Déjà Vu - Réjà Vu : on knowledge-based approaches linking ligand and target information to bioactivity

Westen, G.J.P. van

List of publications

G.J.P. Van Westen, A. Hendriks, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Personalized HIV Treatment Regimen Prediction Employing Proteochemometric Models Generated From Antivirogram Data.* **Submitted**.

G.J.P. Van Westen, R.F. Swier, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Comparative Study and Benchmarking of 13 Amino Acids Descriptors and Applications to Proteochemometric Modeling.* **Submitted**.

G.J.P. Van Westen, O.O. van den Hoven, R. van der Pijl, T. Mulder-Krieger, H. de Vries, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data.* J. Med. Chem. *2012*. **55** (16): 7010-7020.

J.R. Lane, C. Klein Herenbrink, G.J.P. Van Westen, J.A. Spoorendonk, C. Hoffmann, and A.P. IJzerman; *A Novel Nonribose Agonist, LUF5834, Engages Residues That Are Distinct from Those of Adenosine-Like Ligands to Activate the Adenosine A2a Receptor.* Mol. Pharmacol.; *2012*. **81** (3): 475-487.

M.C. Peeters, Q. Li, G.J.P. Van Westen, and A.P. IJzerman; *Three "hotspots" important for adenosine A2B receptor activation: a mutational analysis of transmembrane domains 4 and 5 and the second extracellular loop.* Purinergic Signalling; *2012* **8** (1): 23-38.

G.J.P. Van Westen, J.K. Wegner, P. Geluykens, L. Kwanten, I. Vereycken, A. Peeters, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS One; *2011*. **6** (11): e27518.

G.J.P. Van Westen, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets.* Med. Chem. Commun.; *2011*. **2** (1): 16-30.

M.C. Peeters, G.J.P. Van Westen, Q. Li, and A.P. IJzerman; *Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation.* Trends Pharmacol. Sci.; *2011*. **32** (1): 35-42.

M.C. Peeters, G.J.P. Van Westen, D. Guo, L.E. Wisse, C.E. Müller, M.W. Beukers, and A.P. IJzerman; *GPCR structure and activation: an essential role for the first extracellular loop in activating the adenosine A2B receptor.* The FASEB Journal; *2011*. **25** (2): 632-643.

E. Van der Horst, J.E. Peironcely, G.J.P. Van Westen, O.O. Van den Hoven, W.R.J.D. Galloway, D.R. Spring, J.K. Wegner, H.W.T. Van Vlijmen, A.P. IJzerman, J.P. Overington, and A. Bender; *Chemogenomics Approaches for Receptor Deorphanization and Extensions of the Chemogenomics Concept to Phenotypic Space.* Curr. Top. Med. Chem.; *2011*. **11** (15): 1964-1977.

G.J.P. van Westen, J.K. Wegner, A. Bender, A.P. IJzerman, and H.W.T. van Vlijmen; *Mining protein dynamics from sets of crystal structures using "consensus structures".* Protein Sci.; *2010*. **19** (4): 742-752.

M.R. Doddareddy, G.J.P. van Westen, E. van der Horst, J.E. Peironcely, F. Corthals, A.P. IJzerman, M. Emmerich, J.L. Jenkins, and A. Bender; *Chemogenomics: Looking at biology through the lens of chemistry.* Statistical Analysis and Data Mining; *2009*. **2** (3): 149-160.

Afterword

When pursuing a PhD for four years it is near impossible to describe everything and everyone that made an impact on you on a mere one or two pages. However, there is a large group of people who should be mentioned here as their help, support and co-operation was invaluable for the success of my PhD project.

First and foremost I would like to thank my two promoters Ad and Herman. Without both your vision and scientific input this would all have been a short exercise. Herman, it is incredible what effects a simple email, sent more than six years ago can have. Without the internship at Tibotec I am unsure if I would have continued to pursue a career in science. I think we can safely say that the science described in this thesis is a direct result from the way you supervised me during that time (remember our bi-weekly meetings). Ad I would like to thank you for the invaluable role you played in turning me from a cocky student into a proper scientist. Your patience, wisdom and in particular your skill at communicating one's results to others were something that will benefit me for the rest of my life. While we had clashes at times, I remember with pleasure the discussions we had and plans we made in both your office and mine. Working on a PhD project with two promoters like yourselves has been a very pleasant experience. This brings me to my co-promoter Andreas, whom has been another very significant influence on me these last years. It has been five years since we first met at the Gorlaeus restaurant, I remember our meeting as very energetic and focusing on opportunities and ideas. As a daily supervisor you showed the same qualities. Our day to day contact was always productive and our social outings were pleasant. Where I learned how to do science working with Ad and Herman, you taught me how to be a scientist.

