J. Appl. Environ. Biol. Sci., 7(2S)30-36, 2017

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

The Use of Data Fusion on Multiple Substructural Analysis Based GA Runs

Nor Samsiah Sani

Faculty of Information Science and Technology, National University of Malaysia, Bangi, Selangor, Malaysia

Received: February 21, 2017
Accepted: April30,2017

ABSTRACT

Substructural analysis (SSA) was one of the very first machine learning techniques to be applied to chemoinformatics in the area of virtual screening. Recently, the use of SSA method based on genetic traits particularly the genetic algorithm (GA) was shown to be superior to the SSA based on a naive Bayesian classifier, both in terms of active compound retrieval rate and predictive performance. Extensive studies on data fusion have been carried out on similarity-based rankings, but there are limited findings on the fusion of data obtained from evolutionary algorithm techniques in chemoinformatics. This paper explores the feasibility of data fusion on the GA-based SSA. Data fusion is a method to produce a final ranking list from multiple sets of ranking lists via several fusion rules. Based on the encouraging results obtained using the GA, the application of data fusion to the GA-based SSA weighting schemes are examined in this paper in order to enhance retrieval performance of 2D-based fingerprint predictive method. Our experiments used the MDDR and WOMBAT datasets. The results show that data fusion can indeed enhance retrieval performance of evolutionary techniques further in the case of evolutionary algorithm techniques, and specifically the GA-based SSA.

KEYWORDS: Chemoinformatics, Substructural analysis, Evolutionary Algorithm, Genetic Algorithm, Data Fusion.

INTRODUCTION

Substructural analysis (SSA) is a method under ligand-based virtual screening, pioneered by [5]. The technique is one of the earliest forms of machine learning method used in chemoinformatics. In SSA, it is assumed that each molecule in a dataset is characterized by a series of binary descriptors, most commonly in the form of a 2D fingerprint in which each bit denotes the presence or absence of a substructural feature (often referred to as a fragment). Associated with each such bit is a weight that is a function of the number of active and inactive molecules that have that bit switched on, i.e., that contains the corresponding fragment. This weight reflects the probability that a molecule containing that substructural feature will be active (or inactive); for example, the weight might be the fraction of the active molecules containing that particular fragment. A molecule is then scored by summing (or otherwise combining) the weights of those bits that are set in its fingerprint, the resulting score representing the overall probability that the molecule will be active 1. A major assumption of SSA is that a given substructure can influence the determination of the activity level of a molecule, regardless of the compound in which it occurs. A variety of weighting schemes based on specific relationship-bound equations are available for this purpose. In [10] looked at expanding the SSA method by applying a GA-based weighting scheme to determine the suitable set of fragment weights for any possibility of an upper-bound to the activity prediction of the SSA. From the study, it can be concluded that the GA-based SSA method is superior to the SSA R4 scheme as it successfully managed to provide uplift of active retrieval performance in the top 1% of ranked molecules. Unlike the SSA, the GA-based approach is considered to be an inherently non-deterministic process. High correlation and consistency between multiple GA runs, however, means that the method is reliable and effective as an alternative weighting scheme to the SSA method.

Data fusion is a method of combining the information gained from different sensors to achieve an effective or improved decision, compared to when only a single sensor is considered [7]. This method can be utilized for ligand-based virtual screening. The sensors to be combined are used as functions that score molecules in a database on their likelihood of exhibiting some required biological activity. The combination of different sources of information is already practiced in most human daily activities such as in decision-making processes. A simple example is the use of different sensors in our everyday lives that include our sense of smell, taste, feeling, hearing and seeing. In a more practical sense, for instance, a manager considering the hiring of new employees makes informed decisions based on the different traits of the candidate such as their skills, experience and communication abilities. These traits collectively produce a decision about the candidate's

Corresponding Author: Nor Samsiah Sani, Faculty of Information Science and Technology, National University of Malaysia, Bangi, Selangor, Malaysia, E-mail: norsamsiahsani@ukm.edu.my

eligibility to be hired. Data fusion is increasingly used to combine the outputs of different types of digital or analogue sensors.

