AtlasCBS: A web server to map chemico-biological space

Álvaro Cortés Cabrera[1,2], Antonio Morreale[2], Federico Gago[2] and Celerino Abad-Zapatero[1,3,*]

[1] Departamento de Farmacología, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain.
[2] Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC/UAM), Campus de Cantoblanco, E-28049 Madrid, Spain.
[3] Center for Pharmaceutical Biotechnology, MBRB building, University of Illinois at Chicago, Chicago, IL 60607, USA.

* Celerino Abad-Zapatero. Tel: 312-355-4105; Fax: 312-413-9303; Email: caz@uic.edu

Permanent address: Celerino Abad-Zapatero, Center for Pharmaceutical Biotechnology, MBRB building, University of Illinois at Chicago, Chicago, IL 60607, USA.

ABSTRACT

Given a target-ligand database with chemical structures and associated biological affinities/activities, the AtlasCBS server generates two-dimensional, dynamical, representations of its content based on ligand efficiency indices. These variables allow an effective decoupling of the chemical (angular) and biological (radial) components. BindingDB, PDBBind and ChEMBL databases are currently implemented. Proprietary datasets can also be uploaded and compared. The utility of this atlas-like representation in the future of drug design is highlighted. The web server can be accessed at the following URL: http://ub.cbm.uam.es/atlascbs

INTRODUCTION

The ever growing advances in the fields of structural biology, high-throughput screening and structure-based drug design have resulted in an exponential increase of the information related to targets, ligands, and their complexes that is stored in several databases (i.e., SAR databases: BindingDB, ChEMBL, PDBBind, among others). The vastness of chemical space as it relates to medicinal applications has been recognized[1] and certain tools to aid in navigating it have been introduced[2; 3]. The concept of an atlas-like representation of chemico-biological space (CBS) was introduced recently[4] based on the use of Ligand Efficiency Indices (LEIs) as variables. However, thus far, no friendly tool is available to connect 2D or 3D ligand structures and chemical properties (chemical space), with biological affinity/activity data pertaining to one or more target proteins (biological space). The AtlasCBS server introduced and described here is such a tool.

MATERIAL AND METHODS

1. LEIs and molecular properties calculation

LEIs are calculated as described elsewhere using the formulas contained in Table 1 and in Abad-Zapatero et al[4]. Molecular properties such as atomic masses, number of polar hydrogen atoms, and polar surface area are calculated using the Chemistry Development Toolkit (CDK).

2. Web design

The main web page for the server contains the five tabs shown in figure 1: a) Main: basic information about the server and its purpose, main references, contact information, and access to the main topics covered in the Help tab; b) Map viewer: tools for uploading the data from existing databases and for visualizing and analyzing their content; c) Login: required if the user wants to have private database

access; d) Help: information on how to use the AtlasCBS server; and e) About, data concerning the institutions and people involved in the project. Underneath the server there are three modules that provide all the functionalities for the tabs:

- **Map viewer.** The graphical engine of the server represents data from different sources, allows visualization of the chemical structure, and provides efficient filtering tools to compare and classify the molecules. The users can upload, select and map any data source target available at BindingDB, PDBBind, ChEMBL, or load an external data set (see below). Any set should include the compound name, structural description of the molecules (SMILES strings) and their affinity/activity values (K_d, K_i, or IC_{50}). To generate a map, the user selects X and Y variables from the sets: SEI, BEI; NSEI, NBEI; NSEI with nBEI or mBEI, respectively. Any combination is possible but complementary pairs (LEIs per size and polarity) are recommended. Given a map, molecules can then be selected by clicking on them and their LEIs and 2D chemical structures will appear in two adjacent panels (Figure 1, right panels). Also, the SMILES strings and LEIs values for the whole set of compounds are displayed in a list. In addition, molecules can be chosen by a simple selection method based on "range of values" or using the "Slope" option to choose those that share the same number of polar atoms (in the NSEI-nBEI plane). Selected compounds can be compared or used for similarity searches using molecular fingerprints and Tanimoto coefficients. Other features include: a) mixing and visualization of different data sources at the same time; and b) dynamic scaling of the axis to zoom in/out on particular areas of interest. Finally, it is possible to save and restore any working session. Some of these options could be dependent on the browser's capabilities (Table 2, and figures 1 and 2).

