A method for extraction of the active compound groups which have strong relationship between structure and activity

Takahiro Kishimoto

Shigeto Seno

Yoichi Takenaka

Hideo Matsuda

1 Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University

Keywords: activity prediction, fingerprint, Tanimoto coefficient, clustering

1 Introduction

Huge compound data have been accumulated in the database, and the number of compound data increases more and more. However, almost all compounds have no information about activity. The number of compound data that contains activity information is extremely little with about 1%. Moreover, the cost for the investigation of their activity is enormous by the experiment. Therefore, it is necessary to predict the activity using the computer.

Various compounds have been generated by being continuously modified from prototypic compounds in the development process[1]. The structural changes by this chemical modification add or remove the specific characteristic. It is considered that the structure information contribute significantly to determine activity the compound has. Traditionally, structural tendency was analyzed based on the idea that the compounds with same activity have similar structure. However, most compounds have two or more target proteins and a certain kind of compound have 100 or more target proteins. This fact indicates that not all compounds are similar even if they belong to the same activity group. It is considered that there are the activity groups which have strong relationship between the structure and the activity and the activity group which have no relationship. In the case of the activity group which has strong relationship between structure and activity, it could be easier to computationally predict the activity from structural information. In this study, in order to find the activity which could be predicted more accurately, we propose a method to estimate the strength of relationship between the activity and structure for each activity group.

We investigated the relationships between the structure and the activity using about 4800 drug data in the DRUG BANK database[2].

2 Method

To define similarities between two compounds, we introduced fingerprint and represented each compound structure using fingerprint generated from structure information. We consider the similarity between two fingerprints as the similarity between two compounds. This similarity score is calculated using the Tanimoto coefficient[3].

Several types of clustering were performed for all drug structure information using the distance matrix obtained from the similarity scores. And several patterns of threshold of the parameter were set to define a group of similar drugs.

For each combination of clustering method and parameter value, we took the statistics of distribution of targets for each group of similar drugs. For each target we calculated percentage of compounds act on the target, and ranked targets. Here, the target was higher in rank has the strong correlation between the structure and the activity, and we consider that it is easy to predict the activity by extracting the structural features in the group of drugs.

We consider that extracting such targets is useful for accuracy improvement of the computational activity prediction.
3 Results and Discussions

We used MACCS KEY[4] generated from drug structural information “canonical smiles” as the fingerprint and group average method as the clustering method in this study.

All drug data used in this study were collected from the DRUG BANK database. 4868 drug structure data and 4535 target data.

Each entry of drug data contains the name of drug called “DRUGBANK_GENERIC_NAME”, the ID of drug called “DRUGBANK_ID”, several types of structure information called “mol” and “DRUGBANK_CANONICAL_SMILES” and “DRUGBANK_ISOMERIC_SMILES”, the molecular weight and so on. While each entry of target data contains the target name such as protein the drug actually acts on, the ID of it, “DRUGBANK_ID” of the drug acts on it, the sequence information of it.

Here, we assign target information the drug acts on to each drug information. Table1 shows obtained information as a result.

While Figure1 shows the relationship between a specific target protein and the drug acts on it. In (a), the similarity between the drugs is low and there are few common structures of all drugs. Thus, the correlationship between the structure and the activity is poor. While in (b), the similarity between the drugs is high and there are clearly common structures at first sight. Thus, the correlationship between the structure and the activity is strong. We found that while there are some kinds of target protein drug structure information contribute heavily to determine whether the drug acts on it, there are some kinds of target protein drug structure information seldom contribute. Therefore we consider that exacting the target protein has strong relationship between the structure and activity is very important.

Table1 DRUG BANK

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug entries</td>
<td>4868</td>
</tr>
<tr>
<td>Target entries</td>
<td>4535</td>
</tr>
<tr>
<td>Maximum number of targets one drug have</td>
<td>185</td>
</tr>
<tr>
<td>Average of targets one drug have</td>
<td>2.67</td>
</tr>
<tr>
<td>Maximum number of drugs act on same target</td>
<td>79</td>
</tr>
<tr>
<td>Average of drugs act on same target</td>
<td>2.63</td>
</tr>
</tbody>
</table>

(a) Figure1 The target protein and the drugs act on it

(b)

References