COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 5, Issue: 6, February 2013, e201302009, http://dx.doi.org/10.5936/csbj.201302009

CSBJ

Methods for Similarity-based Virtual Screening

Thomas G. Kristensen [a,#], Jesper Nielsen [a,†], Christian N. S. Pedersen [a,*]

Abstract: Developing new medical drugs is expensive. Among the first steps is a screening process, in which molecules in existing chemical libraries are tested for activity against a given target. This requires a lot of resources and manpower. Therefore it has become common to perform a virtual screening, where computers are used for predicting the activity of very large libraries of molecules, to identify the most promising leads for further laboratory experiments. Since computer simulations generally require fewer resources than physical experimentation this can lower the cost of medical and biological research significantly. In this paper we review practically fast algorithms for screening databases of molecules in order to find molecules that are sufficiently similar to a query molecule.

Mini Review Article

Introduction

Proteins are a class of macromolecules that play some of the most important roles in nature. Proteins have functions as catalysts, in signaling and in structural roles. A protein consists of one chain (or multiple chains) of amino acid residues that fold into a more or less rigid structure that has a biological function. A protein that functions as a catalyst will have a certain place, called the *binding site*, where other molecules will *dock* as the protein performs its function. The binding site is usually an indentation or cave in the structure of the protein. A *ligand* is a molecule that docks with another molecule, such as a protein, to perform some function, see figure 1.

One way to combat a disease is to find a ligand that will dock with a protein important for that disease, and disrupt its normal function. In general one will have a chemical library of molecules that are available for manufacturing. Using computers for predicting the activity of very large libraries of molecules to identify the most promising leads for further laboratory experiments is called *virtual screening*. Simulating the docking between the protein and each ligand on a computer in order search for promising ligands in a library of available molecules requires a lot of computing time and available protein structures.

Instead one may rely on the idea that similar structure leads to similar properties, and predict the properties of a molecule by studying the properties of similar molecules. Hence, if one has identified a ligand that binds to a given target, for example from another medical drug, or observed in nature, one may find other candidate ligands by looking for ligands in a chemical library or database that are similar to the known binder. This similarity- and ligand-based approach to virtual screening works well for the right formalizations of how to represent molecules and quantify their similarity [25]. Due to the size of chemical databases such as PubChem [4] and ChemDB [6], the similarity-baed approach to virtual screening also needs efficient methods for screening a database of molecular representations for molecules that are sufficiently similar to a query molecule. In this paper we review such screening methods for molecules represented as fingerprints or SMILES strings.

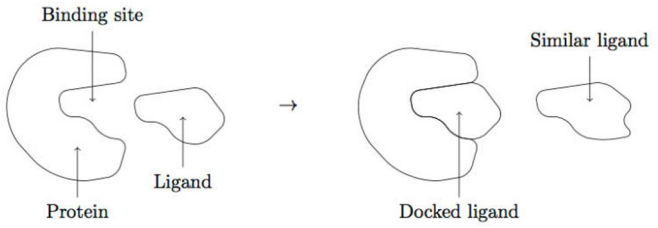

Figure 1. A ligand docking to a protein. Another ligand may dock with the same protein, if it is sufficiently similar.

Representing molecules

It is not immediately obvious how to measure the similarity between two molecules. However, some quite simple measures have proven to be surprisingly good when used for virtual screening [14,22]. For example one might compute a bit-string encoding representative information about the molecules and use the similarity between the bit-strings as a measure of the similarity between the molecules. Such a bit-string for a molecule is denoted a *fingerprint*.

There are many ways to compute the actual fingerprints [5]. One general approach is to select a set of features, each of which a molecule may or may not have. Each feature will then correspond to one bit in the fingerprint, and that bit will be set or not, according to whether the given molecule has the feature [26]. Fingerprints of this form will often be quite long, and with many bits set to zero. To use space more efficiently they can be hashed compressed into shorter fingerprints [1,2,15]. One might also represent a molecule by a *counting vector* of integers, where each integer counts how many

[a]*Bioinformatics Research Center, Aarhus University, C. F. Møllers Allé 8, DK-8000 Aarhus C, Denmark*
#Now employed by Trifork Gmbh
†Now employed by Google Inc

* Corresponding author.
E-mail address: cstorm@birc.au.dk (Christian N. S. Pedersen)

times a certain feature occurs in the molecular. A counting vector allows for a more detailed description of the molecule as a multi-set of features, where a binary fingerprint as introduced above simply describes the molecule as a set of features. However, as counting vectors can easily be converted into binary vectors, for example as illustrated in figure 2, algorithms for handling binary vectors such as the ones reviewed in this paper are also applicable for counting vectors.

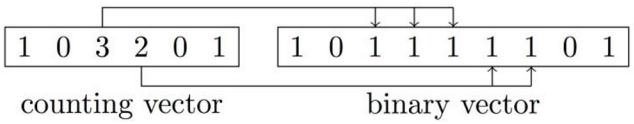

Figure 2. An illustration of converting a counting vector into a binary vector.

