Pergamon

Bioorganic & Medicinal Chemistry 10 (2002) 1037–1041

BIOORGANIC & MEDICINAL CHEMISTRY

Application of Non-Parametric Regression to Quantitative Structure–Activity Relationships

Jonathan D. Hirst,[a,*] T. John McNeany,[a] Trevor Howe[b] and Lewis Whitehead[b]

[a]*School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, UK*
[b]*Novartis Horsham Research Centre, Wimblehurst Road, Horsham, West Sussex RH12 5AB, UK*

Received 29 March 2001; accepted 8 October 2001

Abstract—Several non-parametric regressors have been applied to modelling quantitative structure–activity relationship (QSAR) data. Performances were benchmarked against multilinear regression and the nonlinear method of smoothing splines. Variable selection was explored through systematic combinations of different variables and combinations of principal components. For the training set examined—539 inhibitors of the tyrosine kinase, Syk—the best two-descriptor model had a 5-fold cross-validated q^2 of 0.43. This was generated by a multi-variate Nadaraya–Watson kernel estimator. A subsequent, independent, test set of 371 similar chemical entities showed the model had some predictive power. Other approaches did not perform as well. A modest increase in predictive ability can be achieved with three descriptors, but the resulting model is less easy to visualise. We conclude that non-parametric regression offers a potentially powerful approach to identifying predictive, low-dimensional QSARs.

Introduction

The ability to derive accurate quantitative structure–activity relationships (QSARs) is increasingly important in the search and development of biologically active molecules,[1,2] with the emergence of high throughput screening (HTS) and combinatorial chemistry, where limiting factors are the characterisation of large numbers of molecules and the sparsity of the data. The time and cost of characterisation could be reduced if only the compounds most likely to be highly active were characterised. More informed selection of hits could also be achieved if crude SARs could be rapidly determined. This has motivated many studies on the design of combinatorial libraries and HTS analysis. One study on the development of a heuristic algorithm for reagent picking estimated assay costs at \$0.10 per compound.[3] QSAR provides a powerful technique aimed at delineating the relationships between active and inactive molecules in an assay system. Successful QSARs on HTS data would permit the rapid identification of trends which would result in a more informed lead selection than simply identifying the most active species in a HTS.

Originally, QSAR was restricted to linear regression methods, which limited the complexity of the relationships that could be accurately modelled.[4,5] These traditional QSARs rely on fitting data to linear or parabolic functions in order to find predictive relationships. However, the underlying relationships that occur in complex data sets may be more subtle, and consequently difficult to fit to a conventional polynomial. As computers have become more powerful, it has become increasingly practical to apply non-parametric methods.[6] These statistical methods are explored in many areas of chemistry, from analytical chemistry[7] to the development of non-linear QSARs.[8,9] Non-parametric methods are capable of producing more flexible models, as they are not limited to a specific functional form, such as a polynomial.

In previous studies,[8,9] the utility of non-parametric statistical methods has been investigated in the context of some typical medicinal chemistry data sets, which are small in comparison to contemporary combinatorial libraries or HTS data sets. Whilst some early approaches appeared to offer advantages in terms of speed,[9] it has emerged that potential improvement in predictive accuracy is probably the primary advantage offered by non-parametric statistical methods.[8] The question of applicability to large data sets then naturally arises.

*Corresponding author. Tel.: +44-115-951-3478; fax: +44-115-951-3562; e-mail: jonathan.hirst@nottingham.ac.uk

0968-0896/02/$ - see front matter
PII: S0968-0896(01)00359-5

In this paper, we explore the utility of non-parametric methods to the analysis of a data set of over 900 compounds. This is larger than most data sets used in traditional QSAR studies, although still relatively small in terms of HTS data sets. In this study, we assess the merits of several non-parametric approaches and we explore issues of variable selection. We focus on low-dimensional non-parametric models, partly because they are computationally tractable to generate, but also because we believe that low-dimensional non-parametric models are more readily interpreted, easier to visualise and perhaps more likely to reflect underlying physical affects than high-dimensional models.

