ChemProt-2.0: visual navigation in a disease chemical biology database

Kjærulff, Sonny Kim; Wich, Louis; Kringelum, Jens Vindahl; Jacobsen, Ulrik Plesner; Kouskoumvekaki, Elrini; Audouze, Karine Marie Laure; Lund, Ole; Brunak, Søren; Oprea, Tudor; Taboureau, Olivier

Published in:
Nucleic Acids Research

Link to article, DOI:
10.1093/nar/gks1166

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
ChemProt-2.0: visual navigation in a disease chemical biology database

Sonny Kim Kjærulff¹, Louis Wich¹, Jens Kringelum¹, Ulrik P. Jacobsen¹, Irene Kouskoumvekaki¹, Karine Audouze¹, Ole Lund¹, Søren Brunak¹, Tudor I. Oprea¹,² and Olivier Taboureau¹,³,*

¹Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark, ²Translational Informatics Division, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM 87131-0001, USA and ³Department of Life Sciences, UMR-S973, MTi, University Paris Diderot, F-75013 Paris, France

Received September 14, 2012; Revised October 26, 2012; Accepted October 28, 2012

ABSTRACT

ChemProt-2.0 (http://www.cbs.dtu.dk/services/ChemProt-2.0) is a public available compilation of multiple chemical–protein annotation resources integrated with diseases and clinical outcomes information. The database has been updated to >1.15 million compounds with 5.32 millions bioactivity measurements for 15290 proteins. Each protein is linked to quality-scored human protein–protein interactions data based on more than half a million interactions, for studying diseases and biological outcomes (diseases, pathways and GO terms) through protein complexes. In ChemProt-2.0, therapeutic effects as well as adverse drug reactions have been integrated allowing for suggesting proteins associated to clinical outcomes. New chemical structure fingerprints were computed based on the similarity ensemble approach. Protein sequence similarity search was also integrated to evaluate the promiscuity of proteins, which can help in the prediction of off-target effects. Finally, the database was integrated into a visual interface that enables navigation of the pharmacological space for small molecules. Filtering options were included in order to facilitate and to guide dynamic search of specific queries.

INTRODUCTION

In recent years, there has been a shift from the traditionally secret experimental data kept by the pharmaceutical industry to a more open-access culture in relation to data sharing (1). For this reason, we have been witnessing a steady increase in public repositories of bioactive small molecules such as ChEMBL (2) and PubChem (3). However, as public repositories of bioactive small molecules have only just recently been made available, the problem of how to handle chemical entities is still largely unsolved. Pooling data from small molecule databases poses special problems. Even though standards have been widely adopted to describe genes and proteins (e.g. Ensembl ID, Entrez ID for genes, and UniProt ID for proteins), small molecule identifiers, as well as measures for properties such as biological activities, are not necessarily standardized across different resources (4).

One could claim that the bottleneck in understanding how small molecules perturb biological systems is no longer in the generation, gathering and availability of experimental data but in their organization, presentation and visualization; in other words, in the development of centralized systems that would better enable their exploitation. The problem is not only how to extract data from different (federated) resources, it is also important to provide solutions that facilitate provenance tracking, visualization, uniform and systematic description of data and their integration in ways that can preserve the semantic relationships between the different entities.

Furthermore, the number of failures of drug candidates in advanced stages of clinical trials has increased and the number of submissions for US Food and Drug Administration (FDA) approval has decreased in the last decade. One of the reasons may be our reductionist approach to discovery, whereby a complex system, namely a drug and its metabolites interacting with many proteins across multiple cellular compartments and tissues over time, is reduced to a simplistic ligand–target interaction model. This is probably too crude and emphasizes the
need to look at the effects of compounds on global systems aided by the integration of multiple biological and temporal data sources.

With the emerging fields of chemogenomics (5), systems pharmacology (6) and systems chemical biology (7,8), it becomes feasible to investigate the drug action at different levels from molecular to pathway, cellular, tissues and clinical outcomes (9). For example, it has become apparent that many common diseases such as cancer, cardiovascular diseases and mental disorders are much more complex than initially anticipated, as they are caused by multiple molecular and cellular dysfunctions rather than being the result of a single defect. Therefore, network-centric therapeutic approaches that consider entire pathways rather than single proteins must be investigated (10).

