FRAGMENT REWEIGHTING IN LIGAND-BASED VIRTUAL SCREENING

ALI AHMED ALFAKIABDALLA ABDELRAHIM

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

FEBRUARY 2013

To my beloved father and mother, my wife and my sons

ACKNOWLEDGMENTS

In the Name of Allah, Most Gracious, Most Merciful

All praise and thanks are due to Allah, and peace and blessings be upon his messenger, Mohammed (peace be upon him).

I am indebted to my advisor Professor Dr. Naomie Salim, for the outstanding motivation, guidance, support, and knowledge she has provided throughout the course of this work. She introduced me to the field of chemoinformatics and without her guidance and advice this study would not have been possible. She has been incredibly wise, helpful, understanding, and generous throughout the process. She has truly been a mentor and I owe here my deepest thanks.

I have made many friends during my time in UTM and I thank them for their support and encouragement. Also I am extremely grateful to Dr. Ammar Abdo for his help and knowledge.

A lot of information useful to the work was found via the World-Wide Web; I thank those who made their materials available by means of this medium and those who kindly answered back to my roll-calls of help sent over the World-Wide Web. I am extremely grateful to The Karary University for their generous financial support during this study.

Finally, I would like to thank my parents, wife and my sons, Mohammed and Ayman, for their patience, encouragement, support and understanding.

ABSTRACT

Based on the molecular similarity principle, functionally similar molecules are sought by searching molecular databases for structurally similar molecules to be used in rational drug design. The conventional 2-dimentional similarity methods are the most used methods to measure similarity of molecules, including fragments that are not related to the biological activity of a molecule. The most common methods among the 2-dimentional similarity methods are the vector space model and the Bayesian networks, which are based on mutual independence between fragments. However, these methods do not consider the importance of fragments. In this thesis, four reweighting approaches are proposed to identify the important fragments. The first approach is based on reweighting the important fragments, where a set of active reference structures are used to reweight the fragments in the reference structure. Secondly, a statistically supervised features selection and minifingerprint to select only the important fragments are applied. In this approach, searching is carried out by using sub-fragments that represent the important ones. Thirdly, a similarity coefficient based on mutually dependent fuzzy correlation coefficient is used. The last approach combined the best two out of the three approaches which are reweighting factors and fragment selection based on statistically supervised features selection. The proposed approaches were tested on the MDL Data Drug Report standard data set. The overall results of this research showed that the proposed fragment reweighting approaches outperformed the conventional industry-standard Tanimoto-based similarity search approach.

ABSTRAK

Berdasarkan prinsip persamaan molekul, molekul yang sama fungsi diperolehi dengan mencari molekul yang berstruktur sama dari pangkalan data molekul bagi kegunaan reka bentuk ubat secara rasional. Kaedah persamaan 2-dimensi konvensional telah digunakan secara paling meluas untuk mengukur kesamaan molekul termasuk fragmen yang tidak berkaitan dengan aktiviti biologi sesuatu molekul. Kaedah yang paling biasa digunakan antara kaedah-kaedah persamaan 2-dimensi adalah model ruang vektor dan rangkaian Bayesian yang berasaskan fragmen saling-bebas. Walau bagaimanapun, kaedah-kaedah ini tidak mengambil kira kepentingan fragmen. Dalam tesis ini, empat kaedah bobot semula telah dicadangkan untuk mengenal pasti fragmen-fragmen yang penting. Keadah pertama adalah berdasarkan bobot semula fragmen yang penting, iaitu satu set struktur rujukan aktif telah digunakan untuk bobot semula fragmen dalam struktur rujukan. Kedua, pemilihan ciri terselia secara statistik dan cap jari mini untuk memilih fragmen-fragmen yang penting telah digunakan. Dalam kaedah ini, pencarian dijalankan dengan menggunakan sub-fragmen yang penting. Ketiga, satu pekali persamaan berasaskan koefisien korelasi kabur yang saling bersandar telah digunakan. Kaedah terakhir menggabungkan dua daripada tiga kaedah terbaik iaitu faktor pemberatan semula dan pemilihan fragmen berdasarkan pemilihan ciri terselia secara statistik. Kaedah-kaedah yang dicadangkan telah diuji pada set data piawai MDL Drug Data Report. Keputusan keseluruhan kajian ini menunjukkan bahawa kaedah-kaedah bobot semula fragmen yang dicadangkan mengatasi kaedah piawai konvensional di dalam industri ini iaitu carian persamaan berasaskan Tanimoto.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	-	Two Dimension
3D	-	Three Dimension
ANN	-	Artificial Neural Network
BCI	-	Barnard Chemical Information System
BIN	-	Bayesian Inference Network
BINFS	-	Bayesian Inference Network based on feature selection
BKD	-	Binary Kernel Discrimination
CAS	-	Chemical Abstracts Service
DAG	-	Directed Acyclic Graph
EEFC	-	Atom Type Atom Environment Fingerprint
EHFC	-	Atom Type Hashed Atom Environment Fingerprint
FCFC	-	Functional Class Extended-Connectivity Fingerprint
FEFC	-	Functional Class Atom Environment Fingerprint
FHFC	-	Functional Class Hashed Atom Environment Fingerprint
HTS	-	High Throughput Screening
IR	-	Information Retrieval
LBVS	-	Ligand-Based Virtual Screening
MCS	-	Maximal Common Substructure
MDDR	-	MDL Drug Data Report
MDL	-	Molecular Design Limited
MFPS	-	Minifingerprints
NBC	-	Naïve Bayesian Classifier
NP	-	No Polynomial Time
PCA	-	Principle Component Analysis
QSAR	-	Quantitative Structure-Activity Relationship
RBINRFD	-	Reweighted BIN based on Relevance Feedback
ROSDAL	-	Representation of Organic Structures Description Arranged

		Linearly
SLN	-	Sybyl Line Notation
SMILES	-	Simplified Molecular Input Line System
SOM	-	Self-Organizing Feature Maps
SVM	-	Support Vector Machine
TAN	-	Tanimoto
VS	-	Virtual Screening
WLN	-	Wiswesser Line Notation
WOMBAT	-	World Of Molecular BioActivity

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

Cheminformatics (sometimes spelled as chemo-informatics) is a relatively new discipline, having emerged from several older disciplines such as computational chemistry, computer chemistry, chemometrics, QSAR and chemical information. Cheminformatics is a cross between Computer Science and Chemistry: the process of storing and retrieving information about chemical compounds. The term "chemoinformatics" also referred as Chemoinformatics/Chemiinformatics/Chemical information/Chemical informatics has been recognised in recent years as a distinct discipline in computational molecular sciences [1].

Chemoinformatics was defined by Brown in [2] as:

> "Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization."

