# CONSENSUS METHODS FOR VIRTUAL SCREENING AND CLUSTERING OF CHEMICAL STRUCTURE DATABASES

**Faisal Saeed[1*], Naomie Salim[1,] Mohd Shahir Shamsir[2]**
[1]Faculty of Computing, Universiti Teknologi Malaysia
[2]Faculty of Faculty of Bioscience & Bioengineering, Universiti Teknologi Malaysia, Malaysia
*** Corresponding Author:faisalsaeed@utm.my**

## ABSTRACT

The main combination procedures for chemoinformatics are combinations of multiple molecular descriptors, multiple similarity measures, multiple active reference molecules (queries) and multiple clustering methods. The combination methods provide better results for virtual screening because different descriptors, measures, or queries can provide different sources of evidence. In this paper, we review the recent work of combination methods for chemical structure databases.

## 1. INTRODUCTION

There is a lot of work has been done for the combination procedures in chemoinformatics, which is known as consensus scoring (it is also called data fusion) in both structure-based and ligand-based virtual screening (Salim, 2003; Feher, 2006; Willet, 2006; Hert, 2006; Whittle, 2006; Chen, 2010; Svensson, 2011). Recently, consensus scoring has been considered as a simple way of improving the performance of existing systems for ligand-based virtual screening by fusing or combining the results of two or more screening methods. It is proved that when a large number of searches are averaged, the fused search result provides a high level of consistency which is better than the result obtained by any individual screening method (Willet, 2006).

Consensus scoring is a combination approach where data or decisions that come from or based on multiple sources, about the same set of objects, are combined to improve the quality of decision making under uncertainty about the objects. Many fusion techniques have been applied to molecular similarities which are originally used for combining several independent information retrieval system components to obtain a better performance than the best individual component (Willett, 2000).

The main approaches of combination procedures for chemoinformatics are combining multiple molecular descriptors, multiple similarity measures, multiple active reference molecules (queries) and multiple clustering methods. The combination methods provided better results because different descriptors, measures, queries or clusterings may make use of different sources of evidence. For instance, a particular similarity measure may retrieve active molecules that cannot be obtained by others (Abdo, 2009).

## 2. COMBINATION METHODS

**A. Combining Multiple Similarity Coefficients:** Fusing several similarity measures involves computing the degree of similarity using several types of similarity coefficients, and combining the results using one of the main fusion rules such as MAX, MIN, SUM, etc. The results (output) produced by fusion rule are then used to re-order the molecules to give final ranked output(Abdo, 2009). The early studies of combining multiple similarity coefficients were done by Holliday et al. (Holliday, 2002) and Salim et al. (Salim, 2003). Holiday et al. clustered the similarity coefficients into fewer groups and combined the output of these coefficients using data fusion to improve the results. They found that the data fusion enhances the effectiveness of similarity searching. Similarly, Salim et al. (Salim, 2003) combined several binary similarity coefficients and found that the search performances can be improved by combining coefficients with little extra computational cost. The results indicate that combining coefficients does improve the performance of similarity searches when compared with the use of single measures, in particular the industry standard Tanimoto measure. The optimum number of coefficients to be used in the combination tends to be between two and four with the improvement diminishing at five or more coefficients. However, there was no single combination which gives a consistently high performance for all search types.

More studies were conducted to apply consensus scoring for virtual screening; Hert et al. (Hert, 2006) used the machine learning techniques with consensus scoring in order to enhance the effectiveness of similarity searching for chemical datasets. They demonstrated that consensus scoring is notably more effective than conventional similarity searching for structurally diverse sets of active molecules.

**B. Combining Multiple Reference Structures:** In similarity searching, a query involves the specification of an entire molecule, which is known as target structure, in the form of one or more structural descriptors. Then, the target structure is compared with the corresponding set of descriptors for each compound in the database. After that, a measure of similarity is calculated between the target structure and every molecular structure in the database. Finally, the database structures are sorted based on the results of similarity scores in decreasing order. However, during the last decade, several studies of similarity searching used multiple bioactive reference structures (multiple queries) rather than using one reference structure. In these studies, the performance of combining multiple queries provided noticeable superior results compared to that obtained from the use of a single query (Shemetulskis, 1996; Xue, 2003; Schuffenhauer, 2003).

For instance, Ammar and Salim (Abdo, 2009; Abdo, 2008) introduced a novel method for similarity searching using Bayesian inference network (BIN). In this study, they have compared BIN with other similarity searching methods when multiple bioactive reference molecules are available. Three different 2D fingerprints were used in combination with data fusion and nearest neighbor approaches as search tools. Using this approach, the optimal reference structure for each structure in the database is found. Then, the use of the BIN with optimal reference approach was the most efficient and effective compared to the best conventional similarity methods.

