Neighbourhood Behaviour Studies for Array Chemistry

George Papadatos[1], Val Gillet[1], Peter Willett[1] and Iain McLay[2]

INTRODUCTION: The rationale for many Chemoinformatics applications, such as *virtual screening* and *similarity searching*, is to be found in the *similar property principle* or *similarity principle*[1]. According to it, similar compounds tend to exhibit similar properties, and therefore similar biological activities.

Closely related to the similar property principle is the concept of the *neighbourhood principle* or *neighbourhood axiom*. According to this principle, the compounds of a subset within the same local region ("neighbourhood") of structural space as defined by a molecular descriptor, are more likely to display similar values of a desired property than those of a randomly selected subset of the same size.

Hence, Neighbourhood Behaviour (NB) can be regarded as the ability of structural space to map onto the property space in such a way that neighbouring points in the former are likely to correspond to neighbouring points in the latter (figure 1). Patterson plots are a straight-forward way to visualise NB[2]: **The pairwise dissimilarities among the compounds are plotted against the absolute pairwise differences in property values.** If NB is valid, then the upper left triangle will be sparsely populated ("forbidden area"), indicating that less dissimilar (i.e. similar) compounds show small differences in property values (figure 2).

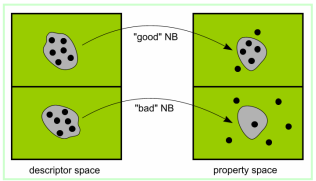


Figure 1

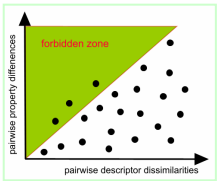


Figure 2

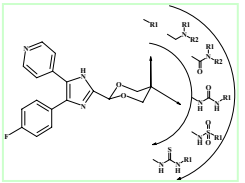


Figure 3

AIM: To date, NB has been applied solely focusing on biological activity in an attempt to identify appropriate descriptors for structure – activity studies. Here we have applied the NB concept to arrays of compounds synthesised during a lead optimisation campaign at GSK[3] (figure 3). In addition to biological activity, we also investigate the relationship between descriptors and a number of properties such as solubility, metabolic stability, permeability and lipophilicity.

METHODOLOGY: The dataset consists of 2971 compounds and their associated experimental data taken from a GSK lead optimisation campaign. As structural descriptors we have used both property vectors and 2D fingerprints and have calculated the pairwise dissimilarities within the dataset using both Euclidean distance and the Tanimoto coefficient respectively.

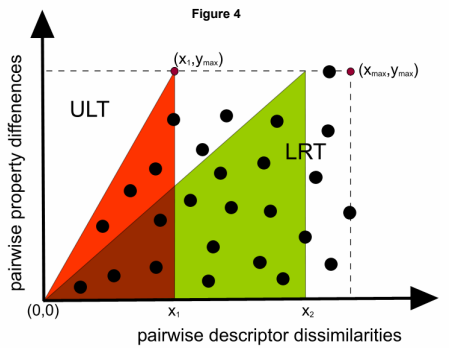
The following algorithm was developed in order to quantify NB: For every data point (x_i,y_i) on the plot a candidate triangle is generated with vertices (0,0), $(x_i,0)$ and (x_i,y_{max}). The optimal triangle is the one with the highest density of points, which combined with the rectangular region to its right forms the lower right trapezoid (LRT). The NB score is then the ratio of the density in the LRT to the density across the whole rectangular region (LRT and the remaining upper left triangular region, ULT). The maximum score is observed when all the data points lie in the lower right triangular half of plot, in which case **NB=2.0** (figure 4).

The statistical significance of the NB calculations was measured with a χ^2 test. N_{LRT} is the number of points observed in the LRT. For the calculation of the expected value n_{LRT}, the y values were scrambled 1000 times among the data points and then we counted the number of scrambled points which fall in the original LRT[4].

$$\text{Density} = \frac{\text{population}}{\text{area}} \qquad \text{NB} = \frac{\text{Density}_{LRT}}{\text{Density}_{LRT \cup ULT}}$$

$$\chi^2 = \frac{(N_{LRT} - n_{LRT})^2}{n_{LRT}}$$

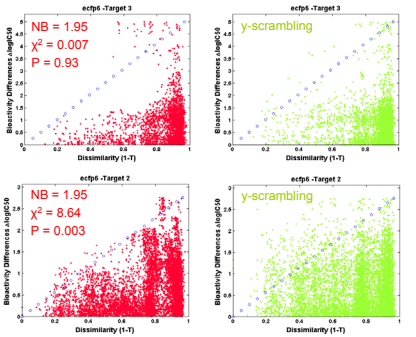
Figure 4



RESULTS

The importance of χ^2

The nature of a descriptor or a dataset can result in spurious NB. For example, in both cases on the right, we used the same descriptor versus bioactivity against two targets. The NB scores are very high and identical. However, scrambling the y values in the first plot results in a similar plot where most of the points still lie within the original trapezoid, indicating a low statistical significance. Conversely, in the second plot y-scrambling gives a very different plot, significantly different to the original one. Such differences can be identified by the corresponding χ^2 scores. A high χ^2 score indicates a gradual "fanning out" of points under the diagonal as we move to the right in the plot.



