ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

Drug screening based on molecular fingerprint similarity

Luo Hong
Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

Abstract: The structural characterization integrity of molecular fingerprints directly impacts the efficiency and accuracy of drug molecule screening. Single molecular fingerprints suffer from limitations such as incomplete structural representation and insufficient generalization capabilities. To address this challenge, this study employs three complementary molecular fingerprints (MACCS, PubChem, Pharmacophore ErG) to construct a multidimensional molecular characterization system. Molecular similarity is calculated to screen the top 5 candidate molecules by similarity ranking. Subsequently, the screened candidate molecules undergo molecular docking validation with the 4mbs receptor to assess binding affinity and drug-like potential. Experimental results demonstrate that the five selected candidate molecules exhibit high binding affinity and excellent docking scores with the 4mbs receptor, while maintaining good consistency with the query molecules in both chemical structure and molecular fingerprint similarity. This study indicates that the proposed method effectively overcomes the limitations of single fingerprint approaches, significantly enhancing the accuracy and robustness of drug molecule screening. This methodology provides efficient and reliable technical support for candidate molecule selection in the drug discovery process.
Keywords: molecular fingerprint; drug screening; similarity

1. Introduction

Drug molecule screening is a core component of new drug development, with its efficiency and accuracy directly determining the R&D cycle, costs, and success rate. With the rapid advancement of computer-aided drug design technology, screening methods based on molecular structural features have become a research hotspot in drug discovery [1-2]. Among these, molecular fingerprinting serves as a central characterization tool for molecular structures. By encoding a molecule's chemical structure, functional groups, and pharmacodynamic characteristics, it provides critical foundations for molecular similarity assessment and activity prediction [3].

However, single molecular fingerprint approaches exhibit significant limitations in structural characterization: some fingerprints focus on capturing local functional groups, struggling to reflect macroscopic structural features; while others concentrate on two-dimensional structural encoding, failing to characterize the three-dimensional distribution of pharmacophores [4]. Consequently, their accuracy and robustness fall short in complex drug molecule screening tasks, struggling to meet the efficiency demands of new drug development. Therefore, constructing a multidimensional, complementary molecular fingerprint characterization system to overcome the limitations of single fingerprints has become a critical breakthrough for enhancing drug molecule screening performance [5].

Currently, multiple molecular fingerprints are widely applied in drug screening research. Among them, MACCS fingerprint [6], PubChem fingerprint [7], and Pharmacophore ErG fingerprint [8] demonstrate promising application potential in different research scenarios due to their unique structural characterization advantages [9]. The MACCS fingerprint, based on the SMARTS pattern, precisely captures local functional group features of molecules. The PubChem fingerprint, with its high dimensionality and broad structural coverage, effectively characterizes macroscopic molecular structures. The Pharmacophore ErG fingerprint encodes the three-dimensional distribution of pharmacophores, compensating for the spatial representation limitations of two-dimensional fingerprints.

Building upon this, this study innovatively combines three complementary molecular fingerprints (MACCS fingerprint, PubChem fingerprint, Pharmacophore ErG fingerprint) to construct a "local-global-interaction" three-

ISSN: 2424-8460(O)
2251-2608(P)

dimensional molecular characterization system. By integrating molecular similarity computation with molecular docking validation techniques, an efficient drug molecule screening method is established. By screening and identifying candidate molecules with the highest molecular similarity and validating them through docking with the 4mbs receptor, the effectiveness and reliability of this screening method are demonstrated. This approach aims to provide efficient and reliable technical support for candidate molecule selection in new drug development, thereby shortening the R&D cycle and reducing development costs.

2. Results

The positive molecule targeting the 4mbs receptor was designated as the reference molecule. All candidate molecules were characterised using the aforementioned three molecular fingerprinting methods, with their fingerprint similarity to the reference molecule calculated. All candidates were ranked based on molecular similarity scores, ultimately selecting the five candidates exhibiting the highest overall similarity to the query molecule as potential high-affinity candidates. Subsequently, molecular docking technology (AutoDock Vina) [10] was employed to dock the candidate molecules and calculate their affinity. The screening process strictly adhered to the following principles: firstly, structural integrity was ensured by excluding candidates with structural defects or obvious chemical stability deficiencies; secondly, molecules structurally identical to the query compound were removed to avoid redundant validation; thirdly, candidate diversity was maintained to guarantee representativeness of the screening results, providing comprehensive sample support for subsequent pharmacodynamic evaluation.

