STM JOURNALS

Research & Reviews: A Journal of
Pharmaceutical Science

ISSN: 2229-7006
Volume 15, Issue 2, 2024
May–August
DOI (Journal): 10.37591/RRJoPS

https://journals.stmjournals.com/rrjops

Review RRJoPS

Quantitative Structure-Activity Relationship in Computer-Aided Drug Design: A Review

Mohd. Wasiullah[1], Piyush Yadav[2], Priyanshu Upadhyay[3], Satish Kumar Yadav[4,*], Manyta Yadav[5]

Abstract

Quantitative Structure-Activity Relationship stands at the forefront of Computer-Aided Drug Design, providing a systematic framework for understanding the relationship between the chemical structure of molecules and their biological activity. The present review delves into the multifaceted realm of quantitative structure-activity relationship methodologies within the landscape of drug discovery. Through an exploration of diverse quantitative structure-activity relationship models, molecular descriptors, validation techniques, and recent advancements, the present article aims to elucidate the pivotal role of quantitative structure-activity relationship in accelerating the development of novel therapeutics. By dissecting the intricacies of quantitative structure-activity relationship analysis and its application in rational drug design, the present review sheds light on the transformative potential of computational modeling in shaping the future of pharmaceutical research. Quantitative structure-activity relationship is a pivotal technique in computer-aided drug design that aids in the rational design of novel drug candidates. The present review provides a comprehensive overview of quantitative structure-activity relationship methodologies, their applications, strengths, limitations, and recent advancements in the field of computer-aided drug design. By exploring various quantitative structure-activity relationship models, datasets, molecular descriptors, and statistical techniques, the present article aims to shed light on the significance of quantitative structure-activity relationship in accelerating drug discovery processes.

Keywords: Receptor, drug designing, biological interaction, molecular docking, modelling, drug discovery, computer-aided drug design, docking, multi-target searching, quantitative structure-activity relationship

***Author for Correspondence**
Satish Kumar Yadav
E-mail: slk.pharma@gmail.com

[1]Principal, Department of Pharmacy, Prasad Institute of Technology, Jaunpur, Uttar Pradesh, India
[2]Academic Head, Department. of Pharmacy, Prasad Institute of Technology, Jaunpur, Uttar Pradesh, India
[3]Research Scholar, Department of Pharmacy, Prasad Institute of Technology, Jaunpur, Uttar Pradesh, India
[4]Associate Professor, Department. of Pharmacy, Prasad Institute of Technology, Jaunpur, Uttar Pradesh, India
[5]Professor, Department of Pharmacy, Prasad Institute of Technology, Jaunpur, Uttar Pradesh, India

Received Date: May 10, 2024
Accepted Date: May 15, 2024
Published Date: July 19, 2024

Citation: Mohd. Wasiullah, Piyush Yadav, Priyanshu Upadhyay, Satish Kumar Yadav, Manyta Yadav. Quantitative Structure-Activity Relationship in Computer-Aided Drug Design: A Review. Research & Reviews: A Journal of Pharmaceutical Science. 2024; 15(2): 55–63p.

INTRODUCTION

Introduction to QSAR and its Significance in Drug Discovery

Quantitative Structure-Activity Relationship (QSAR) is a computational modeling technique that is used to predict the biological activities of chemical compounds based on their structural properties [1]. It establishes a quantitative relationship between the chemical structure of molecules and their observed biological activities, such as binding affinity, enzyme inhibition, or toxicity [2, 3]. By analyzing this relationship, QSAR models assist in identifying promising drug candidates and optimizing their properties before experimental testing (Figure 1) [4].

In the present review, we embarked on a journey into the fundamental principles and applications of QSAR in the context of drug design [1]. We

unravelled the historical evolution of QSAR from its nascent beginnings to its current state-of-the-art methodologies. By dissecting the components of QSAR models, including molecular descriptors, biological activity data, and statistical analysis techniques, we elucidated the intricate mechanisms underlying QSAR analysis [4]. Furthermore, we delved into the nuances of data preparation and dataset selection in QSAR studies, emphasizing the importance of data quality and relevance in model development. We explored the diverse array of QSAR modeling techniques, ranging from traditional linear regression to advanced machine learning algorithms, highlighting their strengths, limitations, and applicability in different contexts [2]. The advent of computer-aided drug design (CADD) has revolutionized the process of drug discovery, offering computational tools and techniques to expedite the identification of potential drug candidates. At the heart of CADD lies QSAR—a discipline that bridges the gap between chemical structure and biological activity through mathematical modelling [1]. QSAR serves as a cornerstone technique, guiding medicinal chemists in the rational design of molecules with desired pharmacological properties [3].

