Deriving a receptor based QSAR model using docking data

Aim

The objective of the study is to demonstrate the utility of using GOLD-derived 3D-parameters along with other external 2D-data to illuminate a quantitative understanding of the SAR of a series of potent serine protease inhibitors. It also serves to highlight the ease with which these various data can be handled and displayed in the fully integrated piece of software, GoldMine.

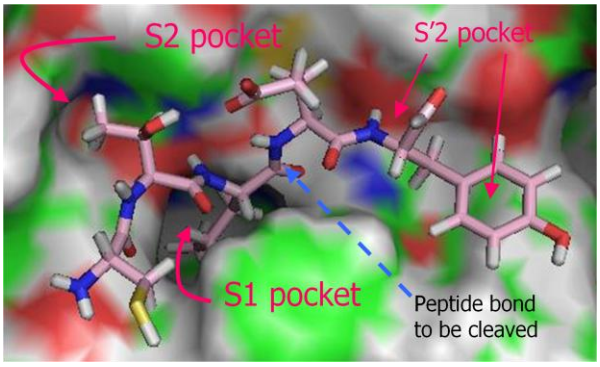
Introduction

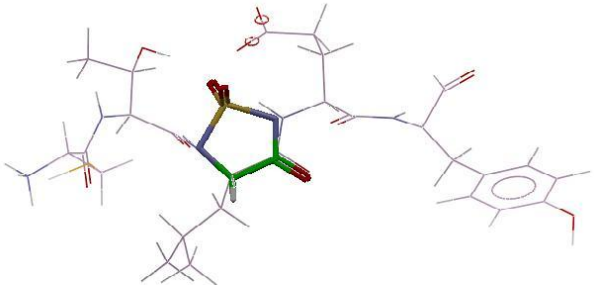
It has long been a problem with all docking routines to find a scoring function that optimizes not only the chances of successfully predicting the binding mode of a putative binder, but which also tries to broadly reflect the potential binding affinity of the ligand with the modelled receptor. When the many approximations made when performing any docking run are taken into account – often of a very coarse nature in order to enable the calculation to be made at all -, it is no surprise that trying to make this correlation is not usually significantly helpful. However, the potential reward if binding affinities could be derived from docking scores is so great that there is never any shortage of attempts or claims to do so. Alternatively, if a significant amount of related SAR data is already known about a system about whose binding mode we might be reasonably confident, then 3D-structural information derived from docking is quite easy to incorporate into a Quantitative Structure activity relationship (QSAR) and this can be significantly and semi-quantitatively helpful when considering future modifications to related structures.

In this study we will demonstrate a receptor-based approach to the QSAR of 1,2,5-thiadiazolidin-3-one dioxide inhibitors of human leukocyte elastase (HLE).

Method

A crystal structure was available of HLE complexed with a natural peptide inhibitor (PDB 1PPF), of which the pentapeptide section binding in the catalytic site is shown below left:

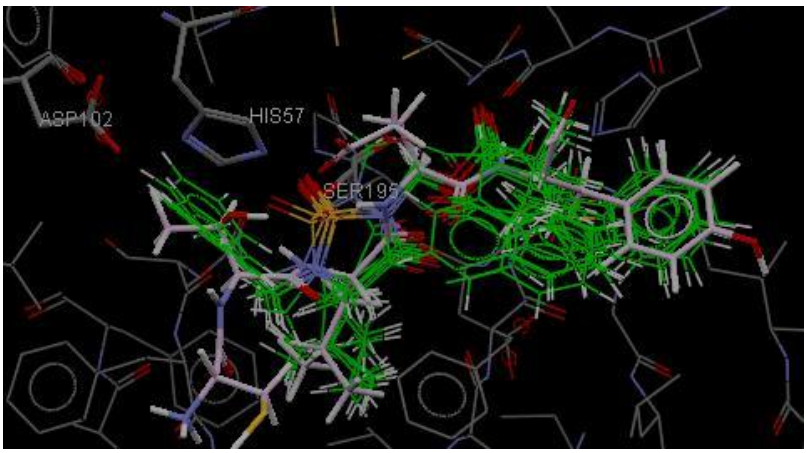




To the right of this is shown the same peptide conformation but with a 1,2,5-thiadiazolidin-3-one dioxide ring superimposed, highlighting how this conformation might be retained in an entropically favoured manner by the incorporation of this scaffold. A series based on this idea has indeed already been prepared and shown to be active by a group at Wichita State University who have published the in-vitro biological data used in this study[1].