Methods

This study examines a series of 910 inhibitors of the tyrosine kinase, Syk; 539 comprise the training set and 371 molecules are in a test set. The dataset contains typical kinase inhibitor like structures, including purines, pyrimidines, indoles, imidazoles, pyrazoles and quinazolines. These are multiply substituted with both aromatic and small functionalised components. As such they all have four or more rotatable bonds and constitute a diverse chemical set. Pending patent applications preclude full disclosure of the exact structures here. Syk kinase is a key component in cell signalling cascades,[10] which are involved in inflammatory processes. Biological activity was taken to be the measured IC_{50} value for each compound. Each molecule was described by 96 physicochemical descriptors computed using the default set from Molecular Simulations Inc. Cerius2 software.[11] Four non-parametric regression techniques, described in more detail below, and (for comparison) multi-linear regression were used to determine QSARs. All possible one-descriptor and two-descriptor models were generated from the 96 descriptors. A limited number of three-descriptor models were explored, in which the descriptors were chosen based on the performance of the one- and two-descriptor models.

In a separate strategy, variable selection was explored using principal component analysis.[12] The first 10 principal components were computed. Each principal component was used to create a single-descriptor model, giving a total of 10 single-descriptor models. Every possible pair of the 10 principal components were used to create a total 45 possible two-descriptor models, and so on, up to a single 10-descriptor model, using all 10 descriptors.

The non-parametric approaches, namely Nadaraya–Watson, shifted Nadaraya–Watson, local linear and MARS, have been discussed in the context of QSAR in detail elsewhere.[8] They are based on estimating the activity of a molecule at a point in a property space, as a moving, local, weighted average of the activities of all the molecules in the data set:

$$\hat{m}(x) = \sum_i w_i(x) y_i \tag{1}$$

The precise form of the estimator is determined by the nature of the weighting function, the so-called regression or smoothing kernel. In the case of the Nadaraya–Watson estimator,[13,14] the kernel at the point x is a Gaussian function:

$$K_x = (2\pi)^{-1/2} e^{-u^2/2h^2} \tag{2}$$

where $u = x - x_i$, for the set of data $\{x_i\}$ and h is the bandwidth of the Gaussian. The bandwidth acts as a smoothing parameter, determining the degree of locality of the regression. This method encounters difficulty in high dimensional spaces, due to the increase in empty space with increasing dimensionality.

One can attempt to circumvent this using additive models.[15] In an additive model, each dimension is treated independently, resulting in a regression function which is the sum of several one-dimensional functions. An additive approach to the regression problem is taken for two methods used in this study. One is the local linear technique,[16] which uses a localised polynomial as the smoothing kernel. The other is based on a shifted Nadaraya–Watson method, a mass recentered smoothing kernel,[17] which appears to perform better for data sets in which there are discontinuities. These methods were employed using our own implementation.[8] Finally, we also assess the more established MARS package,[18] which uses a system of smoothing splines to derive a regression function.

Due to the size of the data set, the computational cost of some of the regression methods was not negligible, although we have not pursued detailed code optimisation. Multilinear regression was the fastest, and the slowest was the multivariate Nadaraya–Watson method. The time requirement of the slower methods precluded a comparative study with all possible three-descriptor models. The predictive ability of the models was assessed on the 539 compounds in the training set using a 5-fold cross-validation approach and measured in terms of a q^2 value:

$$q^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_{i,\mathrm{obsd}} - y_{i,\mathrm{pred}}\right)^2}{\sum_{i=1}^{N} \left(y_{i,\mathrm{obsd}} - \bar{y}_{i,\mathrm{pred}}\right)^2} \tag{3}$$

where N is the number of data points, and obsd refers to observed values and pred to predicted values. A value greater than 0.4 might be considered to be useful. The activities of the independent test set were simply computed using the most predictive model obtained from the non-linear methods.

Results

In Table 1, we compare the predictive ability of the methods on the one- and two-descriptor models, as assessed through cross-validation. We report only the best models, as based on the highest q^2 values. A

randomisation trial, in which the activities of the molecules were randomly assigned, led to no models with any statistical significance, that is q^2 values for all one- and two-descriptor models for the different methods were less than 0.01. In Table 1, we also record the corresponding mean absolute errors (MAE) and Spearman rank correlations from the cross-validation trial. Training set performances were generally higher, as expected. The non-parametric methods perform better than the linear regression. The two-descriptor models are better than the single-descriptor models. Overall, the most accurate cross-validation test set predictions are made by the multi-variate Nadaraya–Watson regressor. The index number identifying the descriptors in the models are given in Table 2. The most predictive two-descriptor model has a q^2 of 0.43 and involves the parameters JX and PHI (Fig. 1), with a low value of JX and a very low value of PHI apparently leading to the most active molecules. JX is a topological index, whose value does not substantially increase with molecule size or the number of rings present.[19] PHI is a molecular flexibility index, based on: (a) the number of atoms, (b) the presence of rings, (c) branching, and (d) the presence of atoms with covalent radii smaller than those of an sp^3 hybridised carbon atom.[20] Higher values of PHI correspond to more flexible molecules. At exceptionally high values there will be a significant entropy factor to be paid, as each rotational bond (the usually accepted figure is 1–3 kcal/mol per rotor) needs to be frozen out on binding. At very low values the molecule might be deemed too stiff to undergo conformational flexing required to fit into the receptor. So the actives (as shown in Fig. 2, for example) appear from the low to intermediate end of this scale. The steep behaviour in the region JX = 1.1–1.3 at PHI = 2–3 may be due to a paucity of data in this region, rather than the nature of the descriptor space.

