Computational Systems Biology

Databases, formats and sources

2013, Valencia

Ignacio Medina
http://bioinfo.cipf.es/imedina
imedina@cipf.es

Head of the Computational Biology Unit
Computational Genomics Institute
Centro de Investigación Príncipe Felipe (CIPF)
Valencia, Spain

Index

- Introduction
- Data formats
- Biological databases

Introduction

Overview of data sources and formats

- Many new data repositories and data formats exist, many small projects and custom formats, **some data sources do not provide standard formats!**
- Biological networks data is split into different repositories, many times you need to parse and join different sources
- Different computation skills needed: RESTful/SOAP web services, XML and XSD, tabular format, RDF, SPARQL, …

Introduction
Computing needs

- XML and schema
  - PyXB, a python schema binding
- Web Services
  - SOAP
  - RESTful
- OBO ontologies
  - SBO ontology
  - MI ontology

Data formats
Overview

- There too many, main data formats include:
  - Simple interaction file (SIF)
  - NNF
  - GML and XGMML
  - SBML
  - BioPAX
  - PSI-MI and PSI-MI tab
  - DOT
- One Graphical language format proposed:
  - SBGN

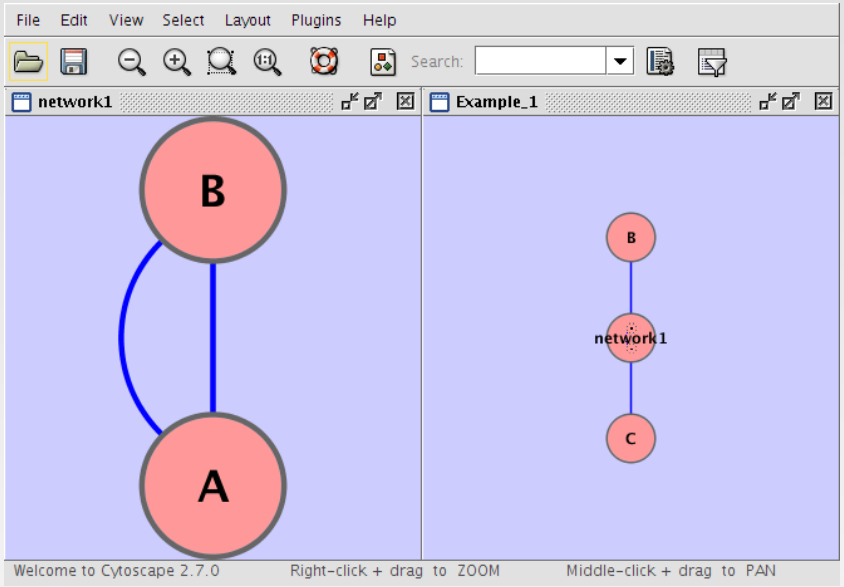
Data formats
SIF format

- ***SIF***: Simple Interaction format
- Very simple format tabulated file with 3 columns:
  - nodeA <*relationship type*> nodeB
    nodeC <*relationship type*> nodeA nodeC
  - Example:
    - node1 typeA node2
      node2 typeB node3 node4 node5
      node0
- The tag *relationship type* can be any string. Commonly:
  - pp ..... protein – protein interaction
    pd ..... protein -> DNA
    pr ..... protein -> reaction
    rc ..... reaction -> compound
    cr ..... compound -> reaction
    gl ..... genetic lethal relationship
    pm ..... protein-metabolite interaction
    mp ..... metabolite-protein interaction
- More info at:
  http://www.cytoscape.org/manual/Cytoscape2_7Manual.html#SIF Format

Data formats
NNF format

- ***NNF***: Nested Network format
- Simple and similar to SIF but allows nested network in a node
  - ***Example***:

```
Example_1     C
Example_1     network1
network1      A      pp      B
network1      B      pp      A
Example_1     C      pp      B
```



- More info and examples at: http://www.cytoscape.org/manual/Cytoscape2_7Manual.html#NNF

Data formats
GML and XGMML format

- GML: Rich graph format. Used by many other network visualization packages
  - GML: http://www.fim.uni-passau.de/en/fim/faculty/chairs/theoretische-informatik/projects.html
- XGMML: the XML evolution of GML
  - Contains information about *interactions, attributes, graphical representation, …*
  - XML schema provided
  - http://wiki.cytoscape.org/XGMML

Data formats
SBML format

- ***SBML***: Systems Biology Markup Language
- It's a XML format to describe biochemical networks
- Web site: http://sbml.org/Documents
- Three different releases are available named '*Levels*'. Levels are intended to coexist .
- Most stable release is ***Level 2 Version 4:*** *http://sbml.org/Documents/Specifications#SBML_Level_2_Version_4*

- XML schema and libraries for many languages provided

Data formats
BioPAX format

- ***BioPAX***: Biological Pathway Exchange format
- Very advance semantic technologies applied
- It's a RDF/OWL-based format to represent biological pathways
  - RDF: Resource Description Framework (http://en.wikipedia.org/wiki/Resource_Description_Framework)
  - OWL: Web Ontology Language (http://en.wikipedia.org/wiki/Web_Ontology_Language)
- Serialized in RDF/XML
- Web site: http://www.biopax.org/