In passing, one might note that simple inspection of the natural substrate in the active site would suggest that optimal occupation of the S1 pocket is likely to have a more strongly beneficial effect on binding than S2 and S'2 substitution in pockets that are either smaller or more exposed than S1.

Nineteen 3D-structures that are S1, S2 and S'2 derivatives of the thiadiazolidinone scaffold were prepared (using the 3D-structure generator CORINA) and docked into the catalytic site of 1ppf using GOLD with GoldScore more or less in default mode (albeit using the 'scaffold constraint' option, followed by a 'rescore' relaxation). Activities and leaving group pKa values were obtained from the original publication[1] (doi:10.1006/abbi.2000.2139).



In addition to saving the standard GoldScore parameters in the GOLD output files, it is possible, using the per-atom score option in GOLD, to break these descriptors into their individual atom-based components throughout the active site. This was all saved in a GoldMine to which was added the external data, in-vitro activity and leaving group pKa. This was done by importing a simple .csv file.

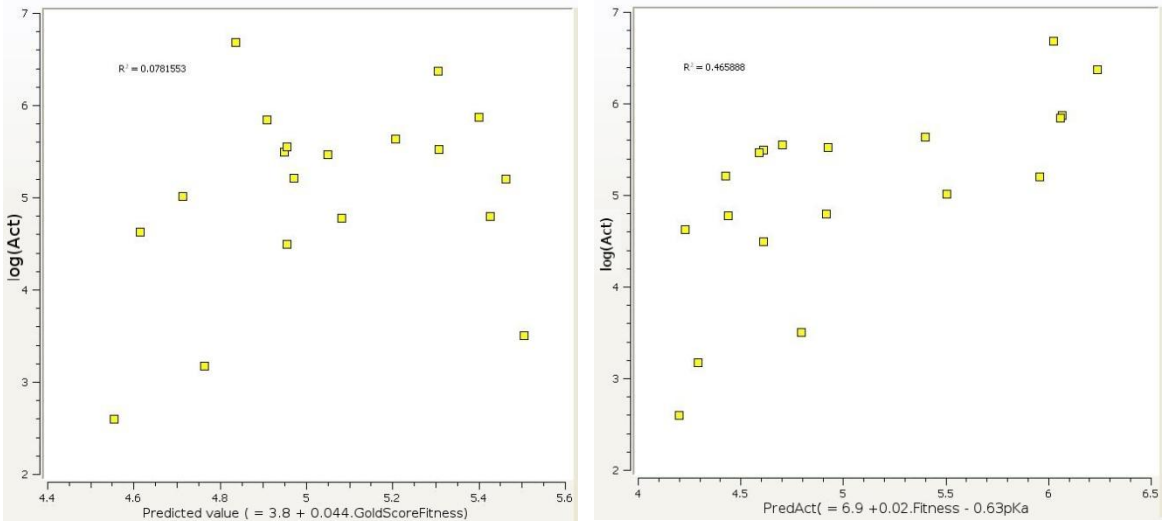
The activity measured was the observed first order rate constant, k_{obs}, for the hydrolysis of a UV-absorbent substrate. This was then quoted as the calculated ratio shown below:

$$k_{inactivn}/K_i = (k_{obs}/[I]).(1+[S]/K_m)$$

Our aim is to model binding affinity and, given that $\Delta G = -RT\ln K_i$, it seems sensible at first to use the GoldMine calculator module to determine $\log(k_{inactivn}/K_i)$, from here on reffered to as log(Act), and try to correlate this with the GoldScore.Fitness value.

