Bioorganic & Medicinal Chemistry Letters 19 (2009) 5861–5863

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

Small kinase assay panels can provide a measure of selectivity

Peter Brandt [a,*], Annika Jenmalm Jensen [b], Jonas Nilsson [c]

[a] Beactica AB, Box 567, SE-751 22 Uppsala, Sweden
[b] MBB-Chemical Biology, Karolinska Institute, SE-171 77 Stockholm, Sweden
[c] i3 Innovus, SE-111 64 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 7 July 2009
Revised 21 August 2009
Accepted 23 August 2009
Available online 27 August 2009

Keywords:
Kinase panels
Ligand selectivity
Promiscuity
Multivariate analysis
PLS
Partial least squares

ABSTRACT

In this Letter, a novel strategy for assessment of ligand promiscuity is described. By using a carefully selected small set of kinases together with multivariate statistical methods, a measure of selectivity can be estimated. This will facilitate an appropriate selection of compounds for further development in lead generation and optimization.

The human genome contains approximately 500 protein kinases that regulate numerous cellular processes.[1,2] The modulation of protein kinase activity has been implicated for the treatment of several diseases such as cancer, inflammation, and metabolic diseases.[3] However, the design of small-molecule kinase inhibitors with optimal selectivity profiles for desired biological activity and safety remains a major challenge.[4–7] This has been further complicated by the fact that several so called 'specific' inhibitors from the past are now proving to be non-selective.[8,9] In the era of HTS, with a philosophy of one-target-one-disease, efforts were focused on finding selective compounds interacting specifically with one target. Today, multitargeted drugs are increasingly common and it is believed that certain diseases, most notably cancers, cannot be effectively treated with single target agents.[10–12] In either case, it is of great need to early on get an idea of the promiscuity of the individual hit/lead compounds. It would be a costly process if one would perform selectivity determinations using large kinase panels for many compounds. In this article is described a method to determine the promiscuity of individual compounds by determining the selectivity on a very limited set of kinases.

A recent publication by Karaman et al. describes a quantitative analysis of kinase selectivity.[5] In their work, they present interaction data for 38 different kinase inhibitors with 290 distinct kinases in the most comprehensive study of kinase selectivity published to date. The data was elegantly displayed in way of interaction maps[13] for each inhibitor. Given the large amount of data available to the authors, they introduced a quantitative parameter called the selectivity score, *S*. In a recent paper by Sciabola et al. this parameter is referred to as the promiscuity score.[14] Graczyk on the other hand, make use of the Gini coefficient, *G*.[15] In the context of this work, $G(c) \approx 1 - S(c)$. $S(c)$ is a value between 0 and 1 and describes the fraction of kinases found to bind with a dissociation constant of less than a given concentration *c*. A high value of *S* would indicate a promiscuous kinase inhibitor. Using $S(3\ \mu M)$, the effect of kinase panel size on apparent selectivity was evaluated.[5] Subpanels ranging from 20 to 288 kinases were randomly selected and apparent $S(3\ \mu M)$ were calculated. The conclusion was stated as 'small assay panels do not provide a robust measure of selectivity'. This triggered us to explore the data in a more rational way using multivariate analysis techniques, that is, PLS (partial least squares projection to latent structures).[16,17] We hypothesized that a carefully selected smaller set of kinases would be predictive of the selectivity score. To test the hypothesis, all K_d data were extracted from the Karaman publication and the selectivity score was calculated at 3 μM.

Missing pK_d values were set to 4 representing inactive compounds and staurosporine was excluded as an outlier. Thereafter, the data set was partitioned into two equally sized classes, one being the test set of the other. Starting with observations from Class 1, a PLS analysis was performed and the 10 variables most correlated with $S(3\ \mu M)$ were identified. Accordingly, a new PLS model including only these 10 variables was determined. The

* Corresponding author. Tel.: +46 18 56 08 85; fax: +46 18 50 68 08.
E-mail address: Peter.Brandt@beactica.com (P. Brandt).

0960-894X/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2009.08.083

derived model was validated by predicting the selectivity scores of observations from Class 2 and by cross-validation. The procedure was repeated starting with observations from Class 2. Results are reported in Table 1.