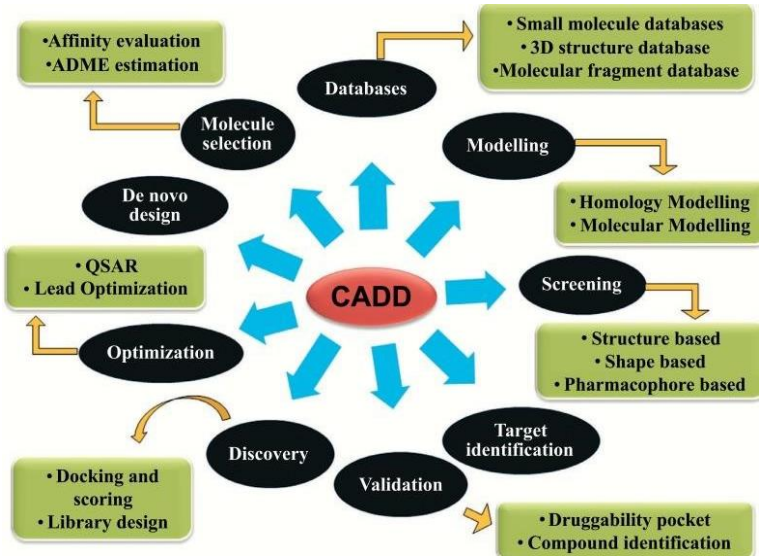

Figure 1. Overview of computer-aided drug design /discovery.

Evolution of QSAR in the Context of CADD

QSAR has undergone significant evolution since its inception in the 1960s, paralleling advancements in computational methodologies and chemical informatics [2]. Initially developed as a tool for understanding the structure-activity relationships of small molecules, QSAR has gradually integrated into CADD workflows. Today, QSAR plays a central role in rational drug design by enabling the prioritization of compounds with desired pharmacological profiles from large chemical libraries [5].

Brief Overview of Computer-Aided Drug Design Approaches

CADD encompasses a range of computational techniques aimed at accelerating the drug discovery process. In addition to QSAR, CADD methodologies include molecular docking, molecular dynamics simulations, virtual screening, and de novo design [3]. These techniques leverage computational algorithms and molecular modeling to predict ligand–receptor interactions, optimize drug candidates, and explore chemical space efficiently. QSAR serves as a complementary tool within the broader framework of CADD, offering insights into the structure-activity relationships of molecules and guiding lead optimization efforts [2].

FUNDAMENTALS OF QSAR

Definition and Concept of QSAR

QSAR is a mathematical modeling approach used to establish a quantitative relationship between the chemical structure of molecules and their biological activity or other properties [6]. The concept behind

QSAR is based on the assumption that the biological activity of a molecule is determined by its physicochemical properties and structural features [7]. QSAR models aim to predict the activity of new compounds based on their structural similarity to known active compounds, thereby guiding the design and optimization of drug candidates [4] (Figure 2).

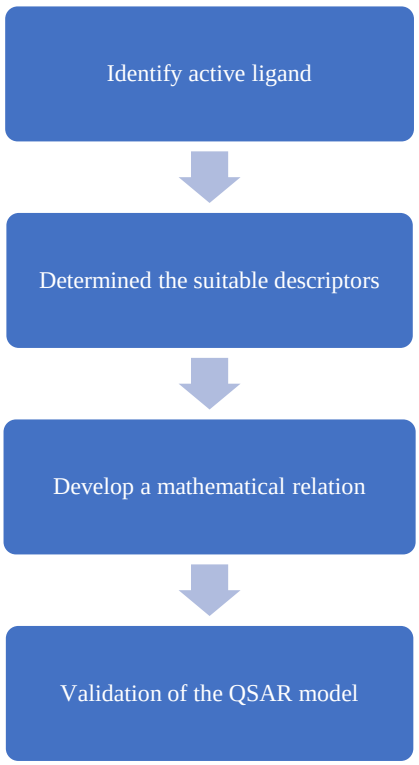

Figure 2. Definition and Concept of QSAR.

