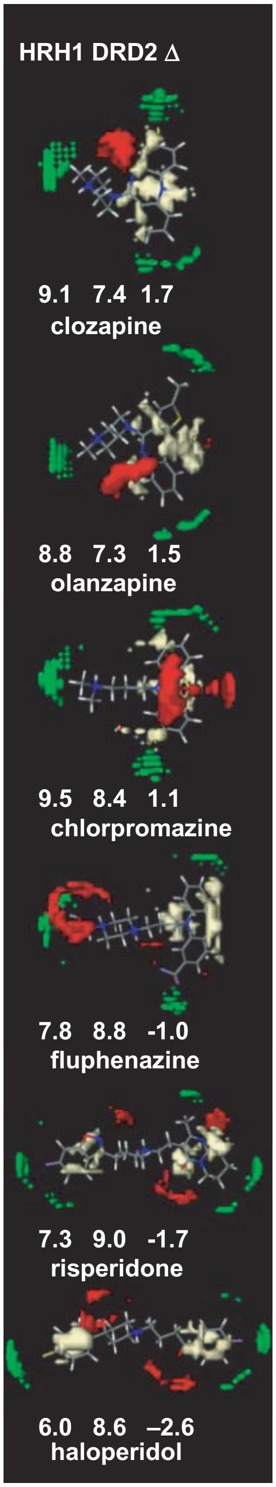

Fig. 3. Graphical representation of MIFs for some HRH1/DRD2 and some DRD2/HRH1 selective compounds. Shown are (from top to bottom) clozapine, olanzapine, chlorpromazine, fluphenazine, risperidone and haloperidol. The DRY MIFs are represented by beige, N1 by red and TIP by green. Indicated are also the $\mathrm{pK_i}$ values for each respective structure’s binding to the HRH1 and DRD2 receptors calculated from the proteochemometric model. The computed HRH1/DRD2 selectivity ($\Delta = \mathrm{pK_{iHRH1}} - \mathrm{pK_{iDRD2}}$) of the compounds is also indicated.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018

Cross-terms may be obtained by multiplication of each ligand descriptor with each receptor descriptor. Correlation by PLS allows one to use a multitude of descriptors while tolerating their mutual colinearity. However, computing cross-terms directly from the descriptors of the current dataset would have resulted in almost 300 000 new variables, which would make further analysis highly resource consuming and, in fact, to our knowledge impossible with currently available academic or commercial software. We elected therefore to use PCA preprocessing, which allows one to keep the number of variables lower than the number of objects in a dataset, without compromising the information content of it.

We here applied PCA separately to ligand and receptor descriptors, which was followed by computations of cross-terms between ligand and receptor PCs. In fact, preliminary modeling using smaller subsets of descriptors revealed that the scores and calculated/predicted *Y* values of PLS models resulting from this type of preprocessing were identical to those of the models based on original descriptors and cross-terms thereof. However, differences would appear if one scaled the PCs relatively to each other, scaled cross-terms relatively to each other or used higher order cross-terms. Such procedures would give a risk for chance-correlations and subsequent deteriorations of models and were therefore avoided.

In earlier proteochemometrics studies the astronomic number of cross-terms was avoided by using simple descriptors (e.g. a limited number of binary descriptors), which is possible in simplified cases (i.e. for simple datasets). However, conventional QSAR has proven that the use of numerous descriptors is required for a thorough representation of the structures/properties of organic compounds. This is mandatory when using, e.g. GRID, CoMFA and GRIND. A proper description of complex biological macromolecules is obviously not an easier task and would require numerous structural descriptors or a large number of descriptors derived from sequence monomers (Kastenholz *et al*., 2000; Lapinsh *et al*., 2002a).

In our present study we aimed at representing compounds and receptors with descriptors that relate to major determinants for receptor–ligand interactions. The structures of organic compounds were characterized by GRIND descriptors. These describe the ability of compounds to interact with different MIF probes mutually located at varying distance ranges. An advantage of using GRIND descriptors is that they do not require the molecules to be spatially aligned with each other when creating a data matrix from series of compounds. In order to overcome some earlier shortages of the GRIND descriptors a recently developed molecular shape field (TIP) was used along with the MIFs (see Fontaine *et al*., 2004). The usefulness of the new molecular shape descriptors was indeed confirmed since the TIP–DRY and N1–TIP correlograms were among the most important for explaining ligand–receptor interactions. However, creation of several models based on different combinations of MIFs showed that not all MIFs were relevant, thus allowing us to find molecular interactions of importance for organic amine GPCR interaction affinity.