Data fusion has been successfully used in different fields such as medicine, defense and information retrieval. The findings on combining more than one query in the field of information retrieval were first discussed by [1]. The study on data fusion was carried out in two different projects at Rutgers University and the Virginia Technology Institute. Together, these projects found that fusing the multiple queries is far more effective in increasing search performance, yielding better retrieval rates than using a single query.

In virtual screening, many studies related to data fusion have been carried out, especially regarding similarity searching. Similarity search is based on three main components: the molecule representation used to describe the molecular structures, the weighting scheme used to compute the score of a particular compound structure to produce compound rankings, and the similarity coefficient used to calculate the degree of similarity between the reference molecule and the database molecules. Furthermore, data fusion in similarity searching can be divided into similarity fusion and group fusion. Similarity fusion is the combination of scores gathered from multiple similarity measures by using a single reference structure for searching a chemical database [20]. For instance, the data fusion ranking is obtained by combining three rankings from different similarity coefficients, for example tanimoto, dice and cosine. Several studies on similarity fusion were carried out by fusing different similarity coefficients in a similarity search [6, 16, 17]. The group fusion approach fuses rankings produced from different reference structures by using the same similarity coefficient and molecular representation [8]. Group fusion can utilize either similarity scores or rankings [22]. For instance, assuming one type of 2D descriptor such as the MDL fingerprints, the similarities between reference structure and other structures in the database are measured using the Tanimoto coefficient. They are then ranked in descending order based on their similarity score.

Comparing the two fusion techniques, similarity fusion tends to perform better than group fusion when the actives are strongly clustered structurally [21]. By contrast, group fusion is best employed when the actives are structurally diverse [9]. Numerous studies have compared these two data fusion techniques in similarity searching. Other studies have found that group fusion is effective as a general approach in similarity searching [2, 8, 20, 24]. Based on the encouraging results obtained using GA in the recent work [10], the application of data fusion to the GA-based SSA weighting schemes are examined in this paper in order to enhance the retrieval performances of 2D-based fingerprint predictive method. Extensive studies on data fusion have been carried out on similarity-based rankings, but there is still a lack of findings on data fusion using genetic algorithm techniques in chemoinformatics.

RESULTS AND DISCUSSION

Three large datasets were used for the evaluation of the methods in question. The datasets used are as follows: (i) the MDL Drug Data Report database (MDDR); (ii) the World of Molecular Bioactivity database and (WOMBAT). The MDDR and WOMBAT datasets used here are described in detail by [6]: the MDDR dataset contains eleven activity classes and 102,514 molecules while the WOMBAT dataset contains 14 activity classes and 138127 molecules.

The molecules from MDDR and WOMBAT datasets were characterized via dictionary-based fingerprints known as the MDL fragment description. The MDL structural keys used in this study was originally developed for a substructure search [14]. The MDL keys consist of 166 bit keysets based on 166 publicly available MDL MACCS structural keys. The structural keys are important fragments listed in a dictionary used to encode molecules in a bit-string. Each bit is associated with a structural key and it denotes the presence or absence of one of the keys or substructure. The MDL fingerprints were used to identify the combination of fragment weights to generate the best possible ranking of the molecules in a database. The MDL fingerprints were generated using SciTegic's Pipeline Pilot software to produce structural descriptors or fragments for all compounds. Pipeline Pilot protocols were used to retrieve the MDL fingerprints from the MDDR database. The protocol involves the process of converting a Daylight SMILES notation found in the property list to a molecular representation. The MDL public key fingerprint component was then used to convert molecules into 166-bit MDL fingerprints, denoting the present fragments as '1' and the absent fragments as '0' in each of the 166 fragments.