Furthermore, I would also like to thank my MSc students, Olaf, Bart, Remco, Alwin and Marysa, and BSc student Tanja, for showing me the value of teaching and the crash course 'management in crisis situations' on a weekly basis. Your contributions have been invaluable and it is fun to see the direction each of you chose to go in after your biopharmaceutical sciences master. Also I should like to thank all the people from Medicinal Chemistry Leiden. Working at this department always felt very comfortable, but it was only at (inter)national conferences that I fully realized how fortunate I was to have worked in a group such as MedChem. The direct (and regular!) contact between experts with a diverse background (informatics, chemistry and bioassays) has been enriching. In particular I would like to thank Hans, Maris, Jaco and Laura for their expertise and help and Thea, Henk and Rianne for their massive contribution in performing all experimental work. Without you guys this thesis would also not have been completed.

I also would like to thank the members of my fraternity and in particular my co-founders Martin, Eric, Rob, Pouce, Caspar, Michel, Maarten, Gunn, GJ, Thijs en Jan (sorry Pouce I just cannot get myself to put Tom here). You guys showed me that indeed you can do anything you want; it's just a simple matter of perseverance. Furthermore, of all social circles I know you are one of the few where anybody can really be himself. In particular I would like to mention Michel. I have spent a total of 34 weeks receiving intravenous antibiotics these last years and you have joined me on nearly all of these trips to Amsterdam. Thank you for that, while you tend to systematically depreciate these actions (and any action that shows caring) it meant a lot to me. Thanks also for my friends with whom I enjoyed a 'cartoon avond' every week, Michael, Michel, Bastiaan and Dirk. These evenings were a very pleasant distraction from the life of a PhD student and a good place to brag and let off steam. Perhaps we can continue these events upon my return to the Netherlands. Special thanks go to Juriaan Bakx for almost 8 years of friendship and for showing me that you can actually do more valuable things than just playing videogames if you have a knack for computers. Our interests are overlapping to a large degree (ASOT) but also you were 'my guy on the inside' in the evil world of semi-corporate ICT at Leiden University. Another group of people that should be mentioned here are the 'Hufters'. The semi-regular gathering we enjoyed was always a good chance to speak to others 'in the trenches of PhD life'. I do hope we can keep the yearly outing alive as it does not get old (rather we do though).

Finally, I would like to thank both my parents for always supporting me and for, even though they disagreed on a course of action or had previous experience, allowing me to make my own mistakes to learn from. I would like to thank my mother for unconditional support but also for giving me her true opinion in every case. It is only now that I have just become a parent myself that I can truly value the way you have always been there. I can now fully understand why you asked the questions you did to my 3rd grade teacher. I would also like to thank my father, the role of a father can be difficult for some, but having had a very good example I must say this task befalls me easier than I feared beforehand. Also it was you who taught me the most valuable lesson of all when presenting something 'the audience doesn't know what you originally planned to say and hence will not judge you for that what you forgot to mention, just on what you actually presented'

Finally, I should like to mention those people who are most important and dear to me Afke, Max and Iris. Thank you for standing by me during everything that has happened over the last years. These last years have by no means been easy, not the PhD but my health was the most difficult hurdle. Thank you for understanding, supporting but mostly for the happy times we have together.

Curriculum Vitae

Gerard van Westen was born on March 28th 1983 in Leiden. He went to high school at the Stedelijk Gymnasium Leiden and graduated in 2001. In that same year he started his education at the Leiden/ Amsterdam Center for Drug Research (LACDR) in the undergraduate biopharmaceutical sciences. His first master internship was at the department of biopharmaceutics of the LACDR under supervision of Dr. Lutters and Prof. Dr. Biessen. Here he characterized the inhibition of P-Selectine using a peptide ligand. His second intership was at Tibotec BVBA (now Janssen pharmaceutica) in Mechelen (BE). Under supervision of Prof. Dr. Van Vlijmen and Dr. Wegner he started on a cheminfomatics and computational drug design project. During this internship Prof. IJzerman was his tutor at the LACDR.

