Online chemical modeling environment: database

Sergii Novotarskyi, Iurii Sushko, Robert Körner, Anil Kumar Pandey and Igor V. Tetko

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH) Institute of Bioinformatics & Systems Biology (HMGU)

Washington D.C., The 238th ACS National Meeting, 16 August 2009
Layout

- Motivation
- Design & features
  - Browsers
  - Data introduction and manipulation
  - Data quality control
  - Hierarchy of users
  - Models
- Conclusions
Motivation

Properties of molecules
- Data are lost after publication of an article
- The original sources of data are difficult to track
- The conditions of experiments are frequently not provided
- The conversion between different units is error prone
- Current databases do not allow community correction of errors
- The tracking of changes by users) is required

Models
- Most published models are never used
- Implementation can be as difficult as new model development
- Different implementations can produce different results*

*See our talk at COMP, Tuesday, August 18 at 16:10-16:40 pm and poster at MEDI, Wednesday, August 19, paper id 1294701, 19:00-21:00
Data structure: behind the scene
Database schema
Simplified overview

Property

\[ \log P = 0.5 \]

Melting Point = 100 °C

Condition

Temperature, pH, species, tissue, method

Tags

Toxicology, Biology, Partition coefficient.

Structure

Benzene. Urea. ...

Data Record

Manipulation

Editing

Organization

Working sets

Introducer

Bill G., Sergey B.

Date of modification

Informationsystem

Article

Garberg, P

“In vitro models for ...”
Database elements: browsers

- Browser of records
  - Filtering options (fragments, name, article, properties)
  - Evidences (experimental, non-experimental)
  - Tracking of changes
  - Hidden/public records (hierarchy of users)
  - Tracking & verification of names
- Browser of properties
- Browser of conditions
- Browser of units
- Browser of articles
  - Browser of journals
- Browser of molecules
- Browser of tags
- Browser of models
Database schema

Another simplified overview

Molecules

- log(IC50-1) = 2.02 -log (mmol/L)
- Temperature = 25.0

Names

Evidences

Properties

- log(IC50-1) (concentration) 1093 records
- LogPsurf (dimensionless) 21 records
- LogPsurf(ion) (dimensionless) 21 records
- LogP (dimensionless) 35 records

Conditions

- species (dimensionless)
- Temperature (temperature)
- dose (concentration)
- Concentration (concentration)

Units

- log(mmol/L) (concentration)
- -log(mg/l) (concentration)
- nM (concentration)
- -log (mmol/l) (concentration)

Journals

- Journal of chemical information and computer sciences
- Chemico-biological interactions
- Neurochemistry international

Helmholtz Zentrum münchen
German Research Center for Environmental Health
User interface icons

- Edit current record (item, article, unit, etc.)
- Delete current record
- Most places — open record-specific submenu, sometimes — view profile
- Open a wiki page with additional explanations
- Send a message to the user
- Download data in XLS format
- Select item
Conclusions: the features, that make our database different

- "Wiki" approach to data handling
  users can add, modify and delete data

- Mandatory reference to an origin of information
  each record in a database should contain a reference to a source (article, book, proceeding or even personal communications), where the data were published

- Storing rich information
  we store measurement conditions to increase data quality

- Several tools to support decision making
  integration with other web-services (validation of molecule names against PubChem database, automatic fetching of article information from PubMed), duplicate records management

- Aimed at model building
  convenient to build training sets from data - filter by property, article and export data either to internal modelling tools or download as Excel file
Our Team
Acknowledgement

Other team members

Mr S. Brandmayer  
Dr. M. Rupp  
Dr. C. Höfer  
Dr. S. Scheek

Contributors

Dr. P. Ertl (Novartis)  
Dr. F. Csizmadia (ChemAxon)  
Prof. J. Gasteiger (Molecular Networks GmbH)  
Prof. R. Todeschini (University Milano-Bicoca)  
Prof. A. Varnek (University of Strasbourg)  
Dr. Vasyl Kovalishyn (IBPC, Kiev)  
Dr. Volodymyr Prokopenko (IBPC, Kiev)

Directors

Prof. G. Wess  
Prof. H.W. Mewes

Open source providers

CDK  
InChi  
APACHE  
and others

BMBF GO-Bio grant 0313883 CADASTER FP7 EU project  
Germany-Ukraine bilateral grant

Thank you for your attention!