The kinases most relevant for explaining $S(3\ \mu M)$ in the model based on Class 1 data are AMPK-alpha1, CLK2, DLK, ITK, JAK2, LATS1, MAP4K1, MERTK, MST2, and MYLK2. For the model based on Class 2, the kinases used are AMPK-alpha1, AMPK-alpha2, IKK-epsilon, MAP4K1, MAP4K3, MARK1, MARK2, MARK3, MARK4, and MST1. These kinases are moderately promiscuous and bind between 5 and 13 ligands below 3 μM (average in the data set is 7.2, median is 6 and data is ranging from 0 to 23 ligands).

The overlap between the two models is only 2 kinases, namely AMPK-alpha1 and MAP4K1, yet, the predictive power of the models is high (Table 1 and Fig. 1). This suggests a redundancy in the data set, or, in other words, some kinases show similar binding profiles to the 38 inhibitors. To verify this, the 10 most important kinases used above were excluded and new models were built as above. The result from this exclusion study clearly shows that predictive models can be calculated using several different combinations of kinases (Table 1).

A more orthogonal selection of kinases would probably reduce the risk of erroneous predictions of novel kinase inhibitors and improve the predictive power of the model. To illustrate this, a selection of the 40 most important kinases were made according to the variable importance parameter in Simca.[17] This would serve as a good starting point for such an analysis. In this case, pK_d data for inhibitors in both classes were used. The top 40 kinases represent the kinase groups AGC, CAMK, CK1, STE, TK, TKL, and various ungrouped kinases. Of particular notice, the CMGC group is not represented and thus, this group of kinases add little information on the overall promiscuity of a kinase inhibitor. From the top 40 kinases a diverse subset of 10 kinases could be selected in various ways. Perhaps most intuitive would be to select kinases distributed over all branches of the human kinome. The selection of, for example, the kinases MAP4K3, FGFR1, AMPK-alpha1, LATS1, SRPK2, DLK representing the groups STE, TK, CAMK, AGC, CMGC, TKL could serve as a minimal panel for an initial assessment of the selectivity of a kinase inhibitor. This results in a predictive model with a R^2 of 0.91 and a Q^2 of 0.88. The internal prediction of $S(3\ \mu M)$ within this model[18] is depicted in Figure 2.

Using all data from the paper by Karaman et al. to produce a model leaves us with no test set for assessing the predictivity of the model. A useful, although less rigorous measure of the predictive ability of the model is Q^2 used as a quality measure of internal validation. In this work, Q^2 has been calculated by sevenfold cross-validation. A model with perfect predictability has Q^2 equals to one while a model with Q^2 equals to zero cannot predict better than random. Hence, the model presented here with $Q^2 = 0.88$ is convincing and the predictive ability is valid. There is always a risk of overfitting and making the model less able to extrapolate. In order to make sure that these results are true and not caused by pure chance several models were calculated where the order of the y values, that is, $S(3\ \mu M)$ were scrambled.[19] The majority of all these models have R^2 and Q^2 close to zero indicating that the present model is stable and not caused by chance.

As a final validation of the work presented in this article, we used a data set published by Bamborough et al.[6] In this article, screening data are presented for 17 kinase inhibitors tested at 10 μM on 203 kinases. 12 of these were identified among the kinases found to be important in assessing S in our analyses above.[20] By transforming the pK_d at 3 μM concentration used in the calculation of $S(3\ \mu M)$ to %control used in the data set of Bamborough, we could calculate an approximate $S(3\ \mu M)$ using all 203 kinases. Doing the same calculation of $S(3\ \mu M)$, using instead the set of the overlapping 12 kinases, we found a very good correlation. This is shown in Figure 3 ($R^2 = 0.72$). The result shows that the kinases identified in this work as being important predictors of ligand promiscuity are indeed useful also in other settings.

Table 1
Statistics from the PLS analyzes

	Training set			
	Class 1	Class 2	Class 1†	Class 2†
R^2	0.92	0.91	0.95	0.92
Q^2	0.91	0.88	0.94	0.89
RMSEP	0.09	0.08	0.06	0.08

† The 10 most important kinases excluded and modeling repeated as before.

