SHAPE-BASED TWO DIMENSIONAL DESCRIPTOR FOR SEARCHING MOLECULAR DATABASE

HENTABLI HAMZA

UNIVERSITI TEKNOLOGI MALAYSIA

SHAPE-BASED TWO DIMENSIONAL DESCRIPTOR FOR SEARCHING MOLECULAR DATABASE

HENTABLI HAMZA

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

Faculty of Computing
Universiti Teknologi Malaysia

MAY 2014

This research study is dedicated to my wife, to my mother, and in the memory of my father

”If we knew what it was we were doing, it would not be called research, would it?” Albert Einstein

Research is to see what everybody else has seen, and to think what nobody else has thought Albert Szent-Gyorgyi

ACKNOWLEDGEMENT

Alhamdulillah, all praises to Allah S.W.T., The Most Greatest and The Most Merciful for His guidance and blessing, because without him I cant finished this research. I also wish to express my gratitude to my thesis supervisor, Prof. Dr. Naomie Salim for her enthusiastic guidance, valuable help, encouragement and patience for all the aspect of the thesis progress. Her numerous comments, criticisms and suggestion during the preparation of this project are greatly appreciated. Also, for her patience on problems that occurred during the process of completing this thesis.

Furthermore, special thanks go to Dr. Faisal Saeed, for many days of scientific discussions, this work would not have been possible without his support. Many thanks goes to the people who supported me in write this thesis, especially amer tawfiq, Dr.Ali ahmed, Dr.Djamel. I also would like to thank all my friends that give me support and help during the writing of the thesis. Their support and help always give me the motivation and energy I need to finish the thesis. My appreciation also extended to all academic and non-academic member of the Faculty of Computing for their warm heart cooperation during my stay in Universiti Teknologi Malaysia.

My heartfelt acknowledgement are expressed to my family, especially My mother and Wife, without their guidance, support, encouragement and advises I may never have overcome this long journey in my studies. When I felt down, their love always give me the strength to face the challenges of the research. I would also like to thank people that directly or indirectly help me in finishing the thesis. Thank you very much.

ABSTRACT

Biological functions of compounds can be predicted from similarity of their chemical structures to discover new compounds for drug development. Molecular similarity can also be used to infer unknown functions and side effects of existing drugs. A multitude of molecular similarity methods based on different molecular representations have been used to perform virtual screenings. The molecules are transformed into descriptors to create a chemical database which allows mathematical manipulation and searching of the chemical information contained in the molecules. In this research, a new Shape based Descriptor of Molecule (SBDM) was developed based on the 2-dimensional shape of a chemical compound. The outline shape of a molecule is split into parts that are related in graph connectivity. The first atom in the molecule is determined using the Morgan algorithm. The molecular features, such as atom name, bond type, angle and rings are represented using specific symbols based on some specification rules. Subsequent atoms are scanned in a clockwise direction with respect to the first atom. The scan is repeated until the first atom is reached again. Two similarity measures were used to evaluate the performance of the molecular descriptors, which are the Basic Local Alignment Search Tool (BLAST) and the Tanimoto coefficient. The performance of the SBDM is compared with six standard molecular descriptors. Simulation of virtual screening experiments with the MDL Drug Data Report database show the superiority of the shape-based descriptor, with 19.32 % and 34.13 % in terms of average recall rates for the top of 1 % and 5 % retrieved molecules, respectively, compared to the six standard descriptors mentioned earlier.