In order to perform data fusion, it was necessary to extract the ranking output of the ten runs of the GA for each activity class to be fused. Based on the encouraging results obtained using the GA3, the application of data fusion to the GA-based SSA weighting schemes are examined in this paper in order to enhance retrieval performance of 2D-based fingerprint predictive method. Results of the GA ten runs are summarized in three important results which are (i) the enrichment factor of active molecules in the top 1% (ii) the mean and standard deviation of the number of actives in the top 1% for the ten GA runs and (iii) the mean correlation and standard deviation between 166 weights using Pearson correlation coefficient for the ten GA runs. Based on results, high correlation values were observed, in which the mean correlation of Pearson's r recorded a

minimum of 0.74 and on average circa 0.78. Some classes were also observed to record a mean Pearson's r as high as 0.86. From the mean and standard deviation values for the ten GA runs, it can be seen that there is a high degree of consistency of the number of actives were retrieved in the top 1% of the ranked data. Also, it indicates that the variation of actives retrieved in the top 1% is less dispersed.

Five types of fusion rules namely the SUM, MAX, MED, MIN and RKP rules were identified. Most rules were first discussed by [1], however, the RKP rule was initially described by [12]. These rules are presented in Table 1. In the table, d_j denotes an individual compound listed in the sets of machine learning technique rankings, ML_i $\{d_j\}$ which consists of n GA rankings. Observing the first fusion rule, SUM computes the mean value of the compound scores or ranks in the list. In this case, this is achieved by aggregating all the scores of each database structure, then dividing the score by n. For the MAX, MIN and MED fusion rules, the scores for each database structure d_j are computed by taking the largest, the smallest and the middle score (or median) in the n rankings respectively. The final rule used for consensus scoring is known as the RKP fusion rule, whereby a compound d_j score is computed by adding the reciprocal of the non-zero scores after the ranking is truncated to a certain percentage p; for instance, 100% (i.e. the whole database), 50%, 5% and 1%. Notably, the formula of the RKP rule is measured by using the rank position of each molecule to be fused as used by [12] in text retrieval.

Table 1: Fusion rules

Fusion Rule	Formula
SUM	$\frac{1}{n}\sum_{i=1}^{n} ML_i(d_j)$
MAX	$\max\{ML_1(d_j), ML_2(d_j), \ldots ML_i(d_j), ..\}$
MED	$\text{med}\{ML_1(d_j), ML_2(d_j), \ldots ML_i(d_j), ..\}$
MIN	$\min\{ML_1(d_j), ML_2(d_j), \ldots ML_i(d_j), ..\}$
RKP	$\sum_{i=1}^{p} \frac{1}{ML_i(d_j)}$

Several studies have reported on success of similarity fusion using the SUM fusion rule in applications of similarity searching [6]. Other studies have reported the MAX rule to be the best fusion rule for group fusion in similarity searching [8, 11, 24]. Several comparisons on consensus scoring were also reported with applications in docking [13, 26] and in 2D and 3D similarity searching [27]. In [4] reported the RKP fusion to be the most effective fusion rule for combining multiple document rankings from an information retrieval. In [2] found that group fusion can even be superior to similarity fusion.

Following the fusion rules criteria, it was determined that two variables could be used in the computation of the GA-based fusion scores: (1) the score of compounds which is the sum of GA weights or (2) the ranking of compounds in the ten sets of GA runs. The first four rules, SUM, MAX, MED and MIN were used to fuse the ten sets of GA runs using both score-based and rank-based data. For the RKP equation, the rule is applicable only on the fusing of n sets of ranks. Hence, the RKP rule was applied with rank-based data only. For this study, p value was set at 100% which otherwise means that the whole database of ranked outputs were fused. In total, nine fusion rules were employed in which a number of the rules are based on ranking information of the data with the rules listed as rank RKP, rank max, rank sum, rank med and rank min. The other fusion rules are based on the scoring information of compounds in a dataset where these fusion rules are referred to as score max, score sum, score med and score min.