After obtaining his master in Septermber 2007 he started as a PhD student in November of that year at the department of Medicinal Chemistry at the LACDR. In the period between November 2007 and March 2012 he performed the work described in this thesis. His PhD project was a cooperation between the LACDR and Tibotec, was funded by Tibotec, and his supervisors were Prof. Dr. IJzerman, Prof. Dr. Van Vlijmen and Dr. Bender.

During his PhD research he presented his work on multiple (inter)national conferences. He was invited as a speaker multiple times among others to the Dutch FIGON days (Lunteren, 2011) and to the Molsoft usergroup meeting (San Diego, 2012). At these conferences he was awarded multiple times for presentations and poster presentations.

He currently works as a postdoc at the European Bioinformatics Institute (part of the EMBL) in the ChEMBL group, headed by Dr. Overington, in Cambridge (UK). Gerard is married and has 2 children.

Curriculum Vitae

Gerard van Westen werd geboren op 28 maart 1983 te Leiden en groeide daar ook grotendeels op. Zijn middelbare schoolopleiding volgde hij aan het Stedelijk Gymnasium te Leiden, alwaar hij in 2001 zijn eindexamen afrondde. Vervolgens begon hij in datzelfde jaar aan de WO opleiding biofarmaceutische wetenschappen aan het Leiden/Amsterdam Center for Drug Research (LACDR). Zijn propedeuse behaalde hij in 2003 en zijn eindonderzoek startte hij in 2005 met een stage aan de afdeling Biofarmacie. Hier legde hij zich onder begeleiding van Dr. Lutters en Prof. Dr. Biessen toe op het karakteriseren van de inhibitie van P-Selectine door middel van een peptide ligand. In 2006 begon hij zijn tweede stage, welke plaats vond onder begeleiding van Prof. Dr. Van Vlijmen en Dr. Wegner bij het toenmalige Tibotec BVBA in Mechelen (BE, huidig Janssen Pharmaceutica) waarbij Prof. IJzerman zijn tutor aan het LACDR was. Deze stage richtte zich meer op cheminfomatics en computational drug design.

Na het behalen van zijn doctoraal examen in september 2007 begon hij in november van dat jaar als promovendus aan de afdeling Medicinal Chemistry van het LACDR. Tot maart 2012 verrrichtte hij het onderzoek wat in dit proefschrift beschreven staat onder begeleiding van Prof. Dr. IJzerman, Prof. Dr. Van Vlijmen en Dr. Bender. Zijn promotieproject was een samenwerking tussen het LACDR en Tibotec en werd gefinancierd door Tibotec.

Gedurende zijn promotie heeft hij op meerdere (inter)nationale en congressen zijn werk gepresenteerd. Hij is meermaals gevraagd als spreker onder andere op de FIGON dagen (2011) en op de usergroup meeting van Molsoft in San Diego (2012). Daarbij heeft hij meerdere prijzen gewonnen voor zowel presentaties en poster presentaties.

Vanaf mei 2012 werkt hij als postdoc aan het European Bioinformatics Institue (onderdeel van het EMBL) in de ChEBML groep, geleid door Prof. Dr. Overington, in Cambridge (VK). Gerard is getrouwd en heeft twee kinderen.

Appendix

Abbreviations

1D	–	One dimensional
2D	–	Two dimensional
3D	–	Three dimensional
AA	–	Amino Acid
ACE	–	Angiotensin-Converting Enzyme
AIDS	–	Acquired Immuno Deficiency Syndrome
AVG	–	Antivirogram
CCO	–	Clinical cut-off
CCP	–	Correctly Classified Percentage
CHO	–	Chinese Hamster Ovary
CPU	–	Central Processing Unit
CV	–	Cross Validation
DLV	–	Delavirdine
DT	–	Decision Tree
EL	–	Extracellular Loop
FASGAI	–	Factor Analysis Scales of Generalized Amino acid Information
FC	–	Fold Change
FN	–	False Negative
FP	–	False Positive
GP	–	Gaussian Processes
GPCR	–	G Protein-Coupled Receptor
GRIND	–	Grid Independent Descriptors
HAART	–	Highly Active Anti-Retroviral Therapy
HIV	–	Human Immunodeficiency Virus
LE	–	Ligand Efficiency
kNN	–	k-Nearest Neighbor
LOO	–	Leave-One Out
LOSO	–	Leave-One-Sequence Out
MCC	–	Matthews Correlation Coefficient
MHC	–	Major Histocompatibility Complex
MIPS	–	Million Instructions Per Second
NPV	–	Negative predictive value