Following screening, the structures of the five candidate molecules and the query molecule were visualized and compared. This enabled an intuitive analysis of their structural similarity, focusing on the consistency of local functional groups, macroscopic structures, and potential pharmacodynamic characteristics. This analysis laid the foundation for interpreting the subsequent molecular docking validation results.

Figure 1 presents a two-dimensional schematic of the six molecules most closely related to the reference molecule within the molecular fingerprint representation domain. We employed PyMOL [11] to generate molecular docking structures based on the binding affinities of these five molecules with the 4mbs protein, simultaneously annotating their binding affinities with 4mbs. Following our screening steps, the selected candidate molecules exhibit binding affinities to 4mbs ranging from -11.8430 to -10.4830, all highly consistent with the reference molecule's affinity value of -12.2120. Observation of these selected molecules reveals significant structural and functional group similarities, with the leading molecule differing from the reference molecule by only one hydrogen atom. Our approach embeds candidate and reference molecules into a representative feature space using three complementary molecular fingerprint representations. By distinguishing compounds in a chemically meaningful manner, it further demonstrates the ability to learn chemically meaningful representations and enables drug screening to a certain extent.

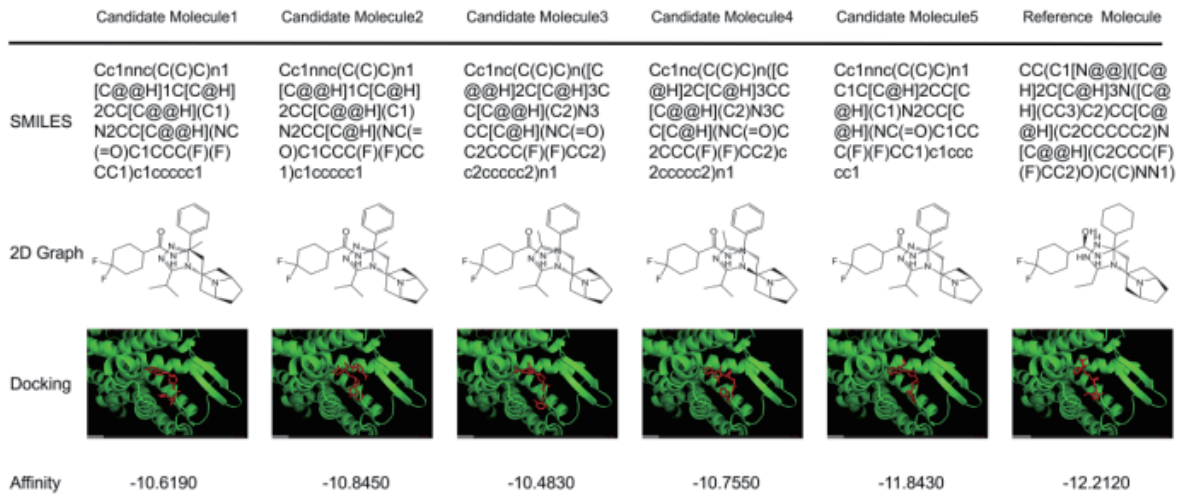

Figure 1. Docking scores and structures of the top five candidate molecules.

ISSN: 2424-8460(O)
2251-2608(P)

3. Materials and methods

3.1. Molecular fingerprints representation

We use here three complementary fingerprints MACCS, PubChem and Pharmacophore ErG fingerprints, as detailed below.

MACCS Fingerprint: A fingerprint based on a substructure key using SMARTS mode. MACCS contains most atomic properties, bond properties, and atomic neighborhoods in different topological separations, which is implicative for drug discovery. We chose the short variant of the 166 bits for this study.

PubChem Fingerprint: An 881-bit fingerprint based on a substructure key with broad chemical structure coverage.

Pharmacomigroup ErG fingerprint: Use the extended reduce graph (ErG) method and pharmacokine type node descriptions applied to encode molecular properties.