In evaluating the GA-based fusion search method, the effectiveness criteria was stressed on quantifying whether the actual output meets the desired output or otherwise. Area under the curve (AUC) also known as receiver operating characteristics (or ROC) [3] is a standard evaluation method used in machine learning experiments. It is however less suitable for virtual screening evaluations as it only considers the full ranking of a database. In fact, methods of virtual screening require only the analysis of a small fraction of the molecules that occurs at the top of the ranking to be considered for further biological screening [19]. Rather than using AUC values, the screening performance was hence measured by the number of actives for the top 1% of the ranked test set (i.e. 1% enrichment value).

The output of the ten GA-based SSA weighting schemes for each activity class were combined and employed for data fusion, using the nine fusion rules mentioned earlier. The enrichment factor of actives retrieved in the top 1% obtained by GA-based fusion based on various fusion rules. The highest values are shown as lightly shaded. Each row of one of the sections of the tables corresponds to a single activity class and lists the total number of actives in the test set; the mean enrichment factor of actives retrieved in the top 1% using a combination of ten runs of the GA weighting schemes and the remaining rows show the results of the data fusion procedures. The fusion results are listed and compared against the mean of the ten GA runs results for benchmarking. The highest values are shown as lightly shaded.

The outcome indicates three observations. Firstly, it was observed that the performance of the GA based fusion is seen to be more effective than the mean GA results for all activity classes in all three databases. Based on enrichment factor of actives in the top 1% for all activity classes, the differences between data fusion methods to GA-based SSA is often very small. On average, there is about one to fifteen active compounds retrieval differences recorded. Secondly, when comparing both the mean of ten runs of the GA and the RKP method in each activity class, it is interesting to note that the RKP rule performs better than these mean runs in all cases. Thirdly, in a comparison of all the nine fusion rules, it was observed that the score min and score med rules jointly yielded the worst performances in recall rate for most classes from both databases using the GA-based methods.

Kendall's W analysis the impact of the GA-based fusion were studied further by employing the Kendall's W test of statistical significance to measure the agreement of the fusion rules performance in all three databases. This coefficient provides a means of quantifying the degree of association between k variables or k sets of rankings of similar objects. Accordingly, Kendall's W calculates the agreements between rankers as it evaluates and ranks a number of subjects according to particular characteristics. The concept is that n subjects are ranked (0 to n-1) by each of the rankers, and the statistic evaluates how much the rankers agree with each other. Kendall's W ranges from 0 to 1, where 0 indicates no agreement and 1 indicates complete agreement.

Specifically, the weighting schemes from each database are ranked in decreasing order of effectiveness of virtual screening for a specific activity class. This is repeated for each class so that there are e.g. 11 rankings for the MDDR dataset. The degree of agreement between the rankings in the top 1% of the ranked compounds is measured by calculating the Kendall Coefficient of Concordance, W. This coefficient provides a means of quantifying the degree of association between sets of rankings of the same objects [25]. If there is an agreement between the rankings of the weighting schemes, it can be concluded that there is a statistical significant result for the null hypothesis, H_0. This predicts the probability that the rankings are not associated, and can thus be rejected. In this analysis, 0.001, 0.01, 0.1 and 0.5 were selected as the significance level. Therefore, if the probability p value is equal to or less than 0.001, it is then necessary to reject the null hypothesis and then can give overall ranking. However, if the p value is more than 0.5, then the computed results are considered insignificant. The equation that has been used to compute the degree of variance among the ranks is given by Equation (1):

$$W = \frac{12 \sum R_i^2 - 3k^2 \times N (N+1)^2}{k^2 \times N (N^2-1)} \quad (1)$$

where k is the number of ranks, for example, 11 activity class in MDDR dataset, N is the number of objects being run; for example 9 weighting schemes were evaluated in this study and is the sum of the squares sums of ranks for each of the N objects.