NB	–	Naïve Bayesian
NN	–	Neural Net
NNRTI	–	Non-Nucleoside Reverse Transcriptase Inhibitor
NRTI	–	Nucleoside Reverse Transcriptase Inhibitor
NtRTI	–	Nucleotide Reverse Transcriptase Inhibitor
PCA	–	Principal Component Analysis
PCM	–	Proteochemometric
PC	–	Principal Component
PDB	–	Protein Data Bank
PI	–	Protease Inhibitor
PLFP	–	Protein-Ligand Fingerprint
PLS	–	Partial Least Squares
PPV	–	Positive predictive value
ProtFP	–	Protein Fingerprint
QSAR	–	Quantitative Structure-Activity Relationship
QSAM	–	Quantitative Sequence-Activity Modeling
RF	–	Random Forest
RMSE	–	Root Mean Squared Error
ROC	–	Receiver / Operator Characteristic
RS	–	Rough Set
RT	–	Reverse Transcriptase
Sens	–	Sensitivity
Spec	–	Specificity
SEM	–	Standard Error of the Mean
SVM	–	Support Vector Machines
TM	–	Trans Membrane
TN	–	True Negative
TP	–	True Positive
TEA	–	Two Entropy Analysis
VHSE	–	Vectors of Hydrophobic, Steric, and Electronic properties
VIP	–	Variable importance projection
VSS	–	Variable subset selection
Wt	–	Wild type

Glossary

Classification	–	Subtype of machine learning that predicts membership of any class present in the training set as output variable.
Compound	–	Chemical substance consisting of two or more different elements that can be separated into simpler substances by chemical reactions.
Data mining	–	The process of pattern discovery in large unsorted data sets.
Descriptor	–	Machine learning interpretable way of describing a compound or target.
External validation	–	Validation procedure for a statistical model based on pre-partitioning the data set with observations into two subdivisions ('Training set' and 'Test set'). Subsequently a statistical model is trained on the training set and validated on the test set.
False negative	–	A compound tested active but inactive according to a model.
False positive	–	A compound tested inactive but active inactive according to a model.
Floating point number	–	Computerized form of scientific notation consisting of the product of a mantissa and a power of 10 with the exponent expressed as an integer. However, floating point numbers can also use base 2, base 8, base 10 and base 16, where base 2 is the most common.
Internal validation/ Cross validation	–	Validation procedure for a statistical model based on partitioning of the training set into n subdivisions. Subsequently a statistical model is trained on n-1 subdivisions and validated on the remaining subdivision. This process is repeated n times.
Ligand	–	Compound that has been shown to bind to a protein of interest.
Negative predictive value	–	In classification, true negative compounds as fraction of the total of compounds inactive according to a model.

Positive predictive value	–	In classification, true positive compounds as fraction of the total of compounds active according to a model.
Regression	–	Subtype of machine learning that predicts a floating point number as output variable based on observed values found in the training set.
Sensitivity	–	In classification, true positive compounds as fraction of the total of compounds active according to experiments.
Specificity	–	In classification, true negative compounds as fraction of the total of compounds inactive according to experiments.
Small molecule	–	Organic compound with a molecular weight of under 500 Dalton
Target	–	Protein of interest in a medicinal chemistry project.
Training set	–	Collection of data points defined as observed examples to capture the distinction between desired compounds (e.g. ligands for a protein) and undesired compounds.
True negative	–	Compound tested inactive and inactive according to a model.
True positive	–	Compound tested active and active according to a model.
Test set	–	Collection of data points used to validate a trained statistical model before production usage.