Cheminformatics is indeed a legitimate new field in which chemistry and computer sciences strongly intersect. Those employed in this field develop new substances, materials, and processes by organizing, analyzing, and visualizing the information available to them. The present chief application of cheminformatics is in the field of drug discovery, but it is finding increasing acceptance and use in other applied areas of chemistry.

Cheminformaticians often work with massive amounts of data. They construct information systems that help chemists make sense of the data, often attempting to accurately predict the properties of chemical substances from a sample of data. Thus, through the application of information technology, cheminformatics helps chemists organize and analyze known scientific data to assist in the development of novel compounds, materials, and processes. People who work in cheminformatics may concentrate on molecular modelling, chemical structure coding and searching, chemical data visualization, or a number of other areas of specialization. Indeed, the various computer graphics codes for chemical structures that let us both view and search chemical structures via computer were developed by cheminformaticians.

Greg Paris[3] provided the following definition:

> "Chemoinformatics is a generic term that encompasses the design, creation, organization, storage, management, retrieval, analysis, dissemination, visualization and use of chemical information, not only in its own right, but as a surrogate or index for other data, information and knowledge."

Hann and Green [4] suggest that chemoinformatics is simply a new name for an old problem. Many informatic methods and techniques used in chemoinformatics have been studied for many years; however, the broad and general definition was given by Gasteiger [5]as:

“Chemoinformatics is the use of informatic methods to solve chemical problems”.

Virtual screening (VS) is a computational technique used in drug discovery research. Computers are used to quickly search large libraries of chemical structures in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme. Virtual screening process usually starts with a ‘query’ to search the chemical database using one of the virtual screening tools, as the query can be a molecule with a desired biological activity. By using this process the chemist tries to identify other molecules in the database that can be tested in an appropriate assay.

Currently virtual screening has become widely used in computer-based search for novel lead molecules. There are two types of virtual screening approaches: ‘virtual screening by docking’ which deals with the 3D structure of biological targets (proteins or enzymes) and ‘similarity-based virtual screening’, where the structural information of one or more known molecules is used as a structural query. The second approach is the basis of this thesis.

The storage and search for chemical structures and associated information in databases are probably the earliest beginnings of what might be called chemoinformatics. Nowadays, chemoinformatics has attracted much recent prominence as a result of developments in computer power and the methods that are used to synthesize new molecules, followed by tests of their biological activity. These developments have led to a massive increase in the number of chemical compounds and biological information that is available for discovery programmes in pharmaceutical and agrochemical industries.

In this thesis, different fragment-based similarity-based virtual screenings are presented. The background of the problem, objectives, importance of the study, and the scope of this research are discussed in the remainder of this chapter.

1.1 Background of the Problem

There are seven sequential steps in the Drug discovery process: disease selection, target hypothesis, lead compound identification (screening), lead optimization, pre-clinical trials, clinical trials and pharmacogenomics optimization. These steps are carried out sequentially and delays in any of the steps results in delays in the entire process [6].These delays represent bottlenecks.

Previously, the main bottlenecks in drug discovery were the time and cost of finding (making) and testing new chemical entities (NCE). The average cost of creating a NCE in a major pharmaceutical company was estimated at around $7,500/compound [7]. In order to reduce these costs, pharmaceutical companies have had to find new technologies to replace the old traditional “hand-crafted” synthesis and testing NCE approaches. High throughput screening (HTS), combinatorial chemistry (CC) and virtual screening are examples of such technologies.

In response to the increased demand for new compounds by biologists, chemists started using combinatorial chemical technologies to produce more new compounds in shorter time periods. By using HTS, it is possible to test hundreds of thousands of compounds in a short time. Computers can be used to aid this process in a number of ways, such as in the creation of virtual libraries, which can be much larger than their real counterparts.

Recently, chemical search techniques have been called virtual screening; the main idea is that these methods test large number of compounds by computer instead of experience. Virtual screening involves a range of computational tools for searching chemical databases to filter out the unwanted compounds. These tools can be used to reduce drug discovery costs by removing undesired compounds as early as possible and providing only those compounds that have the largest a priori probabilities of activity for conventional biological screening.

Virtual screening approaches can be categorized as structure-based approaches, which can be used if the 3D structure of the biological target is available. Examples of this type of approach are ligand-protein docking and *de novo* design. The second type of approach is ligand-based, which is applicable in the case of the absence of such structural information. Similarity methods and machine learning methods are examples of this type of approach.

Similarity methods are the most common, as well as the simplest and most widely used tools for ligand-based virtual screening tools for ligand-based virtual screening of chemical databases. That is because these methods require just a single known bioactive molecule (the reference or target molecule) as a starting point for database search. Here, the database structures are ranked in decreasing order of similarity with active, user defined, reference structure (query), with the expectation that the nearest neighbours will exhibit as the reference structure.

There are many studies in the literature associated with the measurement of the molecular similarity [4, 8-11]. However, the most common approaches are based on 2D fingerprints, with the similarity between a reference structure and a database structure computed by using an association coefficient such as Tanimoto coefficient [8, 12]. There are many other similarity methods in which the structural similarity between molecules can be computed. The effectiveness of any similarity method has found to vary from one biological activity to another in a way that is difficult to predict [9]. In addition, the use of any two methods has been found to retrieve a different subset of actives from databases, so it is advisable to use several search methods where possible. Current research focuses on three main areas: molecular similarity measures; the analysis of molecular diversity and the design of combinatorial libraries; and the representation and searching of biological macromolecules. Our research group directions focus on consensus clustering and shape-based molecular descriptor [13, 14].

Many studies in chemoinformatics have proved that retrieval models based on inference networks give significant improvements in retrieval performance compared to conventional models[15, 16]. In more recent studies, the Bayesian inference network has been introduced as promising the similarity search approach[17, 18]. The retrieval performance of the Bayesian inference network was observed to improve significantly when multiple reference structures were used or more weights were assigned to some fragments in the molecule structure. Unfortunately, such information is unlikely to be available in the early stages of a drug discovery program when just a single weak lead is available. Unfortunately, such information is unlikely to be available in the early stages of a drug discovery program when just a single weak lead is available. In the literature, there are many methods used to improve Bayesian inference network [19-21].

1.2 Problem Statement

Conventional Bayesian inference network similarity method has two implicit problems. First, it considers all molecular features as equal in importance; therefore all molecular features are used when we calculate similarity measure. Second, all weighting schemes calculate the weight for each feature independently with no relation to all other features [22]. In order to enhance the effectiveness of a retrieved active target, feature reweighting can enhance the recall of similarity measure.