Similarly, Chen et al. (Swift, 2004) used the Bayesian inference network with data fusion and showed that group fusion is most effective when many reference structures are used. They have reported a comparison of 15 different fusion rules and the focus was on parameter-free rules that do not require training data, since such data is unlikely to be available in the early stages of a drug discovery. Extensive searches of the MDDR and WOMBAT databases using the ECFC_4 fingerprints and the Bayesian inference network demonstrated that the group fusion is most effective when many reference structures are used.

**C. Combining Multiple Clustering Methods:** The early work of consensus clustering is conducted by Monti et al. (Monti, 2003) for class discovery and visualization of gene expression data. The methods of consensus clustering represented the consensus across multiple runs of a clustering algorithm to assess the stability of the discovered clusters. These methods were used in conjunction with resampling techniques. The partitions (ensemble) also generated by multiple runs of a clustering algorithm with random initialization (such as K-means), to lower the sensitivity to the initial conditions. In this experiment, the consensus method attempted to produce data partitions that are more robust than the single clustering algorithms.

Similarly, Swift et al. (Swift, 2004) presented that microarray analysis using single clustering methods can suffer from lack of inter-method consistency which is occurred when assigning related gene-expression profiles to clusters. However, they obtained a consensus partition using combining the results of multiple clustering methods to increase the confidence in the analysis of gene-expression. When consensus clustering was coupled with a statistically-based gene functional analysis, it allowed the identification of novel genes and the unfolded protein response in certain B-cell lymphomas.

For chemical databases, Chu et al. (Chu, 2010) evaluated the performance of seven consensus clustering methods using the MDDR and IDAlert datasets and the molecules were represented by the extended connectivity fingerprints (ECFP_4). The seven consensus clustering methods were CC-Pivot and BOK methods (Filkov, 2004; Bertolacci, 20007), Majority Rule (Goder, 2008), Average Linkage and Complete Linkage (Everitt, 2001), Direct and Graph-based (Karypis, 2003). They reported the evaluation of consensus clustering for chemical databases using an approach based on a consensus similarity matrix. The consensus clustering was performed by combining multiple runs of K-means clustering method, and also combining single runs of multiple clustering methods. Chu (Chu, 2012) used different criteria to evaluate the effectiveness of clustering methods, which are the Shannon Entropy (Cover, 2006), F-Measure (Van Rijsbergen, 1979) and Quality Partitioning Index (Varin, 2008), and the results of consensus clustering were compared to that of Ward's method.

Chu et al.( Chu, 2010; Chu, 2012) found that, the Majority Rule method provides consistently worst performances across all numbers of clusters when the F-measure is used and the single best consensus clusterings have shown better performance than Ward's method (4 in 6 times). Using the QPI measure, the direct and graph-based methods were consistently in the leading group and provided better QPI values. Similarly, the Majority Rule method consistently has worst results over all numbers of clusters and 5 in 6 single best consensus clusterings are found to be superior to Ward's method. Using the entropy measure, the Average Linkage method provided the consistently best performance over all numbers of clusters and the most consensus clusterings are found to be superior to Ward's method.

## 3. DISCUSSION

As presented in the above sections, the consensus scoring is widely used for chemoinformatics. Many consensus scoring approaches have been used for similarity searching of chemical databases including combining multiple molecular descriptors, multiple similarity measures and multiple active reference molecules (queries). Many studies used more than one combination to obtain the best fused results. For instance, Ammar (Abdo, 2009) combined multiple molecular descriptors and multiple reference structures for similarity searching. This combination substantially improved the retrieval performance when compared to equivalent similarity method which only used one fusion method.

A recent review has been conducted by Willett (Willett, 2013) , in which two main combinations that have been applied for the fusion of similarity searching are discussed. The first one is to combine a single reference structure (query) using multiple similarity measures; while the second one is to combine multiple reference structures using a single similarity measure. In addition, the combination of different types of virtual screening method and the comparison of supervised fusion with existing screening approaches based on machine learning are expected to improve the effectiveness of data fusion (Willett, 2013).

In many studies, it was reported that the consensus scoring has been widely used for enhancing the effectiveness of similarity searching and it was found that the performance of consensus scoring results is superior to the industry-standard similarity method which is Tanimoto similarity coefficient (Chen, 2010; Abdo, 2009; Holliday, 2002; Ahmed, 2014). In addition, it was found that combining multiple reference structures with multiple similarity measures normally being noticeably superior to similarity fusion where a single reference structure is used (Willett, 2013). Recently, Ahmed et al. (Ahmed, 2014) developed a Condorcet fusion model to enhance the effectiveness of ligand-based virtual screening. The overall results of the proposed method showed that the screening similarity search outperformed the Tanimoto which considered the conventional similarity methods. In addition, there was evidence to suggest that this method, Condorcet fusion at Top100, was more effective for high diversity data sets.

