Computational Biology and Chemistry 48 (2014) 14–20

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

The optimization of running time for a maximum common substructure-based algorithm and its application in drug design

Jian Chen[a,1], Jia Sheng[b,1], Dijing Lv[a], Yang Zhong[a,c], Guoqing Zhang[b,*], Peng Nan[a,*]

[a] *Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200433, PR China*
[b] *Shanghai Center for Bioinformation Technology, Shanghai 200235, PR China*
[c] *Institute of Biodiversity Science and Geobiology, Tibet University, Lhasa 850000, China*

ARTICLE INFO

Article history:
Received 6 August 2013
Received in revised form 11 October 2013
Accepted 11 October 2013

Keywords:
Running time
MCS
Optimization

ABSTRACT

In the field of drug discovery, it is particularly important to discover bioactive compounds through high-throughput virtual screening. The maximum common substructure-based (MCS) algorithm is a promising method for the virtual screening of drug candidates. However, in practical applications, there is always a trade-off between efficiency and accuracy. In this paper, we optimized this method by running time evaluation using essential drugs defined by WHO and FDA-approved small-molecule drugs. The amount of running time allocated to the MCS-based virtual screening was varied, and statistical analysis was conducted to study the impact of computation running time on the screening results. It was determined that the running time efficiency can be improved without compromising accuracy by setting proper running time thresholds. In addition, the similarity of compound structures and its relevance to biological activity are analyzed quantitatively, which highlight the applicability of the MCS-based methods in predicting functions of small molecules. 15–30 s was established as a reasonable range for selecting a candidate running time threshold. The effect of CPU speed is considered and the conclusion is generalized. The potential biological activity of small molecules with unknown functions can be predicted by the MCS-based methods.

1. Introduction

Drug discovery is a highly complex and multidisciplinary process whose goal is to identify new drugs. Furthermore, it is time consuming and costly to search for new biologically active compounds using traditional high-throughput screening (Carnero, 2006). With the development of organic synthesis in recent years, more and more compounds have been synthesized and traditional high-throughput screening can no longer meet the needs of novel drug discovery (Dobson, 2004). Therefore, it is in urgent need of developing the quantity of drug-like compounds through computer-based predictive methods, such as similarity study and cluster analysis used in identification of drug-like compounds, and druggability prediction of small molecules using computational filtration technology and models (Bender and Glen, 2004; Cheng et al., 2007). Computer-Aided Drug Design (CADD) has become an integral part of drug discovery and development since 1980s (Marshall, 1987; Richards, 1994). Currently, high-throughput screening methods have been developed into integrated approaches that combine both computer-based *in-silico* and traditional *in vitro* screening to reduce the time of R&D and increase the success rate of drug discovery projects (Engels and Venkatarangan, 2001; Shen et al., 2003; Sirois et al., 2005).

There are a number of computational methods for comparing the similarity of two chemical structures (Cao et al., 2008; Quintus et al., 2009). Methods based on substructure and superstructure relationships are commonly used in structure similarity studies of CADD. These types of methods use substructure related to a particular biological activity as the keyword to search an existing database, and the resulting compounds that contain the substructure may share similar activities. However, this comparison strategy is too strict, and does not provide quantitative scores, which leaves the problem of meaningful ordering of the computational results unaddressed.

An alternative type of approach, based on structural descriptors, is also commonly used in structural similarity search and activity prediction. Using these methods, structure descriptors are used to represent chemical structures and generate a quantitative measure of structural similarity. It is worth noting that not structure alignment, but rather comparison of structure

* Corresponding authors. Tel.: +86 21 65642957; fax: +86 21 65642468.
E-mail addresses: gqzhang@scbit.org (G. Zhang), nanpeng@fudan.edu.cn (P. Nan).

[1] These authors contributed equally to this work.

1476-9271/$ – see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.compbiolchem.2013.10.003

descriptors, is used to calculate the similarity score. Examples of the commonly used structural descriptor-based methods include fingerprint, atom pair, *etc.* (Carhart et al., 1985; Chen and Reynolds, 2002). Although in practice this type of approach is simple and effective, one of the major drawbacks is that they cannot directly reflect the overall structural similarity between two compounds.

Maximum common substructure (MCS) refers to the largest substructure shared by two compounds. First, MCS of drug structures is very likely to be the key structural element related to their activity. Second, this method allows the common part of a pair of chemical structures to be easily visualized. In addition, it avoids the disadvantages of many traditional methods, such as an overly strict comparison strategy or being limited to global similarity.