In order to enhance the effectiveness of Bayesian inference network similarity method, the aim of this research is to develop a ligand-based similarity method based on Bayesian network and reweighted fragments and 2D fingerprints to search large chemical databases to retrieve compounds with the most similar biological activity to the reference structure. This method applies four different approaches to fragment reweighting; the first approach is based on fragment

reweighting factors; fragment reweighting is the process of adding new weight to the original weight in order to improve retrieval performance in information retrieval systems[6]. Turbo Similarity Searching (TSS) and relevance feedback [23, 24] are two examples of reweighting fragments or features in Ligand-based virtual screening. The second is the implementation of the idea of reweighting in terms of sub-fragments which apply two techniques: selecting the important fragment and using the idea of Minifingerprint, the main idea of Minifingerprint is to limit or reduce features or fragments and correctly identify the percentage of compounds with similar biological activity. The third approach develops a novel of fuzzy correlation coefficient based on mutual dependence between fragments, while the last approach is combination of first two approaches.

1.3 The Research Question

The main research question is:

Can reweighted molecular fragments or features positively effect and increase the retrieval recall of Bayesian Inference Network.?

Thus, the following issues will need to be addressed in order to answer the main research question stated above:

- Can we develop fragment reweighting using reweighting factors and relevance feedback to improve the retrieval recall of Bayesian Inference Network?

- Can we identify important sub-fragments using a supervised statistical feature selection model and minifingerprints to improve the retrieval recall of Bayesian Inference Network?

- Can we develop a novel fuzzy correlation coefficient based on mutual dependence between molecular fragments?

- Is effectiveness of the proposed approaches better than conventional Bayesian Inference Network virtual screening model?

1.4 Objectives of the Research

The main goal of this research is to develop a similarity-based virtual screening approach using reweighted fragments and Bayesian Inference Network, with the ability to improve the retrieval effectiveness and provide an alternative to existing tools for ligand-based virtual screening.

To achieve this goal, the following objectives have been set:

- To investigate reweighting factor and relevance feedback for use in similarity calculations to enhance the retrieval effectiveness of Bayesian Inference Network model.

- To determine the retrieval performance of the reweighted fragment Bayesian Inference Network model for molecular similarity searching.

- To investigate the selected of important fragments based on feature selection and minifingerprints for molecular similarity searching when 2D fingerprint and several reference structures are available.

- To investigate a novel similarity based virtual screening for molecular similarity searching based on mutual dependence between fragments for molecular similarity searching.

- To combine the different methods of fragment reweighting.

- To compare the retrieval performance of reweighted fragments and fuzzy correlation coefficient with conventional similarity methods.

1.5 Importance of the Study

The similarity principle states that structurally similar molecules will exhibit similar physicochemical and biological properties [8, 11, 12, 25, 26],which has become the basis for many rational drug design efforts. In fact, the observation that common fragments lead to similar biological activities can be quantified from database analysis [27]. This concept leads to the term molecular similarity, which has become widely used in chemical literature [8, 11, 12].

Over the past last decade, technological advances in synthesis and high throughput screening have increased the capability to synthesize large libraries of compounds and the capability to screen hundreds of thousands of compounds in a short time. These developments increase the necessity for the application of computer based methods for compound selection and evaluation. In addition,

increases in computer power have enabled similarity applications to be performed on very large databases of compounds.

The development of new drugs is both time consuming and cost-intensive, where the estimated cost for discovering and bringing a new drug to the market costs at around $7,500/ compound , taking an average of 12 to 13 years [28]. This is due to the high failure rates in the later stages of drug development.

1.6 Scope of the Study

This study will focus on 2D fingerprint-based similarity methods. These methods are used to quantify the degree of structural resemblance between a pair of molecules characterised by 2D fingerprints. These methods are applied with binary and non-binary 2D fingerprints.

In addition, this study focuses on the different approaches of fragment reweighting methods. Typically, four different approaches are used to enhance the effectiveness of molecular retrieval. Reweighting factor is used to reweight the input query fragment weights. A statistical supervised feature selection model is applied to select only the important fragments that will be used later in similarity calculation; the study also develops a novel fuzzy correlation similarity method based on mutual dependence between fragments.

The similarity approaches in this study evaluated a large dataset derived from MDL Drug Data Report (MDDR) database [29], where single and multiple reference structures are available. The performance of this method is evaluated against the

performance of conventional 2D similarity methods (Tanimoto and conventional Bayesian inference network).

1.7 Thesis Outline

This thesis consists of seven major parts, excluding the introductory chapter. While the first two parts describe the background as well as the previously published work in the field of molecular similarity, the third part describes the research methodology for the work in this thesis. Finally, the last four parts present the algorithmic details of the reweighting fragment virtual screening method.

Chapter 2, *Molecular Similarity*, begins with an overview of computer representations of chemical structures and various types of searching mechanisms offered by chemical information systems. In the third section, we present molecular representations which can be employed for molecular similarity searching as well as for molecular analysis and clustering. Here, we also describe in detail the 2D fingerprint-based similarity methods and different types of similarity coefficients. This chapter discusses the implementation of machine learning techniques to molecular similarity. Similarity searching in text database has been reviewed in this chapter. We conclude with a discussion and summary of the applicability of the mentioned methods to molecular similarity searching and the best ways to improve the performance of these methods.

Chapter 3, *Research Methodology*, describes the overall methodology adopted in this research to achieve the objectives of this thesis. In that part, we try to give a general picture about each phase in our research framework. In this chapter, also we discuss the implementation reweighting fragment techniques to molecular

similarity. We give an overview of the relevant feedback and query expansion methods that are used in molecular similarity searching. Ligand-based virtual screening based on sub-fragments is also reviewed in this chapter. Here, we discuss two methods of selecting sub-fragments, using either supervised feature selection algorithm to select the important fragments, or using the idea of minifingerprint, which can be considered an unsupervised feature selection method. In addition, the implementation of reweighting factor for reweighting molecular fragments has been addressed. The implementation of fuzzy correlation coefficient has also been introduced. We conclude this chapter with a discussion and summary.

Chapter 4, *Similarity-based Virtual Screening using Reweighted Fragments*, describes the fragment reweighting methods as an enhancement to a virtual screening tool. Here, we present a novel approach to molecular similarity searching recall problems using various reweighting methods and approaches. This approach works with a multiple reference structure and a single fingerprint. At the end of this chapter, an evaluation of the results of this approach is presented.

Chapter 5, *Similarity-Based Virtual Screening Using Sub-Fragments*, describes the similarity searching problem which occurs when the molecular fragments are too numerous but may contain important active parts that consists of very important fragments. This chapter describes supervised and unsupervised approaches ways to select for important fragments. In the results and discussion section, the results are presented and discussed.

