An Upper Bound to the Effectiveness of Substructural Analysis Methods

Nor Sani, John Holliday and Peter Willett

Information School, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK.

1. Introduction

Substructural analysis (SSA) was one of the first machine learning techniques applied in chemoinformatics. SSA is a method that uses compound's biological activity and molecular structure information for the prediction of biological activity. The relationship between substructures and activity state for every compound of a given biological activity class are established using weights calculated via a weighting scheme. A variety of weighting schemes are available for this purpose[2]. The approach, first described some 40 years ago[1], is very closely related to a naïve Bayesian classifier (NBC)[3], a machine learning method that has become very popular in the last few years with its availability in the Pipeline Pilot system software [4].

This poster reports on recent work to identify an upper-bound to the effectiveness of SSA methods using a genetic algorithm (GA) we developed for the calculation of fragment weights based on 2D fingerprints.

2. Experimental Details

2.1 The Comparison of existing SSA weighting schemes

Experimental Procedure

For the first part of our study, we present an updated analysis on the effectiveness of established SSA weighting schemes to measure the predictive performance of a given biological activity class. Various weighting schemes are readily available in SSA, however, in this study, we evaluated nine schemes that performed well in the comparative study of previous work [2,3].

For the experiments, we used two difference datasets which are the MDL Drug Data Report (MDDR) and the World of Molecular Bioactivity database (WOMBAT), these comprising of 11 and 14 activity classes respectively. Both retrospective and predictive experiments were conducted for our analysis, with the general SSA workflow described in Figure 1 below:

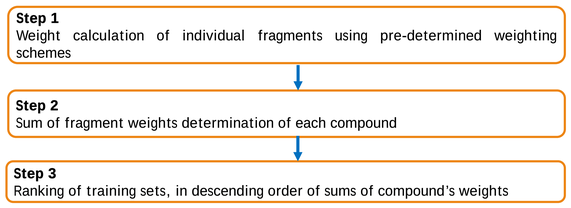

Figure 1: Methodology of the SSA procedures conducted for the retrospective / predictive analysis

For the retrospective experiment, the training set is made up of the entire dataset of either MDDR or WOMBAT, amounting to 102,540 and 138,127 compounds, respectively, this comprising both the active and inactive compounds of the given biological activity. For the predictive experiments, we generated two sets of predictive studies (dubbed Predictive_1 and Predictive_2 sets), each containing 10% randomly selected actives of a given activity and 10% of the inactives. The compounds in Predictive_1 sets are also unique from the Predictive_2 ones. The test-set, is comprised of the remaining 90% of the dataset.

Experimental Results

From the experiments, we observed the most effective weighting scheme across the eleven MDDR and fourteen WOMBAT bioactivity classes to be the R4 weight, followed by R3, R1 and R2. Then follows AVID, WT2, WT1, SAF and subsequently SAS as the worst performers. Such results are shown in Figure 2 (a and b).

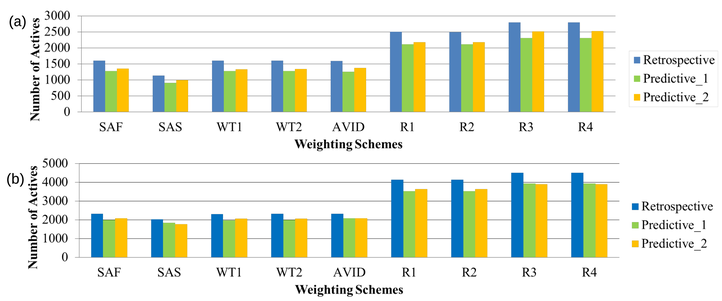

Figure 2: Comparison of SSA weighting schemes based on the number of actives retrieved in the top 1% ranking of activity classes from the two databases; (a) MDDR database and (b) WOMBAT database

2.2 SSA-based Genetic Algorithm (GA) for the Prediction of Biological Activity

Experimental Procedure

In this section of the work, we seek to evaluate any possibility of uplift in the approach of a GA on fragment weighting determination when compared to the R4 weight (originally carried out by Robertson and Sparck Jones in the context of text search engines) and the Pipeline Pilot NBC.

