cando.py: Open source software for analyzing large scale drug-protein-disease data

William Mangione [1], Zackary Falls [1], Gaurav Chopra [2] and Ram Samudrala [1,*]
[1]Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, 14120, United States;
[2]Department of Chemistry, Purdue Institute for Drug Discovery, Integrated Data Science Institute, Purdue University, West Lafayette, IN, 47907, United States

*To whom correspondence should be addressed.

Abstract

Motivation: Elucidating drug-protein interactions is essential for understanding the beneficial effects of small molecule therapeutics in human disease states. Common drug discovery methods focus on optimizing the efficacy of a drug against a single biological target of interest. However, evidence supports the multitarget theory, i.e., drugs work by exerting their therapeutic effects via interaction with multiple biological targets. Analyzing drug interactions with a library of proteins can provide further insight into disease systems while also allowing for prediction of putative therapeutics.
Results: We present the CANDO Python package for analysis of drug-proteome and drug-disease relationships. This package allows for rapid drug similarity assessment, most notably via the bioinformatic docking protocol in which protein interactions can be quickly scored for thousands of compounds. The platform can be benchmarked through a variety of protocols to determine how well drugs are related to each other in terms of the indications/diseases for which they are approved. Drug predictions are generated through consensus scoring of the most similar compounds to drugs known to treat a particular indication.
Availability: The CANDO Python package is available on GitHub at https://github.com/ram-compbio/CANDO, through the Conda Python package installer, and at http://compbio.org/software/.
Contact: ram@compbio.org

1 Introduction

Drugs and small molecule compounds exert therapeutic effects via the perturbation of multiple macromolecules, especially proteins. Growing evidence suggests small molecule drugs interact with multiple proteins to enact cellular changes, contrary to the "magic bullet" philosophy often practiced in drug discovery [19, 7, 6]. Therefore, interpreting the totality of protein interactions for drugs provides greater insight into their therapeutic functions, with the potential of more fruitful drug discovery. In addition, drug repurposing has emerged as a valuable alternative to traditional drug discovery pipelines, potentially easing the burden associated with clinical trial failures due to adverse events and lack of efficacy [20, 13, 12, 11].

We have developed the Computational Analysis of Novel Drug Opportunities (CANDO) platform for analysis of drug interactions on a proteomic scale, adhering to multitarget drug theory [10, 16, 15, 9, 5]. An overview of the platform is provided in Figure 1. CANDO version 2 (v2) is comprised of a library of 14,610 protein structures extracted from the Protein Data Bank, 2,162 human-approved drugs from DrugBank, and 2,178 indications/diseases from the Comparative Toxicogenomics Database (CTD), encompassing 18,709 drug-indication associations [18, 4, 1]. The platform relates small molecules based on their computed interactions with all protein structures, known as an interaction signature, then assesses a drug repurposing accuracy based on how similar drug-proteomic signatures are for those drugs approved to treat the same indications. The hypothesis is that drugs with similar interaction signatures will have similar behavior in biological systems and will therefore be useful against the same indications.

Here, we present the CANDO Python package for easier handling of these drug-protein interaction signatures with the ultimate goal of making novel putative drug candidate generation easy and accessible. The package reads in a matrix of pre-computed interaction scores with any number of proteins, which is then benchmarked. Novel compound-protein interaction signatures are quickly computed and added to the matrix using our bioinformatic docking protocol (detailed in the Supplementary Information), allowing for direct comparison to other drug signatures in the platform. The package can also read in any drug-drug similarity/distance matrix computed using any method the user desires, which may be benchmarked or used for drug-disease association prediction.