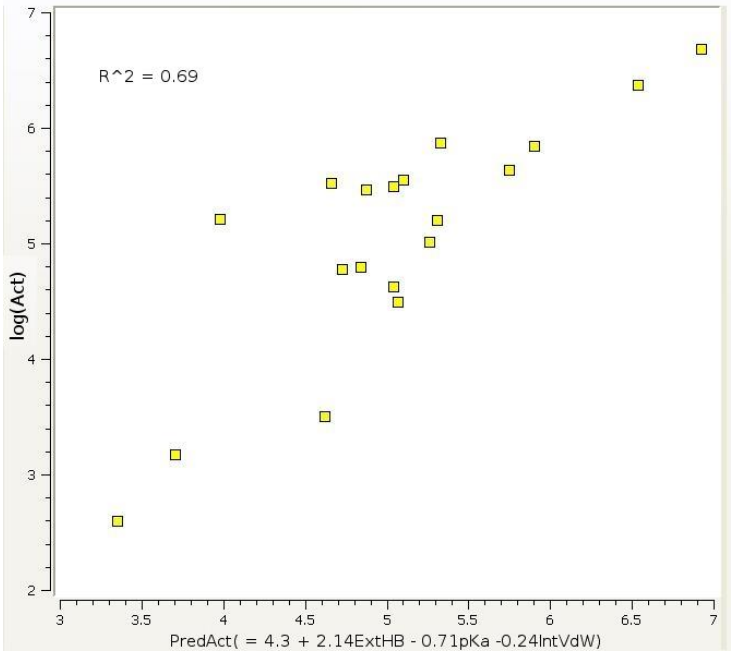
Results

This correlation is carried out in the Regression module of GoldMine. A very poor relationship was determined, $R^2 = 0.08$, and, whilst the plot below left (produced in the regression module) may hint at a relationship to the very optimistic, there is certainly no useful data to help realistically guide any future synthesis.



Left hand side model: log(Act) = 3.8 + 0.044GoldScore.Fitness
Right hand side model: log(Act) = 6.9 + 0.02GoldScore.Fitness – 0.63pKa

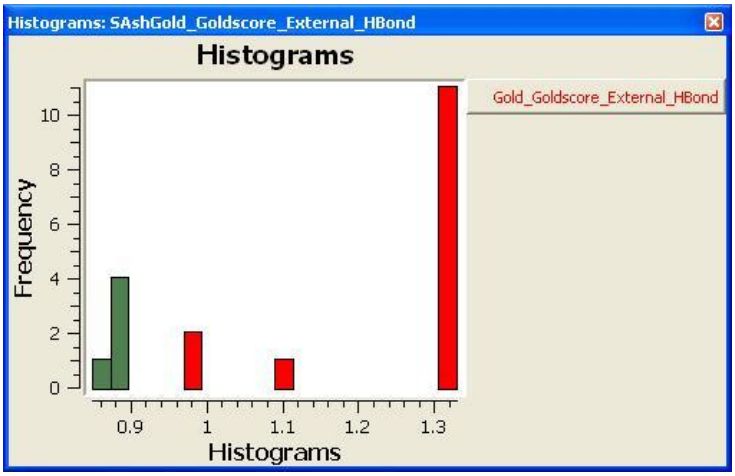
However, the original paper also notes that the acidity of the putative S'2 leaving-group affects the inhibitory potency. If this pKa is included in the regression along with the GoldScore.Fitness, then the corresponding plot looks a little different and perhaps a little more useful. And if rather than using the fixed relationship between the component terms of the GoldScore scoring function (which have been generalized for use with any and every receptor) one considers these terms individually in this receptor-specific regression then the corresponding plot changes further.



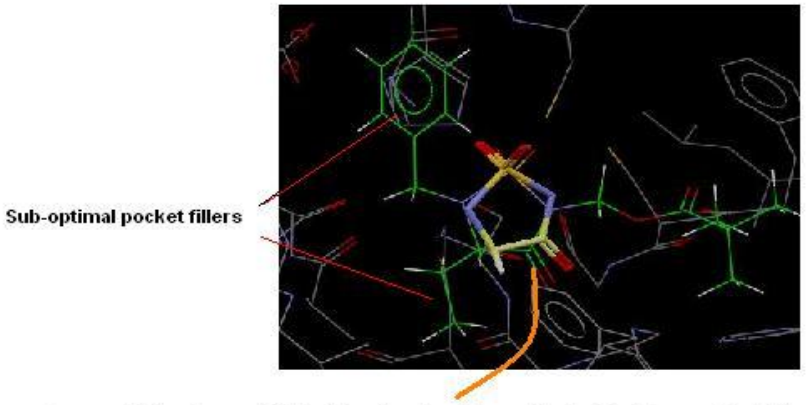
Model: log(Act) = 4.3 + 2.1ExtHB – 0.7pKa – 0.24IntVdW

With an F-to-enter ratio of 3.5 this QSAR model seems statistically reasonable and the model covers ~ 4 orders of magnitude of activity. Whilst it is easy within GoldMine to randomly split the data into teaching/training sets of any size, it was found to be impossible with only 19 observations to do this in such a way that the results for both sets were still significantly significant. However, it was interesting to note how remarkably robust the QSAR model remained all the way down to a 50% split in the data set.

