Bioinformatics Advance Access published June 2, 2016

CART – a chemical annotation retrieval toolkit

Samy Deghou,[1,‡] Georg Zeller,[1,‡] Murat Iskar,[1,‡] Marja Driessen,[1] Mercedes Castillo,[1] Vera van Noort,[1,2] and Peer Bork[1,3,*]

[1] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[2] Centre of Microbial and Plant Genetics, KU Leuven, Leuven, Belgium
[3] Max Delbrück Centre for Molecular Medicine, Berlin, Germany

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Data on bioactivities of drug-like chemicals is rapidly accumulating in public repositories, creating new opportunities for research in computational systems pharmacology. However, integrative analysis of these data sets is difficult due to prevailing ambiguity between chemical names and identifiers and a lack of cross-references between databases.

Results: To address this challenge, we have developed CART, a **C**hemical **A**nnotation **R**etrieval **T**oolkit. As a key functionality, it matches an input list of chemical names into a comprehensive reference space to assign unambiguous chemical identifiers. In this unified space, bioactivity annotations can be easily retrieved from databases covering a wide variety of chemical effects on biological systems. Subsequently, CART can determine annotations enriched in the input set of chemicals and display these in tabular format and interactive network visualizations, thereby facilitating integrative analysis of chemical bioactivity data.

Availability: CART is available as a Galaxy web service (cart.embl.de). Source code and an easy-to-install command line tool can also be obtained from cart.embl.de.

Contact: bork@embl.de

Understanding the effects of chemicals, in particular small organic molecules, on biological systems is fundamental to research in pharmacology, toxicology, chemical biology and related fields. Bioactivities of chemicals can be investigated at various scales analyzing drug-associated readouts, such as protein interactions, cellular phenotypes, toxicity or side effects (Iskar *et al.*, 2012). Owing to the development of high-throughput screening technologies, bioactivity data for large chemical libraries has rapidly accumulated in recent years and is increasingly becoming available in public repositories (see Table 1). While this has created tremendous opportunities for research that aims to integrate these heterogeneous data sets in order to gain a better systemic understanding of chemical effects, in practice such efforts are severely impeded by disparities in data representation. In particular, unambiguous identification of chemicals across databases can be difficult, because a myriad of synonyms and trade names exist for many chemicals, and even controlled nomenclature and structural descriptions are sometimes ambiguous, similar to the problem of mapping between various gene, transcript and protein nomenclatures, now overcome by many bioinformatics tools (Huang *et al.*, 2009, among others). To address the persisting need in chemoinformatics, we here present CART, a **C**hemical **A**nnotation **R**etrieval **T**oolkit. In solving the chemical name-matching problem, CART aims at integrating bioactivity annotations across various databases to provide functional annotation and enrichment analysis for chemicals. Thereby CART can identify coherent functional themes, analogous to gene ontology annotation tools, such as DAVID (Huang *et al.*, 2009). This makes CART useful, e.g. for the automatic characterization of hits derived from chemical screens (Rihel *et al.*, 2010, for instance). Also in other contexts, annotating chemicals with various biological effects is becoming an important task, which has so far largely required expert manual annotation, but can be greatly simplified by CART.

1 APPROACH

The first component of CART consists of matching user-provided chemical names to a comprehensive dictionary of synonyms, serving as a reference space for disambiguation to unique chemical identifiers (Figure 1). To improve matching sensitivity over exact synonym look-up, we additionally implemented an approximate text matching method based on the Apache Lucene search engine (http://lucene.apache.org/) and heuristics such as the conversion between salt (e.g. salicylate) and acid form (salicylic acid, see Suppl. Note 1 for details). CART also offers the possibility to match structural chemical identifiers, SMILES and InChI keys, via exact string matching. Taken together, these search capabilities go beyond what existing tools currently offer (see Suppl. Table 1).