The success of using consensus scoring for similarity-based virtual screening and the success of combining multiple clusterings in many areas such as machine learning, applied statistics, pattern recognition and bioinformatics have motivated researchers to apply the combination methods on chemical structures clustering, which is known as consensus clustering.

The basic concept of consensus clustering is to cluster a set of objects by finding a clustering partition that agrees as much as possible with a set of individual clusterings, which is known as an ensemble. In other words, the goal of consensus clustering is to find a consensus partition that try to optimally summarize an ensemble and improve the quality of clustering compared with individual clustering methods. The consensus clustering has been used in in many areas for many applications including clustering of categorical data, detecting outliers and improving clustering robustness (Gionis, 2007). In addition, combining multiple individual clusterings provides a framework for knowledge reuse, which can be used to exploit the powerful of existing knowledge that is implicated in multiple clusterings (Strehl, 2002).

The comparison of combinatorial clustering methods has been done by Rivera-Borroto et al. (Rvera- Borroto, 2011) and the results on the relative performance of clustering algorithms are encouraged because they provided three mathematically and algorithmically clustering methods with a similar performance as Ward's algorithm. Comparison of the best five clustering methods with machine learning techniques indicates, on median scores, that they performed similarly to supervise classifiers but they were outperformed by the consensus of these functions.

Based on the experiments of on consensus clustering in (Saeed, 2013), Saeed et al. reported that the graph-based and hypergraph-based methods could significantly outperform the Ward's method when using the QPI measure for ALOGP descriptor, while they provide inferior performances to that of Ward's method using ECFP_4. To improve the performance of the combination methods, the cumulative voting-based aggregation algorithm CVAA was developed (Saeed, 2012) and it was shown that it is the method of choice among consensus clustering methods.

The performance of the CVAA consensus clustering significantly outperforms the Ward's and graph-based consensus methods using all used evaluation measures. In addition, it was shown that voting-based consensus clustering can perform well when the partitions are generated by a single run of multiple individual clusterings that use Jaccard coefficient in the ensemble generation process. Moreover, the adaptive cumulative voting-based aggregation algorithm A-CVAA was developed in (Saeed, 2013) to overcame the order dependent limitation of the CVAA, such that the ensemble partitions were sorted based on the mutual information associated with each partition (using entropy measure) in order to obtain a unique consensus partition. The experiments show that the A-CVAA can improve the effectiveness of combining multiple clusterings of chemical structures.

Recently, Saeed et al ( Saeed, 2013; Saeed, 2014) developed the weighted voting-based consensus clustering (WCVAA) compared with Ward's method and other consensus clusterings. The W-CVAA overcomes the main limitations of the CVAA by using a pre-defined ordering for the ensemble partitions (to solve the ordering dependent problem), and by assigning different weights for the individual clusterings that generate the ensemble. The evaluation of the W-CVAA method on different descriptors and using different criteria suggests that the W-CVAA consensus method can deliver significant improvements for the effectiveness of chemical structure clustering.

## 4. CONCLUSIONS

In this paper, we review the recent work of consensus methods for chemoinformatics. Many studies reported the interesting results of combining multiple molecular descriptors, multiple similarity measures, multiple active reference molecules (queries) and multiple clustering methods. The combination of different methods could improve the performance of virtual screening and clustering methods. Future work should investigate of using new and different consensus methods for chemoinformatics applications.

## ACKNOWLEDGMENTS

This work is supported by Research Management Centre (RMC) at Universiti Teknologi Malaysia under Research University Grant (PAS) and Cost Center No. Q.J130000.2728.01K73.

**REFERENCES**

Abdo, A., and Salim, N. Bayesian inference network for molecular similarity searching using 2D fingerprints and multiple reference structures. Jurnal Teknologi Maklumat, 20(3), 2008, 1-13.

Abdo, A., Similarity-Based Virtual Screening Using Bayesian Inference Network For Searching Chemical Database. Universiti Teknologi Malaysia. PhD Thesis, 2009.

Ahmed, A., Saeed, F., Salim, N., & Abdo, A, Condorcet and borda count fusion method for ligand-based virtual screening. Journal of Cheminformatics, 6(1), 2014, 1-10.

Bertolacci, M. and Wirth, A., Are approximation algorithms for consensus clustering worthwhile? Proc. Seventh SIAM ICDM, 2007, 437-442.

Chen, B., Mueller, C., Willett, P., Combination Rules for Group Fusion in Similarity-Based Virtual Screening. Molecular Informatics, 29, 2010, 533-541

Chu, C-W, Holliday, J., Willett, P., Combining multiple classifications of chemical structures using consensus clustering. Bioorganic & Medicinal Chemistry 2012, 20(18): 5366–5371.

