Bias Data Fusion with Turbo Similarity Searching

Jenny Chen[1], John Holliday[1] and John Bradshaw[2]

Similarity searching is perhaps the simplest tool available for lignad-based virtual screening of chemical databases, requiring just a single known bioactive molecule, the reference or target structure, as the starting-point for a database search. The most common similarity search involves the use of a simple association coefficient, normally the Tanimoto coefficient, with a 2D fragment bit-string representation of molecular structure. More recently, data fusion in similarity searching has emerged which uses more than one coefficient to evaluate the similarity between the target structure and the database structures. In addition, using multiple reference structures such as turbo similarity searching with group fusion has also been studied. In our experiments, we try to combine bias data fusion using the four coefficients concluded from earlier work with turbo similarity searching (Hert, J. et al., 2005; Hert, J. et al., 2006). Both the MIN and MAX fusion rules are experimented.

In our earlier work we selected 20 queries of varying sizes and used these as queries to find the best coefficient for each of the 20 retrieved active size ranges. The result is shown as below.

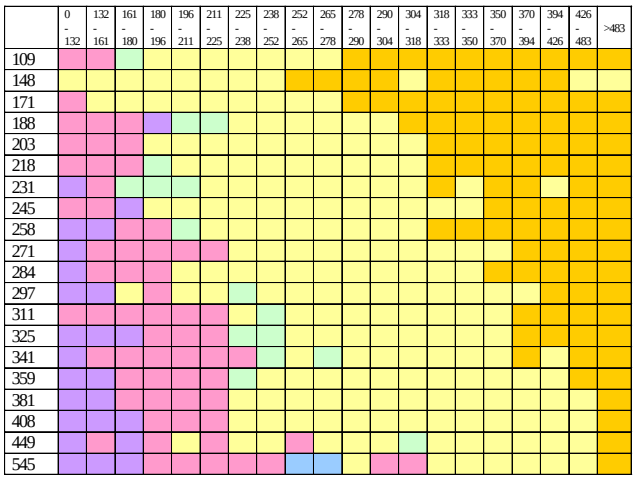

Figure 1. The results of machine learning approach for finding the best coefficients for data fusion.

Figure 1 demonstrates that when performing a similarity search, Forbes is the best coefficient when retrieving the actives that are much smaller than the query size, Simple Matching is the best coefficient when retrieving the actives that are generally smaller than the query size, Tanimoto is the best coefficient when retrieving the actives that are similar to the query size and Russell-Rao is the best coefficient when retrieving actives that are larger than the query size. In general, the choice of best coefficient can be found from this figure when the query size is known and the size of the retrieved actives is determined.

Due to their complementary effects in similarity searching, as shown in Figure 1, the Forbes, Simple Matching, Tanimoto and Russell-Rao coefficients are used in our following bias data fusion experiments. These experiments are combined with turbo similarity searching using both MIN and MAX rules.

For our experiments, a weighting scheme is applied to the four coefficients in steps of 0.25 from 0.0 to 1.0. The program runs through all possibilities of combination of the weights from the four coefficients and, for each of the combination, an initial MIN or MAX data fusion is carried out followed by turbo search (Appendex 1). The turbo similarity searching is repeat and continued until no improvement in retrieval performance is seen. The best-performing combination of the weights of the four coefficients is taken as the preferred the combination to use for that particular query size and for that particular class. This is the training part of the experiments.

A total of 55 queries from 11 classes are used for both MIN and MAX fusion in the training stage. The average improvement rate of the MIN bias data fusion with turbo similarity searching is 26.7% over Tanimoto with the best individual improvement rate of 151% over Tanimoto. The average improvement rate of the MAX bias data fusion with turbo similarity searching is 33% over Tanimoto with the best individual improvement rate of 284% over Tanimoto.

Table 1 contains some of the best combinations of weights from the MAX experiment for some queries. Figure 2 is the size distributions of four classes where the blue line is the size distribution of MDDR database.