3D structures of compounds were created by the Corina unit and the geometry was optimized by the cosmic utility of the Tsar 3.3 software package. The conversion and optimization process was very fast, taking only a few seconds per molecule. This contrasted to our previous approach (Lapinsh *et al*., 2002b) where large conformational ensembles of each molecule were obtained by a time-consuming simulated annealing procedure. Accordingly the current approach could potentially be used to apply on large combinatorial libraries or to screen large compound databases. Moreover, Corina seems to provide some advantage since in our previous study the hydrophobic moieties tended to bunch together in the 3D structures for flexible molecules, while Corina creates extended low energy conformations that are close to the X-ray determined structures (c.f. also Sadowski and Gasteiger, 1994). In fact, for the present dataset simulated annealing produced inferior models compared with Corina generated structures (Lapinsh, M. and Wikberg, J.E.S., unpublished data). However, further studies on the dependence of proteochemometrics modeling to the approach for 3D modeling of compounds using broader datasets are warranted, to allow generalizations and to pinpoint the best method to be used for particular datasets.

The GPCRs in the present study were encoded by three *z*-scales of the amino acids of the transmembrane regions of the receptors. These *z*-scales represent the major differences in physicochemical properties of amino acids and would be the ones that primarily determine the ligand interactions with the receptors. Thus, overall using the present approaches for ligand and protein descriptions our technology affords predictive models without the need for ligand docking and accurate protein 3D structures.

The present dataset included 31 receptor subtypes representing five amine GPCR families. An advantage of the present data matrix was that it had quite few missing values. Otherwise affinities for weak binders are often omitted in scientific reports. For sake of mathematical modeling ‘positive’ and ‘negative’ interaction data are equally important. Thus, also for this reason the present model was improved compared with the earlier model where the interaction matrix contained about 30 percent values missing in a systematic fashion (Lapinsh *et al*., 2002b).

The validity of our model was assessed not only by the conventional q^2 estimate, generally used in QSAR, but also by predicting affinity for ligands and receptors entirely excluded from the model. The high values of the q^2lig, q^2rec and q^2ext estimates obtained herein indicate clearly the validity of our present modeling approach.

A clear advantage of the present modeling approach compared with the earlier one is that it gives fully interpretable models. In the previous study only a rough summary of the importance of different types of MIFs was possible to obtain (Lapinsh *et al*., 2002b), while the current approach allows unambiguous assessment of the importance of each descriptor for the interaction affinities. We here used this feature to assess the contribution of each GRIND for the compounds’ overall affinity for the amine GPCRs, as well as for receptor subtype selectivity. Such analysis of a proteochemometric model may provide an experimentalist with suggestions how to modify a compound chemically in order to improve its selectivity and to afford a new compound with a desirable affinity profile.

In conclusion we have here shown how proteochemometrics can be adapted for the analysis of large-scale datasets yielding models which are straightforward to interpret in a chemical sense.

ACKNOWLEDGEMENTS

This work was financially supported by the Swedish VR (621-2002-4711).

Conflict of Interest: Wikberg declares that he holds stocks in Genetta Soft, a vendor of bioinformatics software.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018

REFERENCES

Baldwin,J.M. *et al*. (1997) An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol*., **272**, 144–164.

Baroni,M. *et al*. (1993) Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat*., **12**, 9–20.

Bikker,J.A. *et al*. (1998) G-Protein coupled receptors: models, mutagenesis, and drug design. *J. Med. Chem*., **41**, 2911–2927.

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

Fontaine,F. *et al*. (2004) Incorporating molecular shape into the alignment-free GRid-INdependent descriptors. *J. Med. Chem*., **47**, 2805–2825.

Geladi,P. and Kowalski,B.R. (1986) Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1–17.

Jacoby,E. *et al*. (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.

Kastenholz,M.A. *et al*. (2000) GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem*., **43**, 3033–3044.

Lapinsh,M. *et al*. (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta*, **1525**, 180–190.

Lapinsh,M. *et al*. (2002a) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci*., **11**, 795–805.

Lapinsh,M. *et al*. (2002b) Proteo-chemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharm*., **61**, 1465–1475.

Lapinsh,M. *et al*. (2005) Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes. *Mol. Pharm*., **67**, 50–59.

Pastor,M. *et al*. (2000) GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem*., **43**, 3233–3243.

Prusis,P. *et al*. (2001) PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand–receptor interactions. *Biochim. Biophys. Acta*., **1544**, 350–357.

Prusis,P. *et al*. (2002) Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. *Protein Eng*., **15**, 305–311.

Sadowski,J. and Gasteiger,J. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci*., **34**, 1000–1008.

Sandberg,M. *et al*. (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.

Wikberg,J.E.S. *et al*. (2003) Melanocortin receptors: ligands and proteochemometrics modeling. *Ann. N. Y. Acad. Sci*., **994**, 21–26.

Wikberg,J.E.S., Lapinsh,M. and Prusis,P. (2004) Proteochemometrics: a tool for modeling the molecular interaction space. In Kubinyi,H. and Müller,G. (eds), *Chemogenomics in Drug Discovery—A Medicinal Chemistry Perspective*. Wiley-VCH, Weinheim, pp. 289–309.

Wold,S. *et al*. (1987) Principal component analysis. *Chemom. Intell. Lab*., **2**, 37–52.

Wold,S. (1995) PLS for multivariate linear modeling. In van de Waterbeemd,H. (ed.), *Chemometric Methods in Molecular Design*. VCH, Weinheim, Germany, pp. 195–218.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/23/4289/195272
by guest
on 29 July 2018