Chu, C-W., Clusterings for 2D Chemical Strictures. University of Sheffield. PhD Thesis, 2010.

Cover, T.M., and Thomas, J. A., Elements of Information Theory, 2nd edn., Wiley Series in Telecommunications and Signal Processing (Wiley-Interscience), 2006.

Everitt, B.S., Landau, S., Leese, M., Cluster Analysis. 4th edition, Edward Arnold. London, 2001.

Feher, M., Consensus Scoring for Protein-Ligand Interactions. Drug Discovery Today, 11, 2006, 421-428.

Filkov, V., and Skiena, S., Integrating microarray data by consensus clustering, International Journal on Artificial Intelligence Tools, 13(4): 2004, 863-880.

Gionis, A., Mannila, H., Tsaparas, P., Clustering aggregation. ACM Trans. Knowl. Discovery Data 2007, 1, Article 1.

Goder, A., and Filkov, V., Consensus Clustering Algorithms: Comparison and Refinement. In: Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX), 2008, San Francisco, USA, 109-117.

Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A., 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, 2006, 462-470.

Holliday, J. D., Hu, C. Y., and Willett, P., Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Combinatorial chemistry & high throughput screening, 5(2), 2002, 155-166.

Karypis, G, CLUTO: A Clustering Toolkit, Release 2.1.1: University of Minnesota, 2003.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T., Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning, 52(1-2), 2003, 91-118.

Rivera- Borroto, O, Arrero-Ponce, Garc a-de la Vega and Grau- balo, Comparison of combinatorial clustering methods on pharmacological datasets represented by machine learning-selected real molecular descriptors. Journal of chemical information and modeling, 51(12), 2011, 3036-3049.

Saeed, F., Ahmed, A., Shamsir, . S., & Salim, N, Weighted voting-based consensus clustering for chemical structure databases, Journal of computer-aided molecular design, 1-10. 2014.

Saeed, F., Salim, N., & Abdo, A. Consensus methods for combining multiple clusterings of chemical structures. Journal of chemical information and modeling, 53(5), 2013, 1026-1034.

Saeed, F., Salim, N., Abdo, A, Information theory and voting based consensus clustering for combining multiple clusterings of chemical structures, Journal of Molecular Informatics. 2013.

Saeed, F.,Salim, N., Abdo, A, Voting-based consensus clustering for combining multiple clusterings of chemical structures, Ournal of Cheminformatics, 4(37), 2012.

Saeed, F.,Salim, N., Abdo, A., Hentabli, H, Graph-based consensus clustering for combining multiple clusterings of chemical structures, Journal of Molecular Informatics, 32 (2), 2013, 165-1788.

Salim, N., Holliday, J.D., Willett, P., Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. Journal of Chemical Information and Computer Sciences, 43, 2003, 435 – 442.

Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E., Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. Journal of Chemical Information and Computer Sciences, 43(2), 2003, 391-405.

Shemetulskis, N.E., Weininger, D., Blankley, C.J., Yang, J.J., and Humblet, C., Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets. Journal of Chemical Information and Computer Sciences, 36(4), 1996, 862-871.

Strehl, A. and Ghosh, J., Cluster Ensembles:A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research, 3, 2002, 583-617.

Svensson, F., Karlén, A., and Sko ld, C.,Virtual screening data fusion using both structure-and ligand-based methods. Journal of Chemical Information and Modeling, 52(1), 2011, 225-232.

Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., and Kellam, P., Consensus clustering and functional interpretation of gene-expression data. Genome biology, 5(11), 2004, R94.

Van Rijsbergen, C.J., Information Retrieval, London Butterworth, 1979.

Varin, T., Saettel, N., Villain, J., Lesnard, A., Dauphin, F., Bureau, R., Rault, S. J. 3D Pharmacophore, hierarchical methods, and 5-HT4 receptor binding data. Journal of Enzyme Inhibition and Medicinal Chemistry, 23, 2008, 593−603.

Whittle, M., Gillet, V.J., Willett, P., and Loesel, J., Analysis of data fusion methods in virtual screening: similarity and group fusion. Journal of chemical information and modeling, 46(6), 2006, 2206-2219.

Willet, P., Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion. QSAR & Combinatorial Scienc, 25, 2006, 1143−1152.

Willett, P., Combination of Similarity Rankings Using Data Fusion. Journal of Chemical Information and Modeling, 2013.

Willett, P., Fusing similarity rankings in ligand-based virtual screening. Computational and Structural Biotechnology Journal, 2013. 5.

Willett, P., Textual and chemical information processing: different domains but similar algorithms. Information Research, **5**(2), 2000.

Xue, L., Godden, J. W., Stahura, F.L., and Bajorath, J., Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys. Journal of Chemical Information and Computer Sciences, 43(4), 2003, 1218-1225.