Tab-Separated Value (TSV) Files from BindingDB

February 26, 2015

Overview

BindingDB provides data files in a spreadsheet-compatible, Tab-separated value (TSV) format. Each row contains information for one binding measurement, that is, for the interaction of one small molecule ligand (also termed a compound or monomer) with one protein target. Each row includes a SMILES string for the ligand, the identity of the target, the measured affinity (usually Ki, IC50, or Kd), the source of the data, and links to related information in other databases. When a piece of information is unavailable, the data cell is left blank.

BindingDB switched from the widely used comma-separated value (CSV) format to TSV because some data fields occasionally contain legitimate commas. These are improperly interpreted as value-separators, unless additional characters are added distinguish them from the commas intended as value-separators. We found that errors often occur during the addition and the subsequent removal of these additional characters. Because no data fields include legitimate Tab characters, the Tab-separated value format avoids these problems.

BindingDB TSVs reference following databases:

ChEBI: www.ebi.ac.uk/chebi/
ChEMBL: www.ebi.ac.uk/chembl/
CSAR: www.csardock.org/
DrugBank: www.drugbank.ca/
IUPHAR_GRAC: www.iuphar-db.org/
KEGG: www.genome.jp/kegg/
PDB: www.pdb.org
PDSP Ki: pdsp.med.unc.edu/pdsp.php
PubChem: pubchem.ncbi.nlm.nih.gov/
PubMed: www.ncbi.nlm.nih.gov/pubmed
UniProtKB: www.uniprot.org/
ZINC: zinc.docking.org/

Columns in a BindingDB TSV file

The following list defines the content of each column, in the order presented. Note that the total number of columns can vary from row to row, because the row ends with a set of columns that repeats for each protein chain (i.e., BindingDB Polymer) of the target. For example, if a protein is a trimer, and if this information was fully captured by the curator, then the last set of columns will occur three times, once for each protein chain.

- **BindingDB Reactant_set_id**. Stable internal BindingDB identifier for this binding reaction.
- **Ligand SMILES.**
- **Standard InChI.**
- **InChI key**
- **BindingDB MonomerID**. Internal BindingDB identifier for this ligand

- **BindingDB Ligand Name.** Human-readable name assigned to this Ligand by whoever curated the data.
- **Target Name Assigned by Curator or DataSource**. Name of protein Target.
- **Target Source Organism According to Curator or DataSource.** Organism associated with the protein Target.
- **Ki (nM)**
- **IC50 (nM)**
- **Kd (nM)**
- **EC50 (nM)**
- **kon (M-1-s-1)**
- **koff (s-1)**
- **pH.**
- **Temp (C)**
- **Curation/DataSource.** Typically one of the following: BindingDB, ChEMBL, PubChem, PDSP Ki, CSAR, PubChem AID, or Deposited by Author
- **Article DOI.** Digital object identifier for the source document, usually a journal article, if available.
- **PMID**. PubMed ID of article, if available.
- **PubChem AID.** PubMed Assay ID, for data drawn from PubChem.
- **Patent Number**. If applicable.
- **Authors.**
- **Institution.** Where the measurement was made. Usually a university or company.
- **Link to Ligand in BindingDB.** Preformatted URL to a query for data for this Ligand within BindingDB.
- **Link to Target in BindingDB.** Preformatted URL to a query for data for this Target within BindingDB.
- **Link to Ligand-Target Pair in BindingDB.** Preformatted URL to a query for data for the Ligand-Target pair in BindingDB.
- **Ligand HET ID in PDB.** If available, the hetero group ID for this ligand in the PDB.
- **PDB ID(s) for Ligand-Target Complex.** If available. Criterion for protein match is 85% sequence identity.

- **PubChem CID.** Compound ID of this Ligand in PubChem.
- **PubChem SID.** Substance ID of this Ligand in PubChem.
- **ChEBI ID of Ligand**
- **ChEMBL ID of Ligand.**
- **DrugBank ID of Ligand.**
- **IUPHAR_GRAC ID of Ligand.**
- **KEGG ID of Ligand.**
- **ZINC ID of Ligand.**
- **Number of Protein Chains in Target** (>1 implies a multichain complex). The following information will be provided, if available, for each chain in the protein.
    - **BindingDB Target Chain Sequence**
    - **PDB ID(s) of Target Chain.** Criterion for a match is 85% sequence identity.
    - **UniProt (SwissProt) Recommended Name of Target Chain.** Criteria for a UniProt match is 100% sequence identity and a matching Source Organism. However, it is not required that the full lengths of the chains match.
    - **UniProt (SwissProt) Entry Name of Target Chain**
    - **UniProt (SwissProt) Primary ID of Target Chain**
    - **UniProt (SwissProt) Secondary ID(s) of Target Chain**. Secondary IDs are obsolete or deprecated. They are provided for the sake of backward compatibility.
    - **UniProt (SwissProt) Alternative ID(s) of Target Chain**. Alternative IDs are Primary or Secondary IDs for chains with 100% matches that are annotated as deriving from a different Source Organism.
    - **UniProt (TrEMBL) Submitted Name of Target Chain**
    - **UniProt (TrEMBL) Entry Name of Target Chain**
    - **UniProt (TrEMBL) Primary ID of Target Chain**
    - **UniProt (TrEMBL) Secondary ID(s) of Target Chain**
    - **UniProt (TrEMBL) Alternative ID(s) of Target Chain**