Nonetheless, it seems a little surprising that this model implies such importance to what are predominantly hb-contacts with the common thiadiazolidinone core. Inspection of the external_hb histogram shows the distribution of this term to be somewhat bipolar:

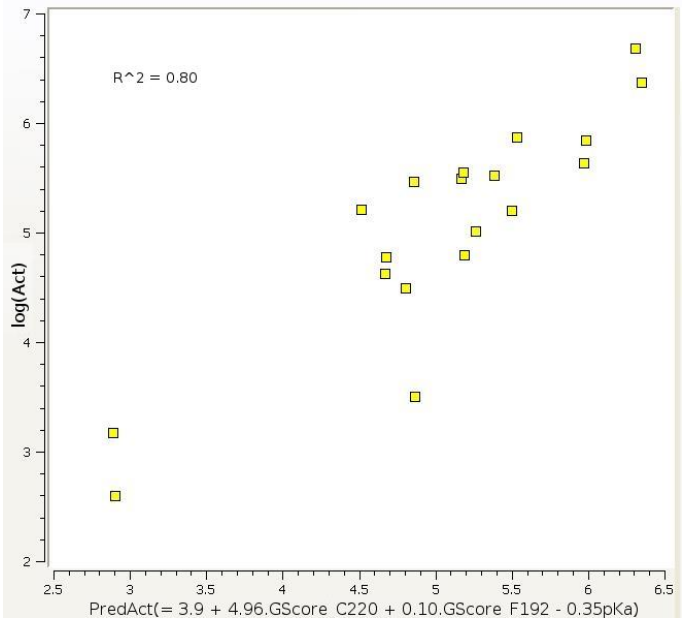


When the 'least active' ligands are selected and viewed, the influence of this common fragment on activity becomes apparent. It turns out that any sub-optimal substitution in either the S1 or the S'2 pocket is sufficient to markedly effect the final location of the scaffold which is in turn affects the value determined for the ligand ext_hb score. Similarly extension of substitution beyond the threonine filled S2 pocket, for example by N-benzylation, induces a movement of the ring which reduces this interaction.



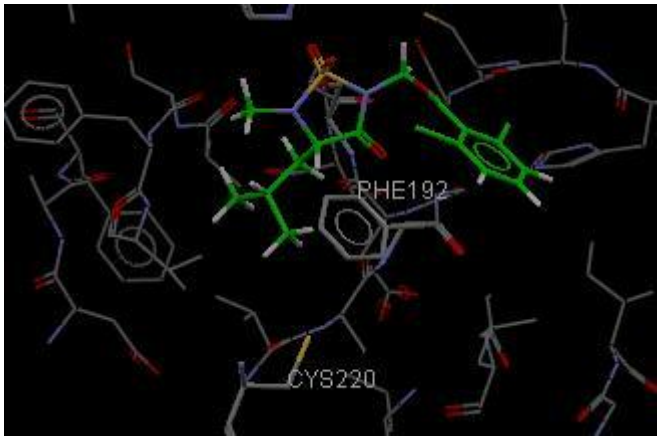
As a result, the ring scaffold rotates to a less than optimal orientation and the lig/recep interaction also falls

Although this explanation is in some measure plausible, it seemed desirable to try and find other possibly more explicit terms to model the observed SAR. We had already saved the per-atom contributions for all the docked conformations and used GoldMine to reassemble these into per-residue contributions (GScore.Cys220, GSCore.Phe192 etc). These were then used as the independent variables in a new MLR regression along with the leaving group pKa term.