2 Methods

2.1 Bioinformatic docking protocol

The compound-protein interaction scores in the CANDO v2 matrix are computed using a bioinformatic docking protocol that compares the structures of query drugs to all ligands known to bind to a given site on a protein [5]. Specifically, the COFACTOR algorithm is used to elucidate potential binding sites on each query protein, which utilizes homology modeling of known PDB structures with available binding site information [14, 21]. COFACTOR output includes a co-crystallized ligand for each potential binding site, which can then be compared to a compound/drug of interest using chemical fingerprinting

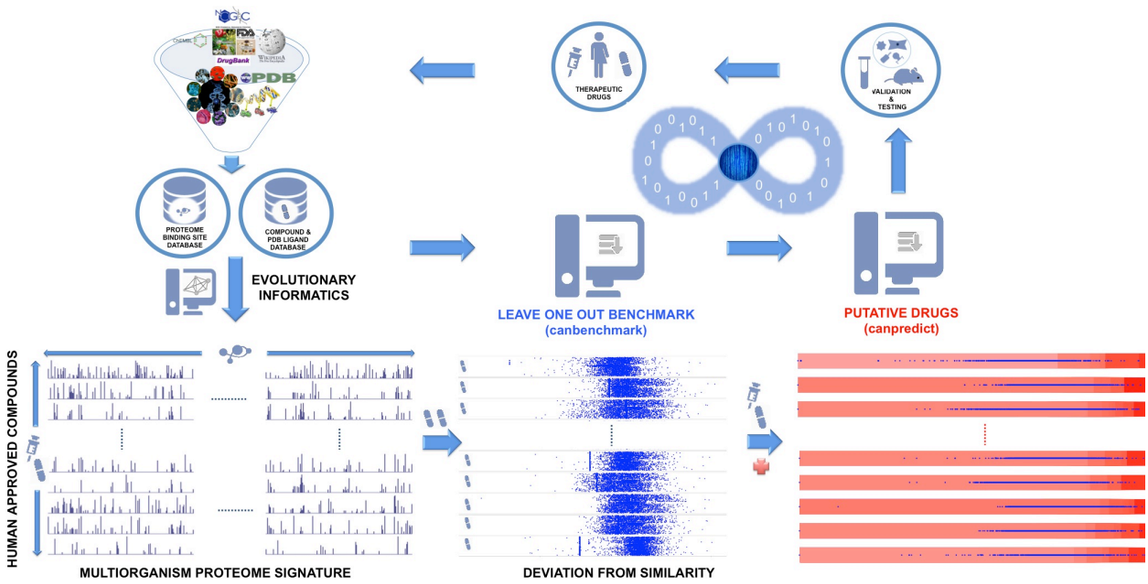

Figure 1: Overview of the CANDO drug discovery and repurposing pipeline. Drugs, small molecule compounds, and protein structures are collected from public databases, most notably DrugBank and the Protein Data Bank. Interactions are computed between all compounds and all proteins using our bioinformatic docking method or virtual docking simulations. The platform is benchmarked using known drug-indication associations as a gold standard, and putative drug candidates are predicted based on the similarity of their proteomic interaction signatures to those of drugs known to treat specific diseases. Validation experiments are conducted with top hits proceeding to clinical trials for approval. All results are fed back into the platform to optimize performance.

methods that binarize the presence or absence of particular molecular substructures. The maximum Tanimoto coefficient, or Jaccard index, between the binary vectors of the query compound and the set of all predicted binding site ligands for a protein serves as a proxy for the binding strength.

2.2 Benchmarking protocol

Each drug/compound is ranked relative to each other based on the similarity of their proteomic signatures, calculated using the root mean square deviation (RMSD) metric. Each drug, therefore, has a ranked list of other small molecule compounds most similar to it. The exact set of proteins to be considered for the RMSD calculation may be modified to allow for more specific queries, which is also true for the canpredict module (discussed below). Other distance metrics, such as cosine distance, may also be used.

The benchmarking module/protocol (canbenchmark) utilizes a hold-one-out scheme to compute an accuracy for each indication. For a given indication, each approved drug is held-out and the most similar compounds (within various cutoffs) are checked to see if they are also approved for the indication (Figure 2). This protocol is run iteratively and averaged across all indications with two or more drugs approved to provide a drug repurposing accuracy at each cutoff. Both the average indication accuracy (described above) and the pairwise accuracy (the weighted average based on the number of compounds approved for the disease) are outputted, as well as the non-zero indication accuracy coverage.