Historical Development of QSAR

The origins of QSAR can be traced back to the mid-20th century when Hansch and Fujita introduced the concept of QSARs in the field of medicinal chemistry [5]. QSAR has undergone significant evolution, driven by advancements in computational chemistry, molecular modeling techniques, and statistical methodologies. Early QSAR models were primarily based on simple linear regression equations, but modern QSAR approaches incorporate sophisticated machine learning algorithms and computational tools for model development and validation [8–11].

Components of QSAR Models

Three Main Components

Molecular Descriptors: These are numerical representations of the chemical structure and properties of molecules, such as molecular weight, size, shape, electronegativity, and hydrophobicity. Molecular descriptors serve as input variables in QSAR models and play a crucial role in capturing the structural characteristics relevant to biological activity (Figure 2) [12].

Biological Activity: This refers to the experimental or computational measurements of the biological response or property of interest, such as binding affinity to a target protein, inhibition potency, or pharmacological effect. Biological activity data are used to train QSAR models and validate their predictive performance.

Statistical Analysis: QSAR models employ statistical techniques to establish quantitative relationships between molecular descriptors and biological activity. Common statistical methods used in QSAR include linear regression, multiple linear regression, partial least squares regression, support vector

machines, neural networks, and random forest analysis. Statistical analysis helps to identify the most relevant molecular descriptors and optimize the predictive performance of QSAR models [7].

3D – QSAR Approaches

In 3D QSAR method, the 3D properties of molecule are evaluated by considering individual substituents or moiety. This is a remarkable tool in design of new drugs. In this method the necessary software and hardware are affordable and easy to use [13] (Figure 3).

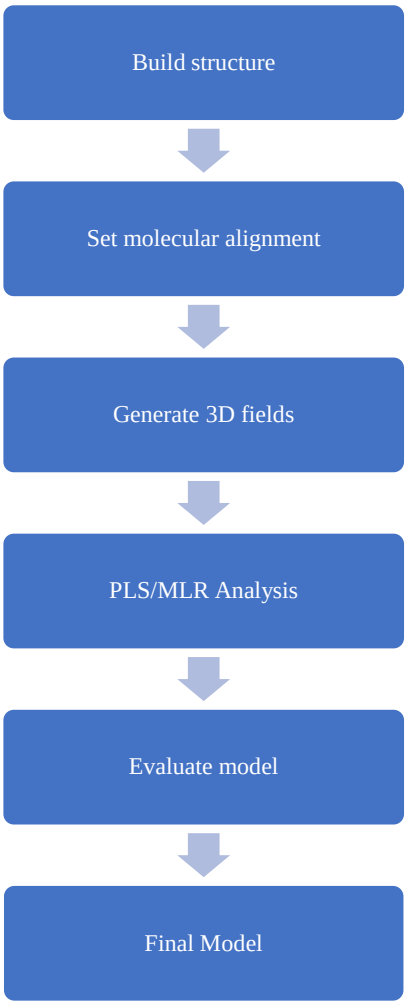

Figure 3. Methodology for 3D QSAR analysis.

Comparative Molecular Field Analysis

In 1987 Cramer developed the predecessor of 3D approaches called Dynamics Lattice Oriented Molecular Modelling System (DYLOMMS). It invites use of principal component analysis (PCA) to extract vectors from the molecular interaction field. After sometimes he modified it by combining the two techniques—GRID and partial least square (PLS)—to develop the powerful 3D-QSAR method known as comparative molecular field analysis (CoMFA). Today CoMFA is a prototype of 3D-QSAR method [12].

Comparative Molecular Similarity Indices Analysis

Comparative molecular similarity indices analysis (CoMSIA) is modified version of CoMFA that focus on ligand-based 3D-QSAR approaches. It is a computational method used in molecular modelling to predict biological activity. It correlates the biological activity of the compound with their molecular features. CoMSIA method analyzes the steric, electric, and hydrophobic property of compound [14].

