OXFORD

Structural bioinformatics

ChemBioServer 2.0: an advanced web server for filtering, clustering and networking of chemical compounds facilitating both drug discovery and repurposing

Evangelos Karatzas[1,2,†]**, Juan Eiros Zamora**[3,†]**, Emmanouil Athanasiadis**[4,5,6]**, Dimitris Dellis**[7]**, Zoe Cournia**[3,*] **and George M. Spyrou**[2,8,*]

[1]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Ilisia, 15784 Athens, Greece, [2]Bioinformatics Department, The Cyprus Institute of Neurology and Genetics, 2370 Nicosia, Cyprus, [3]Biomedical Research Foundation Academy of Athens, 115 27 Athens, Greece, [4]Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK, [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK, [6]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, Cambridge CB2 0AW, UK, [7]Greek Research and Technology Network, S.A., 11523 Athens, Greece and [8]The Cyprus School of Molecular Medicine, 2370 Nicosia, Cyprus

*To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Lenore Cowen

Abstract

Summary: ChemBioServer 2.0 is the advanced sequel of a web server for filtering, clustering and networking of chemical compound libraries facilitating both drug discovery and repurposing. It provides researchers the ability to (i) browse and visualize compounds along with their physicochemical and toxicity properties, (ii) perform property-based filtering of compounds, (iii) explore compound libraries for lead optimization based on perfect match sub-structure search, (iv) re-rank virtual screening results to achieve selectivity for a protein of interest against different protein members of the same family, selecting only those compounds that score high for the protein of interest, (v) perform clustering among the compounds based on their physicochemical properties providing representative compounds for each cluster, (vi) construct and visualize a structural similarity network of compounds providing a set of network analysis metrics, (vii) combine a given set of compounds with a reference set of compounds into a single structural similarity network providing the opportunity to infer drug repurposing due to transitivity, (viii) remove compounds from a network based on their similarity with unwanted substances (e.g. failed drugs) and (ix) build custom compound mining pipelines.

Availability and implementation: http://chembioserver.vi-seem.eu.

Contact: zcournia@bioacademy.gr or georges@cing.ac.cy

1 Introduction

Despite the improvement of available technologies in the pharmaceutical industry, the cost of commercializing a new drug doubles every 9 years (Scannell *et al.*, 2012). Designing novel organic compounds in a systematic fashion is a daunting task as it has been estimated that there can be up to 10^{60} molecules with drug-like properties (Polishchuk *et al.*, 2013). One of the initial stages in drug development is to explore the chemical space using compound libraries that attempt to capture its vastness with a small subset of very diverse molecules. Generating these libraries through exploration of this space is a challenge in itself, and several researchers have tackled the problem through different computational approaches, such as

exhaustive search (Gómez-Bombarelli *et al.*, 2016), genetic algorithms (Virshup *et al.*, 2013) and recently, deep neural networks (Gómez-Bombarelli *et al.*, 2018). Once a sufficiently large and diverse library of compounds is obtained, its components are virtually screened against a desired target to predict their free energy of binding (Lionta *et al.*, 2014). This initial prediction is of paramount importance; in order to save both time and resources the initial library is narrowed down to only the best scoring molecules that are selected for further screening using more detailed computational models, filters and experimental assays. This approach has been demonstrated to enhance the success rate of virtual screening experiments as demonstrated in Lionta *et al.* (2014) and Athanasiadis *et al.* (2012).

One issue related to drug discovery is the problem of specificity. The complexity of a cell is still far beyond the reach of current biomolecular simulations capabilities, while drug targets are never found in isolation. Therefore, a compound that binds with a strong affinity to a specific target could also have other off-target interactions, leading to undesired side effects. This is very often the case for protein families: groups of evolutionarily related proteins that share structural similarities.

On the other hand, already existing drugs might prove useful against a disease outside their initial target spectrum. Drugs with high structural similarity imply similar mode of action against similar targets (Campillos *et al.*, 2008). As highlighted in the study of Zhang *et al.* (2014), drug similarity analytics, including chemical structure similarity, aim to identify candidate drugs, which display similar pharmacological characteristics to the drug of interest. Drug repurposing studies using tools based on drug structural similarity have already been performed (Gottlieb *et al.*, 2011; Li and Lu, 2012). A drug–drug network with nodes linked by their pairwise structural similarities shows direct association of compounds allowing the researcher to either choose or filter out compounds based on these relations, as an additional filtering method.