The Simplified Molecular Input Line Entry Specification (SMILES) [23] is a standard way to encode the two dimensional structure of a molecule in a one dimensional string that has a canonical form such that every molecule can be represented by a unique SMILES. The SMILES string is generated by writing a sequence of letters, one for each atom type, marking branches with parentheses and rings with numerical indexes. As an example, consider the visualization of 3-cyanoanisole in figure 3, which can be represented by the SMILES string "COc(c1)cccc1C\# N". The main path of the molecule is the string "COccccC\# N", the hash mark symbolizing a triple bond. (c1) marks the branch containing just one carbon atom, and the number "1" here and later in the path defines the bond between the two carbon atoms.

Any string of length $n \geq q$ will have exactly n-q+1 substrings of length q. In [21], a substring, of length q, of a SMILES string is called a LINGO. Thus the SMILES string of a ligand can be viewed as a multi-set of LINGOs, which in [21] is called the LINGO profile of the molecule.

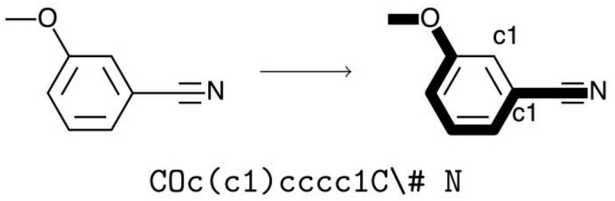

Figure 3. Illustration of a possible SMILES string for 3-cyanoanisole. The primary backbone is highlighted with thick lines. c1 indicates the two points where the ring is merged.

Similarity between molecules

There are of course several ways to quantify the similarity between two sets (or multi-sets) of features, but the *Tanimoto coefficient* has proven very useful [24,26]. If A and B are sets, or multi-sets, of features, then the Tanimoto coefficient, $S_T(A, B)$, is:

$$S_T(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

If A and B are given as two bit-strings, then the Tanimoto coefficient becomes:

$$S_T(A,B) = \frac{|A \wedge B|}{|A \vee B|},$$

where $\wedge$ and $\vee$ are bitwise logical 'and' and logical 'or' respectively, and $|A|$ is the number of bits set to one in the bit-string A. See figure 4 for an example.

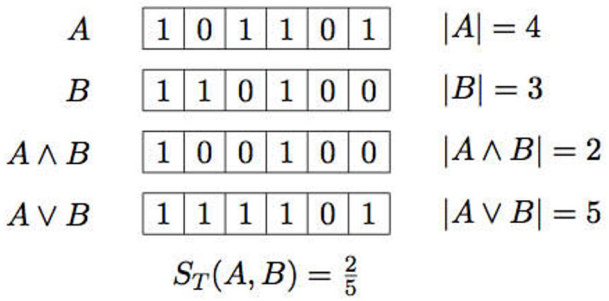

Figure 4. The notation used for bit-strings.

The Tanimoto coefficient as defined above quantifies the similarity between two bit-strings as a number in the interval [0;1], where 0 says that the two bit-strings have no one-bits in common, and 1 says that the two bit-strings are equal. The coefficient is only defined if there is at least one bit set to one in the two bit-strings (i.e. one feature is shared), which is a very reasonable assumption for molecular fingerprints.

Recall, that the LINGO profile of a molecule is the multi-set of LINGOs in its SMILES string. The similarity between two ligands can thus be measured as the Tanimoto coefficient between their LINGO profiles. This measure is called the LINGOsim between the ligands [21].

One of the major motivations for quantifying molecular similarity is to identify molecules for medical drugs. The problem can be formalized as: We are given a database of representations (for example fingerprints or SMILES) of synthesizable molecules, a query molecule A, and a minimal similarity S_{MIN}. The task is then to find all molecules B in the database where $S_T(A, B) \geq S_{MIN}$. This query can of course be performed by a naive screening of the database, where we examine every fingerprint A in the database to compute $S_T(A, B)$. However, due to the typical size of the database, this is not a desirable approach. In the following sections, we review how to perform such queries more efficiently in practice. We first consider the problem for molecules represented as bit-strings (fingerprints), and secondly, for molecules represented as SMILES.

Searching for molecules with similar fingerprints

Given a database of N fingerprints of length n, a query fingerprint A (also of length n), and a minimal similarity S_{MIN}. We want to find all molecules B in the database where $S_T(A, B) \geq S_{MIN}$. In [19] it is observed that since $|A \wedge B| \leq \min(|A|, |B|)$ and $|A \vee B| \geq \max(|A|, |B|)$, then we can upper-bound the similarity between A and B by

$$S_T(A,B) = \frac{|A \wedge B|}{|A \vee B|} \leq \frac{\min(|A|,|B|)}{\max(|A|,|B|)} = \text{COUNT-MAX}\,(A,B).$$