- **Private database manager.** It allows users to upload and process their own datasets in a secure way provided they register and accept the terms of usage of the site (access is granted using a valid e-mail address and a user chosen password). Data are read in through CSV files, with several fields separated by semicolons, which are readily available from common spreadsheets such as those produced by Microsoft Excel or OpenOffice Calc. Compounds can be also added manually as long as the required data are correctly given in the specified format. Users can modify each field interactively and they can also use filtering and searching tools.

- **Help.** It contains and explains the elements and functionalities of the server as well as references to background papers.

Implementation details

The server is organized in three layers: clients, application server and database. Each layer can communicate with the nearest neighbor but not beyond. Respectively, the three layers have been implemented with the following elements for the different components: i) Java, JavaScript, and HTML clients; ii) the Apache Tomcat servlet container; and iii) the MySQL database engine. The front-end is based on HTML, JavaScript and JSP or a Java applet; Java servlets handle the data traffic between the interface and the database, using a Model-View-Controller (MVC) paradigm. The information contained in a current release of BindingDB, PDBBind and ChEMBL was imported into the MySQL server's database by standalone Java programs that also compute the molecular properties and the efficiency indices for the molecules. The user's processes to upload external data are performed on-the-fly, with a special servlet using the CDK and the AJAX technology. The accepted format of the user's external database and examples are described in the help pages: it is based on a semicolon separated values (CSV) file containing: name, smiles, type of affinity variable (K_i, IC_{50}, K_d), and the affinity value (in nanomolar units).

Visualization of the data is based on a Java applet that allows to extend the graphic representation permitting the display of multiple pages simultaneously, to zoom in/out user's predefined areas, and to select compounds using SMARTS strings, similarity, or automatic detection of compound series (as in the NSEI-nBEI plane by the number of polar atoms: the slope each line in the graph). If the user's browser allows, we have also developed a javascript-based web application which avoids the use of Java but implements less features (Table 2, Fig. 1 Figs. 2).

Molecular fingerprints and Tanimoto coefficients

Molecular comparisons are based on the calculation of the CDK fingerprint which encodes the topology of the compounds as bit strings. Bit strings are compared using the Tanimoto coefficient (Tc, Eq. 1) which evaluates, from 0 (no similarity) to 1 (identity), the similarity between the compounds:

$$Tc = a/(a-b-c) \qquad \text{Eq. 1}$$

where a are the bits in common for two compounds, b are the bits activated in the first compounds but not in the second, and c are the bits activated in the second compound but not in the first.

RESULTS AND DISCUSSION

LEIs represent a relatively simple concept that naturally connects the chemistry and the biology via the affinity variable(s) (K_i, IC_{50}, or equivalent). In the suggested unified formulation, LEIs (see Table 1 for definitions) are obtained by weighting the affinity values with molecular properties such as the number of heavy (non-hydrogen, NHA) or polar atoms (NPOL), the polar surface area (PSA), or the molecular weight (MW). The combined use of two complementary LEIs: i) affinity/polarity (K_i combined with NPOL, PSA), x-axis; and ii) affinity/size (K_i combined with NHA, MW), y-axis, permits a very intuitive representation of the database content in a series of Cartesian diagrams constituting an atlas-like representation of CBS. The characteristics and appearance of these plots (efficiency planes) depend on the choice of variables (see Table 1).

Table 1 near here

A very interesting characteristic of these LEIs variables is that pairs of complementary indices can be represented on a Cartesian plane (efficiency plane) that unveils some of the intricacies of CBS in an appealing graphical representation. Although any combination of affinity/polarity (x-coordinate) and affinity/size (y-coordinate) is useful[4], of special interest are the NSEI-nBEI, NSEI-mBEI (x,y) efficiency planes (Table 1 and Fig. 1). In this type of plot, the slope of the line occupied by each target-ligand pair (its angular coordinate) depends only on the chemical composition of the ligand (in this case given by the number of polar atoms). The unique position along the line (the radial coordinate) for each target-ligand pair will depend on the biological affinity/activity of the ligand for the given target[4]. The drug discovery efforts then can be represented in the suggested efficiency planes as 'trajectories' in CBS[4] and optimized trajectories could be devised or proposed in the future based on numerical or statistical criteria[5].