The model shown in Figure 1 is readily visualised and understood. As the number of descriptors increases, the models become less easily understood and there is a concern that statistical artefacts may be dominating genuine physical influences. For the data set under consideration, the predictive ability of three-descriptor models is only marginally superior to that of the two-descriptor models (Table 3), and probably not sufficiently so to compensate the greater complexity of the models. From Table 3, it is also apparent that principal component analysis does not offer any advantage in the modelling of these data. Using five, six or seven of the most significant principal components does not lead to models that are more predictive than the two- or three-descriptor models based on the original descriptors.

The JX and PHI descriptors were calculated for the 371 molecules in the independent test set. These molecules are of a similar chemical class to those in the training set and were subsequently made and tested in the same Syk

Table 1. Comparison of methods (single and two-descriptor models)

Method	Identity of descriptors	Best q^2	r^2	MAE	Spearman rank
Multilinear regression	71	0.174	0.179	0.476	0.370
Multivariate Nadaraya–Watson	19	0.239	0.264	0.438	0.473
Additive local linear	44	0.230	0.250	0.434	0.493
Additive shifted Nadaraya–Watson	44	0.229	0.248	0.436	0.492
MARS	38	0.244	0.256	0.432	0.503
Multilinear regression	64, 77	0.226	0.005	0.456	0.464
Multivariate Nadaraya–Watson	29, 36	0.432	0.542	0.365	0.613
Additive local linear	46, 71	0.331	0.359	0.416	0.557
Additive shifted Nadaraya–Watson	46, 71	0.336	0.359	0.413	0.560
MARS	46, 71	0.355	0.345	0.397	0.584

Table 2. Descriptors featured in most predictive models

Descriptor number	Descriptor
17	Number of hydrogen-bond donor groups
18	AlogP, Ghose and Crippen log of the partition coefficient
19	Mol Ref, Ghose and Crippen molar refractivity
24	IAC total, information of atomic composition
26	V-dist mag, based on vertex distance indices
29	JX, Balaban index
33	Kappa-1-AM, Kier shape index order 1
34	Kappa-2-AM, Kier shape index order 2
36	PHI, molecular flexibility index
38	SC-1, Kier and Hall subgraph count first order index
44	CHI-2, Kier and Hall molecular connectivity index, order 2
46	CHI-3_P, Kier and Hall molecular connectivity index, order 3
56	Zagreb index
64	Jurs-PPSA-3, partial positively charged surface area
71	Jurs-FPSA-3, fractional positively charged partial surface area
77	Jurs-WPSA-3, surface weighted charged partial surface area
83	Jurs-TPSA, total polar surface areas

kinase assay. These data evaluated within the multivariate Nadaraya–Watson model are shown in Figure 2. Although the q^2 value for the model of 0.43 is at the lower limit of predictivity, Figure 2 shows that qualitatively, in the independent test set, most of the positives in the assay are correctly captured. The active compounds cluster mainly in the region predicted to be highly active by the model, as shown in Figure 1. Figure 2 shows that the less active molecules are dispersed across the range of JX and PHI values. The 371 mole-