ABSTRAK

Fungsi biologi sebatian boleh diramal dari persamaan struktur kimia mereka untuk meneroka sebatian baru untuk pembangunan ubat-ubatan. Persamaan molekul juga boleh digunakan untuk menilai fungsi yang tidak diketahui serta kesan sampingan ubat-ubatan sedia ada. Satu kaedah persamaan molekul berdasarkan perwakilan molekul yang berbeza telah digunakan untuk melaksanakan pemeriksaan maya. Molekul-molekul tersebut diubah menjadi penghurai untuk mewujudkan pangkalan data kimia yang membenarkan manipulasi matematik dan pencarian maklumat kimia yang terkandung di dalam molekul-molekul tersebut. Dalam kajian ini, satu Penghurai Molekul baru berasaskan Bentuk (SBDM) dibangunkan berdasarkan bentuk 2-dimensi sebatian kimia. Bentuk rangka molekul dipisahkan kepada bahagian-bahagian yang berkaitan dalam satu hubungan graf. Atom pertama dalam molekul ditentukan dengan menggunakan algoritma Morgan. Ciri-ciri molekul, seperti nama atom, jenis ikatan, sudut dan pergelangan diwakili menggunakan simbol-simbol tertentu berdasarkan beberapa peraturan spesifikasi. Atom-atom yang seterusnya diimbas mengikut arah jam bersandarkan atom pertama. Imbasan ini diulang sehingga atom pertama dicapai semula. Dua ukuran persamaan digunakan untuk menilai prestasi penghurai molekul, iaitu Alat Pencarian Penjajaran Tempatan Asas (BLAST) dan pekali Tanimoto. Prestasi SBDM dibandingkan dengan enam penghurai molekul piawai. Simulasi eksperimen pemeriksaan maya menggunakan pangkalan data Laporan Data Ubatan MDL menunjukkan keunggulan penghurai berdasarkan bentuk, di mana pencapaian 19.32% dan 34.13% dari segi purata kadar ingatan, masing-masing untuk 1% dan 5% molekul tertinggi yang dicapai semula, berbanding dengan enam penghurai piawai yang dinyatakan sebelum ini.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Chemoinformatics is the collection, representation and organization of chemical data to create chemical information in which it can be applied to create chemical knowledge. In pharmaceutical and agrochemical industry, Chemoinformatics has been used for identification of novel compounds with useful and commercially valuable biological properties. The drug discovery process is very complex and it is a multi-disciplinary task with many stages to be performed in a long time. However, the molecule that has the potential to become drugs may cause unexpected long-term side effects and the drug discovery process can take about 12 years and the costs may be up to $350 million per drug [1].

The high costs of bringing a drug into the market have increased the pressure on the pharmaceutical industries. Therefore, attention is given to research and development in order to develop faster and more effective way to produce chemical compounds that can react to the disease and also produce antibodies towards the disease. This has encouraged the study of chemoinformatics and drug discovery as one of the new areas in Malaysia's research and development (R&D) [1].

The primary task of the similarity calculation and molecular searching in chemical database is to find and quantify the relationship between structures of compounds with the hope that the quantity represents similarity in physical, chemical, or biological properties. Thus, an appropriate description (molecular descriptor) of chemical structure for the property to be represented is an essential task. Molecular descriptors are numerical values (or vectors of numbers) that characterize properties of compounds. They can be generated from a machine-readable structure representation like a 2D connection table or a set of experimental or calculated 3D coordinates. Molecular descriptors should be effective (can differentiate between different compounds) and efficient (fast to calculate) by playing a vital role in similarity calculations especially when a large database is used. However, there is a conflict between these two requirements since the most effective descriptors tend to be the least efficient to calculate, and vice versa. Molecular descriptors can be classified into 1D, 2D and 3D descriptors. Molecule representations based on the real shape of the molecules are important because it represents the physical interaction between molecules, especially protein-ligand bindings, instead most of the descriptors are developed based on the molecule features or topological indices.

1.2 Problem Background

In organic chemistry research the chemists synthesize novel compounds that exhibit the desired properties and perform better than existing compounds. The commercial objective is to sell these new and superior compounds directly or in new products containing them. During the process from the initial creation of a new compound until it is commercialized, a lot of data related to it is generated such as physiochemical properties, analytical and toxicological data. All this data needs to be stored and scientists must be able to search it by chemical structure [2].