Among the recent advances in the field of systems chemical biology, servers supporting drug profiling such as STITCH (11), DisGENET (12) or the new database PROMISCUOUS (13) should be mentioned. STITCH3 provides confidence scores that reflect the level of confidence and significance of compound–protein interactions. PROMISCUOUS is a resource focused on drug compounds, including withdrawn and experimental, containing drug–protein interaction and side-effect (SE) information. DisGENET is a comprehensive gene–disease association database focused on the current knowledge of human genetic diseases including Mendelian, complex and environmental diseases.

We have previously reported the development of ChemProt, a disease chemical biology database (14). Compared with other approaches, ChemProt-1.0 offered a high level of integration of chemical and biological data, including internally curated disease-associated protein–protein interactions (PPIs) (15). Here, we present the second release of ChemProt, a resource of annotated and predicted disease chemical biology interactions. ChemProt-2.0 can be accessed at http://www.cbs.dtu.dk/services/ChemProt-2.0/. The present release contains a compilation of over 1 100 000 unique chemicals with biological activity for >15000 proteins. We have added a visual interface that supports user-friendly navigation through the data, biological activities and disease associations. ChemProt-2.0 now enables the user to query the database not solely by chemicals or proteins but also through therapeutic effects, adverse drug reactions and diseases. The similarity ensemble approach (SEA) developed by Keiser et al. (16) has also been implemented, so that protein sequence similarity can be used when examining chemical promiscuity. With these updates, ChemProt-2.0 offers an integrative approach to understand the impact of small molecules on biological systems and contributes to the investigation of molecular mechanisms related to diseases and clinical outcomes. A workflow of the implementation is shown in Figure 1.

**DATA SOURCES**

Chemical–protein interactions data were gathered in June 2012 from updated open-source databases ChEMBL.
ChemProt-2.0. Also, the human disease network de-
(28) and Gene Ontology (29) databases was also down-
of OMIM (25), GeneCards (26), KEGG (27), Reactome
Health Organization, as well as SE data from Dailymed
(24) developed by the World
information from the Anatomical Therapeutic Chemical
co-occurrence of a chemical term and a protein (gene)
chemical–protein relationships from text mining the
modulate gene expression, whereas STITCH provides
data from CTD (23) and STITCH (11). CTD extracts lit-
erature data about environmental chemicals and how they
modulate gene expression, whereas STITCH provides
chemical–protein relationships from text mining the
the co-occurrence of a chemical term and a protein (gene)
term in MEDLINE abstracts. Clinical outcomes were of
special interest in this version and we decided to include
information from the Anatomical Therapeutic Chemical
( ATC ) Classification System (24) developed by the World
Health Organization, as well as SE data from Dailymed
(http://dailymed.nlm.nih.gov/dailymed/).

From a biological perspective, we updated our internal human interactome platform to reach 14421 genes inter-
acting through 507 142 unique PPIs. The updated version
of OMIM (25), GeneCards (26), KEGG (27), Reactome
(28) and Gene Ontology (29) databases was also down-
loaded (June 2012), curated and integrated in
ChemProt-2.0. Also, the human disease network de-
veloped by Goh et al. (30) was integrated, allowing associa-
tion of proteins to disease categories.

PREDICTIONS AND METHODS

Based on the assumption that compounds sharing similar
structure have potential similar bioactivities, we encoded
the chemical structure with two different types of finger-
prints: the 166 MACCS key which encode the presence or
absence of some predefined substructural or functional
groups (31) and the FP2 fingerprints computed with
OpenBABEL (32). Chemical similarity between two com-
pounds is quantitatively assessed using the Tanimoto
coefficient. By including the SEA method (16), one
can also predict potential new targets for a compound.
For the internal development of SEA, compounds with an
activity value <100 µM were considered (only IC50, EC50,
Potency, AC50, Ki values were used). Furthermore, to
complete the set of active protein ligands, annotated com-
pound–protein interactions from CTD, DrugBank and
PharmGKB were also included, together with annotated
protein–compound in the STITCH database. For this
dataset, the raw similarity score, i.e. the sum of ligand pair
wise Tanimoto coefficients based on the FP2 fingerprint,
is 0.44. All proteins with more than five bioactive ligands
were considered.

In addition, for all protein targets, we operated under
the assumption of promiscuity, i.e. proteins with high-
sequence similarity may share similar functions and may
be targeted by the same compound (likely with different
bioactivities). Protein sequences were obtained from
Uniprot (33), and sequence comparisons were computed
using BLASTP (34). The similarity of two sequences was
assessed using an E-score, an expectation value related to
the probability that sequence similarity between two
proteins is not achieved by random chance (34). We
filtered the output and proteins with an E-value <$10^{-10}$
as default are depicted.