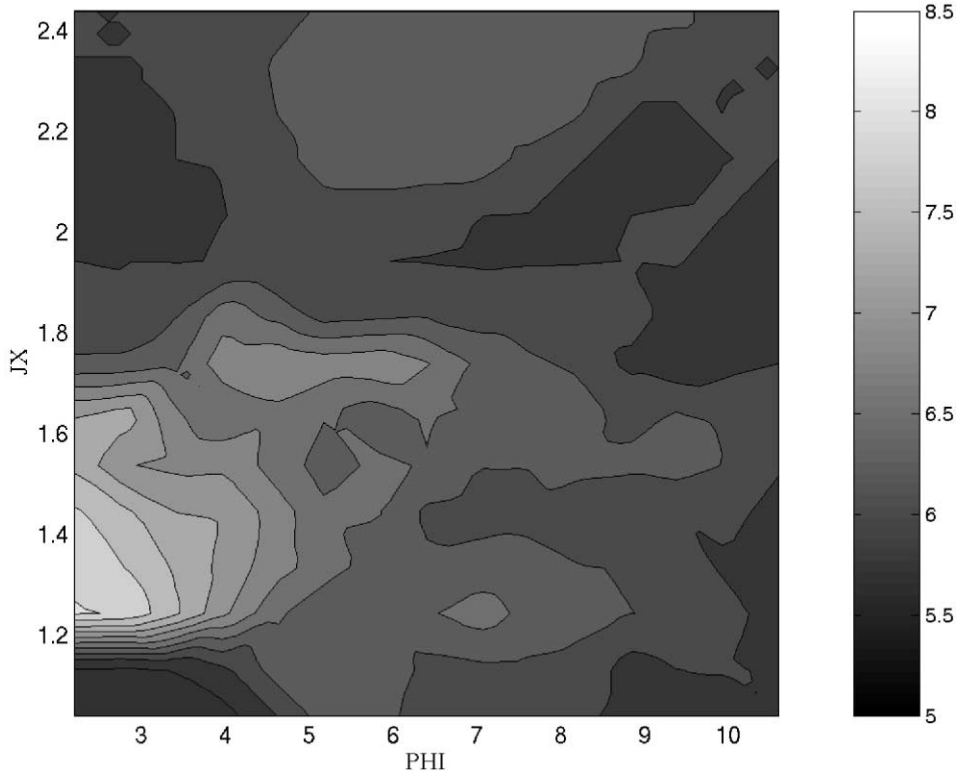

Figure 1. Most predictive two-descriptor model for inhibitors of syk tyrosine kinase. Grey-scale shows activity as $-\log_{10}$ IC_{50}.

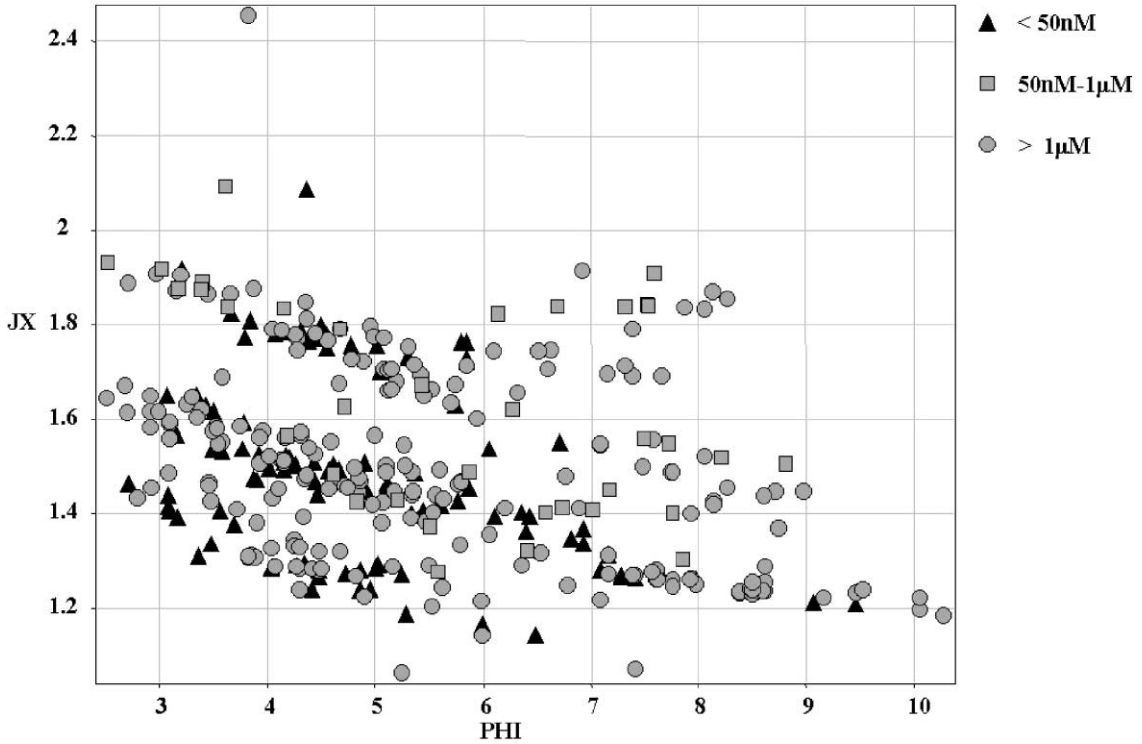

Figure 2. Independent test set of 371 molecules (each represented by a dot) in the PHI-JX descriptor space of the multivariate Nadaraya–Watson model shown in Figure 1.