Chapter 6, *Fuzzy Correlation Coefficient for Similarity-Based Virtual Screening*, describes a new approach for solving the similarity searching problem when different 2D fingerprints and multiple reference structures are available. This chapter describes using current correlation coefficients and introduces a novel correlation coefficient based on mutual dependence between molecular fragments. In the results and discussion section, the FCC results are presented and discussed.

Chapter 7, *Combination of reweighting fragment approaches*, this chapter describes a new approach of fragment reweighting by combining reweighting factors and fragment selection approaches. At the end of this chapter, an evaluation of the results of this approach is presented and compared with all previous reweighting approaches as well as the standard similarity measures.

Chapter 8, *Conclusion and Future Work*, is the last chapter, which discusses and concludes the overall works of this thesis highlights the findings and contribution made by this study and provides suggestions and recommendations for future research.

1.8 Summary

In this chapter, we give a broad overview of the problems involved in the molecular similarity. This chapter serves as an introduction to the research problem set out earlier in this thesis. The goal, objectives, the scope, and the outline of this thesis are also presented.

REFERENCES

1. Begam, B.F. and J.S. Kumar, A Study on Cheminformatics and its Applications on Modern Drug Discovery. Procedia Engineering, 2012. 38: p. 1264-1275.
2. Brown, F., Chemoinformatics, what it is and how does it impact drug discovery. *Annual Reports in Medicinal Chemistry*, 1998. **33**: 375-384.
3. Paris, G. Augest 1999 Meeting of the American Chemical Society http://www.warr.com/warrzone2000.html.
4. Maldonado, A.G., et al., Molecular similarity and diversity in chemoinformatics: from theory to applications. Molecular diversity, 2006. 10(1): p. 39-79.
5. Gasteiger, J. and Enge, T., *Chemoinformatics: A Textbook*. Weinheim: Wiley-VCH: 2003.
6. de Castro, P.A.D., et al., Query expansion using an immune-inspired biclustering algorithm. Natural Computing, 2010. 9(3): p. 579-602.
7. Augen, J.,The evolving role of information technology in the drug discovery process. Drug discovery today **7**(5), 315-323 (2002).
8. Willett, P., Barnard, J.M., Downs, G.M., Chemical similarity searching. Journal of Chemical Information and Computer Sciences **38**(6), 983-996 (1998).
9. Sheridan, R.P., Kearsley, S.K., Why do we need so many chemical similarity search methods? Drug discovery today **7**(17), 903-911 (2002).
10. Nikolova, N., Jaworska, J., Approaches to measure chemical similarity–a review. QSAR & Combinatorial Science **22**(9-10), 1006-1026 (2003).

11. Bender, A., Glen, R.C., Molecular similarity: a key technique in molecular informatics. Org. Biomol. Chem. **2**(22), 3204-3218 (2004).
12. Johnson, M.A.a.M., G. M., Concepts and Application of Molecular Similarity:. John Wiley & Sons, New York (1990).
13. Faisal Saeed, Naomie Salim, Ammar Abdo, Hamza Hentabli, Combining Multiple Individual Clusterings of Chemical Structures Using Cluster-Based Similarity Partitioning Algorithm, AMLTA, 322(4), 276-284, DOI 10.1007/978-3-642-35326-0_28 (2012).
14. Hamza Hentabli, Naomie Salim, Faisal Saeed, Ammar Abdo, LWDOSM: Language for Writing Descriptors of Outline Shape of Molecules, AMLTA, 322(4), 247-256, DOI 10.1007/978-3-642-35326-0_25 (2012).
15. Willett, P., Textual and chemical information processing: Different domains but similar algorithms. Information Research **5**(2) (2000).
16. Willett, P., Chemoinformatics: an application domain for information retrieval techniques. In: 2004, pp. 393-393. ACM.
17. Abdo, A., Salim, N., Similarity-Based Virtual Screening with a Bayesian Inference Network. ChemMedChem **4**(2), 210-218 (2009).
18. Abdo, A., Chen, B., Mueller, C., Salim, N., Willett, P., Ligand-based virtual screening using bayesian networks. Journal of chemical information and modeling **50**(6), 1012-1020 (2010).
19. Abdo, A., Salim, N., New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening. Journal of chemical information and modeling (2011).
20. Abdo, A., Salim, N., Bayesian inference network significantly improves the effectiveness of similarity searching using multiple 2D fingerprints and multiple reference structures. QSAR & Combinatorial Science **28**(11-12), 1537-1545 (2009).
21. Abdo, A., Saeed, F., Hamza, H., Ahmed, A., Salim, N., Ligand expansion in ligand-based virtual screening using relevance feedback. Journal of computer-aided molecular design, 1-9 (2012).
22. Vogt, M., Wassermann, A.M., Bajorath, J., Application of Information—Theoretic Concepts in Chemoinformatics. Information **1**(2), 60-73 (2010).

23. López-Pujalte, C., Guerrero-Bote, V.P., de Moya-Anegón, F., Genetic algorithms in relevance feedback: a second test and new contributions. Information processing & management **39**(5), 669-687 (2003).
24. Taktak, I., Tmar, M., Hamadou, A., Query Reformulation Based on Relevance Feedback. Flexible Query Answering Systems, 134-144 (2009).
25. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., Weinberger, L.E., Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. Journal of Medicinal Chemistry **39**(16), 3049-3059 (1996).
26. Martin, Y.C., Kofron, J.L., Traphagen, L.M., Do structurally similar molecules have similar biological activity? Journal of Medicinal Chemistry **45**(19), 4350-4358 (2002).
27. Holliday, J.D., Salim, N., Willett, P., On the magnitudes of coefficient values in the calculation of chemical similarity and dissimilarity. In: 2005, pp. 77-95. ACS Publications.
28. DiMasi, J.A., Hansen, R.W., Grabowski, H.G., The price of innovation: new estimates of drug development costs. Journal of health economics **22**(2), 151-185 (2003).
29. Symyx Technologies. MDL drug data report. http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp. Accessed October 20, 2011.
30. Leach, A.R., Gillet, V.J., An Introduction to Chemoinformatics. Kluwer Academic Publishers, London (2003).
31. Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M., Watson, D.G., The development of versions 3 and 4 of the Cambridge Structural Database System. Journal of Chemical Information and Computer Sciences **31**(2), 187-204 (1991).
32. Ricketts, E.M., Bradshaw, J., Hann, M., Hayes, F., Tanna, N., Ricketts, D.M., Comparison of conformations of small molecule structures from the Protein Data Bank with those generated by Concord, Cobra, ChemDBS-3D, and Converter and those extracted from the Cambridge Structural Database. Journal of Chemical Information and Computer Sciences **33**(6), 905-925 (1993). doi:doi:10.1021/ci00016a013