NB studies for chemical arrays – the role of the descriptors

14 arrays were selected from the initial dataset with an average size of 42 compounds. For each one the NB and χ^2 scores were calculated using 8 descriptors and biological activity data. In 5 arrays, the χ^2 values across the descriptors were relatively high, indicating meaningful structure-activity relationships and thorough exploration of the chemical space around the seed compound. The NB scores were averaged over the 5 arrays and are shown in the bar graph on the right (figure 5). SciTegic's 2D fingerprints seem to perform consistently better, while MDL Public Keys and the physicochemical properties vector perform the worst.

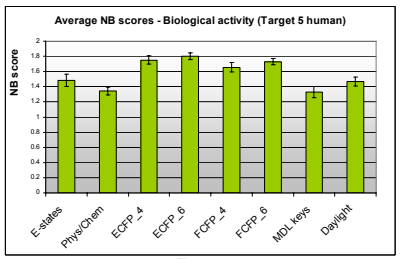


Figure 5

Biological Activity	Size	min χ^2
Target 1	433	44.1
Target 2	119	4.6
Target 5 human	600	6.3
Target 5	600	6.5
Target 6	600	3.8

Table 1

Property	Size	min χ^2
Solubility	420	4.5
Permeability	571	19.8

Table 2

NB studies across several properties and targets

Besides arrays, we carried out experiments with the whole dataset using bioactivity across a number of targets as well as other properties (tables 1 and 2). In cases where size > 1000 compounds, a random sample of 600 compounds was taken. Again, the results were filtered based on the values of the χ^2 test, in order to minimise the probability of spurious neighbourhood behaviour. Hence, the NB scores are statistically significant for a 95% confidence level. ECFP and FCFP circular fingerprints were consistently better. Surprisingly, they also behave as well as for the arrays across physicochemical properties such as solubility and permeability (figure 6). However, in case of lipophilicity and metabolic stability, the χ^2 values were consistently low, and the results were hence not included.

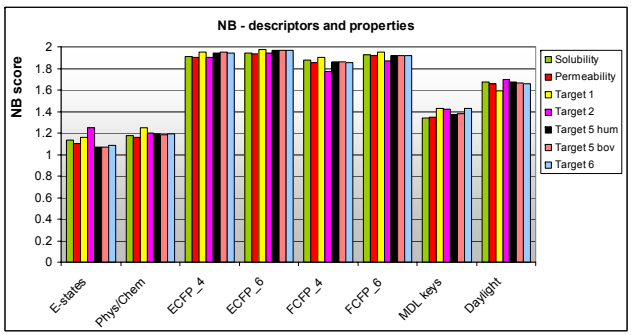


Figure 6

CONCLUSIONS

A method has been developed aiming to assess datasets and descriptors for NB. Initial results illustrate that ECFP fingerprints performed better across all datasets, arrays and properties in a statistically significant manner.
Future work will focus on applying the algorithm to more arrays using a set of diverse descriptors. Moreover, the role of χ^2 will be further explored.

[1] Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S10 2TN, UK
[2] GlaxoSmithKline, Gunnels Wood Road, Stevenage, SG1 2NY, UK

Acknowledgements
This work is funded by GlaxoSmithKline and EPSRC

References
1. Johnson, M.A. & Maggiora, G.M. (eds.) (1990). *Concepts and Application of Molecular Similarity*. New York: Wiley and sons.
2. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. & Weinberger, L.E. (1996). *Journal of Medicinal Chemistry*, **39** (16), 3049-59.
3. McLay, I.M., Halley, F., Souness, J.E., McKenna, J., Benning, V., Birrell, M., Burton, B., Belvisi, M., Collis, A. & Constan, A. (2001). *Bioorganic and Medicinal Chemistry*, **9** (2), 537-554.
4. Dixon, S.L. & Merz, K.M. (2001). *Journal of Medicinal Chemistry*, **44** (23), 3795-3809.