Table 3. Comparison of methods

Method	Identity of descriptors	Best q^2	r^2	MAE	Spearman rank	Best q^2 from PCA	Number of components
Multilinear regression	17, 56, 26	0.273	0.286	0.437	0.537	0.246	6
Multivariate Nadaraya–Watson	29, 36, 83	0.456	0.542	0.354	0.661	0.402	5
Additive local linear	29, 34, 18	0.375	0.416	0.391	0.600	0.324	7
Additive shifted Nadaraya–Watson	24, 29, 39	0.381	0.310	0.386	0.601	0.334	7
MARS	17, 56, 33	0.383	0.390	0.391	0.595	0.336	8

cules contain 30 compounds active below 50 nM, giving a hit rate of 8.1%. The multivariate Nadaraya–Watson model predicts 117 compounds to be active below 50 nM, of which 11 actually are. This corresponds to a hit rate of 9.4%, representing a modest enrichment.

Conclusions

In this study, we have illustrated the utility of non-parametric regressors in the analysis of structure–activity data. We have shown that such methods can be applied to data sets of a reasonable size, in our example $\sim$900 compounds, and our approach of focusing on accurate, low-dimensional models would be readily extensible to much larger data sets. The familiar approach of using principal component analysis appears not be the most profitable strategy for the data set under consideration. The greater accuracy of the non-parametric regressors over multi-linear regression is clear. The data do not afford any very obvious, simple relationships and are perhaps representative of typical real world data. Nevertheless, the best non-parametric models achieve cross-validated q^2 values over 0.4, and this model has been proven to have some predictive utility.

We have explored more complicated models in a relatively limited fashion, constrained by the computational cost of the methods and the combinatorial explosion of models involving larger numbers of descriptors. Strategies to sample higher-dimensional models, such as genetic algorithms as have been employed elsewhere,[21] warrant further investigation, but the issues of statistical significance and interpretability need to be borne in mind.

Acknowledgements

We thank the Nuffield Foundation for financial support. We thank Dr. Pere Constans for useful discussions and his work on code development. We thank the Royal Society for financial support.

References and Notes

1. Kubinyi, H. *Drug Des. Today* **1997**, *2*, 457.
2. Hirst, J. D. *Curr. Opin. Drug Discov. Dev.* **1998**, *1*, 28.
3. Good, A. C.; Lewis, R. A. *J. Med. Chem.* **1997**, *40*, 3926.
4. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962**, *194*, 178.
5. Hansch, C. A. *Acc. Chem. Res.* **1969**, *2*, 232.
6. Efron, B.; Tibshirani, R. *Science* **1991**, *253*, 390.
7. Sekulic, S.; Seasholtz, M. B.; Kowalski, B. R.; Lee, S. E.; Holt, B. R. *Anal. Chem.* **1993**, *65*, 835A.
8. Constans, P.; Hirst, J. D. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 452.
9. Hirst, J. D. *J. Med. Chem.* **1996**, *39*, 3526.
10. Chu, D. H.; Morita, C. T.; Weiss, A. *Immun. Rev.* **1998**, *165*, 167.
11. *Cerius2*; Molecular Simulations Inc.: San Diego, CA, USA.
12. Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984.
13. Nadaraya, E. A. *Theory Probability Appl.* **1964**, *10*, 186.
14. Watson, G. S. *Sankhya Ser. A* **1964**, *26*, 359.
15. Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models*; Chapman and Hall: New York, 1990.
16. Stone, C. J. *Ann. Stat.* **1977**, *5*, 595.
17. Mammen, E.; Marron, J. S. *Biometrika* **1997**, *84*, 765.
18. Friedman, J. H. *Ann. Stat.* **1991**, *19*, 1.
19. Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
20. Hal, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling; *Reviews in Computational Chemistry II*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; p 367.
21. So, S.-S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.