Mapping to this chemical reference space facilitates subsequent retrieval of bioactivity annotations (Table 1, Suppl. Note 2). This allows for easy, multi-facetted annotation of chemical libraries,

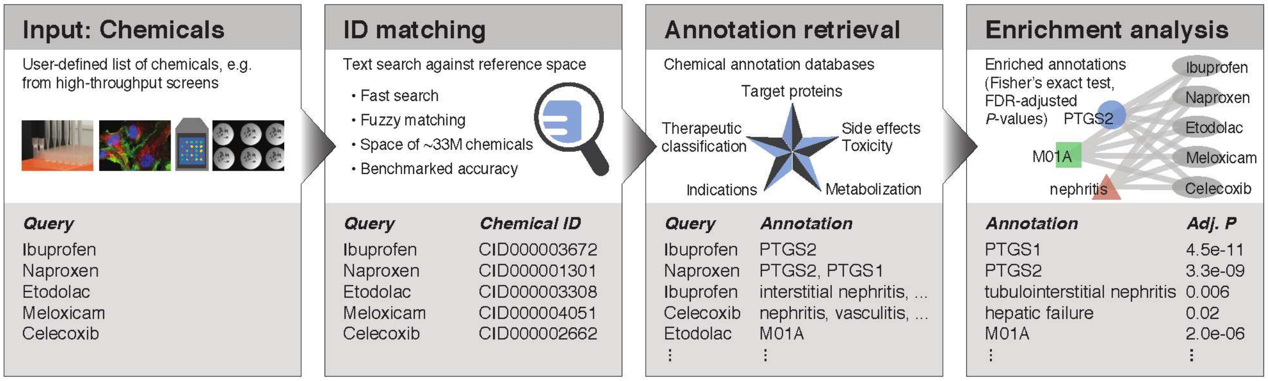
Table 1. Chemical bioactivity databases available through CART.

Bioactivity	Database	Size^a	Reference
Molecular targets	STITCH	221,724 / 9,015	stitch.embl.de
	TTD	11,340 / 1,120	bidd.nus.edu.sg/group/cjttd
	DrugBank	853 / 147	www.drugbank.ca
Metabolization	DrugBank	396 / 64	www.drugbank.ca
Therapeutic class.	ChEMBL	1,118 / 1,538	www.ebi.ac.uk/chembl/ftc
	ATC	2,515 / 924	www.whocc.no/atc
Drug side effects	SIDER	1,309 / 4,130	sider.embl.de
Toxicity	DrugMatrix	742 / 22	ntp.niehs.nih.gov/drugmatrix

^{a} Annotated chemicals / annotation terms, see Suppl. Fig. 3 and Suppl. Note 2.

[*] to whom correspondence should be addressed: bork@embl.de.

[‡] these authors contributed equally.


Input: Chemicals
User-defined list of chemicals, e.g. from high-throughput screens
ID matching
Text search against reference space
• Fast search
• Fuzzy matching
• Space of ~33M chemicals
• Benchmarked accuracy
Annotation retrieval
Chemical annotation databases
Target proteins
Therapeutic classification
Side effects Toxicity
Indications
Metabolization
Enrichment analysis
Enriched annotations (Fisher's exact test, FDR-adjusted P-values)
PTGS2
M01A
nephritis
Ibuprofen
Naproxen
Etodolac
Meloxicam
Celecoxib

Query
Ibuprofen
Naproxen
Etodolac
Meloxicam
Celecoxib

Query	Chemical ID
Ibuprofen	CID000003672
Naproxen	CID000001301
Etodolac	CID000003308
Meloxicam	CID000004051
Celecoxib	CID000002662

Query	Annotation
Ibuprofen	PTGS2
Naproxen	PTGS2, PTGS1
Ibuprofen	interstitial nephritis, ...
Celecoxib	nephritis, vasculitis, ...
Etodolac	M01A
⋮	⋮

Annotation	Adj. P
PTGS1	4.5e-11
PTGS2	3.3e-09
tubulointerstitial nephritis	0.006
hepatic failure	0.02
M01A	2.0e-06
⋮	⋮

Fig. 1. Typical CART workflow including chemical name matching, annotation retrieval and enrichment analysis. The lower panels contain a toy example of non-steroidal anti-inflammatory (NSAID) compounds and show excerpts of how these are matched and annotated by CART, the rightmost panel displays a (partial) enrichment network; PTGS – prostaglandin-endoperoxide synthase targets, M01A – ATC code for NSAIDs, Adj. P – FDR-corrected P-value, nephritis and vasculitis are NSAID-associated side effects. See Suppl. Note 3 and Suppl. Fig. 4 for an application of CART to hits from a drug screen.

synonym retrieval, which is useful e.g. for text mining, and the identification of bioactivities that are enriched in the user-provided input. Statistical significance for these enrichments is established using Fisher's exact test with FDR correction for multiple testing.