There have been a number of theoretical studies on MCS. Despite the fact that it has long been seen as a promising method in comparing structural similarity between compounds, MCS has drawn far less attention than others, mostly due to the high intrinsic complexity of MCS computation (Conte et al., 2004). Searching for MCS is a very computationally intensive task, and the running time complexity increases exponentially with the number of atoms in the structure of the two molecules. Therefore, for molecules with large structures, the running time that is required to compute the MCS structure can be prohibitively long.

The MCS toolkit developed by Cao et al. is an effective and easy-to-use implementation of an efficient MCS algorithm, and it provides a new avenue for the prediction of bioactive compounds (Cao et al., 2008). In Cao's MCS toolkit, a time-based cap can be applied to the MCS computation, and when the cap is reached and the optimal MCS has not been found, the computation is terminated and the best result searched is returned. In practice, there has always been a trade-off between computational speed and the accuracy of the result. What should be the optimal running time threshold? In this paper we will perform running time evaluation on this MCS algorithm to optimize this method.

All ligand-based activity-prediction methods are built on a commonly accepted principle: the similar property principle (Johnson, 1990), which is based on the experimental observation that structurally similar compounds often exhibit similar physical and chemical properties and biological activities. Based on this principle, many quantitative similarity measures have been proposed to effectively represent the similarity between the compound structures and to predict their potential biological activities. In our study, a structural score of MCS will be characterized and its relationship to biological functions will be explored.

2. Methods

2.1. Data collection and model building

The source code for the MCS-based algorithm developed was obtained from Dr. Cao, University of California Riverside. The structural similarity between two small-molecule compounds was represented by the MCS coefficient:

$$\text{MCS coefficients} = \frac{\text{MCS}(G1, G2)}{\min(G1, G2)}$$

where $\text{MCS}(G1, G2)$ is the number of atoms in the MCS of two small molecules, and $\min(G1, G2)$ is the number of atoms in the smaller molecule of $G1$ and $G2$.

As shown in Table 1, the structural and functional information on the small molecule drugs used in the computation was from essential drugs defined by WHO (Laing et al., 2003) and the FDA-approved small-molecule drugs published by Drugbank (Knox et al., 2011).

Table 1
Data source table.

Data source	Number
WHO essential drug	291
FDA-approved small molecule drug	1405

2.2. Running time evaluation

To perform running time evaluation, we first conducted tests using WHO essential drugs as the training set. These drugs were divided into pairs and the solver included in the MCS toolkit was used to calculate the score; the running time threshold was set as 5 s, 15 s, 30 s, 45 s, 60 s and 80 s, respectively.

We performed the analysis in two ways: first, we calculated the average MCS score at different running time thresholds; and second, assuming that the MCS score calculated at 80 s was final, we computed the difference between the MCS scores obtained at different running time thresholds and the final MCS score. With the confidence interval set to 95%, an independent samples *t*-test was then performed.

Based on the above calculations, we increased the statistical sample size, and paired the FDA-approved small-molecule drugs published on Drugbank website for comparison. Running time threshold was set to 30 s, and the distribution of all computation running times was analyzed. For molecules whose computational running time exceeded 30 s, a new running time threshold was set and the impact of the computation running time cap on the result was evaluated.

2.3. Calculation running time of the different CPU

The data above is based on our server, whose CPU is an Intel® Xeon® E5620 CPU with 2.40 GHz processor. As running time depends on the computer's CPU speed, can the running time threshold set as 30 s be long enough for different machines? To evaluate this, we used two additional machines to evaluate the effect of CPU speed. The CPU of two additional machines are Intel® Xeon® E5620 CPU with 2.00 GHz processor, and AMD® Opteron® 244 with 1.75 GHz processor, respectively. We randomly selected 10 molecule pairs with computational running times longer than 30 s.

2.4. Application in drug design

In drug design, the function prediction of chemical compound is very important. In order to explore the relationship between structural similarity and function, we performed statistical analyses on FDA-approved small-molecule drugs. First, we sampled molecule pairs from the drug set for similarity comparisons, with the computation running time capped at 30 s. Next, the function of each small molecule was labeled. We calculated the proportions of molecule pairs that shared the same functions at MCS score intervals of 1–0.95, 0.95–0.9, 0.9–0.85, 0.85–0.8, respectively. As drugs may have multiple functions, we considered two small molecules in a pair as having the same function if they share one common function when compared.

Furthermore, by studying the listed drugs, we can find their common substructures. Designing compounds on this basis is likely to find more effective and safer drugs.