The chromosome for the GA is a vector containing *n* real numbers, where the *i*-th element is the fragment weight for the *i*-th bit in the fingerprint. The fitness function for the GA is the number of active molecules that occur in the top 1% of a ranked training sets based on the *N* weights representing the chromosomes. The GA is run for a pre-set number of generations or until the weights have stabilized, thus providing an estimate of the best possible SSA weights that can be obtained using that training-set. The resulting weights can then be applied to a separate test-set. A summary of the workflow is described below in Figure 3:

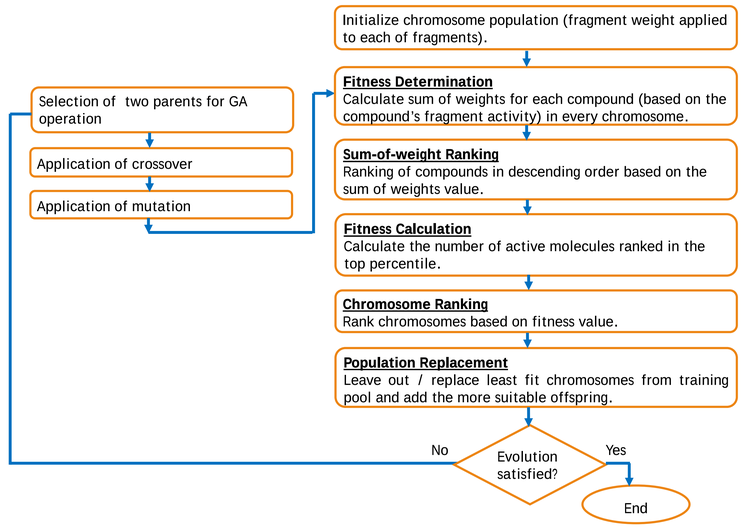

Figure 3: The flowchart of the Genetic Algorithm approach to Substructural Analysis

The basic idea is illustrated using a training-set containing three molecules $M_{1\text{-}3}$ (Table 1a), each represented by a 2D fingerprint encoding the presence or absence of five fragments $F_{1\text{-}5}$. In the example, the population contains six randomly initialized chromosomes, $C_{1\text{-}6}$, as shown in Table 1(b). Each chromosome is then used to compute the sum-of-weights for every molecule, as shown in Table 1(c). Considering the case of chromosome C_1 for example, the sums-of-weights for M_1, M_2 and M_3 are 3, 6 and 1 respectively, this then followed by the ranking of the training set: $M_2 > M_1 > M_3$. C_2 yields the ranking: $M_3 > M_2 > M_1$, and so on.

Molecule	F_1	F_2	F_3	F_4	F_5
M_1	0	1	0	0	1
M_2	1	0	0	1	0
M_3	0	0	0	1	1

(a)

Chromosome	W_1	W_2	W_3	W_4	W_5
C_1	6	2	7	0	1
C_2	4	3	1	8	5
C_3	9	9	3	6	7
C_4	1	7	5	1	3
C_5	8	4	8	2	8
C_6	5	8	4	7	2

(b)

Chromosome	M_1	M_2	M_3
C_1	3	6	1
C_2	8	12	13
C_3	16	15	13
C_4	10	2	4
C_5	12	10	10
C_6	10	12	9

(c)

Table 1: The GA operation based on a population containing three molecules, with six chromosomes created at initialisation

Parameterization Analysis

We used MDDR-based RNN and COX activity classes for extensive experimentation and optimisation of the suitable GA parameters, subsequently used for the remaining activity classes over both databases. From the experiments, we found that the best GA parameters were the following: (1) *population size* of *200*; (2) *a maximum* of *500 GA iterations*; (3) *parents selection* using *the roulette wheel method*; (4) *crossover rate* of *0.95*; (5) *mutation rate* of *0.01*; (6) *crossover method* using *one-point crossover;* and *weights in the range -100 to +100.*

This parameter set was then subsequently applied for all the other activity classes over both databases, conducted for the predictive_1 and predictive_2 training sets. Subsequently, the obtained predictive GA-based weighting values were applied and evaluated in the test set.