The chemist performs a chemical reaction and enters the relevant chemical structures, the exact procedure and outcome into this electronic lab notebook (ELN) so that other chemists can find it and repeat the reaction if necessary. After this new compound is synthesized, a new entry for it is generated in the chemical registration system. Unfortunately there is not much scientific literature concerning how and what such registration systems do [2]. The main purpose is to assign each compound a unique in-house reference number that then can be used to cross-reference it. There are many commercial software solutions for ELNs and chemical registrations such as those from PerkinElmer (CambridgeSoft), Accelrys, ChemAxon or Dotmatics to name a few. However these solutions may be costly especially if they also require a commercial relational database management system (RDBMS). An exception is

ChemAxon Compound Registration [2] [3] [4] which runs on MySQL. Another issue is that they are closed-source and if such systems are used in the context of scientific experiments, the experiment is not fully reproducable by other scientists unless they also have access to a license for the same system. Because these systems need to be highly configurable, they are also very complex. Configuring and administrating such a system requires a lot of very specific expertise. Consequentially, it can be advantageous to build your own solution especially for organizations that already have an in-house software development department [5, 6, 7]. A custom solution does not have to be highly configurable as it can be tailored to your needs. This includes administrative interfaces for user management and specific tasks like informing all users about new features. Such tools can greatly reduce administrative overhead [2].

The foundation of a chemical information system is the ability to represent molecules in a computer and to compare a molecule's structure with another. Molecular comparison has been used in the early chemical information systems, e.g. structure and substructure searching [8, 9]. Structure searching involves searching a chemical database for a particular query structure with aims to retrieve all molecules with an exact match to the query structure whereas substructure searching retrieves all molecules that contain the query structure as a subgraph. The equivalence (similarity) between two structures can be achieved by using a graph and subgraph isomorphism algorithms. Isomorphism algorithms are time consuming because it is a combinatorial problem. Various isomorphism algorithms have been developed for efficient performance but they are still too slow for large chemical databases [10, 11, 12, 13], However, structure and substructure searching were later complemented by another searching mechanism called similarity searching.

Similarity searching methods may be the simplest tools for ligand based virtual screening. The basic idea in similarity searching is the similar property principle, which states that structurally similar molecules will exhibit similar physicochemical and biological properties [14]. Over the years, many ways of measuring the structural similarity of molecules have been introduced [15, 16]. 2D similarity methods can be divided into two classes, the first class is the graph-based similarity methods and the second class is the fingerprint-based similarity methods. The graph-based similarity methods directly compare the molecular structures with each other and identify the similar (or common) substructures. These methods relate parts of one molecule to parts of the other molecule, they generates a mapping or alignment between molecules. Maximum common sub-graph method (MCS) is an example of the graph-based similarity methods. Another example of the graph-based similarity methods is the feature trees. The feature trees were introduced by Rarey and Dixon [17], which are the most abstract way of representing a molecule by means of a graph. A feature tree represents hydrophobic fragments and functional groups of the molecule and the way these groups are linked together. Each node in the tree is labelled with a set of features representing chemical properties of the part of the molecule corresponding to the node. The comparison of feature trees is based on matching subtrees of two feature trees onto each other. Feature trees allow similarity searching to be performed against large database, when combined with a fast mapping algorithms [18]. However, the most common similarity approaches use molecules characterized by fingerprints that encode the presence of fragments features in a molecule. The similarity between two molecules is then computed using the number of sub-structural fragments that is common to a pair of structures and a simple association coefficient [19].