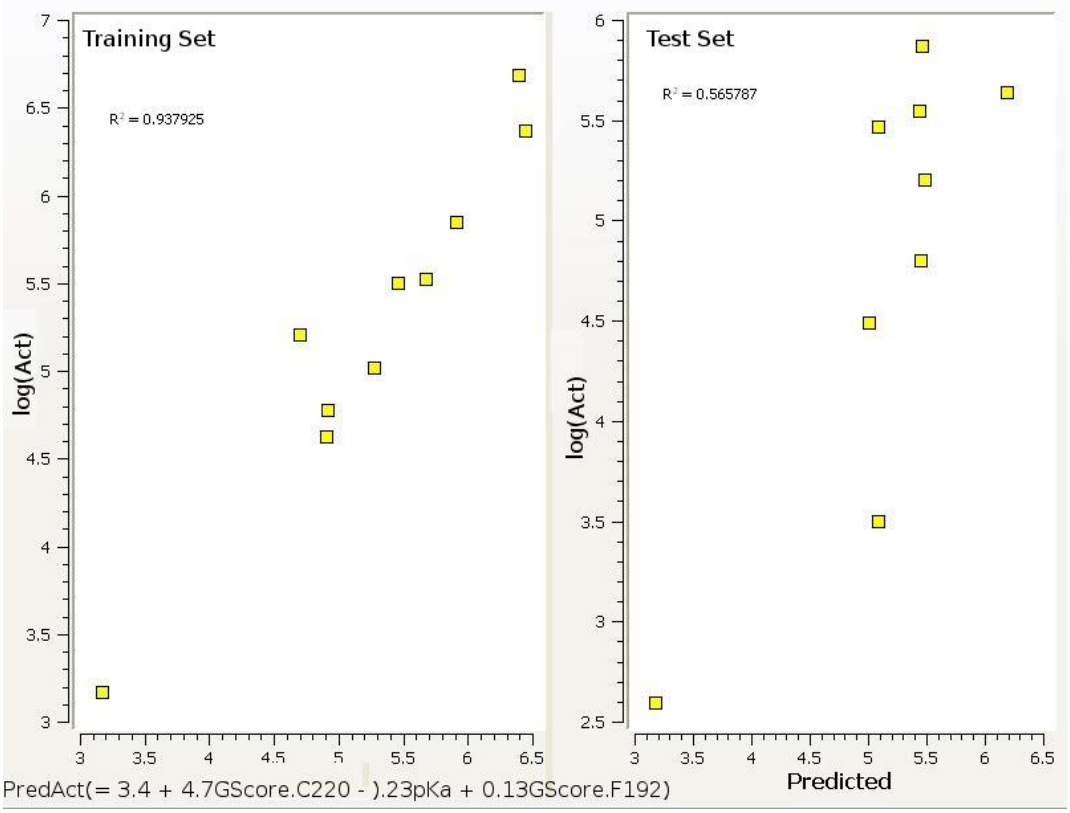


Model: log(Act) = 3.9 + 4.96Gold.Score.Fitness.CYS220 + 0.10GoldScore.Fitness.PHE192 – 0.35pKa

The resulting model (with better F-test) once again highlights the importance of optimally filling the S1 and S'2 pockets and also the dependence on the acidity of the putative leaving group in this latter region. The position of the two residues selected relative to the two pockets can be seen below.



As stated previously the small number of data points available suggests splitting the data set into training and test sets is dangerous and reduces the statistical level of significance excessively. However in this instance, although the plot for the teaching set shows distinct signs of overfitting, the QSAR coefficients still seem very robust and the plot for the test shows it to be just about useful. Certainly, the model seems sufficiently stable to lend some credence to the original 3-parameter model.



Model: log(Act) = 3.4 + 4.7Gold.Score.Fitness.CYS220 + 0.13GoldScore.Fitness.PHE192 – 0.23pKa

Conclusions

All the operations noted above were undertaken within GoldMine. Clearly there are other ways of carrying out all these steps involving calculation and display of parameter properties and distribution. The same applies to the selection of differing subsets and stepwise MLR, but we feel the availability of all this functionality within a single integrated piece of software allows these processes to be carried out with an interactive ease that makes the completion of a rigorous job even more likely.

One should note this is a preliminary communication. So, for example, the 19 conformations that were the preferred solutions that are the basis of our present analysis nearly all contain a common ligand-receptor clash that we would not normally consider reasonable when making a final selection. This is undoubtedly because the original conditions of scaffold constraint and the subsequent relaxation using 'rescore' alone were too stringent. Were we to repeat the process, this could certainly be improved.

References

1. W.C.Groutas et al, Bioorg and Med Chem, **8,** 1005-1016, (2000)

Products

GOLD – an accurate and reliable protein-ligand docking program

GoldMine – a dynamic data analysis tool for post-processing docking results enabling users to maximise the value of docking results

Hermes – CCDC's life science visualiser, used by GOLD, GoldMine, Relibase+ and SuperStar

For further information please contact Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK. Tel: +44 1223 336408, Fax: +44 1223 336033, Email: admin@ccdc.cam.ac.uk