MOLECULAR DESCRIPTORS

Types of Molecular Descriptors

Molecular descriptors are numerical representations of chemical compounds that capture various physicochemical and structural properties [12]. They can be classified into different types based on the level of molecular detail they represent:

1D Descriptors: Also known as constitutional descriptors, these descriptors encode basic molecular properties such as molecular weight, number of atoms, and chemical composition. Examples include molecular formula, number of hydrogen bond donors and acceptors, and atom counts for different chemical elements [9].

2D Descriptors: These descriptors capture two-dimensional structural information about molecules, such as connectivity, topology, and geometrical features [11]. Common 2D descriptors include molecular fingerprints, fragment counts, topological indices (for example, Wiener index, Zagreb index), and graph-based descriptors.

3D Descriptors: These descriptors represent three-dimensional structural characteristics of molecules, including molecular shape, spatial orientation, and stereochemistry. Examples of 3D descriptors include molecular volume, surface area, molecular conformational flexibility, and pharmacophore features [13].

Calculation Methods and Software Tools for Molecular Descriptors

Several computational methods and software tools are available for the calculation and generation of molecular descriptors. These include [13]:

Quantum Chemical Calculations: Quantum chemical methods, such as density functional theory (DFT) and semi-empirical methods, can be used to calculate molecular properties and descriptors based on electronic structure theory [12].

Chemical Informatics Software: Various software packages and libraries, such as RDKit, Open Babel, and ChemAxon, provide functionalities for calculating a wide range of molecular descriptors from chemical structures.

Online Databases and Web Servers: Online databases and web servers, such as PubChem, ChEMBL, and ChemSpider, offer tools for retrieving molecular descriptors and properties for large compound libraries [13].

Importance of Molecular Descriptors in QSAR Studies

Molecular descriptors play a crucial role in QSAR studies by encoding the structural and physicochemical properties of molecules into numerical features that can be used for quantitative modelling [15]. The selection of appropriate molecular descriptors is essential for capturing the relevant structural features that influence the biological activity or property of interest. By characterizing molecular structures in a quantitative manner, descriptors enable the development of predictive QSAR models that can guide the rational design of new drug candidates and optimization of lead compounds [16]. Moreover, molecular descriptors facilitate the interpretation of QSAR models by identifying key structural determinants of biological activity and guiding medicinal chemistry efforts towards compounds with desired pharmacological profiles [15, 17, 18].

DATA PREPARATION AND DATASET SELECTION

Data Collection and Curation

Data collection is a critical step in QSAR studies, involving the gathering of experimental or computational data on molecular structures and their corresponding biological activities or properties. The following considerations are important in the data collection and curation process [19, 20]:

Data Sources: QSAR datasets can be sourced from various sources, including literature databases, public repositories (for exampke, PubChem, ChEMBL), proprietary databases, and in-house experimental measurements [11].

Data Quality: Ensuring the quality and reliability of the data is essential to the success of QSAR modeling. Data quality control measures may include filtering out erroneous or inconsistent data points, removing duplicates, and standardizing data formats [10].

Data Annotation: Proper annotation of the collected data, including the chemical structures, biological activities, assay conditions, and experimental protocols, is crucial for reproducibility and interpretation of QSAR models [21].

Considerations in Dataset Selection for QSAR Studies

The selection of an appropriate dataset is crucial for the development of robust and reliable QSAR models [14]. Several factors should be considered when choosing a dataset for QSAR analysis, including:

Relevance to the Research Objective: The dataset should be relevant to the specific research question or target of interest. For example, if the goal is to predict the binding affinity of ligands to a particular protein target, the dataset should consist of compounds with known binding affinities for that target.

Diversity and Coverage: A diverse and representative dataset covering a wide range of chemical structures and biological activities is essential for building generalizable QSAR models. Care should be taken to ensure adequate coverage of relevant chemical space.

Data Size: The size of the dataset can impact the performance and generalization ability of QSAR models. Larger datasets provide more training samples and can lead to more robust models, but they may also require more computational resources for analysis.

Data Balance: Imbalance in the distribution of active and inactive compounds in the dataset can affect the performance of QSAR models, particularly in binary classification tasks. Balancing techniques such as oversampling or undersampling may be applied to address this issue [18].