1.3 Problem Statement

The shape similarity between two molecules can be determined by comparing the shapes of those molecules; find the overlap volume between them and then use similarity measure (e.g. Tanimoto) to calculate the similarity between the molecules. However, most of the works in shape-based similarity approaches depended on the 3D molecular shape [20, 21]. Recently, shape-based similarity approaches have been used more. The shape comparison program Rapid Overlay of Chemical Structures (ROCS) [22] is used to perceive similarity between molecules based on their 3D shape. The objective of this approach is to find molecules with similar bioactivity to a target molecule but with different chemotypes, i.e., scaffold hopping. However, a disadvantage of 3D similarity methods is that the conformational properties of the molecules should be considered and therefore these methods are more computationally intensive than methods based on 2D structure representation. The complexity increases considerably if conformational flexibility is taken into account. There are many 2D structure representations in a numerical form integer or real. The simplest 2D descriptors are based on simple counts of features such as hydrogen donors, hydrogen bond acceptors, ring systems (such as aromatic rings) and rotatable bonds, whereas the complex 2D descriptors are computed from complex mathematical equations such as 2D fingerprints and topological indices. Topological indices are integer or real value numbers (single value) that represent the constitution of the molecules and can be calculated from the 2D graph representation of molecules and may contain additional property information about the molecule [23, 24, 25]. They characterize molecular structures according to their size, degree of branching and overall shape where the structural diagram of molecules is considered as a mathematical graph, but not the contour of molecule shape. In this thesis, we introduced a new shape-based molecular descriptor that have been inspired by research in information retrieval on the use of contour based shape descriptor for image retrieval systems [26, 27, 28]. Shape-based molecular descriptor is a new method to obtain rough description of 2D molecular structure from its 2D outline shape of its 2D diagram. shape-based molecular descriptor is a textual descriptor that allows rigorous structure specification by the use of a very small and natural grammar.

1.4 Research Question

This study will focus on the following questions:

1. How can we develop new descriptor based on 2D shape of molecule to be used in similarity searching?

2. How can this descriptor be modeled to represent the molecule in fingerprint format?

3. How can graph theory be applied to canonical representation of molecule for similarity searching?

1.5 Research Aim

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, and health researches. The molecules are transformed into numbers, allowing some mathematical treatment of the chemical information contained in the molecule. The aim of this research is to propose and develop a new descriptor for chemical compounds based on their 2D shape used in similarity searching and molecular retrieval.

1.6 Objectives

The objectives of this research are as follows:

1. To develop a new descriptor based on 2D molecular structure shape for similarity searching.

2. To develop a new fingerprint descriptor for shape based representation of chemical compound.

3. To develop canonical representation based on 2D shape of molecule and graph theory for molecular similarity searching.

1.7 Research Scope

In order to achieve the objectives stated above, the scope of this study is limited to the following:

1. Develop the new descriptor based on the shape of molecules extracted from 2D connection table representation.

2. To evaluate the new descriptor and compared with the six standard descriptors (ALOGP, MACCS, EPFP4, CDKFP, PCFP, GRFP), then evaluate using Tanimoto coefficient.

3. The database aimed to be used in this study is only limited to chemical data from MDL Drug Data Report (MDDR).

1.8 Significance and justification of study

The field of molecular descriptors is strongly interdisciplinary and involves a mass of different theories. For the definition of molecular descriptors, knowledge of algebra, graph theory, information theory, computational chemistry, theories of organic reactivity and physical chemistry is usually required. This study is going to produce a descriptor based on shape of molecules, which can be used in similarity searching in chemical database for drug design and discovery.

1.9 Organization of the report

This thesis will be organized into seven Chapters. Chapter 2 will describe the relevant literature review, the fundamentals of chemoinformatics from the molecule descriptors and the molecule searching. Chapter 3 introduces the methodology that is used to build up the proposed descriptor. Chapter 4 introduces the development of the first model labeled as ”Language for Writing Descriptor or Outline Shape of Molecule (LWDOSM)”. The development of the second model 2D Fingerprint Descriptors of Outline Shape of Molecules using the stepwise fragmentation of the first language is presented in Chapter 5 and finally Chapter 6 presents the third and the last model of our work shape-based molecular descriptor SBDM. Chapter 7 discusses and concludes this thesis, highlights the contribution and finding of this work, and provide suggestions and recommendations for future research.

1.10 Summary

The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment. This Chapter gives an overview of the problems involved in the representation of chemical database that store the molecular structures. It highlights the importance of the problem, especially for similarity searching and drug discovery.

REFERENCES

1. Barry Werth. *Billion Dollar Molecule: The Quest for the Perfect Drug*. SimonandSchuster. com, 1995.

2. Norbert Haider. Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. *Molecules*, 15(8):5079–5092, 2010.

3. Bing Xia, Zheng-Fu Tai, Yu-Cheng Gu, Bang-Jing Li, Li-Sheng Ding, and Yan Zhou. Mymoldb: A micromolecular database solution with open source and free components. *Journal of Computational Chemistry*, 32(13):2942–2948, 2011.