33. Willett, P.,Similarity and Clustering in Chemical Information Systems. John Wiley Sons, Inc., (1987).
34. Wiswesser, W.J., A Line-Formula Chemical Notation. (1954).
35. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences **28**(1), 31-36 (1988).
36. Weininger, D., Weininger, A., Weininger, J.L., SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of Chemical Information and Computer Sciences **29**(2), 97-101 (1989).
37. Barnard, J.M., Jochum, C.J., Welford, S.M., A Universal Structure/Substructure Representation for PC-Host Communication. Chemical Structure Information Systems. Interfaces, Communications, and Standards **400**, 76-81 (1989).
38. Ash, S., Cline, M.A., Homer, R.W., Hurst, T., Smith, G.B., SYBYL line notation (SLN): A versatile language for chemical structure representation. Journal of Chemical Information and Computer Sciences **37**(1), 71-79 (1997).
39. Tarjan, R.E., Algorithms for Chemical Computation. American Chemical Society: Washington D.C (1977).
40. Morgan, H.L., The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. (1965).
41. Garey, M.R., Johnson, D.S., The rectilinear Steiner tree problem is NP-complete. SIAM Journal on Applied Mathematics, 826-834 (1977).
42. Barnard, J.M., Substructure searching methods, Old and new. Journal of Chemical Information and Computer Sciences **33**(4), 532-538 (1993).
43. Carhart, R.E., Smith, D.H., Venkataraghavan, R., Atom pairs as molecular features in structure-activity studies: definition and applications. Journal of Chemical Information and Computer Sciences **25**(2), 64-73 (1985).
44. Willett, P., Winterman, V., Bawden, D., Implementation of nearest-neighbor searching in an online chemical structure search system. Journal of Chemical Information and Computer Sciences **26**(1), 36-41 (1986).
45. Willett, P. (ed.) Similarity Searching in Chemical Structure Database. (2003).
46. Todeschini, R., Consonni, V., Handbook of molecular descriptors, vol. 79. Wiley-vch, (2008).

47. Downs, G.M., Willett, P., Fisanick, W., Similarity searching and clustering of chemical-structure databases using molecular property data. Journal of Chemical Information and Computer Sciences **34**(5), 1094-1102 (1994).

48. Sadowski, J., Kubinyi, H., A scoring scheme for discriminating between drugs and nondrugs. Journal of Medicinal Chemistry **41**(18), 3325-3329 (1998).

49. Byvatov, E., Fechner, U., Sadowski, J., Schneider, G., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. Journal of Chemical Information and Computer Sciences **43**(6), 1882-1889 (2003).

50. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1. Advanced drug delivery reviews **46**(1-3), 3-26 (2001).

51. Lipinski, C.A., Drug-like properties and the causes of poor solubility and poor permeability. Journal of pharmacological and toxicological methods **44**(1), 235-249 (2000).

52. Dixon, S.L., Merz Jr, K.M., One-dimensional molecular representations and similarity calculations: methodology and validation. Journal of Medicinal Chemistry **44**(23), 3795-3809 (2001).

53. Dittmar, P., Farmer, N., Fisanick, W., Haines, R., Mockus, J., The CAS ONLINE search system. 1. General system design and selection, generation, and use of search screens. Journal of Chemical Information and Computer Sciences **23**(3), 93-102 (1983).

54. Barnard, J. M. and Downs, G. M., Chemical Fragment Generation and Clustering Software&sect. *Journal of Chemical Information and Computer Sciences*, 1997. **37**(1): 141-142.

55. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G., Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences **42**(6), 1273-1280 (2002).

56. Hodes, L., Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching. Journal of Chemical Information and Computer Sciences, (1976).

57. Willett, P., A screen set generation algorithm. Journal of Chemical Information and Computer Sciences **19**(3), 159-162 (1979).

58. Nilakantan, R., Bauman, N., Dixon, J.S., Venkataraghavan, R., Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. Journal of Chemical Information and Computer Sciences **27**(2), 82-85 (1987).

59. Daylight. Daylight Chemical Informtion Systems, Inc. http://www.daylight.com/.

60. Unity. Tripos Inc. http://www.tripos.com/.

61. Flower, D.R., On the properties of bit string-based measures of chemical similarity. Journal of Chemical Information and Computer Sciences **38**(3), 379-386 (1998).

62. Downs, G.M.a.W., P., Similarity Searching in Databases of Chemical Structures. Reviews in Computational Chemistry. (2007).

63. Bajorath, J., Integration of virtual and high-throughput screening. Nature Reviews Drug Discovery **1**(11), 882-894 (2002).

64. Bender, A., Glen, R.C., A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. Journal of chemical information and modeling **45**(5), 1369-1375 (2005).

65. Wang, N., DeLisle, R.K., Diller, D.J., Fast small molecule similarity searching with multiple alignment profiles of molecules represented in one-dimension. Journal of Medicinal Chemistry **48**(22), 6980-6990 (2005).

66. Willett, J.: Similarity and clustering in chemical information systems. John Wiley & Sons, Inc., (1987).

67. Willett, P., Winterman, V., Bawden, D., Implementation of nonhierarchic cluster analysis methods in chemical information systems: Selection of compounds for biological testing and clustering of substructure search output. Journal of Chemical Information and Computer Sciences **26**(3), 109-118 (1986).

68. Wilton, D.J., Harrison, R.F., Willett, P., Delaney, J., Lawson, K., Mullier, G., Virtual screening using binary kernel discrimination: analysis of pesticide data. Journal of chemical information and modeling **46**(2), 471-477 (2006).

69. Hall, L.H., Kier, L.B., Issues in representation of molecular structure: the development of molecular connectivity. Journal of Molecular Graphics and Modelling **20**(1), 4-18 (2001).

70. Kier, L.B., Hall, L.H.: Molecular connectivity, intermolecular accessibility and encounter simulation. Journal of Molecular Graphics and Modelling **20**(1), 76-83 (2001).

71. Randić, M., The connectivity index 25 years after. Journal of Molecular Graphics and Modelling **20**(1), 19-35 (2001).

72. Wiener, H., Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. Journal of the American Chemical Society **69**(11), 2636-2638 (1947).

73. Wiener, H., Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure. Surface Tension, Specific Dispersion, and Critical Solution Temperature in Aniline. The Journal of Physical Chemistry **52**(6), 1082-1089 (1948).

74. Kier, L. B. and Hall, L. H., *Molecular connectivity in structure-activity analysis*. New York: John Wiley: 1986.

75. Balaban, A.T., Ciubotariu, D., Medeleanu, M., Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. Journal of Chemical Information and Computer Sciences **31**(4), 517-523 (1991).