2.3 Putative drug candidate generation (prediction) protocol

Aside from benchmarking, the ranked lists of most similar compounds to each drug can be investigated as potential novel treatments. A consensus scoring approach is utilized where for each drug associated with a specific indication, the number of times a particular drug shows up within a certain cutoff of each list is counted. The prediction module/protocol (canpredict) then ranks the top compounds by their consensus scores, removing those predictions that are already approved for the disease. Figure 3A shows the output for human immunodeficiency virus using a cutoff count of 25 compounds. The drug at rank 3, the chemotherapeutic decitabine (which is not currently approved against HIV), has shown considerable anti-HIV activity both *in vitro* and *in vivo* [3, 2]. Modifying the protein signatures to only include three HIV proteins present in the non-redundant PDB set results in the prediction of 2 members of the 4-hydroxycoumarin class, dicoumarol and phenprocoumon, which have shown anti-HIV as well [8, 17] (Figure 3B). For diseases that lack approved treatments, compounds with the largest interaction score sums across all proteins or a specified subset of proteins are outputted.

2.4 Machine learning module

The CANDO package also provides support for several machine learning algorithms that learn more complex relationships within the proteomic vectors relative to the RMSD calculation. The currently supported algorithms include support vector machines (SVMs), 1-class SVMs, random forests, and logistic regression. This module is benchmarked using a protocol similar to canbenchmark; for a given indication, each approved drug is held-out while the model is trained on all other drugs approved for the indication in an iterative fashion. An equal number of neutral samples are chosen as negative samples during training (except in the case of 1-class SVMs), which represent drugs/compounds not associated with the indication. Random seeds can be set to guarantee the same compounds are used in training for reproducibility. The recall is calculated based on the number of times the held-out drug is positively

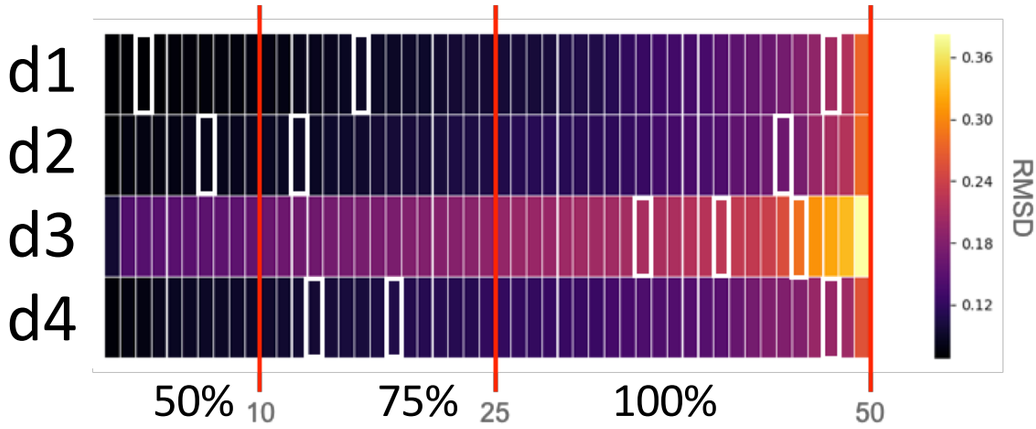

Figure 2: Benchmarking protocol. Example of benchmarking a disease with four drugs approved at cutoffs 10, 25, and 50. The white outlines represent other drugs approved for that disease, so in the list of similar compounds to d1, that means either d2, d3, or d4. Only d1 and d2 have drugs approved for the same indication within their top 10 most similar compounds, resulting in a 50% top 10 accuracy. The CANDO platform can be used to rigorously evaluate computational drug discovery and repurposing methods.