It has been shown in initial retrospective studies that the optimized ligand(s) within a series of analogues typically map in the upper, right hand, quadrant of the efficiency planes (i.e, NSEI and nBEI or similar), where both variables are maximized[4; 5; 6]. Therefore, in the near future, it is theoretically possible to envisage an automatic procedure that optimizes ligand efficiencies by replacing some molecular fragments and evaluates the LEIs of the new candidates in an iterative fashion, until the best possible ligand(s) is(are) found[7]. It is precisely here that the future power of this methodology and graphical representation should be apparent: as a graphical and numerical guide in the search for better drug candidates.

Figure 1 and 2 near here

Other combinations of the variables defined in Table 1 can be selected and used in pairs to build up an “electronic book” composed of different “Cartesian maps” (i.e., pages or efficiency planes), depicting complete views or selected regions of CBS at different scales, constituting what we refer to as AtlasCBS, an atlas-like representation of CBS (Fig. 1). This representation is visual and dynamic by nature and would be extremely useful to help navigation through the ‘vastness of chemical space’[1]. To our knowledge, this is the first time a tool is presented that graphically displays, and naturally maps and classifies in a user-friendly way, the information stored in these ligand-receptor databases in terms of LEIs. The AtlasCBS web server that we introduce here, besides representing the chemico-biological content of BindingDB, PDBBind, and ChEMBL databases, also allows interested users to upload, map and compare their own data. The web server can be used to explore the CBS of known targets and/or visualize proprietary datasets, allowing an easy comparison between different sources. There are currently two versions of the AtlasCBS tool. A more stable version is at the EBI Hinxton campus server (https://wwwdev.ebi.ac.uk/chembl/atlascbs/intro.jsp) and a development version can be accessed at http://ub.cbm.uam.es/atlascbs; both require registration only for secure and confidential access.

Some applications of the AtlasCBS concept have been reviewed lately by Abad-Zapatero and Blasi[4; 6] in several domains of the drug discovery process. For example: the analysis and comparison of the contents of different databases (mapping of drugs vs. non-drugs), polypharmacology, fragment-based ligand design strategies, drug discovery trajectories and others. The AtlasCBS server presented here permits exploring these important areas of drug discovery dynamically on-line for the first time. We wish to encourage the drug discovery community to use this tool so that it can be improved in the future.

However, the most interesting applications will be those for which LEIs are used to guide the drug discovery process prospectively. In a recent study[8], Blasi and collaborators devised a workflow to obtain better drug candidates towards the transthyretin carrier protein (TTR) combining LEIs, pharmacophoric search, and docking. Briefly, a retrospective NSEI-nBEI map was built first with some known binders, to select the most appropriate candidate for further improvement. Second, the core structure of the selected compound was used as a pharmacophore to inquire a database of commercially available compounds. Those fulfilling the pharmacophore were submitted to docking and the best 80 selected based on the score. Third, these scores where transformed into estimated K_is for ‘theoretical’ LEIs calculations. Finally, a prospective map was built and the 12 ‘most efficient’ compounds (having the highest values of NSEI-nBEI) were selected for experimental tests of activity and pharmacokinetic behavior (results are still pending). Considering that initial retrospective studies[5; 6; 7] suggest that compounds with maximal efficiencies are likely to be good candidates for further development, we propose that the above strategy and the use of the AtlasCBS server would be useful to the drug-discovery community. This could set the basis for a more rigorous, numerically and efficiency based, drug-discovery paradigm[7].

CONCLUSION

In summary, we present here an effective tool aimed at facilitating the drug discovery process by providing an atlas-like representation of the CBS using LEIs as variables. This allows the graphical visualization of database contents as pages in a map-like environment, with different variables and scales, which can be easily navigated to examine the efficiency of existing and prospective molecules in size and polarity. We propose that the atlas representation can be extremely useful in several areas of drug discovery, including guiding new design strategies in the future and to optimize candidates in hit-to-lead campaigns.