List of figures

Figure 1.1: The concept of molecular similarity....11
Figure 1.2: The concept of protein similarity....13
Figure 1.3: Data growth and processing power growth....14
Figure 1.4: Minimal feature width on integrated circuits since 1971 until 2010....15
Figure 1.5: Simplified schematic overview of a bioinformatics project....17
Figure 1.6: Simplified schematic overview of a cheminformatics project....19
Figure 1.7: Validation parameters used in classification based QSAR or PCM models....22
Figure 1.8: Validation plots in QSAR or PCM validation....23
Figure 1.9: Electron density from PDB structure 3EML....25
Figure 1.10: The different computational data analysis methods mentioned in this thesis....27
Figure 2.1: An example of the applicability domain concept....37
Figure 2.2: The difference between QSAR and PCM....38
Figure 2.3: Possibilities of PCM on a hypothetical dataset....40
Figure 2.4: A single PCM could also potentially be used to model both allosteric and orthosteric 46
Figure 2.5: Conversion of physicochemical properties of amino acids into a protein descriptor....52
Figure 2.6: Principal components 1 and 2 of the PCA analysis which resulted in the Z-scales....54
Figure 3.1: The approach used to characterize descriptor set distances and similarities....86
Figure 3.2: Principal components resulting from the PCA on 58 AAindices....94
Figure 3.3: Comparison of the distances between individual AA pairs....95
Figure 3.4: Principal component analysis of the distances between the different descriptor sets....97
Figure 3.5: The average performance in the ACE inhibitors 70-30 validation experiments....99
Figure 3.6: The average performance in the GPCR 70-30 validation experiments....100
Figure 3.7: The average performance in the GPCR LOSO validation experiments....101
Figure 3.8: The average performance in the NNRTIs 70-30 validation experiments....103
Figure 3.9: The average performance in the NNRTIs LOSO validation experiments....105
Figure 3.10: The average rank of the descriptor sets in the bioactivity benchmarks....107
Figure 4.1: The binding site we used to define the target similarity visualized in structure 3EML. ...118
Figure 4.2: Principal component analysis of the similarity in target space....119
Figure 4.3: Principal component analysis of ligand chemical space....121
Figure 4.4: Cross validation plot of the final model....125
Figure 4.5: Typical dose response curve obtained during the in vitro model validation....129
Figure 4.6: Flowchart of the work we performed....133

Figure 5.1: Graphical representation of the NNRTI dataset. 151
Figure 5.2: The binding site used in our models. 154
Figure 5.3: Model performance in the prospective experimental validation. 157
Figure 5.4: Extension of the applicability domain to target space. 161
Figure 5.5: Performance of PCM in leave-one-sequence-out experiments. 163
Figure 5.6: Example structures that where included in the model. 165
Figure 5.7: Overview of the contribution of mutations present at all individual residue positions. . 166
Figure 5.8: Overview of the contribution of the different chemical substructures. 168
Figure 6.1: Model internal validation. 185
Figure 6.2: The model performance in the LOSO experiments 187
Figure 6.3: Performance of PCM based models compared with sequence based models 190
Figure 6.4: Model interpretation, known mutations that lead to NNRTI (cross) resistance. 193
Figure 6.5: Model interpretation, mutations leading to drug specific resistance. 196
Figure 6.6: Model performance predicting the Stanford University data set. 199
Figure 7.1: The backbone of the NNRTI pocket, colored by the changes in average B-factor. 219
Figure 7.2: The backbone of the NNRTI pocket, colored by the average residue displacement 220
Figure 7.3: Overview of the changes occurring at the catalytic site as a result of DNA binding 221
Figure 7.4: Difference between surfaces that represent low conservation and high conservation. .. 224
Figure 7.5: The consensus binding pocket. 226
Figure 7.6: Consensus surfaces visualizing all HB locations. 227
Figure 8.1: How computational methods can be integrated in existing research projects. 252
Figure 8.2: Overlap between public and proprietary databases. 253
Figure A1: The structure of asperin and caffein. 267
Figure A2: The concept of similarity is output variable dependant. 268
Figure A3: Unknown situations for our model. 269

List of tables

Table 2.1: List of applications of PCM modeling 42
Table 2.2: Modeling techniques previously used 58
Table 3.1. Descriptor sets included. 80
Table 3.2. Principal Components Resulting from the AAindex selection. 85
Table 3.3. The Data Sets Used for the Bioactivity Benchmarks. 87
Table 3.4. Overall Descriptor Set Ranking. 108
Table 4.1. Structures of the newly identified human adenosine receptor ligands. 127
Table 5.1. Sequence information of the RT sequences in the data set 152
Table 5.2. Performance of different methods in experimental validation 158
Table 5.3. Best performing compounds (per sequence and overall) 169
Table 5.4. Worst performing compounds (per sequence and overall) 170
Table 6.1: Performance of PCM compared to sequence only models 191
Table 6.2: Novel resistance conferring mutations derived from the dataset (NNRTI). 194
Table 6.3: Novel resistance conferring mutations derived from the dataset (NRTI). 195
Table 6.4: Novel resistance conferring mutations derived from the dataset (PI). 195
Table 6.5: Personalized prediction examples for isolates not present in the original data set. 201
Table 6.6: Description of the data set used in the current study (Obtained from Virco). 203
Table 7.1: Volumetric information relating consensus structure size to the size of NNRTIs. 225
Table 7.2: Summary of the PDB structures that were used in all analyses. 230