A) CANPREDICT with 14k proteins

```
32 compounds found for MESH:D015658 --> HIV Infections
Generating top 10 compound predictions...

rank    score   id      name
1       8       683     enprofylline
2       7       1732    cobicistat
3       7       1099    decitabine
4       7       1793    uridine_triacetate
5       7       1660    sofosbuvir
6       7       253     cidofovir
7       7       736     emtricitabine
8       6       738     quinapril
9       6       554     moexipril
10      6       766     repaglinide
```

B) CANPREDICT with 3 HIV proteins

```
32 compounds found for MESH:D015658 --> HIV Infections
Generating top 10 compound predictions...

rank    score   id      name
1       5       57      pyridoxine
2       5       41      pyridoxal
3       4       132     cladribine
4       4       646     aciclovir
5       4       156     dicoumarol
6       4       2044    glycol_salicylate
7       4       798     phenprocoumon
8       4       807     netilmicin
9       4       736     emtricitabine
10      3       738     quinapril
```

Figure 3: Examples of canpredict for illustrative purposes. A) Canpredict output for human immunodeficiency virus (HIV) with a compound cutoff of 25 using all proteins. The drug at rank 3, decitabine, has shown considerable evidence of anti-HIV activity in the literature. B) Canpredict output for HIV with a compound cutoff of 25 using only HIV proteins. The protein signature is automatically filtered to include only PDB chains or UniProt IDs matching a user-inputted list of UniProt IDs. The 4-hydroxycoumarin class drugs at ranks 5 and 7, dicoumarol and phenprocoumon, respectively, have also shown anti-HIV in the literature. Thus the CANDO platform enables prediction of novel leads and therapeutics for particular indications from multiple perspectives.

classified as approved for the indication. The user may also make predictions for novel or non-associated compounds after training the classifier on all approved drugs for a particular indication.

2.5 Development and implementation

The CANDO software is available in Python 2.7, 3.6, and 3.7. It is available for installation via the Python Anaconda installer. The source code, API document, and a Jupyter Notebook tutorial are available on GitHub at https://github.com/ram-compbio/CANDO as well as on http://compbio.org/software/.

3 Discussion and conclusion

The multitarget approach to drug discovery is vastly unexplored and shows promise for identifying novel treatments for various diseases. The CANDO Python package allows users to investigate drug-protein interactions on a proteomic scale, moving away from the single target philosophy. The multitarget approach is conducive for understanding drug behavior holistically, which will allow for better understanding of the therapeutic effects these small molecules exert on biological systems. We anticipate that broader use of this platform will inform researchers about potential lead compounds that may be therapeutic for specific indications, leading to accelerated drug discovery. In addition to the bioinformatic docking matrix described above, a compound-protein interaction matrix generated using our state of the art docking program CANDOCK with predicted binding energies will be available for use shortly; as well, the platform can accept a compound-compound matrix generated by any method (virtual docking, molecular fingerprinting, etc.). Also, a webserver hosted on compbio.org that will feature many of the functionalities described is under development.

Acknowledgements

The authors would like to acknowledge Matthew Hudson, Manoj Mammen, and Jim Schuler for their contributions and testing of the CANDO software.

Funding

This work was supported in part by a National Institute of Health Director's Pioneer Award (DP1OD006779), a National Institute of Health Clinical and Translational Sciences Award (UL1TR001412), a National Library of Medicine T15 Award (T15LM012495), a National Cancer Institute/Veterans Affairs Big Data-Scientist Training Enhancement Program Fellowship in Big Data Sciences, and startup funds from the Department of Biomedical Informatics at the University at Buffalo.

References

[1] Helen M Berman et al. "The protein data bank, 1999–". In: *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*. Springer, 2006, pp. 675–684.

[2] Christine L Clouser, Steven E Patterson, and Louis M Mansky. "Exploiting drug repositioning for discovery of a novel HIV combination therapy". In: *Journal of virology* 84.18 (2010), pp. 9301–9309. ISSN: 0022-538X.

[3] Christine L Clouser et al. "Activity of a novel combined antiretroviral therapy of gemcitabine and decitabine in a mouse model for HIV-1". In: *Antimicrobial agents and chemotherapy* 56.4 (2012), pp. 1942–1948. ISSN: 0066-4804.