ChemBioServer (Athanasiadis *et al.*, 2012) is a very successful application that has been continuously supported by our Groups and is gaining attention from the scientific community (for the last 11 months, from July 2018 to June 2019, it has an average of 8749 hits per month). We have updated the initial version of this server with (i) a functionality that re-ranks virtual screening results based on ensemble docking screenings, i.e. screening the same compound library against different protein members of the same family, selecting only those compounds that score high for the protein of interest, (ii) a group of networking tools in order to allow researchers to create networks of compounds and provide useful network metrics, (iii) a functionality that infers potential drug repurposing based on structural similarity and (iv) a filtering functionality to filter out compounds that are similar to unwanted substances (e.g. failed drugs of a clinical trial).

2 Application

In this section, we describe the updates in ChemBioServer 2.0.

2.1 Filtering

The 'Filtering' section of ChemBioServer 2.0 allows researchers to browse and filter compounds based on intra-ligand steric clashes, unwanted toxicophores and desirable or undesirable chemical moieties or physicochemical properties. In this update, the functionality 'Re-ranking for Ensemble Docking' has been added to this group of actions. Very often users need to select compounds that rank high for their target of interest but low for evolutionarily related proteins with similar binding sites (e.g. in a set of protein kinases) in order to avoid potential side effects. Thus, they employ cross-docking virtual screening in multiple receptor structures to identify compounds that will be predicted to bind only to the receptor of interest and not to receptors of the same protein family (Amaro *et al.*, 2018). ChemBioServer 2.0 can post-process cross-docking results and automatically re-rank virtual screening output to reveal compounds that rank high for the protein of interest in seconds. To accomplish this, first, the user uploads virtual screening results for the target(s) of interest using the 'Upload target file(s)'. Multiple file upload is allowed as users may choose to dock a chemical library in multiple conformations of a given protein. In the next step, the user can upload virtual screening results in SDF format, including docking scores, for protein structures of the same family. The chemical library used for virtual screening should be the same for all protein structures. ChemBioServer 2.0 then re-ranks and generates a filtered list of compounds that rank high for the target of interest and low for undesired targets (based on the provided docking scores).

The re-ranking algorithm is equipped with three compound selectivity methods for the target protein: automatic, manual or based on minimum desired docking score difference of the compound set.

In all three methods, the user has to specify the minimum number of compounds that should be retrieved from the re-ranking procedure. The automatic method detects high-scoring docked compounds for the target of interest that have a low docking score for the undesired protein targets. It thus starts by defining low and high docking score cutoffs as the top 1% best scoring compounds for the target(s) and the top 1% worst scoring compounds for the rest of the proteins, respectively. These cutoffs are iteratively relaxed using 1% increment until the minimum number of user selected compounds meets the filter conditions. The manual method provides more flexibility, as the user manually specifies the low and high docking scores as cutoffs and a direct search is performed. The third method provides an alternative way to define compound specificity for a given protein target. Often, the absolute values of docking scores as cutoffs might not be as important as the actual predicted free energy difference (docking score) between the compounds for each protein. The larger this difference, the more selective the compounds will be. Therefore, with the 'Score Difference' selection from the Method Selection tab the user can specify a desired level of energy difference, and the program will proceed in a similar fashion to the automatic procedure. It will start by defining the top 1% lowest scoring compounds for the target protein and the second cutoff will be set above by the given score difference. While the number of compounds that pass this filter is below the minimum number of compounds specified, the low energy cutoff will be gradually increased by 1% steps, and the high energy cutoff will always be at least above the set score difference (in kcal/mol). These two last methods are not guaranteed to succeed, as there might be no compounds that meet the selection criteria defined by the user. In such case, the program will fall back to the automatic method. Filtered compounds are available for download in CSV format. The algorithm uses the Pandas Python package API7. One of the three methods can be chosen and corresponding input boxes appear. The input files are stored in the server and analyzed by calling a Python script through PHP. Results are stored for a week and a link to download them is presented to the user before executing the analysis.

2.2 Clustering

ChemBioServer 2.0 still features the two clustering methods that were initially included under the 'Clustering' labeled section; hierarchical and affinity propagation clustering. Both methods return structural clusters of the input compounds to the users together with their distance matrix as well as a graphical visualization. The affinity propagation clustering also returns exemplar compounds for each cluster.

