BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 12 2012, pages 1661–1662
doi:10.1093/bioinformatics/bts249

Structural bioinformatics

Advance Access publication April 26, 2012

DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets

Adrià Cereto-Massagué[1], Laura Guasch[1], Cristina Valls[1], Miquel Mulero[1], Gerard Pujadas[1,2] and Santiago Garcia-Vallvé[1,2,*]

[1]Grup de Recerca en Nutrigenòmica, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus de Sescelades, C/Marcel.lí Domingo s/n, 43007 Tarragona and [2]Centre Tecnològic de Nutrició i Salut (CTNS), TECNIO, CEICS, Camí de Valls 81-87, 43204 Reus, Catalonia, Spain

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Decoys are molecules that are presumed to be inactive against a target (i.e. will not likely bind to the target) and are used to validate the performance of molecular docking or a virtual screening workflow. The Directory of Useful Decoys database (http://dud.docking.org/) provides a free directory of decoys for use in virtual screening, though it only contains a limited set of decoys for 40 targets.To overcome this limitation, we have developed an application called DecoyFinder that selects, for a given collection of active ligands of a target, a set of decoys from a database of compounds. Decoys are selected if they are similar to active ligands according to five physical descriptors (molecular weight, number of rotational bonds, total hydrogen bond donors, total hydrogen bond acceptors and the octanol–water partition coefficient) without being chemically similar to any of the active ligands used as an input (according to the Tanimoto coefficient between MACCS fingerprints). To the best of our knowledge, DecoyFinder is the first application designed to build target-specific decoy sets.

Availability: A complete description of the software is included on the application home page. A validation of DecoyFinder on 10 DUD targets is provided as Supplementary Table S1. DecoyFinder is freely available at http://URVnutrigenomica-CTNS.github.com/DecoyFinder

Contact: santi.garcia-vallve@urv.cat

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 4, 2012; revised on April 10, 2012; accepted on April 23, 2012

1 INTRODUCTION

Ligand enrichment is a key metric for assessing the performance of molecular docking or virtual screening workflows. It involves measuring the ability of a method or procedure to discriminate between active and inactive compounds. However, sufficient amounts of inactive compounds are generally not available for such testing; thus, decoys (i.e. molecules that are presumed to be inactive against the examined target) are commonly used for this purpose (Kirchmair *et al.,* 2008). To avoid bias and to ensure that the enrichment is not simply due to physical differences between active and decoy compounds, decoys should exhibit physical properties (e.g. molecular weight and calculated log $P-$values) that are similar to active compounds, while still being chemically distinct from them (Huang *et al.*, 2006). The largest publicly accessible database of decoys is the Directory of Useful Decoys (DUD; Huang *et al.*, 2006; Irwin, 2008), which is available at http://dud.docking.org/. The DUD contains known active and decoy compounds for 40 target proteins and is currently the gold standard for benchmarking virtual screening and molecular docking algorithms. However, the DUD only contains decoys for a small set of protein targets and has several limitations, such as the possibility of identifying a larger decoy set and the risk of overfitting (i.e. inadvertently tuning algorithms and score functions to perform well on a single benchmark; Irwin, 2008; Wallach and Lilien, 2011). To overcome these limitations, we have created an application called DecoyFinder that selects, for a collection of active ligands of a protein target, a set of decoys from a database of compounds. To the best of our knowledge, DecoyFinder is the first application that is designed to build target-specific decoy sets.

2 PROGRAM OVERVIEW

2.1 Input files

The input files that are used by DecoyFinder contain a set of active molecules (called queries) for a particular target and additional files containing a set of molecules (called potential decoys) from which decoys will be selected. These files can be in sdf, mol or any other format that is recognized by OpenBabel (http://openbabel.org; O'Boyle *et al.*, 2011), including compressed files. For the potential decoy set, the program is able to directly use subsets of the ZINC database (Irwin and Shoichet, 2005) and provides the option, if enabled, to store these subsets as cache files and use them several times. To avoid bias, when reading the potential decoy files and to enable the acquisition of different decoy sets when DecoyFinder is re-run, potential decoy files are read in a different random order each time. In addition, it is possible to use a third file input option to submit files containing a set of known decoy molecules or decoys that have been previously selected (called known decoys) using the 'add new decoys' function. These known decoy compounds will not be re-evaluated to determine whether they are decoys, but will be considered when searching for new decoys and will be included in the resulting decoy set.

*To whom correspondence should be addressed.