4. Bruno Bienfait and Peter Ertl. Jsme: a free molecule editor in javascript. *Journal of cheminformatics*, 5(1):1–6, 2013.

5. John J Irwin and Brian K Shoichet. Zinc-a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

6. Stuart Armstrong, Garrett M Morris, Paul W Finn, Raman Sharma, Loris Moretti, Richard I Cooper, and W Graham Richards. Electroshape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *Journal of computer-aided molecular design*, 24(9):789–801, 2010.

7. Bahram Hemmateenejad, Saeed Yousefinejad, and Ahmad Reza Mehdipour. Novel amino acids indices based on quantum topological molecular similarity and their application to qsar study of peptides. *Amino acids*, 40(4):1169–1183, 2011.

8. Craig and Ebert. Eleven years of structure seraching using the skf (smith,

kline and french) fragmentaton codes. *J. Chem. Doc.*, 9:141–149, 1969.

9. G. G. Vander Stouw, P. M. Elliott, and A. C. Isenberg. Automated conversion of chemical substance names to atom-bond connection tables. *Journal of Chemical Documentation*, 14(4):185–193, 1974.

10. P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, 38:983–994, 1998.

11. ito Kei, Tanaka Nobuya, and Fujita Shinsaku. Development and application of xym2mol system for convertingstructural data by xym notation into connection tables. *Journal of Computer Chemistry, Japan*, 4(2):79–88, 2005.

12. Eleanor M. Ricketts, John Bradshaw, Mike Hann, Fiona Hayes, Neil Tanna, and David M. Ricketts. 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*, 33:905–925, 1993.

13. William Fisanick, Kevin P. Cross, and Andrew Rusinko. Similarity searching on cas registry substances. 1. global molecular property and generic atom triangle geometric searching. *Journal of Chemical Information and Computer Sciences*, 32:664–674, 1992.

14. P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines, and J. Mockus. The cas online search system. 1. general system design and selection, generation, and use of search screens. *Journal of Chemical Information and Computer Sciences*, 23:93–102, 1983.

15. Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25:64–73, 1985.

16. Ash and E Janet. *Communication, storage, and retrieval of chemical information*. E. Horwood; Halsted Press, 1985.

17. R.E. Tarjan. Graph algorithms in chemical computation. *R.E. Christoffersen. American Chemical Society Symposium Series*, 46:1–20, 1977.

18. H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5:107–113, 1965.

19. M. Garey and D. Johnson. The rectilinear steiner tree problem is np-complete. *SIAM Journal on Applied Mathematics*, 32(4):826–834, 1977.

20. Barnard and M. John. Substructure searching methods: Old and new. *Journal of Chemical Information and Computer Sciences*, 33:532–538, 1993.

21. P.M. Dean and R. A. Lewis. *Molecular Diversity in Drug Design*. Kluwer,Amsterdam, 1999.

22. P. Willett. Structural biology in drug metabolism and drug discovery. *Biochemical Society Transactions*, 31(part 3), 2003.

23. P. Willett. Similarity-based virtual screening using 2d fingerprints. *Drug discovery today*, 11(23):1046–1053, 2006.

24. Renxiao Wang and Shaomeng Wang. How does consensus scoring work for virtual library screening? an idealized computer experiment. *Journal of Chemical Information and Computer Sciences*, 41:1422–1426, 2001.

25. P. H. A. Sneath and R. R. Sokal. *Numerical taxonomy. The principles and practice of numerical classification*. San Francisco, 1973.

26. Theodosios Pavlidis. Algorithms for shape analysis of contours and waveforms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-2(4):301–312, july 1980.

27. H. Freeman and A Saghri. Generalized chain codes for planar curves. *Proceedings of the 4th International Joint Conference on Pattern Recognition*, pages 701–703, 1978.

28. H Freeman. On the encoding of arbitrary geometric configurations. *Electronic Computers, IRE Transactions on*, EC-10:260 –268, june 1961.

29. Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. Wiley New York, 1990.

30. Amos Tversky. Features of similarity. *Psychological review*, 84:327–338, 1977.

31. Dennis H Rouvray. Definition and role of similarity concepts in the chemical and physical sciences. *Journal of chemical information and computer sciences*, 32(6):580–586, 1992.

32. Thomas R Hagadone. Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *Journal of chemical information and computer sciences*, 32:515–521, 1992.

33. David Weininger. 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.

34. David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

35. Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. 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.

36. Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, 1947.

37. Balaban and T. Alexandru. Applications of graph theory in chemistry. *Journal of Chemical Information and Computer Sciences*, 25(3):334–343, 1985.

38. Richard A. Lewis, Jonathan S. Mason, and Iain M. McLay. Similarity measures for rational set selection and analysis of combinatorial libraries:the diverse property-derived (dpd) approach. *Journal of Chemical Information and Computer Sciences*, 37:599–614, 1997.

39. Dmitrii Filimonov, Vladimir Poroikov, Yulia Borodina, and Tatyana Gloriozova. Chemical similarity assessment through multilevel neighborhoods of atoms: Definition and comparison with the other descriptors. *Journal of*

Chemical Information and Computer Sciences, 39:666–670, 1999.

40. Sibson R Jardine, N. Mathematical taxonomy. *John Wiley and Sons, NY*, 1971.

41. Zdenek Hubalek. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57:669–689, 1982.

42. D. Ellis, J. Furner-Hines, and P Willett. Measuring the degree of similarity between objects in text-retrieval systems. *Perspect. Inf. Manage.*, 3:128–137, 1994.

43. G. W. Adamson and J. A Bush. A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.*, 15:55–70, 1975.

44. Zhang Dengsheng and Lu Guojun. Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1–19, 2004.

45. Bradley D. Christie, Burton A. Leland, and James G. Nourse. Structure searching in chemical databases by direct lookup methods. *Journal of Chemical Information and Computer Sciences*, 33:545–547, 1993.

46. W. Fisanick, A. H. Lipkus, and A Rusinko. Similarity searching on cas registry substances. 2. 2d structural similarity. *J. Chem. Inf. Comput. Sci.*, 34:130–140, 1994.

47. John Figueras. Substructure search by set reduction. *Journal of Chemical Documentation*, 12:237–244, 1972.

48. Edward H. Sussenguth. A graph-theoretic algorithm for matching chemical structures. *Journal of Chemical Documentation*, 5:36–43, 1965.

49. Annette Von Scholley. A relaxation algorithm for generic chemical structure screening. *Journal of Chemical Information and Computer Sciences*, 24:235–241, 1984.

50. David J. Wild and Peter Willett. Similarity searching in files of

three-dimensional chemical structures. alignment of molecular electrostatic potential fields with a genetic algorithm. *Journal of Chemical Information and Computer Sciences*, 36:159–167, 1996.

51. A Barth. Status and future development of reaction databases and online retrieval systems. *J. Chem. Inf. Comput. Sci.*, 30:384–394, 1990.

52. P. Sheridan Robert and K. Kearsley Simon. Why do we need so many chemical similarity search methods? *Drug Discovery Today*, 7(17):903–911, 2002.

53. Nina Nikolova and Joanna Jaworska. Approaches to measure chemical similarity a review. *QSAR & Combinatorial Science*, 22(9-10):1006–1026, 2003.

54. AnaG. Maldonado, J.P. Doucet, Michel Petitjean, and Bo-Tao Fan. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity*, 10:39–79, 2006.

55. Matthias Rarey and J.Scott Dixon. Feature trees: A new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design*, 12:471–490, 1998.

56. P. J. Hansen and P. C Jurs. Chemical applications of graph theory. *J. Chem. Ed.*, 65:574–586, 1988.

57. J. A. GRANT, M. A. GALLARDO, and B. T. PICKUP. A fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape. *Journal of Computational Chemistry*, 17:1653–1666, 1996.