The significance of the W was computed using a X^2 distribution with a degree of freedom df =N-1 for which the equation (Equation 2):

$$X^2 = k(N-1)W \quad (2)$$

If the size of the samples is larger (N >7), then the chi square and the probability p values were identified by referring to the chi square distribution table. Otherwise, the table of critical values was used to identify the probability [18]. Whenever W is larger than the critical values, this result would be considered significant and thus the null hypothesis would be rejected.

For the GA-based fusion, the results obtained from Kendall's W analysis were analyzed. The rankings are determined based on the enrichment factor of actives in the top 1% (or 1% cut-off value). The rankings were listed in decreasing order.

Kendall's W analysis of the fusion rules in MDDR classes is calculated, in which the total computed value of W is 0.46. The significance of this value was tested using X^2 distribution, giving a value of 45.38 for X^2 at a significance level of $p < 0.01$. The analysis therefore suggests the following ranking:

Rank RKP > Score Max > Rank Max > Score Sum > Rank Sum > Rank Min > Score Med > Rank Med > Score Min > Mean GA

Similar to the MDDR case, the results for the fusion rules in WOMBAT-based classes were calculated. The value obtained for W is computed as 0.30 and the significance of the X^2distribution is valued at 42.26 at a significance level of $p < 0.01$. The following ranks the fusion rules, from the best to worst performing ones:

Rank RKP > Rank Max > Score Max > Score Sum > Rank Sum > Score Min > Rank Min > Score Med > Rank Med > Mean GA

In essence, at a significance level of $p < 0.01$, the Rank RKP was found to be the best performing rule for the GA-based fusion in MDDR datasets and WOMBAT sets. The rest of the fusion rules exhibit mixed results across all three databases. The worst performing method can be seen in the mean GA and score med, which were consistently placed in the lower tier of the ranking position.

Table 2 highlighted fusion rules based on the mean rank for all three databases (i.e., MDDR and WOMBAT). Here, it was possible to obtain the following observations for GA-based fusion Table 2 whereby the calculated value $W = 0.94$ which gives a value of $X^2 = 25.40$. The results denote a significant value at $p < 0.01$. Subsequently, the best overall ranking of data fusion using GA-based fusion is as follows:

Rank RKP > Rank Max > Score Max > Score Sum > Rank Sum > Rank Min > Score Med > Rank Med > Score Min > Mean GA

Table 2: Kendall's W analysis for the top 1% based on the average of enrichment factor actives in the top 1% of the GA-based SSA from the MDDR and WOMBAT databases

Fusion	Databases		Mean	Rank
Rules	MDDR	WOMBAT	Rank	Position
Rank RKP	6.95	6.50	6.73	1
Rank Max	6.36	6.29	6.33	2
Score Max	6.64	5.29	5.97	3
Score Sum	5.73	5.25	5.49	4
Rank Sum	5.64	5.00	5.32	5
Rank Min	3.95	3.82	3.89	6
Score Med	3.50	3.46	3.48	8
Rank Med	2.95	3.43	3.19	7
Score Min	2.00	3.89	2.95	9
Mean GA	1.27	2.07	1.67	10

It was concluded that when comparing all the nine fusion rules using the ten runs of GA-based SSA methods in all MDDR and WOMBAT activity classes, the RKP rule was found to perform better than other rules in most cases. This is in agreement with the results reported by [2], whereby they found that RKP is superior to the other rules in group fusion. In contrast to the best fusion rule determined, it was consistently observed that both MED and MIN rules jointly yielded the worst performances in recall rate for most classes from both databases. This occurred when applying fusion on GA-based SSA compounds ranking results.