2.2 Algorithm for decoy selection

The algorithm for decoy selection implemented in DecoyFinder is similar to that used to construct the DUD database (Huang *et al.*, 2006; Irwin, 2008) and other benchmarks (Wallach and Lilien, 2011). MACCS fingerprints (Durant *et al*., 2002) and five physical descriptors are calculated for each active and potential decoy molecule using the OpenBabel toolbox (O'Boyle *et al.*, 2011). The Tanimoto coefficients between the MACCS fingerprints of each potential decoy and active molecule and between the potential decoys are then calculated. For each active molecule included in the query, DecoyFinder selects a set of decoys (36 when the default program options are used) from either the ZINC database or any set of molecules that is used as an input. Molecules are considered to be decoys if the following conditions are met:

- They are similar to the active molecule according to five physical descriptors: molecular weight, the number of rotational bonds, total hydrogen bond donors (HBDs), total hydrogen bond acceptors (HBAs) and the octanol–water partition coefficient (log *P*). Thus, the decoy compounds exhibit physical properties that are similar to active compounds, which prevent bias and ensure that the enrichment is not simply due to physical differences between the active and decoy compounds. Using the default program options, the physical descriptors of a decoy are considered to be similar to those of an active ligand if the following conditions are met: (1) the molecular weight is within 25 Da of the active ligand; (2) they contain the same number ± 1 of rotational bonds and HBDs, and the same number ± 2 of HBAs and (3) the log *P*-value is within 1.0 of the active ligand. These constraint values can be relaxed in cases where a full decoy set cannot be generated or would take too much time to complete.
- The Tanimoto coefficients between a potential decoy and each of the active molecules are not greater than a defined threshold (with the default set to 0.75). Thus, decoys are chemically different from any of the active molecules of the query.
- The Tanimoto coefficients between a potential decoy and previously selected decoys are not greater than a defined threshold (with the default set to 0.9). This reduces the incidence of analogous structures between decoys and the bias of analogue or trivial enrichment when decoys are used in a virtual screening workflow validation (Irwin, 2008).

As a validation, an analysis of the performance of the decoys obtained with DecoyFinder when using GlideSP to score actives and decoys for 10 DUD targets can be found in Supplementary Table S1.

2.3 Output

The output of DecoyFinder is an sdf file containing the decoy molecules for a specific target and a Comma-separated values (CSV) file that contains information regarding the sdf file and the decoy search options. When a full decoy set cannot be generated, the program displays a warning message and redirects the output to the input screen of the 'add new decoys' option. Thus, the user can attempt to complete the decoy set by either using a different library of potential decoy compounds or relaxing the constraints used.

3 IMPLEMENTATION AND SYSTEM REQUIREMENTS

DecoyFinder has been developed as a python graphical user interface (GUI) application. It has the following dependences:

- Version 4.6 or higher of Nokia's Qt framework (http://qt.nokia.com). DecoyFinder uses this framework for its GUI.
- OpenBabel (http://openbabel.org) version 2.3.0 or higher with python bindings (O'Boyle *et al.*, 2008, 2011). Prior versions contained a bug that prevented DecoyFinder from working. OpenBabel is a powerful cheminformatics toolkit that we use to parse molecule files and calculate molecular properties.
- Python version 2.6 or higher (but lower than version 3.0).
- Python Qt bindings: either PySide 1.0 or higher or PyQt4.

A version of DecoyFinder for Ubuntu 10.10 (and newer versions), another one for Fedora 16 and a Windows version that includes all the dependencies, as well as the source code and several tools (e.g. a Wiki, documentation and a bug tracking system), are available at http://URVnutrigenomica-CTNS.github.com/DecoyFinder.

ACKNOWLEDGEMENTS

This manuscript has been edited by American Journal Experts.

Funding: 'Ministerio de Educación y Ciencia', Spanish Government [AGL2008-01310 and AGL2011-25831], and ACC1Ó program from 'Generalitat de Catalunya' [TECCT11-1-0012].

We acknowledge support from the Generalitat de Catalunya through grant XRQTC.

Conflict of Interest: none declared.

REFERENCES

Durant,J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.,* **42**, 1273–1280.

Huang,N. *et al.* (2006) Benchmarking sets for molecular docking. *J. Med. Chem.*, **49**, 6789–6801.

Irwin,J.J. (2008) Community benchmarks for virtual screening. *J. Comput. Aided Mol. Des.*, **22**, 193–199.

Irwin,J.J. and Shoichet,B.K. (2005) ZINC: a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, **45**, 177–182.

Kirchmair,J. *et al.* (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? *J. Comput. Aided Mol. Des.*, **22**, 213–228.

O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.,* **2**, 5.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminf.,* **3**, 33.

Wallach,I. and Lilien,R. (2011) Virtual decoy sets for molecular docking benchmarks. *J. Chem. Inf. Model.*, **51**, 196–202.