ACKNOWLEDGEMENT

The assistance of the ChEMBL group in facilitating the data extraction for the AtlasCBS database is greatly appreciated. This work was supported by grants from Comunidad Autónoma de Madrid (S-BIO-0214-2006 and S2010-BMD-2457 to F.G. and A.M.), Fundación Severo Ochoa (AMAROUTO program to A.M.) and Ministerio de Educación (FPU AP2009-0203 to A. C. and SAB2010-0037 to C. A-Z).

REFERENCES

[1] C. Lipinski and A. Hopkins, Nature, 432 (2004) 855.
[2] T.I. Oprea and J. Gottfries, J Comb Chem, 3 (2001) 157.
[3] P. Watson, M. Verdonk and M.J. Hartshorn, J Mol Graph Model, 22 (2003) 71.
[4] C. Abad-Zapatero, O. Perisic, J. Wass, A.P. Bento, J. Overington, B. Al-Lazikani and M.E. Johnson, Drug Discov Today, 15 (2010) 804.
[5] S. Christmann-Franck, D. Cravo and C. Abad-Zapatero, Molecular Informatics, 30 (2011) 137.
[6] C. Abad-Zapatero and D. Blasi, Molecular Informatics, 30 (2011) 122.
[7] C. Abad-Zapatero, Expert Opinion in Drug Discovery, 2 (2007) 469.
[8] D. Blasi, G. Arsequel, G. Valencia, J. Nieto, A. Planas, M. Pinto, N.B. Centeno, C. Abad-Zapatero and J. Quintana, Molecular Informatics, 30 (2011) 161.
[9] C.A. Lipinski, J Pharmacol Toxicol Methods, 44 (2000) 235.

TABLE AND FIGURES LEGENDS

Table 1. Definitions of Ligand Efficiency Indices used in AtlasCBS*

Name	Definition
BEI	pK_i, pK_d, pIC_{50}/MW[a]
SEI	pK_i, pK_d, pIC_{50}/PSA[b]
NSEI	pK_i/NPOL[c]
NBEI	pK_i/NHA[d]
nBEI	$-\log_{10}[K_i/NHA]$
mBEI	$-\log_{10}[K_i/MW]$

[a]MW: Molecular weight (in kDa). $pK_i = -\log_{10} K_i$
[b]PSA: Polar surface area scaled to 100 $Å^2$.
[c]NPOL: Number of polar atoms (N and O).
[d]NHA: Number of heavy atoms (non-hydrogen).

*Examples of possible efficiency planes represented and available in the AtlasCBS server and a brief description of their characteristics and appearance:

1. SEI, BEI (x,y). Slope of the lines 10*(PSA/MW). No intersect.
2. NSEI, NBEI (x,y). Slope of the lines NPOL/NHA: a rational number. No intersect.
3. NSEI, nBEI (x,y). Slope of the lines NPOL, intersect $\log_{10}$(NHA).
4. NSEI, mBEI (x,y). Slope of the lines NPOL, intersect $\log_{10}$(MW).

Any combination is possible but pairs of variables related to affinity/polarity (x) and affinity/size (y) are strongly recommended. With this choice of variables the polarity or physico-chemical characteristics (as given by PSA/MW or NPOL/NHA) of the chemical compounds increases counterclockwise as indicated above, mimicking a graphical representation of the variables considered in Lipinski's Rule of Five (Ro5)[6; 9].

Table 2. Viewers characteristics.

Characteristic	Java applet viewer	Javascript/HTML viewer
SMART filtering	✓	-
Similarity search	✓	-
Simple filters	✓	✓
Polarity highlighting	✓	-
Labels	✓	✓
Mixed sources	✓	✓
Save session	-	✓
Dynamic scale change	✓	✓

Figure. 1. Typical Javascript example screen presented by the AtlasCBS server. The upper left tabs correspond to the main pages of the server: Main, Map Viewer, Login (for private access), Help, and About (see text). The left graphical panel within the page represents a typical efficiency plane (NSEI vs. nBEI) for the 807 entries found in ChEMBL for Angiotensin II receptor type I with Kd affinity values. Each point in the plane represents a target-ligand pair. The angular coordinate (NPOL in this case) corresponds to the number of polar (N,O) atoms of the ligand increasing counterclockwise (NPOL=3-12). The radial coordinate corresponds to the affinity of the ligands towards the target. Different colors correspond to different chemical series or different subsets of data. Top right panel shows the different options for the management of the session at the AtlasCBS server. Most importantly, the choice of Cartesian axes (x,y) that determines the appearance of the efficiency planes. Lower right panel shows the compounds selected. The values of the LEIs variables are shown also in this window as well as the SMILES strings of the corresponding compounds. The chemical structure of one of the selected compounds (Valsartan) is shown in the lowest right hand panel and is also annotated on the left map with black lettering. Multiple windows can be displayed simultaneously to compare pages in the 'AtlasCBS' with different variables or scales, as in a real life atlas.