Based on the Kendall's W analysis performed, the GA-based data fusion was deemed more effective than the mean of ten GA-runs in all classes of the databases. Thus, it was necessary to use the Wilcoxon signed rank test to quantify the significance of the difference between the performance of data fusion and the mean of the ten GA runs. To conduct the test, the enrichment factor results in the top 1% from both the mean of ten runs GA against the best data fusion rule were observed. A measure of significance following W is measured by referring to the table of critical values for the Wilcoxon test (i.e. Wcritical). It is necessary to refer to the table of critical values of W. This serves to gauge the level of rejection of the test statistics in order to arrive at the alternate hypothesis. Using the information of the number of differences, N; a probability value with the lowest value of the significance level of 0.01, rejects the null hypothesis H0, if the value of W is less than or equal to the critical value of Wcritical [15].

In the Wilcoxon signed rank test, if two scores of any pair are equal (i.e. there is no difference between the two compared entities), then such pairs are discarded from the analysis. These were consequently ignored. A null hypothesis(Ho) is defined as where the median difference is zero. This means that our default assumption is that both results of weighting schemes are significantly identical. The alternate hypothesis (H1) is defined as the median difference being positive at a significance level of $p = 0.01$ [15]. For the case of the MDDR database and comparing the mean of ten GA runs against the RKP method, the Wilcoxon signed rank test showed a value of $W = 0$ and the critical value of W for $N = 11$ at $p < 0.01$ is 5. In the case of WOMBAT, the value of $W = 3$ and the critical value of W for $N = 14$ at $p < 0.01$ is 12. Overall, the data fusion results appear to be significant when compared against mean results of the GA runs. Hence, it can be concluded that rank RKP fusion rules provide good enhancement of recall rates when compared with the mean of individual GA.

CONCLUSION

This paper described the investigation of data fusion, which sought to combine retrieval results from multiple, individual GA-based SSA results for each activity class. From the experiment and various analyses performed, it can be concluded that, the data fusion was found to perform better in each activity class from the three databases utilizing the rank RKP. By contrast, the fusion rules MED (i.e., rank med and score med) and MIN (i.e., rank min and score min) showed the worst performances relative to the three MDD and WOMBAT

databases. It was also found that for the comparison of data fusion to the mean of ten runs of the GA, the difference was found to be significant for the GA-based fusion. This is in agreement with the results reported by [25]. It was found that the diversity of the relevant documents in the ranked list of documents can affect the performance of data fusion. Better performance of data fusion is more likely with a higher rate of diversity in the fused input.

The GA is essentially a robust and non-deterministic process. Hence, data fusion can be used as a deterministic measure to produce a single, unified outcome. It was found that the fusion of multiple rankings of the GA-based SSA produced a significant improvement in the final ranking results with easy implementation. These conclusions confirm that the data fusion approach SSA is found to be highly effective technique in enhancing the retrieval performance of SSA specifically for the GA-based SSA. The findings of this experiment could be used to help the standard practice of data fusion in virtual screening [23], and to guide further enhancement in SSA.