With respect to SEs, 988 small molecule drugs were
matched against 174 SE as described (35). Term frequency
vectors compiled from Dailymed were integrated in
ChemProt-2.0 and proteins associated to each drug are
then depicted.

VISUAL INTERFACE

In ChemProt-2.0, a visual interface was implemented to
facilitate the visualization of the results using HTML 5
and JavaScript. The core of the interface has been
designed in the form of a heatmap. The chemical–
protein associations are depicted in a pie-chart heatmap
where each pie corresponds to the database from which we
gathered the information. Hovering over the pie-charts
with the pointer, activity values are then displayed. The
user can select different display settings (circles, fill and
rectangles). A valuable feature is the handling of multiple
activities that have been gathered for a given compound–
target pair by selecting ‘All’ values. A color spectrum from
blue (low activity) to red (strong activity) is used to
indicate the activity (Figure 2). It is also possible to
select a specific database or/and a specific activity type
and define a range of activities (threshold) of interest in
order to optimize the query. Results from the SEA
approach are also integrated in the ‘Activity Type’.

The compound query is always shown in the first column
followed by similar compounds (sorted in descending order of
similarity) whereas the protein queried is depicted in
the first row. To optimize the display, the heatmap is limited to
a section of 100 rows × 100 columns. If the chemical–
protein matrix is larger, we have included an arrow
feature (→) that allows the user to upload the next 100
data items for both axes. The user has still the possibility
to view the data in a table format and to download the
results in a flat-file format. In the table format, display
mode the user can dynamically sort and group the activities
according to compound, target, species, activity type, etc.

A second heatmap that depicts protein–disease
categories is also integrated, which suggests proteins that
may be involved in diseases. Next to it, the ‘Diseases’ link
redirects the user to the disease-associated proteins
complex around the selected protein. A new, dynamic
interface has been implemented, where the proteins
associated to a biological term are shown when highlight-
ing the term of interest (Figure 3).

APPLICATIONS

The ChemProt-2.0 database interface is accessible freely
online. In addition to the chemical and protein search that
was previously implemented, the user can search by
diseases, ATC codes and SEs. For example, the query
‘epilepsy’ returns 2662 compounds active on 13 proteins
associated to this disease. Similarly, looking for the SE
‘hallucinations’, 15 drugs (with the term frequency
Some of these drugs (ropinirole, pergolide, amantadine and pramipexole) are used for the treatment of Parkinson diseases, by affecting the dopaminergic and serotonergic systems. Interestingly, visual hallucinations are symptoms of the Parkinson’s disease and perturbing the serotonergic system could help to alleviate these symptoms (36). Another interesting aspect is that these drugs affect several proteins associated to ‘Bone’ and osteoporosis disease. For example, there is a possible association between the polymorphism of the serotonin transporter (HTT) and the development of osteoporosis (37). Some of these drugs bind to HTT and could thus be potentially investigated for drug repurposing.
Many diseases seem not to be the result of a single defect but are rather caused by multiple molecular and cellular abnormalities. Therefore, observations of a drug effect not only at the molecular level but also at cellular and systems levels should guide therapeutic strategies for the development of better and safer drugs. ChemProt-2.0 offers the possibility of interrogating multiple layers of information by linking chemically induced biological perturbations to disease and phenotype. We believe with the advances in proteomics, metabolomics and other – omics sciences, combined with next-generation sequencing technologies, we will no longer evaluate the bioactivity of a chemical solely at the molecular level, but rather we will investigate biomedical knowledge with the integration of genetic polymorphisms and clinical effects (38).

**FUNDING**

The Innovative Medicines Initiative Joint Undertaking, OPENPHACTS (to L.W., J.K.); Innovative Medicines Initiative Joint Undertaking, eTOX (to U.P.J., O.T.); OPENPHACTS (to L.W., J.K.); Innovative Medicines Initiative Joint Undertaking, eTOX (to U.P.J., O.T.); OPENPHACTS (to L.W., J.K.); Innovative Medicines Initiative Joint Undertaking, eTOX (to U.P.J., O.T.). Funding for open access charge: The EU-IMI project OPENPHACTS.

**Conflict of interest statement.** None declared.

**REFERENCES**