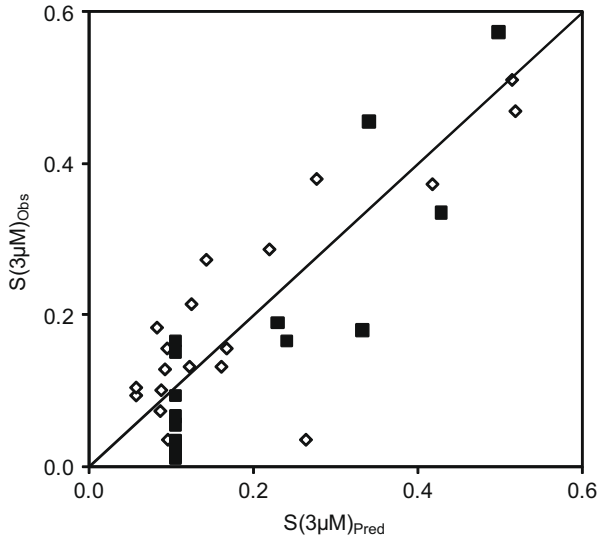

Figure 1. Filled squares represent predictions of Class 2 with a model based on Class 1 data. Open diamonds represents predictions of Class 1 with a model based on Class 2 data.

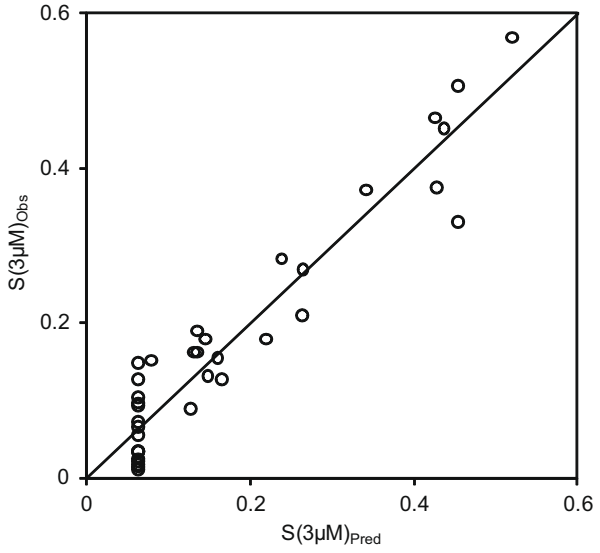

Figure 2. Internal predictions of $S(3\ \mu M)$ within a model based on data from 6 kinases each representing a branch of the human genome.

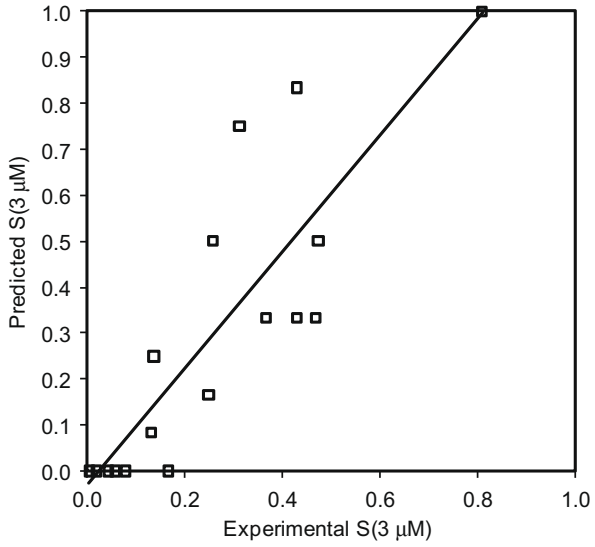

Figure 3. Correlation between experimental approximate $S(3\ \mu M)$ using all kinases and predicted $S(3\ \mu M)$ using only 12 kinases ($R^2 = 0.72$).

As an alternative approach to design a small panel of kinases, a diverse selection based on the loading plot from a principal component analysis (PCA) could be made.

