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

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

[1]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Ilisia, 15784 Athens, Greece

[2]Biomedical Research Foundation Academy of Athens, 4 Soranou Ephessiou, 115 27 Athens, Greece

[3]Department of Haematology, University of Cambridge, Cambridge, UK

[4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK

[5]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, Cambridge, UK

[6]Greek Research and Technology Network, S.A., 7 Kifissias Avenue, 11523 Athens, Greec

[7]The Cyprus Institute of Neurology and Genetics, 6 International Airport Avenue, 2370 Nicosia, Cyprus

[8]The Cyprus School of Molecular Medicine, 6 International Airport Avenue, 2370 Nicosia, Cyprus

[*]Zoe Cournia (zcournia@bioacademy.gr) and George M. Spyrou (georges@cing.ac.cy) are the corresponding authors.

[**] The first two authors have equal contribution.

Abstract

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 chemical compounds, (iii) explore compound libraries for lead optimization based on perfect match substructure 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. The updated web server is available in the URL: http://chembioserver.vi-seem.eu/

Keywords: structure-based drug design, virtual screening, re-ranking, compound networking

Introduction

Despite the improvements in available technologies to the pharmaceutical sector, 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 this chemical space using 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 very recently, deep neural networks (Gómez-Bombarelli, et al., 2018). Once a sufficiently large and diverse library of compounds is obtained (typically thousands of molecules), its components are virtually screened against a desired target to predict their energy and site of interaction (Lionta, et al., 2014). This initial prediction is of paramount importance in order to save both time and money, as the initial library is narrowed down to only the best scoring molecules that are selected for further screening using more detailed computational models and experimental assays.

One issue related to drug discovery is the problem of specificity. The complexity of a cell is still far beyond the reach of current simulation capabilities, and the real targets of drugs are never in isolation. Therefore, a compound that shows a strong affinity for a target could also have many off-target interactions, leading to undesired secondary 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. As it is highlighted in the study of Zhang et al., drug similarity analytics, including chemical structure similarity, aim to find drugs, which display similar pharmacological characteristics to the drug of interest (Zhang, et al., 2014). Drug repurposing studies and tools based on drug structural similarity have been already made (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 virtual screening 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 it has an average of 8749 hits per month). We have updated the initial version of this server with (a) a functionality that re-ranks virtual screening results based on 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, (b) a group of networking tools in order to allow researchers to create networks of compounds and provide useful network metrics, (c) a functionality that infers potential drug repurposing based on structural similarity, (d) a filtering functionality to filter out compounds that are similar to unwanted substances (e.g. failed drugs).

Application

Filtering

The “Filtering” section of ChemBioServer 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 “Docking Re-ranking” 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. 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. 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. Next, the user uploads virtual screening results of the same chemical library that has been performed in protein structures users want to filter against. Again, multiple file upload is allowed. ChemBioServer 2.0 then re-ranks compounds and outputs to the user those 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 methods to define compound selectivity 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% increments until the minimum number of compounds desired by the user 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 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 a case, the program falls back to the automatic method. After the filtered compounds are obtained in a data frame, they are written to an Excel file, which is available for download. This format was chosen to make it more accessible to a general scientific user base with no knowledge of programming. The algorithm uses the Pandas Python package API7, conveniently allowing for data processing. The linchpin of this library is the Data Frame object, which is used to store data in memory by reading CSV files. These objects support Boolean indexing and have multiple methods implemented in C, which are faster than conventional Python 'for' loops. One of the three methods can be chosen and corresponding input boxes appear dynamically using JavaScript. The input files are stored in the server and analyzed by calling a Python script through PHP. The results are stored for 24 hours and a link to download them is displayed after successful finishing of the analysis.

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.

Networking

The “Networking” section of ChemBioServer features all similarity-based network-related actions that have been added 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. Usually, researchers need to discover new uses for existing drugs against diseases, hence lowering the cost of new drug creation (i.e. drug repositioning). Structural drug repurposing is a form of drug repositioning where predicted drugs target the same proteins as drugs structurally similar to them. For this reason, fast screening of drug lists is important in order to bring together test molecules with seemingly suitable substances based on their similarity. On the other hand, chemical substances 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” realise the aforementioned needs. In “Structural Similarity Network Visualization” the user uploads an sdf file and after choosing a similarity metric between “Tanimoto”, “Euclidean”, “Cosine”, “Dice” and “Hamming” and a value cutoff threshold for the edges (based on similarity values) can visualize the network and download the similarity matrix between all input compounds. This matrix is returned through the call of the function calcDrugFPSim from the Rcpi package which calculates the drug molecules' similarity derived from their molecular fingerprints. The 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”, “Betweenness” and “Strength” 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 and paints the two groups of compounds in different colors, as well as allows the user to download the initial similarity matrix between the compounds 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 and then 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 and then compounds similar to ones from the test set (second edge threshold input) are removed, together with their edges, from the network. 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.

Acknowledgments

George M. Spyrou holds the Bioinformatics ERA Chair Position funded by the European Commission Research Executive Agency (REA) Grant BIORISE (Num. 669026), under the Spreading Excellence, Widening Participation, Science with and for Society Framework. Evangelos Karatzas is a PhD student in the National and Kapodistrian University of Athens. His doctoral thesis is being funded by the IKY (State Scholarships Foundation) scholarship, funded by the Action "Strengthening Human Resources, Education and Lifelong Learning", 2014-2020, co-funded by the European Social Fund (ESF) and the Greek State. Juan Eiros Zamora acknowledges funding from PRACE as a member of its Summer of HPC program. Zoe Cournia would like to acknowledge funding from the European Union’s Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No. 785907 (Human Brain Project SGA2).

Conflict of Interest: none declared.

References

Athanasiadis, E., Cournia, Z. and Spyrou, G. ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery. *Bioinformatics* 2012;28(22):3002-3003.

Gómez-Bombarelli, R., *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials* 2016;15(10):1120.

Gómez-Bombarelli, R., *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* 2018;4(2):268-276.

Gottlieb, A., *et al.* PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* 2011;7(1):496.

Li, J. and Lu, Z. A new method for computational drug repositioning using drug pairwise similarity. In, *2012 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE; 2012. p. 1-4.

Lionta, E., *et al.* Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry* 2014;14(16):1923-1938.

Polishchuk, P.G., Madzhidov, T.I. and Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* 2013;27(8):675-679.

Scannell, J.W., *et al.* Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery* 2012;11(3):191.

Virshup, A.M., *et al.* Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *Journal of the American Chemical Society* 2013;135(19):7296-7303.

Zhang, P., *et al.* Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings* 2014;2014:132.