Data formats

PSI-MI and PSI-MI TAB format

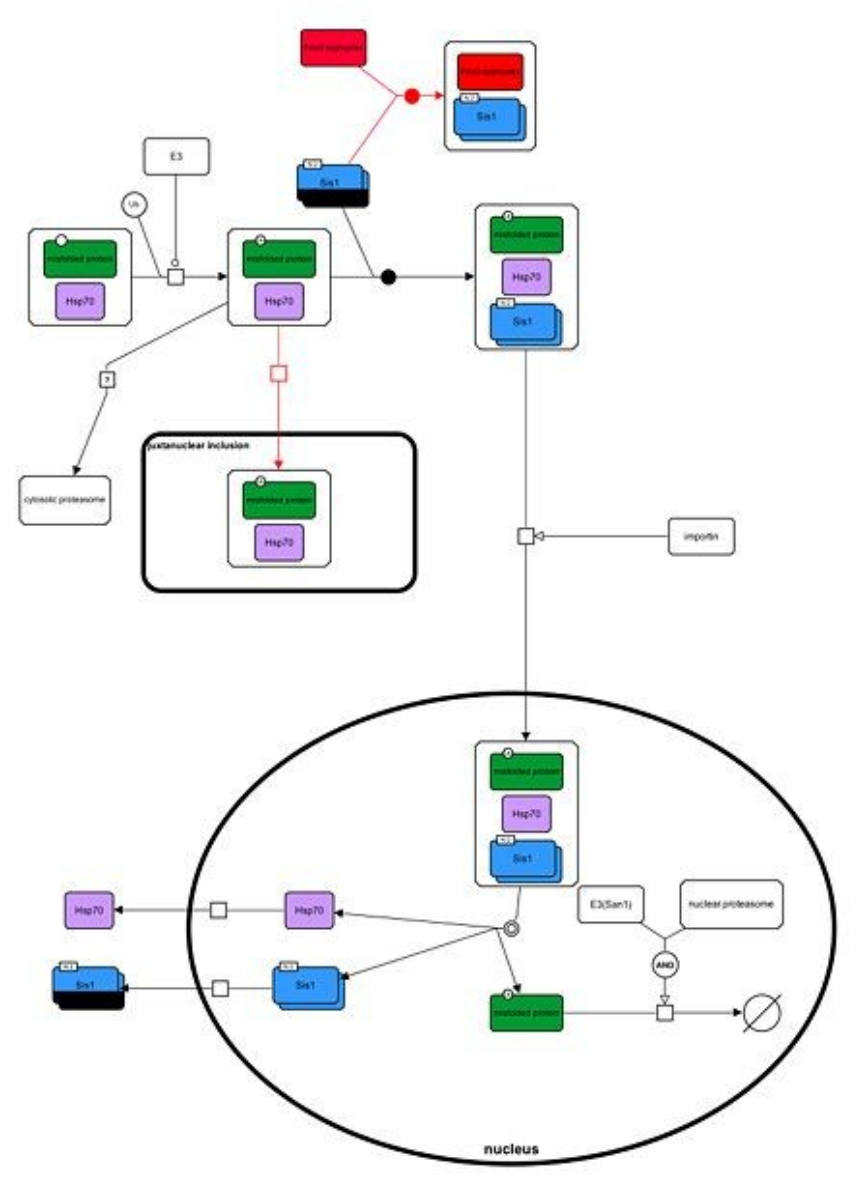
- ***PSI-MI***: Proteomic Standard Initiative – Molecular Interaction. Initiative founded at the HUPO meeting in 2002
- Web site at http://www.psidev.info/
- Two main standards:
  - PSI-MI XML: http://www.psidev.info/mif
  - PSI-MI TAB: ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/README
- MI onotology: http://www.obofoundry.org/cgi-bin/detail.cgi?id=psi-mi

Data formats
DOT format

- DOT format is used in the generic Graphviz package
- Specification:
  - http://www.graphviz.org/content/dot-language
  - Demos: http://api.graphviz.org/webdot/demo.html
- Need to install graphviz?
  - sudo apt-get install graphviz

Data formats
SBGN format

- ***SBGN***: Systems Biology Graphical Notation
- It's an effort to standardize the graphical notation use in biological networks
- Web site: http://www.sbgn.org/Main_Page



Biological Databases
Overview

- Many different sources with different type of information
- Many times the database you need does not exist... so? Just create it yourself
- Very few initiatives to join and standardize databases. Many times you will need to merge a few of them
- Types
  - Protein-protein interaction (PPI) databases
  - Pathways
  - Regulatory
  - Co-expression
  - Functional databases

Biological Databases
Protein-protein interaction databases

- Main PPI databases
  - IntAct: http://www.ebi.ac.uk/intact/
  - Mint: http://mint.bio.uniroma2.it/mint/Welcome.do
  - HPRD: http://www.hprd.org/
  - ...

Biological Databases

Pathway databases

- Some main databases:
  - KEGG: http://www.genome.jp/kegg/
  - Reactome: http://www.reactome.org/
  - BioCyc: http://biocyc.org/
- Important! Different type of data
  - In silico vs in vivo
- Others:
  - MetaCyc: http://metacyc.org/

Biological Databases

Regulatory databases

- Some data respositories
  - MiRTarBase: http://mirtarbase.mbc.nctu.edu.tw/
  - Ensembl and Encode data... not a real database

Biological Databases

Co-expression databases

- Main database
  - Expression Atlas at EBI: http://www.ebi.ac.uk/gxa/ and data at ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/gxa/

Biological Databases

Functional databases

- Two main data sources:
  - STRING: http://string-db.org/
  - Genemania: http://www.genemania.org/