58. Jennings and Mike Tennant. Selection of molecules based on shape and electrostatic similarity: Proof of concept of electroforms. *Journal of Chemical Information and Modeling*, 47(5):1829–1838, 2007.

59. G. Madhavi Sastry, Steven L. Dixon, and Woody Sherman. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *Journal of Chemical Information and Modeling*, 51(10):2455–2466, 2011.

60. Kirstin Moffat, Valerie J. Gillet, Martin Whittle, Gianpaolo Bravi, and Andrew R. Leach. A comparison of field-based similarity searching methods: Catshape, fbss, and rocs. *Journal of Chemical Information and Modeling*, 48(4):719–729, 2008.

61. Jayashree Srinivasan, Angelo Castellino, Erin K. Bradley, John E. Eksterowicz, Peter D. J. Grootenhuis, Santosh Putta, and Robert V. Stanton. Evaluation of a novel shape-based computational filter for lead evolution: Application to thrombin inhibitors. *Journal of Medicinal Chemistry*, 45(12):2494–2500, 2002.

62. Thomas S. Rush, J. Andrew Grant, Lidia Mosyak, and Anthony Nicholls. 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.

63. Wendy A. Warr. Representation of chemical structures. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(4):557–579, 2011.

64. B Kier Lemont and H Hall Lowell. Molecular connectivity: intermolecular accessibility and encounter simulation. *Journal of Molecular Graphics and Modelling*, 20(1):76–83, 2001.

65. H. Lowell, Hall, and B. Kier Lemont. Issues in representation of molecular structure: The development of molecular connectivity. *Journal of Molecular Graphics and Modelling*, 20(1):4–18, 2001.

66. Randi Milan. The connectivity index 25 years after. *Journal of Molecular Graphics and Modelling*, 20(1):19–35, 2001.

67. Thierry Kogej, Ola Engkvist, Niklas Blomberg, and Sorel Muresan. Multifingerprint based similarity searches for targeted class compound selection. *Journal of Chemical Information and Modeling*, 46(3):1201–1213, 2006.

68. Slimane Larabi, Saliha Bouagar, FelixMiguel Trespaderne, and Eusebiodela-Fuente Lopez. Lwdos: Language for writing descriptors of outline shapes. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, volume 2749

of *Lecture Notes in Computer Science*, pages 1014–1021. Springer Berlin Heidelberg, 2003.

69. Slimane Larabi. Textual description of shapes. *Journal of Visual Communication and Image Representation*, 20(8):563–584, 2009.

70. Ammar Abdo and Naomie Salim. Similarity-based virtual screening using bayesian inference network: Enhanced search using 2d fingerprints and multiple reference structures. *QSAR & Combinatorial Science*, 28(6-7):654–663, 2009.

71. Ammar Abdo and Naomie Salim. Similarity-based virtual screening with a bayesian inference network. *ChemMedChem*, 4(2):210–218, 2009.

72. Forty Liz, Smith Daniel, Jones Lisa, Jones Ian, Caesar Sian, Fraser Christine, Gordon-Smith Katherine, and Craddock Nick. Identifying hypomanic features in major depressive disorder using the hypomania checklist (hcl-32). *Journal of Affective Disorders*, 114(13):68–73, 2009.

73. He Ningning, Wang Xiaoqi, Kim Nayoung, Lim Jong-Seok, and Yoon Sukjoon. 3d shape-based analysis of cell line-specific compound response in cancers. *Journal of Molecular Graphics and Modelling*, 43(0):41–46, 2013.

74. Li Guo-Bo, Yang Ling-Ling, Shan Feng, Zhou Jian-Ping, Huang Qi, Xie Huan-Zhang, Li Lin-Li, and Yang Sheng-Yong. Discovery of novel mglur1 antagonists: A multistep virtual screening approach based on an {SVM} model and a pharmacophore hypothesis significantly increases the hit rate and enrichment factor. *Bioorganic & Medicinal Chemistry Letters*, 21(6):1736–1740, 2011.