76. Balaban, A.T., Using real numbers as vertex invariants for third-generation topological indexes. Journal of Chemical Information and Computer Sciences **32**(1), 23-28 (1992).

77. Balaban, A.T., Local versus global (ie atomic versus molecular) numerical modeling of molecular graphs. Journal of Chemical Information and Computer Sciences **34**(2), 398-402 (1994).

78. Randi , M., Wilkins, C.L., Graph theoretical approach to recognition of structural similarity in molecules. Journal of Chemical Information and Computer Sciences **19**(1), 31-37 (1979).

79. Hall, L.H., Kier, L.B., Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. Journal of Chemical Information and Computer Sciences **35**(6), 1039-1045 (1995).

80. Kellogg, G.E., Kier, L.B., Gaillard, P., Hall, L.H., E-state fields: Applications to 3D QSAR. Journal of computer-aided molecular design **10**(6), 513-520 (1996).

81. Randi , M.: On unique numbering of atoms and unique codes for molecular graphs. Journal of Chemical Information and Computer Sciences **15**(2), 105-108 (1975).

82. Kearsley, S.K., Sallamack, S., Fluder, E.M., Andose, J.D., Mosley, R.T., Sheridan, R.P., Chemical similarity using physiochemical property descriptors. Journal of chemical information and computer sciences **36**(1), 118-127 (1996).

83. Sheridan, R.P., Miller, M.D., Underwood, D.J., Kearsley, S.K., Chemical similarity using geometric atom pair descriptors. Journal of chemical information and computer sciences **36**(1), 128-136 (1996).

84. Lewis, R.A., Mason, J.S., McLay, I.M., Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. Journal of Chemical Information and Computer Sciences **37**(3), 599-614 (1997).

85. Xue, L., Godden, J.W., Bajorath, J., Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. Journal of Chemical Information and Computer Sciences **40**(5), 1227-1234 (2000).

86. Xue, L., Stahura, F.L., Godden, J.W., Bajorath, J., Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. Journal of Chemical Information and Computer Sciences **41**(2), 394-401 (2001).

87. Xue, L., Godden, J.W., Stahura, F.L., 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), 1218-1225 (2003).

88. Kogej, T., Engkvist, O., Blomberg, N., Muresan, S., Multifingerprint based similarity searches for targeted class compound selection. Journal of chemical information and modeling **46**(3), 1201-1213 (2006).

89. Vogt, M., Bajorath, J., Bayesian Screening for Active Compounds in High-dimensional Chemical Spaces Combining Property Descriptors and

Molecular Fingerprints. Chemical Biology & Drug Design **71**(1), 8-14 (2008).

90. Rarey, M., Dixon, J.S., Feature trees: a new molecular similarity measure based on tree matching. Journal of computer-aided molecular design **12**(5), 471-490 (1998).
91. Rarey, M., Stahl, M., Similarity searching in large combinatorial chemistry spaces. Journal of computer-aided molecular design **15**(6), 497-520 (2001).
92. Böhm, H.J., Flohr, A., Stahl, M., Scaffold hopping. Drug discovery today: Technologies **1**(3), 217-224 (2004).
93. Schneider, G., Neidhart, W., Giller, T., Schmid, G., "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. Angewandte Chemie International Edition **38**(19), 2894-2896 (1999).
94. Barker, E.J., Buttar, D., Cosgrove, D.A., Gardiner, E.J., Kitts, P., Willett, P., Gillet, V.J., Scaffold hopping using clique detection applied to reduced graphs. Journal of chemical information and modeling **46**(2), 503-511 (2006).
95. Jenkins, J.L., Glick, M., Davies, J.W., A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. Journal of Medicinal Chemistry **47**(25), 6144-6159 (2004).
96. Rush III, T.S., Grant, J.A., Mosyak, L., Nicholls, A., A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. Journal of Medicinal Chemistry **48**(5), 1489-1495 (2005).
97. Brown, R.D., Martin, Y.C., Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. Journal of Chemical Information and Computer Sciences **36**(3), 572-584 (1996).
98. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A., Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. Journal of chemical information and computer sciences **44**(3), 1177-1185 (2004).
99. Brown, R.D., Martin, Y.C., The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. Journal of Chemical Information and Computer Sciences **37**(1), 1-9 (1997).

100. Matter, H., Pötter, T., Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. Journal of Chemical Information and Computer Sciences **39**(6), 1211-1225 (1999).

100. Matter, H., Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. Journal of Medicinal Chemistry **40**(8), 1219-1229 (1997).

102. Whittle, M., Willett, P., Klaffke, W., van Noort, P., Evaluation of similarity measures for searching the dictionary of natural products database. Journal of Chemical Information and Computer Sciences **43**(2), 449-457 (2003).

103. Holliday, J.D., Salim, N., Whittle, M., Willett, P., Analysis and display of the size dependence of chemical similarity coefficients. Journal of Chemical Information and Computer Sciences **43**(3), 819-828 (2003).

104. Holliday, J.D., Hu, C., Willett, P., Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Combinatorial Chemistry &# 38; High Throughput Screening **5**(2), 155-166 (2002).

105. Pearl, J., Probabilistic reasoning in intelligent systems, networks of plausible inference. Morgan Kaufmann Publishers Inc., (1988).

106. Cramer, R.D., Redl, G., Berkoff, C.E., Substructural analysis. Novel approach to the problem of drug design. Journal of Medicinal Chemistry **17**(5), 533-535 (1974). doi:doi:10.1021/jm00251a014.

107. Robertson, S.E., Jones, K.S., Relevance weighting of search terms. Journal of the American Society for Information Science **27**(3), 129-146 (1976).

108. Ormerod, A., Willett, P., Bawden, D., Comparison of Fragment Weighting Schemes for Substructural Analysis. Quantitative Structure-Activity Relationships **8**(2), 115-129 (1989).

109. Ormerod, A., Willet, P., Bawden, D., Further Comparative Studies of Fragment Weighting Schemes for Substructural Analysis. Quantitative Structure-Activity Relationships **9**(4), 302-312 (1990).

110. Harper, G., Bradshaw, J., Gittins, J.C., Green, D.V.S., Leach, A.R., Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. Journal of Chemical Information and Computer Sciences **41**(5), 1295-1300 (2001). doi:doi:10.1021/ci000397q.

111. Xia, X., Maliski, E.G., Gallant, P., Rogers, D., Classification of Kinase Inhibitors Using a Bayesian Model. Journal of Medicinal Chemistry 47, 4463-4470 (2004).

112. Bender, A., Mussa, H.Y., Glen, R.C., Reiling, S., Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. Journal of Chemical Information and Computer Sciences 44, 170-178 (2004).