Data from MAX Experiment				
Query Size	For. Wt.	SM Wt.	Tan Wt.	Rus. Wt.
Renin Inhibitor				
90	0.0	0.0	0.75	0.25
135	0.0	0.5	0.25	0.25
Angiotensin II AT1 Antagonist				
90	0.0	0.0	0.75	0.25
135	0.0	0.75	0.0	0.25
5HT1A Agonist				
90	0.25	0.0	0.75	0.0
138	0.0	1.0	0.0	0.0
5HT Reuptake Inhibitor				
90	0.25	0.25	0.5	0.0
139	0.5	0.5	0.0	0.0

Table 1. Some examples of best combination of weights for the four coefficients on query size around 90 bits and 135 bits from the MAX bias data fusion with turbo similarity searching.

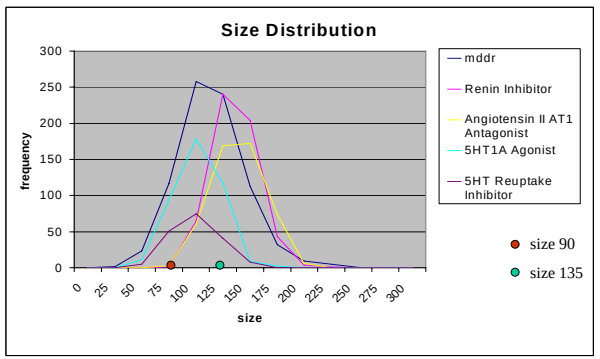

Figure 2. Size distributions of four classes where the dark blue line is the size distribution of MDDR database.

From Figure 2, the query of size 90 bits is far smaller than the size distribution of Renin Inhibitor; hence, in Table 1, this query needs the weight of 0.75 for Tanimoto and 0.25 for Russell-Rao where Tanimoto and Russell-Rao retrieve medium and large sizes of actives when comparing with the query size. For the same query size 90 bits for class 5HT1A Agonist, the query is slightly smaller than its class size distribution; hence, this query needs the weights of 0.25 for Forbes and 0.75 for Tanimoto since Forbes and Tanimoto retrieve small and medium sizes of actives when comparing with the query size. The same methodology can be applied to the larger query size at around 135 bits. This phenomenon explains that the best weight and the choice of coefficients for MAX bias data fusion with turbo similarity searching depend on the relationship between the query size and the position of its class size distribution.

In conclusion, bias data fusion using the four coefficients, Forbes, Simple Matching, Tanimoto and Russell-Rao, in combination with turbo similarity searching, it is possible to perform better than the industrial standard, Tanimoto. In addition, the training results from the MAX bias data fusion with turbo search clearly illustrate the relationships among the query size, its class size distribution and the best weights for the 4 coefficients.

The future research will continue with the testing part of the experiments. Selecting actives that have the same sizes as the training queries as the testing queries, and using only the best combination of the weights from the training results for that particular query size and that particular class to carry out MIN and MAX bias data fusion with turbo search. This is to see if a combination of weights of the coefficients can be repeated used within the same class and same query size.

Appendix 1. The procedures for carry out Data Fusion with Turbo Similarity Searching:

Input the reference structure R
Perform data fusion using specific weights for each of the coefficients to give a sorted database *SD (0)*
Identify the k NNs of R from the top of the list *SD (0)*
  Compute the similarity of *NN (i)* with every molecule in D
  Sort D in decreasing order of the calculated simliarty values to give a sorted database *SD (i)*
Fuse the sorted lists *SD (0)* – *SD (k)* to give the final out put from the trubo simialrity search

Reference

Hert, J, Willett, P, Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information. J. Med. Chem. 2005, 48, 7049-7054.

Hert, J., Willett, P., and Wilton, D. J. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity seraching. J. Chem. Inf. Model. 2006, 46, 462-470.

Salim, N., Holliday, J. D., Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. J. Chem. Inf. Comput. Sci. 2003, 43, 435-442.

1 Department of Information Studies, University of Sheffield, Sheffield, S1 4DP. UK

2 Daylight Chemical Information Systems Inc., Cambridge, UK