75. T. Maalouf Fadi, Brent David, Clark Luke, Tavitian Lucy, Munnell McHugh Rebecca, J. Sahakian Barbara, and L. Phillips Mary. Neurocognitive impairment in adolescent major depressive disorder: State vs. trait illness markers. *Journal of Affective Disorders*, 133(3):625–632, 2011.

76. T. Maalouf Fadi, Clark Luke, Tavitian Lucy, J. Sahakian Barbara, Brent David, and L. Phillips Mary. Bias to negative emotions: A depression state-dependent marker in adolescent major depressive disorder. *Psychiatry Research*, 198(1):28–33, 2012.

77. Z Shi, XH Ma, C Qin, J Jia, YY Jiang, CY Tan, and YZ Chen. Combinatorial support vector machines approach for virtual screening of selective multi-target serotonin reuptake inhibitors from large compound libraries. *Journal of Molecular Graphics and Modelling*, 32:49–66, 2012.

78. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A Schuffenhauer. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.*, 46:462–475, 2006.

79. David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

80. Ye Hu, Eugen Lounkine, and Jurgen Bajorath. Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function. *ChemMedChem*, 4(4):540–548, 2009.

81. Andreas Bender, Jeremy L Jenkins, Meir Glick, Zhan Deng, James H Nettles, and John W Davies. improve retrieval rates in virtual screening and define orthogonal bioactivity space. *Journal of chemical information and modeling*, 46(6):2445–2456, 2006.

82. Dragos Horvath. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *Journal of medicinal chemistry*, 40:2412–2423, 1997.

83. Dik-Lung Ma, Daniel Shiu-Hin Chan, and Chung-Hang Leung. Drug repositioning by structure-based virtual screening. *Chemical Society Reviews*, 42(5):2130–2141, 2013.

84. Claudio N Cavasotto. Homology models in docking and high-throughput docking. *Current topics in medicinal chemistry*, 11(12):1528–1534, 2011.

85. Peter Ripphausen, Britta Nisius, Lisa Peltason, and Jurgen Bajorath. Quo vadis, virtual screening? a comprehensive survey of prospective applications. *Journal of medicinal chemistry*, 53(24):8461–8467, 2010.

86. Tiejun Cheng, Qingliang Li, Zhigang Zhou, Yanli Wang, and Stephen H

Bryant. Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS journal*, 14(1):133–141, 2012.

87. Herve Abdi. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510, 2007.

88. Miroslav Chraska. *Metody pedagogickeho vyzkumu*. Grada Publishing as, 2007.

89. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

90. Chris Oehmen and Jarek Nieplocha. Scalablast: A scalable implementation of blast for high-performance data-intensive bioinformatics analysis. *Parallel and Distributed Systems, IEEE Transactions on*, 17(8):740–749, 2006.

91. Christopher S Oehmen and Douglas J Baxter. Scalablast 2.0: rapid and robust blast calculations on multiprocessor systems. *Bioinformatics*, 29(6):797–798, 2013.

92. Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas Madden. Blast: architecture and applications. *BMC bioinformatics*, 10(1):421–429, 2009.

93. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

94. Chun Wei Yap. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474, 2011.

95. Ammar Abdo, Beining Chen, Christoph Mueller, Naomie Salim, and Peter Willett. Ligand-based virtual screening using bayesian networks. *Journal of Chemical Information and Modeling*, 50(6):1012–1020, 2010.

96. Robert D. Brown and Yvonne C. Martin. Use of structurea activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences*, 36:572–584, 1996.

97. Hans Matter and Thorsten Peter. Comparing 3d pharmacophore triplets and 2d fingerprints for selecting diverse compound subsets. *Journal of Chemical Information and Computer Sciences*, 39:1211–1225, 1999.

98. S. C. Basak, V. R. Magnuson, G. J. Niemi, and R. R Regal. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.*, 19:17–30, 1988.

99. David Vidal, Michael Thormann, and Miquel Pons. Lingo, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *Journal of Chemical Information and Modeling*, 45(2):386–393, 2005.

100. Matthias Rarey and Martin Stahl. Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design*, 15:497–520, 2001.