Among these, MACCS can effectively capture local functional groups (such as carboxyl groups and pyridine groups). PubChem can supplement large-scale features such as macrocycles and stereocenters. ErG can encode 3D pharmacophore distributions, addressing the limitations of the previous two 2D fingerprints. Therefore, the combination of the three achieves a three-level complementary approach of "local-global-interaction," significantly enhancing the accuracy, robustness, and generalization capabilities of drug discovery task

Combining these three into a hybrid fingerprint to equation 1:

$$FP=CONCAT(FP_{MACCS}, FP_{PubChem}, FP_{Pharmacophore\ ErG}) \quad (1)$$

The fingerprints vector was inputted into the artificial neural network (ANN) to obtain the following representation equation 2:

$$V=W\cdot FP+b \quad (2)$$

3.2. Molecular fingerprints similarity calculation

Molecular similarity calculation is a core step in candidate molecule screening, based on the principle that "similar structures exhibit similar activities." That is, candidate molecules structurally similar to the query molecule are more likely to possess comparable pharmacodynamic activities. This study calculates similarity coefficients between the query molecule and each candidate molecule using three molecular fingerprinting methods. The widely adopted Tanimoto coefficient serves as the similarity evaluation metric. This coefficient effectively measures the overlap between two binary vectors (molecular fingerprints), offering high computational accuracy and efficiency, making it suitable for similarity analysis of high-dimensional fingerprints.

The Tanimoto coefficient is calculated using Equation 3:

$$T(A,B)=\frac{|A\cap B|}{|A\cup B|} \quad (3)$$

Here, A and B represent the fingerprint vectors of the query molecule and candidate molecule, respectively. $|A\cap B|$ denotes the number of bits where both vectors are "1" (shared substructures/pharmacological features), $|A\cup B|$ denotes the number of positions where at least one vector is "1" (all unique substructures/pharmacological features). The Tanimoto coefficient ranges from 0 to 1. A coefficient closer to 1 indicates higher structural similarity between molecules, while a coefficient closer to 0 indicates lower structural similarity.

References

[1] Natesh Singh, Philippe Vayer, Shivalika Tanwar, et al. Drug discovery and development: introduction to the general public and patient groups. Frontiers in Drug Discovery, 3, 2023. ISSN 2674-0338. doi:10.3389/fddsv.2023.1201419. URL https://www.frontiersin. org/journals/drug-discovery/articles/10.3389/fddsv.2023.1201419.

[2] Antonio Lavecchia, Carmen Di Giovanni. Virtual screening strategies in drug discovery: a critical review. Current medicinal chemistry, 20(23):2839–2860, 2013.

[3] Lixia Yao, James A Evans, Andrey Rzhetsky. Novel opportunities for computational biology and sociology in drug discovery: corrected paper. Trends in biotechnology, 28(4):161–170, 2010.

[4] Amol B Deore, Jayprabha R Dhumane, Rushikesh Wagh, et al. The stages of drug discovery and development process. Asian Journal of Pharmaceutical Research and Development, 7(6):62–67, 2019.

[5] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, et al. How to improve r&d productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery, 9(3):203–214, 2010.2. Wang, L. et

ISSN: 2424-8460(O)
2251-2608(P)

al. Conformational space profiling enhances generic molecular representation for AI-powered ligand-based drug discovery. Adv. Sci. 11, e2403998 (2024).

[6] Durant, J.L., Leland, B.A., Henry, D.R, et al.Reoptimization of MDL Keys for Use in Drug Discovery. J. Chem. Inf. Comput. Sci. 2002, 42, 1273–1280.

[7] Stiefl, N., Watson, I.A., Baumann, K., et al. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. J Chem Inf Model 2006, 46, 208–220.

[8] Bolton, E.E., Wang, Y., Thiessen, P.A., et al. PubChem: Integrated Platform of Small Molecules and Biological Activities. In Annual Reports in Computational Chemistry, 2008; Vol. 4, pp. 217–241.

[9] Cai, H., Zhang, H., Zhao, D., et al. FP-GNN: A Versatile Deep Learning Architecture for Enhanced Molecular Property Prediction. Brief Bioinform 2022, 23, bbac408.

[10] Eberhardt, J., Santos-Martins, D., Tillack, A.F., et al. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model 2021, 61, 3891–3898.

[11] DeLano, W.L., Scientific, D., Carlos, S. PyMOL: An Open-Source Molecular Graphics Tool. CCP4 Newsl Protein Crystallogr 2002, 40, 82–92.