113. Klon, A.E., Glick, M., Thoma, M., Acklin, P., Davies, J.W., Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. Journal of Medicinal Chemistry 47(11), 2743-2749 (2004).

114. Ribeiro, B.A.N., Muntz, R., A belief network model for IR. In: 1996, pp. 253-260. ACM.

115. 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(2), 462-470 (2006).

116. Nidhi, Glick, M., Davies, J.W., Jenkins, J.L., Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. Journal of Chemical Information and Modeling 46, 1124-1133 (2006).

117. Gasteiger, J., Zupan, J., Neural networks in chemistry. Angewandte Chemie International Edition in English 32(4), 503-527 (1993).

118. Schneider, G., Wrede, P., Artificial neural networks for computer-based molecular design. Progress in biophysics and molecular biology 70(3), 175-222 (1998).

119. Kövesdi, I., Dominguez Rodriguez, M.F., Ôrfi, L., Náray Szabó, G., Varró, A., Papp, J.G., Mátyus, P., Application of neural networks in structure–activity relationships. Medicinal research reviews 19(3), 249-269 (1999).

120. Winkler, D.A., Neural networks as robust tools in drug lead discovery and development. Molecular biotechnology 27(2), 139-167 (2004).

121. Lobanov, V., Using artificial neural networks to drive virtual screening of combinatorial libraries. Drug Discovery Today: BIOSILICO 2(4), 149-156 (2004).
122. Winkler, D.A., Burden, F.R., Application of neural networks to large dataset QSAR, virtual screening, and library design. METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA- **201**, 325-368 (2002).
123. Prasad Y, J., Bhagwat, S.S., Simple neural network models for prediction of physical properties of organic compounds. Chemical Engineering & Technology **25**(11), 1041-1046 (2002).
124. Yan, A., Application of self-organizing maps in compounds pattern recognition and combinatorial library design. Combinatorial Chemistry &# 38; High Throughput Screening **9**(6), 473-480 (2006).
125. Schneider, P., Tanrikulu, Y., Schneider, G., Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. Current medicinal chemistry **16**(3), 258-266 (2009).
126. Taskinen, J., Yliruusi, J., Prediction of physicochemical properties based on neural network modelling. Advanced drug delivery reviews **55**(9), 1163-1183 (2003).
127. Kohonen, T., The self-organizing map. Proceedings of the IEEE **78**(9), 1464-1480 (1990).
128. Devillers, J., Neural networks in QSAR and drug design, vol. 2. Academic Pr, (1996).
129. Ajay, Bemis, G.W., Murcko, M.A., Designing libraries with CNS activity. Journal of Medicinal Chemistry **42**(24), 4942-4951 (1999).
130. Balakin, K.V., Tkachenko, S.E., Lang, S.A., Okun, I., Ivashchenko, A.A., Savchuk, N.P., Property-based design of GPCR-targeted library. Journal of Chemical Information and Computer Sciences **42**(6), 1332-1342 (2002).
131. Balakin, K.V., Lang, S.A., Skorenko, A.V., Tkachenko, S.E., Ivashchenko, A.A., Savchuk, N.P., Structure-based versus property-based approaches in the design of G-protein-coupled receptor-targeted libraries. Journal of Chemical Information and Computer Sciences **43**(5), 1553-1562 (2003).
132. Viswanadhan, V.N., Mueller, G.A., Basak, S.C., Weinstein, J.N., Comparison of a neural net-based QSAR algorithm (PCANN) with hologram-and multiple linear regression-based QSAR approaches: application to 1, 4-

dihydropyridine-based calcium channel antagonists. Journal of Chemical Information and Computer Sciences **41**(3), 505-511 (2001).

133. Hemmateenejad, B., Akhond, M., Miri, R., Shamsipur, M., Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1, 4-dihydropyridines (nifedipine analogous). Journal of Chemical Information and Computer Sciences **43**(4), 1328-1334 (2003).
134. Yasri, A., Hartsough, D., Toward an optimal procedure for variable selection and QSAR model building. Journal of Chemical Information and Computer Sciences **41**(5), 1218-1227 (2001).
135. So, S.S., van Helden, S.P., van Geerestein, V.J., Karplus, M., Quantitative structure-activity relationship studies of progesterone receptor binding steroids. Journal of Chemical Information and Computer Sciences **40**(3), 762-772 (2000).
136. Frank, R., Winkler, D.A., New QSAR methods applied to structure-activity mapping and combinatorial chemistry. Journal of Chemical Information and Computer Sciences **39**(2), 236-242 (1999).
137. Frank, R., Ford, M.G., Whitley, D.C., Winkler, D.A., Use of automatic relevance determination in QSAR studies using Bayesian neural networks. Journal of Chemical Information and Computer Sciences **40**(6), 1423-1430 (2000).
138. Winkler, D.A., Burden, F.R., Bayesian neural nets for modeling in drug discovery. Drug Discovery Today: BIOSILICO **2**(3), 104-111 (2004).
139. Yang, Z.R., Biological applications of support vector machines. Briefings in bioinformatics **5**(4), 328-338 (2004).
140. Burbidge, R., Trotter, M., Buxton, B., Holden, S., Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers & chemistry **26**(1), 5-14 (2001).
141. Warmuth, M.K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., Lemmen, C., Active learning with support vector machines in the drug discovery process. Journal of Chemical Information and Computer Sciences **43**(2), 667-673 (2003).

142. Jorissen, R.N., Gilson, M.K., Virtual screening of molecular databases using a support vector machine. Journal of chemical information and modeling **45**(3), 549-561 (2005).

143. Zhao, C., Zhang, H., Zhang, X., Liu, M., Hu, Z., Fan, B., Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology **217**(2), 105-119 (2006).

144. Geppert, H., Horváth, T., Gärtner, T., Wrobel, S., Bajorath, J., Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. Journal of chemical information and modeling **48**(4), 742-746 (2008).

145. Wassermann, A.M., Geppert, H., Bajorath, J., Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. Journal of chemical information and modeling **49**(3), 582-592 (2009).

146. Paetz, J., Schneider, G., A neuro-fuzzy approach to virtual screening in molecular bioinformatics. Fuzzy sets and systems **152**(1), 67-82 (2005).

147. Paetz, J., Descriptor vector redesign by neuro-fuzzy analysis. Soft Computing-A Fusion of Foundations, Methodologies and Applications **10**(4), 287-294 (2006).

148. Horvath, D., Mao, B., Neighborhood behavior. Fuzzy molecular descriptors and their influence on the relationship between structural similarity and property similarity. QSAR & Combinatorial Science **22**(5), 498-509 (2003).