[4] Allan Peter Davis et al. "The comparative toxicogenomics database: update 2017". In: *Nucleic acids research* 45.D1 (2016), pp. D972–D978. ISSN: 0305-1048.

[5] Zackary Falls et al. "Exploration of interaction scoring criteria in the CANDO platform". In: *BMC Research Notes* 12.1 (2019), p. 318. ISSN: 1756-0500. DOI: 10.1186/s13104-019-4356-3. URL: https://doi.org/10.1186/s13104-019-4356-3.

[6] Jeremy A Horst et al. "Computational multitarget drug discovery". In: *Polypharmacology* (2012), pp. 236–302.

[7] M L Bolognesi. "Polypharmacology in a single drug: multitarget drugs". In: *Current medicinal chemistry* 20.13 (2013), pp. 1639–1645. ISSN: 0929-8673.

[8] Sneh Lata et al. "HIV-1 Rev downregulates Tat expression and viral replication via modulation of NAD (P) H: quinine oxidoreductase 1 (NQO1)". In: *Nature communications* 6 (2015), p. 7244. ISSN: 2041-1723.

[9] William Mangione and Ram Samudrala. "Identifying protein features responsible for improved drug repurposing accuracies using the CANDO platform: Implications for drug design". In: *Molecules* 24.1 (2019), p. 167.

[10] Mark Minie et al. "CANDO and the infinite drug discovery frontier". In: *Drug discovery today* 19.9 (2014), pp. 1353–1363. ISSN: 1359-6446.

[11] Manish D Paranjpe, Alice Taubes, and Marina Sirota. "Insights into computational drug repurposing for neurodegenerative disease". In: *Trends in pharmacological sciences* (2019). ISSN: 0165-6147.

[12] Azam Peyvandipour et al. "A novel computational approach for drug repurposing using systems biology". In: *Bioinformatics* 34.16 (2018), pp. 2817–2825. ISSN: 1367-4803.

[13] S. Pushpakom et al. "Drug repurposing: progress, challenges and recommendations". In: *Nat Rev Drug Discov* (2018). ISSN: 1474-1784 (Electronic) 1474-1776 (Linking). DOI: 10.1038/nrd.2018.168. URL: https://www.ncbi.nlm.nih.gov/pubmed/30310233.

[14] Ambrish Roy, Jianyi Yang, and Yang Zhang. "COFACTOR: an accurate comparative algorithm for structure-based protein function annotation". In: *Nucleic acids research* 40.W1 (2012), W471–W477. ISSN: 0305-1048.

[15] James Schuler and Ram Samudrala. "Fingerprinting cando: Increased accuracy with structure and ligand based shotgun drug repurposing". In: *bioRxiv* (2019), p. 591123.

[16] Geetika Sethi, Gaurav Chopra, and Ram Samudrala. "Multiscale modelling of relationships between protein classes and drug behavior across all diseases using the CANDO platform". In: *Mini reviews in medicinal chemistry* 15.8 (2015), pp. 705–717. ISSN: 1389-5575.

[17] Suvit Thaisrivongs et al. "Structure-based design of HIV protease inhibitors: 4-hydroxycoumarins and 4-hydroxy-2-pyrones as non-peptidic inhibitors". In: *Journal of medicinal chemistry* 37.20 (1994), pp. 3200–3204. ISSN: 0022-2623.

[18] David S Wishart et al. "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic acids research* 46.D1 (2017), pp. D1074–D1082. ISSN: 0305-1048.

[19] Li Xie et al. "Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir". In: *PLoS computational biology* 7.4 (2011), e1002037. ISSN: 1553-7358.

[20] Jaswanth Yella et al. "Changing trends in computational drug repositioning". In: *Pharmaceuticals* 11.2 (2018), p. 57.

[21] Chengxin Zhang, Peter L Freddolino, and Yang Zhang. "COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information". In: *Nucleic acids research* 45.W1 (2017), W291–W299. ISSN: 0305-1048.