Preprocessing Techniques

Before building QSAR models, it is essential to preprocess the dataset to ensure its quality and suitability for analysis. Common preprocessing techniques include:

Normalization: Scaling the values of molecular descriptors to a common scale to prevent biases due to differences in magnitude [7].

Outlier Detection: Identifying and removing outliers or erroneous data points that may skew the analysis or model predictions [22].

Imputation: Handling missing values in the dataset through techniques such as mean imputation, median imputation, or predictive imputation.

Proper data preparation and dataset selection are crucial steps in QSAR studies, laying the foundation for the development of accurate and reliable predictive models. Careful consideration of these factors can help ensure the validity and applicability of QSAR models in drug design and discovery [23].

QSAR MODELING TECHNIQUES

Overview of QSAR Modeling Approaches

QSAR modeling encompasses a diverse range of approaches, including statistical methods, machine learning algorithms, and hybrid approaches that combine multiple techniques [20]. Each approach has

its advantages and limitations, and the choice of modeling technique depends on factors such as the nature of the dataset, the complexity of the relationship between molecular descriptors and biological activity, and the desired predictive performance of the model [2].

Commonly Used Algorithms

Several algorithms are commonly employed in QSAR modeling, each with its unique characteristics and applicability:

Linear Regression: Linear regression is a classical statistical technique used to model the linear relationship between independent variables (molecular descriptors) and a dependent variable (biological activity). It is simple, interpretable, and suitable for datasets with linear relationships [14].

Random Forest: Random forest is an ensemble learning method that combines multiple decision trees to improve predictive performance. It is robust to overfitting, handles non-linear relationships effectively, and provides feature importance rankings.

Support Vector Machines: Support vector machine (SVM) is a supervised learning algorithm that constructs a hyperplane in a high-dimensional space to separate data points into different classes. SVMs are effective for modeling complex relationships and are particularly useful in binary classification tasks [22].

Neural Networks: Neural networks are computational models inspired by the structure and function of the human brain. They consist of interconnected nodes organized into layers and are capable of learning complex non-linear mappings between inputs and outputs. Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in QSAR modeling tasks.

Gradient Boosting Machines: Gradient boosting machines (for example, XGBoost, LightGBM) are ensemble learning methods that build a series of weak learners sequentially to improve predictive accuracy. They are highly flexible and can capture complex interactions between variables [20].

Strengths and Limitations of Different QSAR Modeling Techniques

Linear regression is simple and interpretable but may not capture complex non-linear relationships [23]. Random forest and gradient boosting methods are powerful for capturing non-linear relationships and handling high-dimensional data but may suffer from overfitting. SVMs are effective for modeling non-linear boundaries and have good generalization performance but may be sensitive to the choice of kernel function and hyperparameters [20]. Neural networks offer flexibility and can learn complex patterns in data but require large amounts of training data and computational resources [21]. The choice of QSAR modeling technique should be guided by the specific characteristics of the dataset and the objectives of the study, with careful consideration of the trade-offs between model complexity, interpretability, and predictive performance [23].

CONCLUSION

In conclusion, QSAR has emerged as a powerful tool in CADD, offering significant advantages in accelerating the drug discovery process. Through the establishment of quantitative relationships between molecular descriptors and biological activity, QSAR models provide valuable insights into the structure-activity relationships underlying drug-target interactions. The present review has provided a comprehensive overview of QSAR methodologies, applications, and recent advancements in the field of CADD. QSAR modeling techniques, ranging from traditional statistical methods to advanced machine learning algorithms, offer diverse approaches for predicting the biological activity of molecules. Each modeling technique has its strengths and limitations, and the choice of approach depends on factors such as the nature of the dataset, the complexity of the relationship between

molecular descriptors and biological activity, and the desired predictive performance of the model. Proper validation of QSAR models is essential to ensure their reliability and generalizability. Cross-validation techniques, such as leave-one-out cross-validation and k-fold cross-validation, provide robust estimates of model performance and help identify potential sources of bias or overfitting. External validation using independent datasets further validates the predictive capabilities of QSAR models and enhances their credibility in real-world applications.