3. Experiment Results

Enrichment Curve

Performance measures of the enrichment curves are highlighted in Figure 4, summarizing three examples of activity classes in which weighting values derived from Predictive_1 studies were applied on the test set. In general, we observed a consistent trend of higher active retrieval rates of the GA in the first 1% rankings for almost every activity class.

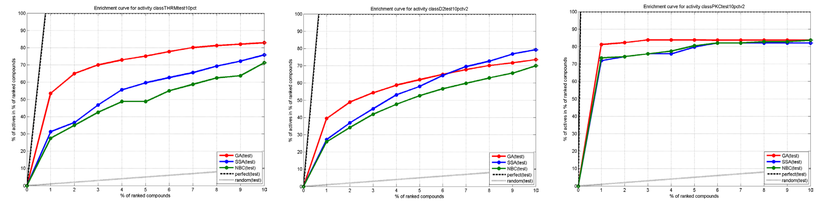

Figure 4: Cumulative recall plots measuring the active compounds retrieval rates for three given weighting schemes, the GA (red curve), the SSA R4 (blue curve) and the NBC (green curve). Weight values are derived from Predictive_1 studies, and applied on a test-set data; (a) MDDR-based THRM activity class plot; (b) WOMBAT-based D2 activity class plot; (c) MDDR-based PKC activity class plot

Kendall-W Analysis

We refer to Table 2 for the overall Kendall W results, the GA-based weighting scheme is consistently at the top for both studies; this is followed by the SSA R4 and then the NBC method. We observed statistically significant differences between the performance of the various weighting schemes, ($p \leq 0.001$). Following the mean ranking, we denote the following SAA weighting schemes measure of performance:

GA > SSA R4 > NBC

Weighting Scheme	MDDR 1%				WOMBAT 1%				Mean	
	Predictive-1		Predictive-2		Predictive-1		Predictive-2			
	Actives	Rank	Actives	Rank	Actives	Rank	Actives	Rank	Actives	Rank
GA	291.45	2.00	281.45	2.00	347.86	2.00	339.14	2.00	314.98	2.00
SSA R4	206.00	0.55	228.91	0.82	273.29	0.82	278.14	0.79	246.59	0.74
NBC	178.87	0.45	187.82	0.18	22.74	0.18	235.46	0.21	206.22	0.26

Table 2: Kendall's W analysis for the top 1% based on the average actives retrieved from the MDDR and WOMBAT databases

Conclusion

We draw two conclusions from these experiments. Firstly, the GA provides a possible non-deterministic method for generating the fragment weights for use in SSA-based virtual screening. Second, and more importantly, the results of the GA in general are observed to be slightly more effective than those obtained from existing, deterministic methods for generating such weights.

References

1.Cramer, R. D.; Redl, G.; & Berkoff, C. E. *Journal of Medicinal Chemistry,* 1974, *17*(5), 533-535.
2.Ormerod, A.; Willett, P.; & Bawden, D. *Quantitative Structure-Activity Relationships,* 1989, *8*(2), 115-129.
3.Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. *Journal of Chemical Information and Modeling* 2006, *46*, 462-470.
4.Rogers, D.; Brown, R. D.; Hahn, M. *Journal of Biomolecular Screening,* 2005, *10*, 682-686.

Acknowledgments

• Information School, University of Sheffield, Sheffield S10 2TN, UK.
• The National University of Malaysia, 43600 UKM, Bangi Selangor, Malaysia.
• Pipeline Pilot software and the MDL Drug Data Report database were provided by Accelrys, Inc.