Our conclusion is that, for a given kinase inhibitor, it is enough to determine selectivity on a small panel of kinases to estimate the degree of promiscuity. This is a parameter is of great importance also to guide the early phases of drug discovery such as the screen-to-hit and lead generation processes. For cost reasons, most companies can not address this issue on a large enough number of compounds this early. This hampers the efficiency of finding a lead compound with a desired profile.[21] Given the strong correlation between $S(3\ \mu M)$ used by Karaman et al. and the Gini coefficient used by Graczyk in a paper were single point data were used,[15] it might be more efficient from a cost performance perspective to run a larger kinase panel at a single point concentration. The authors of this paper would strongly recommend screening companies to use their proprietary data to develop models for prediction of kinase inhibitor promiscuity based on data from smaller kinase panels. This would encourage small and mid-sized pharmaceutical companies to enter the field of kinase inhibitors.

In conclusion, the results from our study show that there is redundant information in data derived from large kinase panel screens and that a proper selection of a carefully designed subset of kinases will provide sufficient information to estimate ligand promiscuity.

References and notes

1. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. *Science* **2002**, *298*, 1912.
2. Hanks, S. K. *Genome Biol.* **2003**, *4*.
3. Cohen, P. *Nat. Rev. Drug Disc.* **2002**, *1*, 309.
4. Sawa, M. *Mini-Rev. Med. Chem.* **2008**, *8*, 1291.
5. Karaman, M. W. et al *Nat. Biotechnol.* **2008**, *26*, 127.
6. Bamborough, P.; Drewry, D.; Harper, G.; Smith, G. K. K.; Schneider, K. *J. Med. Chem.* **2008**, *51*, 7898.
7. Caffrey, D.; Lunney, E.; Moshinsky, D. *BMC Bioinformatics* **2008**, *9*, 491.
8. Fedorov, O.; Marsden, B.; Pogacic, V.; Rellos, P.; Muller, S.; Bullock, A. N.; Schwaller, J.; Sundstrom, M.; Knapp, S. *Proc. Natl. Acad. Sci.* **2007**, *104*, 20523.
9. Bain, J.; Plater, L.; Elliott, M.; Shpiro, N.; Hastie, C. J.; McLauchlan, H.; Klevernic, I.; Arthur, J. S. C.; Alessi, D. R.; Cohen, P. *Biochem. J.* **2007**, *408*, 297.
10. Fernandez, A.; Crespo, A.; Tiwari, A. *Drug Discovery Today* **2008**, *4*, 1.
11. Horuk, R. *Expert Rev. Mol. Med.* **2009**, *11*, e1.
12. Hopkins, A.; Mason, J.; Overington, J. *Curr. Opin. Struct. Biol.* **2006**, *16*, 127.
13. Fabian, M. A. et al *Nat. Biotechnol.* **2005**, *23*, 329.
14. Sciabola, S.; Stanton, R. V.; Wittkopp, S.; Wildman, S.; Moshinsky, D.; Potluri, S.; Xi, H. *J. Chem. Inf. Model.* **2008**, *48*, 1851.
15. Graczyk, P. P. *J. Med. Chem.* **2007**, *50*, 5773.
16. (a) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1; (b) Martens, H.; Næs, T. In *Multivariate Calibration*; John Wiley & Sons, 1992.
17. Simca P+, version 12.0.0.0, Umetrics AB, www.umetrics.com.
18. $S(3\ \mu M) = 0.0129\ pK_d(AAK1) + 0.0252\ pK_d(AMPK\text{-}alpha) + 0.00764\ pK_d(DLK) + 0.0501\ pK_d(FGFR1) + 0.0257\ pK_d(LATS1) + 0.0445\ pK_d(MAP4K3) - 0.601$.
19. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold S. In *Multi- and Megavariate Data Analysis, Principles and Applications*; Umetrics AB, 2001, 489–491.
20. The kinases used were AMPK-alpha1, AMPK-alpha2, CLK2, FGFR1, ITK, MAP4K1, MARK1, MARK2, MARK4, MERTK, MYLK2, and STK3.
21. Ma, H.; Deacon, S.; Horiuchi, K. *Expert Opin. Drug Discovery* **2008**, *3*, 607.