In summary, QSAR represents a cornerstone technique in CADD, offering a data-driven approach to drug design that is poised to shape the future of medicine. By harnessing the power of computational modeling and predictive analytics, QSAR has the potential to accelerate the discovery of novel therapeutics, improve treatment outcomes, and address unmet medical needs in diverse disease areas.

REFERENCES

1. Hou T, and Xu X. Recent development and application of virtual screening in drug discovery: an overview. Current Pharmaceutical Design. 2004; 10(9): 1011–1033p.
2. Yu W, Mac Kerell AD. Computer-aided drug design methods. In: Antibiotics. New York: Humana Press; 2017.
3. Macalino SJY, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. Archives of Pharmacal Research. 2015; 38(9): 1686–1701p.
4. Huang HJ, Yu HW, Chen CY, Hsu CH, Chen HY, Lee KJ, Tsai FJ, Chen CYC. Current developments of computer-aided drug design. Journal of the Taiwan Institute of Chemical Engineers. 210; 41(6): 623–635p.
5. Baig MH, Ahmad K, Roy S, Ashraf JM, Adil M, Siddiqui MH, Khan S, Kamal MA, Provaznik I. Computer aided drug design: success and limitations. Current Pharmaceutical Design. 2016; 22(5): 572–581p.
6. Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. International Journal of Molecular Sciences. 2019; 20(11): 2783p.
7. Kalyaanamoorthy S, Chen YPP. Structure-based drug design to augment hit discovery. Drug Discovery Today. 2011; 16(17–18): 831–839p.
8. Clark DE. What has computer-aided molecular design ever done for drug discovery? Expert Opinion on Drug Discovery. 2006; 1(2): 103–110p.
9. Anderson AC. The process of structure-based drug design. Chemistry & Biology. 2003; 10(9): 787–797p.
10. Wang X, Song K, Li L, Chen L. Structure-based drug design strategies and challenges. Current Topics in Medicinal Chemistry. 2018; 18(12): 998–1006p.
11. Vyas VK, Ukawala RD, Chintha C, Ghate M. Homology modeling a fast tool for drug discovery: current perspectives. Indian Journal of Pharmaceutical Sciences. 2012; 74(1): 1–17p.
12. Kapetanovic IM. Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chemico-Biological Interactions. 2008; 171(2): 165–176p.
13. Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. Proteins: Structure, Function, and Bioinformatics. 2006; 65(1): 15–26p.
14. Melo-Filho CC, Braga RC, Andrade CH. 3D-QSAR approaches in drug design: perspectives to generate reliable CoMFA models. Current Computer-Aided Drug Design. 2014; 10(2): 148–159p.
15. Rackers JA, Wang Z, Lu C, Laurey ML, Lagardere L, Schnieders MJ, Piquemal JP, Ren P, Ponder JW. Tinker 8: software tools for molecular design. Journal of Chemical Theory and Computation. 2018; 14(10): 5273–5289p.
16. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nature Structural Biology. 2002; 9(9): 646–652p.
17. Phillips JC, Zheng G, Kumar S, Kalé LV. NAMD: Biomolecular simulation on thousands of processors. Proceedings of the ACM/IEEE SC 2002 Conference (SC'02); Baltimore, MD, USA; 2002 Nov 16–22.
18. Smith W, Yong CW, Rodger PM. DL_POLY: application to molecular simulation. Molecular Simulation. 2002; 28(5): 385–471p.

19. Prathipati P, Dixit A, Saxena AK. Computer-aided drug design: integration of structure-based and ligand-based approaches in drug design. Current Computer-Aided Drug Design. 2007; 3(2): 133–148p.
20. Schaller D, Sribar D, Noonan T, Deng L, Nguyen TN, Pach S, Machalz D, Barmdez M, Wolber G. Next generation 3D pharmacophore modeling. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2020; 10(4): e1468.
21. Van Drie JH. Generation of three-dimensional pharmacophore models. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2013; 3(5): 449–464p.
22. Vuorinen A, Schuster D. Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. Methods. 2015; 71: 113–134p.
23. Klabunde T, Giegerich C, Evers A. Sequence-derived three-dimensional pharmacophore models for G-protein coupled receptors and their application in virtual screening. Journal of Medicinal Chemistry. 2009; 52(9): 2923–2932p.