149. Muller, K.-R., Ratsch, G., Sonnenburg, S., Mika, S., Grimm, M., Heinrich, N., Classifying 'Drug-likeness' with Kernel-Based Learning Methods. Journal of Chemical Information and Modeling **45**(2), 249-253 (2005). doi:doi:10.1021/ci049737o.

150. Zernov, V.V., Balakin, K.V., Ivaschenko, A.A., Savchuk, N.P., Pletnev, I.V., Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. Journal of Chemical Information and Computer Sciences **43**(6), 2048-2056 (2003).

151. Aitchison, J., Aitken, C.G.G., Multivariate binary discrimination by the kernel method. Biometrika **63**(3), 413-420 (1976).

152. Li, Y., Bayesian Model Based Clustering Analysis: Application to a Molecular Dynamics Trajectory of the HIV-1 Integrase Catalytic Core. Journal of Chemical Information and Modeling **46**, 1742-1750 (2006).
153. Klon, A.E., Glick, M., Thoma, M., Acklin, P., Davies, J.W., Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. Journal of Medicinal Chemistry **47**, 2743-2749 (2004).
154. Klon, A.E., Glick, M., Davies, J.W.: Combination of a Naive Bayes Classifier with Consensus Scoring Improves Enrichment of High-Throughput Docking Results. Journal of Medicinal Chemistry **47**, 4356-4359 (2004).
155. Glick, M., Klon, A.E., Acklin, P., Davies, J.W.: Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier. J. Biomol. Screen **9** (2004).
156. Glick, M., Jenkins, J.L., Nettles, J.H., Hitchings, H., Davies, J.W.: Enrichment of High-Throughput Screening Data with Increasing Levels of Noise Using Support Vector Machines, Recursive Partitioning, and Laplacian-Modified Naive Bayesian Classifiers. Journal of Chemical Information and Modeling **46**, 193-200 (2006).
157. Abdo, A., Salim, N., Ahmed, A.: Implementing Relevance Feedback in Ligand-Based Virtual Screening Using Bayesian Inference Network. Journal of biomolecular screening **16**(9), 1081-1088 (2011).
158. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: 1996, pp. 4-11. ACM.
159. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A.: Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. Journal of medicinal chemistry **48**(22), 7049-7054 (2005).
160. Lv, Y., Zhai, C.X.: Positional relevance model for pseudo-relevance feedback. In: 2010, pp. 579-586. ACM.
161. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to information retrieval, vol. 1. Cambridge University Press Cambridge, (2008).
162. Shemetulskis, N., Weininger, D., Blankley, C., Yang, J., Humblet, C.: Stigmata: an algorithm to determine structural commonalities in diverse

datasets. Journal of chemical information and computer sciences **36**(4), 862-871 (1996).

163. Schuffenhauer, A., Floersheim, P., Acklin, P., Jacoby, E.: Similarity metrics for ligands reflecting the similarity of the target proteins. Journal of chemical information and computer sciences **43**(2), 391-405 (2003).
164. Sheridan, R.P.: The centroid approximation for mixtures: calculating similarity and deriving structure-activity relationships. Journal of chemical information and computer sciences **40**(6), 1456-1469 (2000).
165. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: 1994, pp. 232-241. Springer-Verlag New York, Inc.
166. Robertson, S.E., Walker, S., Hancock-Beaulieu, M.M.: Large test collection experiments on an operational, interactive system: Okapi at TREC. Information processing & management **31**(3), 345-360 (1995).
167. Chen, B., Mueller, C., Willett, P.: Evaluation of a Bayesian inference network for ligand-based virtual screening. Journal of cheminformatics **1**(1), 1-10 (2009).
168. Turtle, H., Croft, W.B.: Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems (TOIS) **9**(3), 187-222 (1991).
169. Callan, J.P., Croft, W.B., Broglio, J.: TREC and TIPSTER experiments with INQUERY. Information processing & management **31**(3), 327-343 (1995).
170. James, C.A.W., D.: Daylight theory manual. Chemical Information Systems (1995).
171. Ogawa, Y., Morita, T., Kobayashi, K.: A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy sets and systems **39**(2), 163-179 (1991).
172. Technologies, S., MDL drug data report. (2011).
173. Pipeline Pilot Basic Chemistry Component collection, SciTegic Inc.
174. Moffat, K., Gillet, V.J., Whittle, M., Bravi, G., Leach, A.R.: A Comparison of Field-Based Similarity Searching Methods: CatShape, FBSS, and ROCS. Journal of Chemical Information and Modeling **48**(4), 719-729 (2008). doi:doi:10.1021/ci700130j.

175. Vogt, M., Bajorath, J.: Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints. Chem. Biol. Drug Des. **71**, 8-14 (2008).

176. Tan, L., Lounkine, E., Bajorath, J.: Similarity searching using fingerprints of molecular fragments involved in protein− ligand interactions. Journal of chemical information and modeling **48**(12), 2308-2312 (2008).

177. Glick, M., Jenkins, J.L., Nettles, J.H., Hitchings, H., Davies, J.W.: Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. Journal of chemical information and modeling **46**(1), 193-200 (2006).

178. Chen, B., Harrison, R.F., Papadatos, G., Willett, P., Wood, D.J., Lewell, X.Q., Greenidge, P., Stiefl, N.: Evaluation of machine-learning methods for ligand-based virtual screening. Journal of computer-aided molecular design **21**(1), 53-62 (2007).

179. Siegel, S. and N.J. Castellan, J.(1988). Nonparametric Statistics for the Behavioral Sciences. McGraw-HiU Book Company, New York.

180. Legendre, P.: Species associations: the Kendall coefficient of concordance revisited. Journal of agricultural, biological, and environmental statistics **10**(2), 226-245 (2005).

181. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artificial intelligence **97**(1-2), 245-271 (1997).

182. Beltrán, N.H., Duarte-Mermoud, M.A., Salah, S., Bustos, M., Peña-Neira, A.I., Loyola, E., Jalocha, J.: Feature selection algorithms using Chilean wine chromatograms as examples. Journal of food engineering **67**(4), 483-490 (2005).

183. Liu H, M.H.: Computational Methods of Feature Selection. book (2008).

184. Ellis, D., Furner-Hines, J., Willett, P.: Measuring the degree of similarity between objects in text retrieval systems. Perspectives in Information Management **3**(2), 128-149 (1993).

185. Holliday, J.D., Hu, C.Y., Willett, P.: Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. Comb. Chem. High Throughput Screen **5**, 155 (2002).

186. Salim, N., Holliday, J., Willett, P.: Combination of fingerprint-based similarity coefficients using data fusion. Journal of Chemical Information and Computer Sciences **43**(2), 435-442 (2003).