2.3 Networking

The 'Networking' section of ChemBioServer 2.0 features all similarity-based network-related actions that have been implemented to this update. Similarity networks present a visualization of the strongest connections between substances based on their structural similarity. Nodes that are close to each other imply similar mode of action in a pharmaceutical setting. Apart from the holistic type of visualization, network analysis offers insights regarding the neighborhood of each node and the topology of the network reveals nodes that may connect distinct subnetworks of compounds, inferring multiple modes of action for some compounds. Moreover, key drug players can be highlighted based on network properties such as degree, strength or betweenness, as structural representatives of a highly connected group of compounds. Often, researchers need to discover new uses for existing drugs against diseases, (i.e. drug repurposing) in order to lower the cost of drug design. Structural drug repurposing identifies chemical similarity of approved drugs with an inhibitor of the desired drug target; these drugs have a high chance to bind to the desired drug target. For this reason, fast screening of drug-like libraries to find chemical similarity with known drugs is important for drug repurposing. On the other hand, drug candidates might be deemed inappropriate for further studies based on structural criteria such as similarity to toxic substances or previously failed drugs from

clinical trials. The similarity edge lists derived from ChemBioServer's networking actions can be further explored via network analytics applications. Five networking functionalities are implemented and labeled 'Structural Similarity Network Visualization', 'Structural Similarity Network Analysis', 'Combine two SDF files in a Network', 'Attach similar-only nodes to Network' and 'Remove nodes from Network, based on similarity'. In 'Structural Similarity Network Visualization' the user uploads an SDF file and can choose a similarity metric between 'Tanimoto', 'Euclidean', 'Cosine', 'Dice' and 'Hamming' and a cutoff value for the edges (based on the resulting similarity values). According to the bibliography, the Tanimoto, Dice and Cosine metrics yield better results than the Euclidean metric regarding cheminformatic similarity calculations (Bajusz *et al.*, 2015). Another study has also deemed the Tanimoto metric superior to the Hamming metric when used for the classification of binary spectra based on similarities (Woodruff *et al.*, 1975). After the inputs are processed, the network is visualized and the similarity matrix between all input compounds can be downloaded. This matrix is returned through the function *calcDrugFPSim* from the Rcpi package, which calculates the drug molecules' similarity derived from their molecular fingerprints. A molecular fingerprint is a series of bits that represent the presence or absence of chemical substructures in a molecule. The molecular fingerprints are extracted from the respective mol structure format types via the *extractDrugMACCS* function. The mol structures are the parsed version of the input SDF or mol files and are calculated via the *readMolFromSDF* function. The output graph is drawn in the user interface via the javascript library vis.js. 'Structural Similarity Network Analysis' uses the same type of input values and the calculated similarity matrix is used as an adjacency matrix in order to create a graph using the igraph package in R. Node metrics 'Degree', 'Strength', 'Transitivity' and 'Eigenvector Centrality' are then presented in a sortable table, after execution.

The 'Combine two SDF files in a Network' action allows the user to test an SDF file against another reference SDF set, coloring the two groups of compounds differently, while allowing users to download the initial similarity matrix of both input sets. In the 'Attach similar-only nodes to Network' tab, a main network is created for the reference set with a given edge threshold, while compounds from the test set are attached to the main network via another edge threshold (e.g. stricter connections). Then, the user can download the upper triangular adjacency matrix of the whole network, as well as the edge list of the reference—test edges. Finally, in the 'Remove nodes from Network, based on similarity' tab, a main network is created for the reference set with a given edge threshold, while compounds similar to ones from the test set (second edge threshold input) are removed from the network, together with their edges. Once again, the user can download the upper triangular adjacency matrix of the new network, as well as the edge list of the reference—test edges that accounted for the removal of the reference nodes.

References

Amaro,R.E. *et al.* (2018) Ensemble docking in drug discovery. *Biophys. J.*, **114**, 2271–2278.

Athanasiadis,E. *et al.* (2012) ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery. *Bioinformatics*, **28**, 3002–3003.

Bajusz,D. *et al.* (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.*, **7**, 20.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Gómez-Bombarelli,R. *et al.* (2016) Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.*, **15**, 1120–1127.

Gómez-Bombarelli,R. *et al.* (2018) Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.*, **4**, 268–276.

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Li,J. and Lu,Z. (2012) A new method for computational drug repositioning using drug pairwise similarity. In: *2012 IEEE International Conference on Bioinformatics and Biomedicine*. pp. 1–4. IEEE.

Lionta,E. *et al.* (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.*, **14**, 1923–1938.

Polishchuk,P.G. *et al.* (2013) Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.*, **27**, 675–679.

Scannell,J.W. *et al.* (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.*, **11**, 191.

Virshup,A.M. *et al.* (2013) Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.*, **135**, 7296–7303.

Woodruff,H. *et al.* (1975) Similarity measures for the classification of binary infrared data. *Anal. Chem.*, **47**, 2027–2030.

Zhang,P. *et al.* (2014) Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits Transl. Sci. Proc.*, **2014**, 132.