In a typical use case, users may want to subject a set of hits resulting from a high-throughput chemical screen to CART analysis. After name matching, the enrichment analysis can be done relative to a user-specified background, in this case the library of all chemicals probed in the screen. Enriched annotations are subsequently retrieved from databases describing chemical effects at various scales, including molecular targets, metabolizing enzymes, functional classifications, indication areas and side effects (Table 1, Suppl. Note 2). The results are visualized as a network linking the input set of chemicals to enriched annotations (Figure 1, Suppl. Note 3, Suppl. Fig. 4). Implemented in Cytoscape.js (Franz *et al.*, 2015), this network can be interactively explored.

The Galaxy (Goecks *et al.*, 2010) front-end of CART enables users to combine individual modules into new workflows, allowing for easy customization and extension of the standard use case described above. Galaxy moreover facilitates reproducibility due to its history and sharing functionalities (Goecks *et al.*, 2010).

2 RESULTS

CART uses a comprehensive chemical reference space of about 167.6 million names and synonyms disambiguated to 33.38 million chemical identifiers based on information from the STITCH database version 4.0 (Kuhn *et al.*, 2014). Matching user-provided chemical names into this reference space is very fast, e.g. processing 1,000 chemicals takes <40 seconds (Suppl. Fig. 1), allowing integrative analyses at a large scale. This is becoming crucial due to the data deluge of publicly available chemical bioactivity data (Wang *et al.*, 2012).

We benchmarked the accuracy of CART's (approximate) name matching algorithm using four datasets, for which a mapping to STITCH or PubChem identifiers already existed and could serve as a gold standard. We found CART's sensitivity to exceed 99% on all benchmarks, while precision ranged between 79 and 98% (Suppl. Fig. 2). As an additional means of ensuring high analysis standards, CART enables the user to interactively curate the automatic name matching results before proceeding further.

Owing to its unified reference chemical space, CART offers seamless integration of user-provided data with a number of databases containing functional annotations of chemicals at various scales (Table 1). These databases vary in scope, as the number of annotated chemicals ranges from >220,000 compounds with known protein interactions (Kuhn *et al.*, 2014; Qin *et al.*, 2014) to a few hundred drugs for which therapeutic classification, metabolization and toxicity information (Croset *et al.*, 2014; Law *et al.*, 2014; Kuhn *et al.*, 2015) is publicly available (Suppl. Fig. 3). However, for a set of 702 well-characterized chemicals, annotations from ≥5 databases are provided (Suppl. Fig. 3). CART's annotation and enrichment functionality is demonstrated on drug sets previously defined in a study (Rihel *et al.*, 2010) that screened chemicals for behavioural effects on zebrafish larvae (Suppl. Note 3 and Suppl. Fig. 4). It revealed coherent themes of drug bioactivities, which could otherwise only be discovered by expert manual annotations (as done in Rihel *et al.* (2010)).

In summary, CART implements a fast and accurate approach for matching chemical names to a comprehensive chemical universe. This facilitates the retrieval of enriched annotations from various databases describing chemical effects on biological systems (Table 1) and their exploration in an interactive network view. CART thus makes integrative analysis of chemical bioactivity data easy even for non-specialists.

ACKNOWLEDGEMENT

We thank Yan Ping Yuan for technical support, Sevi Durdu and Nurlanbek Duishoev for images, and members of the Bork group and the anonymous reviewers for helpful suggestions.

REFERENCES

Croset,S., et al. (2014) The fuctional therapeutic chemical classification system. *Bioinformatics*, **30(6)**, 876-83.

Goecks,J., et al. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, **11(8)**, R86.

Huang,da W., et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4(1)**, 44-57.

Iskar,M., et al. (2012) Drug discovery in the age of systems biology: the rise of computational approaches for data integration, *Curr Opin Biotechnol* **23(4)**, 609-16.

Kuhn,M., et al. (2015) The SIDER database of drugs and side effects. *Nucl Acids Res*, [Epub ahead of print].

Kuhn,M., et al. (2014), STITCH 4: integration of protein-chemical interactions with user data. *Nucl Acids Res*, **42**, 401-7.

Law,V., et al. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*, **42**, 1091-7.

Franz,M., et al. (2015). Cytoscape. js: a graph theory library for visualisation and analysis. *Bioinformatics*, [Epub ahead of print].

Qin,C., et al. (2014) Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res*, **42**, 111-23.

Rihel,J., et al. (2010) Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science*, **327(5963)**, 348-51.

Wang,Y., et al. (2012) PubChem's BioAssay Database. *Nucleic Acids Res*, D400-12.