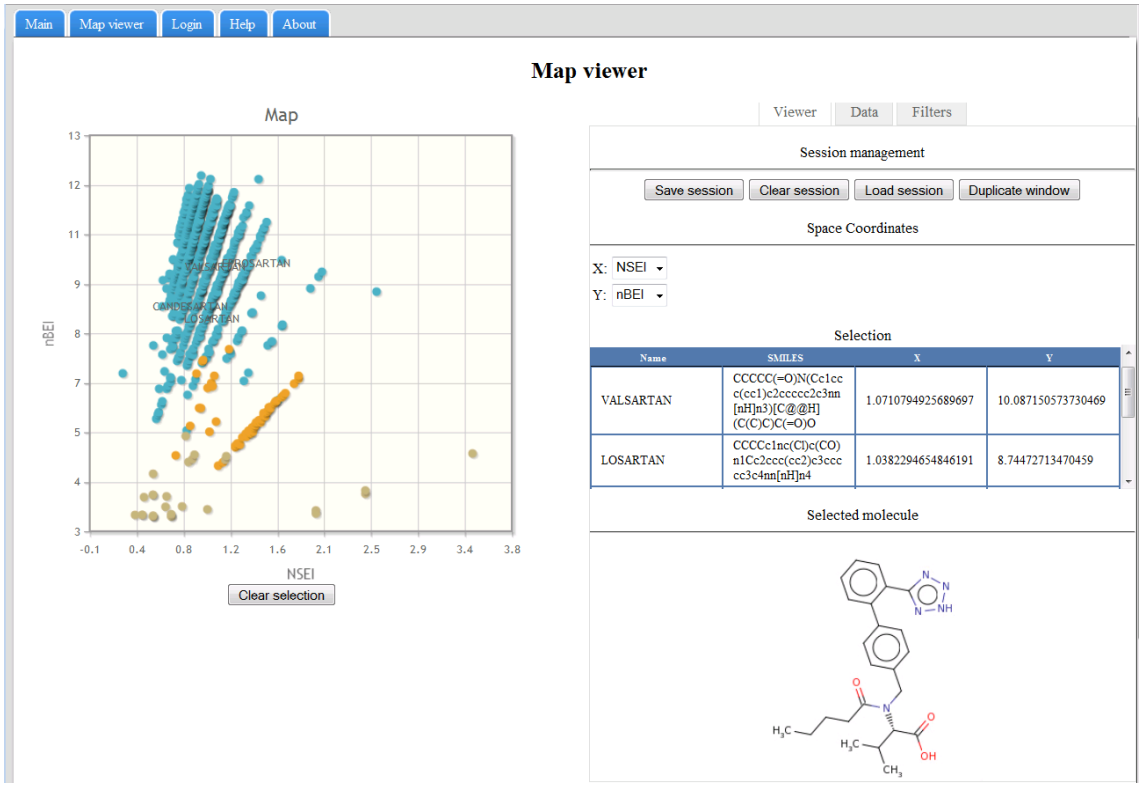

Figure 2. Java applet Map Viewer displaying several classes of anti-AIDS compounds from ChEMBL (target ID=6654). Different colors indicate an increasing number of polar (N,O) atoms (NPOL=2-28). Nevirapine^{TM} is displayed in the window and its position in the map annotated (name and downward arrowhead) in the corresponding line (NPOL=5), increasing counterclockwise from NPOL=2. The radial coordinates is given by the measured affinity towards the corresponding target. Different measurements for different targets (i.e., wild type vs. mutant(s)) will all map along the same line. The thickness of the lines depends on the number of compounds with the same NPOL number but different number of NHA atoms (log_{10} (NHA)). See Table 1.