REFERENCES

1. Belkin, N.J., P. Kantor, E.A. Fox and J.A. Shaw, 1995. Combining the Evidence of Multiple Query Representations for Information Retrieval. Information Processing and Management, 31 (3): 431-448.
2. Chen, B., C. Mueller and P. Willett, 2010. Combination Rules for Group Fusion in Similarity□Based Virtual Screening. Molecular Informatics, 29(6□7): 533-541.
3. Clark, R.D. and D.J. Webster-Clark, 2008. Managing Bias in ROC Curves. Journal of Computer-Aided Molecular Design, 22 (3-4): 141-146.
4. Cormack, G.V., C.L. Clarke and S. Buettcher, 2006. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In the Proceedings of the 200632nd International ACM SIGIR Conference on Research and Development In Information Retrieval, pp: 758-759.
5. Cramer, R.D., G. Redl and C.E. Berkoff, 1974. Substructural Analysis. A Novel Approach to the Problem of Drug Design. Journal of Medicinal Chemistry, 17 (5): 533-535.
6. Gardiner, E.J., V.J. Gillet, M. Haranczyk, J. Hert, J.D. Holliday, N. Malim, Y. Patel, C.M.R. Ginn, P. Willett and J. Bradshaw, 2000. Combination of molecular similarity measures using data fusion. In: Perspectives in Drug Discovery and Design (ed G. Klebe) pp. 1-16. Springer, Amsterdam.
7. David L. Hall and Sonya A.H. McMullen, 2004. Mathematical techniques in multisensor data fusion. Artech House.
8. Hert, J., P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, 2004. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. Organic and Biomolecular Chemistry, 2 (22): 3256-3266.
9. Hert, J., P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, 2006. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. Journal of Chemical Information and Modeling, 46 (2): 462-470.
10. Holliday, J.D., N. Sani and P. Willett, 2015. Calculation of Substructural Analysis Weights Using a Genetic Algorithm. Journal of Chemical Information and Modeling, 55 (2): 214-221.
11. Nasr, R.J., S.J. Swamidass and P.F. Baldi, 2000. Large Scale Study of Multiple-Molecule Queries. Journal of Cheminformatics, 1 (1): 1-19.
12. Nuray, R. and F. Can, 2006. Automatic Ranking of Information Retrieval Systems Using Data Fusion. Information Processing and Management, 42 (3): 595-614.
13. Oda, A., K. Tsuchida, T. Takakura, N. Yamaotsu and S. Hirono, 2006. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes. Journal of Chemical Information and Modeling, 46 (1): 380-391.
14. Olah, M., C. Bologa and T.I. Oprea, 2004. An Automated PLS Search for Biologically Relevant QSAR Descriptors. Journal of Computer-Aided Molecular Design, 18 (7-9): 437-449.
15. R. Lyman Ott and Micheal T. Longnecker, 2015. An introduction to statistical methods and data analysis. Nelson Education.
16. Salim, N., J. Holliday and P. Willett, 2003. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. Journal of Chemical Information and Computer Sciences, 43 (2): 435-442.
17. Sheridan, R.P. and S.K. Kearsley, 2002. Why Do We Need So Many Chemical Similarity Search Methods? Drug Discovery Today,7 (17): 903-911.

18. S. Siegel and N. John Castellan, 1988. Nonparametric statistics for the behavioral sciences. McGraw-Hill Book Company.
19. Truchon, J.F. and C.I. Bayley, 2007. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. Journal of Chemical Information and Modeling, 47 (2): 488-508.
20. Whittle, M., V.J. Gillet, P. Willett, A. Alex and J. Loesel, 2004. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. Journal of Chemical Information and Computer Sciences, 44 (5): 1840-1848.
21. Whittle, M., V.J. Gillet, P. Willett and J. Loesel, 2006. Analysis of Data Fusion Methods in Virtual Screening: Similarity and Group Fusion. Journal of Chemical and Information Modelling, 46 (6): 2206-2219.
22. Willett, P.,2013. Combination of Similarity Rankings Using Data Fusion. Journal of Chemical Information and Modeling, 53 (1): 1-10.
23. Willett, P., 2009. Turbo Similarity Searching: Effect of Fingerprint and Dataset on Virtual-Screening Performance. Statistical Analysis and Data Mining, 2 (2): 103-114.
24. Williams, C., 2006. Reverse Fingerprinting, Similarity Searching by Group Fusion and Fingerprint Bit Importance. Molecular Diversity, 10 (3): 311-332.
25. Wu, S. and C. Huang, 2014. Search Result Diversification via Data Fusion. In the Proceedings of the 2014 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp: 827-830.
26. Yang, J.M., Y.F. Chen, T.W. Shen, B.S. Kristal and D.F. Hsu, 2005. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. Journal of Chemical Information and Modeling, 45 (4): 1134-1146.
27. Zhang, Q. and I. Muegge, 2006. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. Journal of Medicinal Chemistry, 49 (5): 1536-1548.