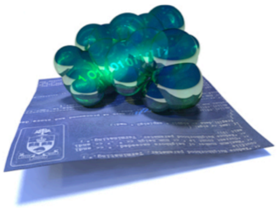
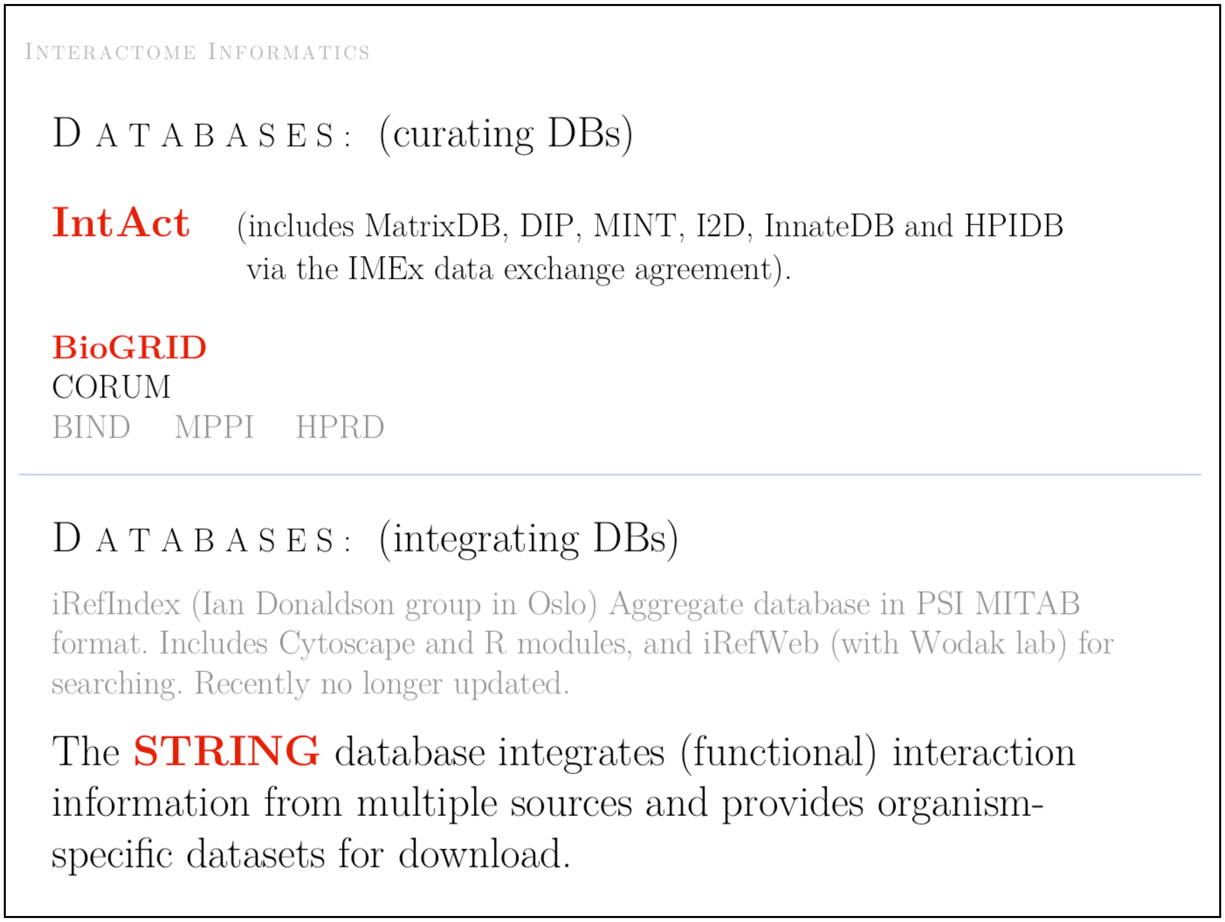
A
Bioinformatics
Course

PPI Databases



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



INTERACTOME INFORMATICS

DATABASES: (curating DBs)

IntAct (includes MatrixDB, DIP, MINT, I2D, InnateDB and HPIDB via the IMEx data exchange agreement).

BioGRID
CORUM
BIND MPPI HPRD

DATABASES: (integrating DBs)

iRefIndex (Ian Donaldson group in Oslo) Aggregate database in PSI MITAB format. Includes Cytoscape and R modules, and iRefWeb (with Wodak lab) for searching. Recently no longer updated.

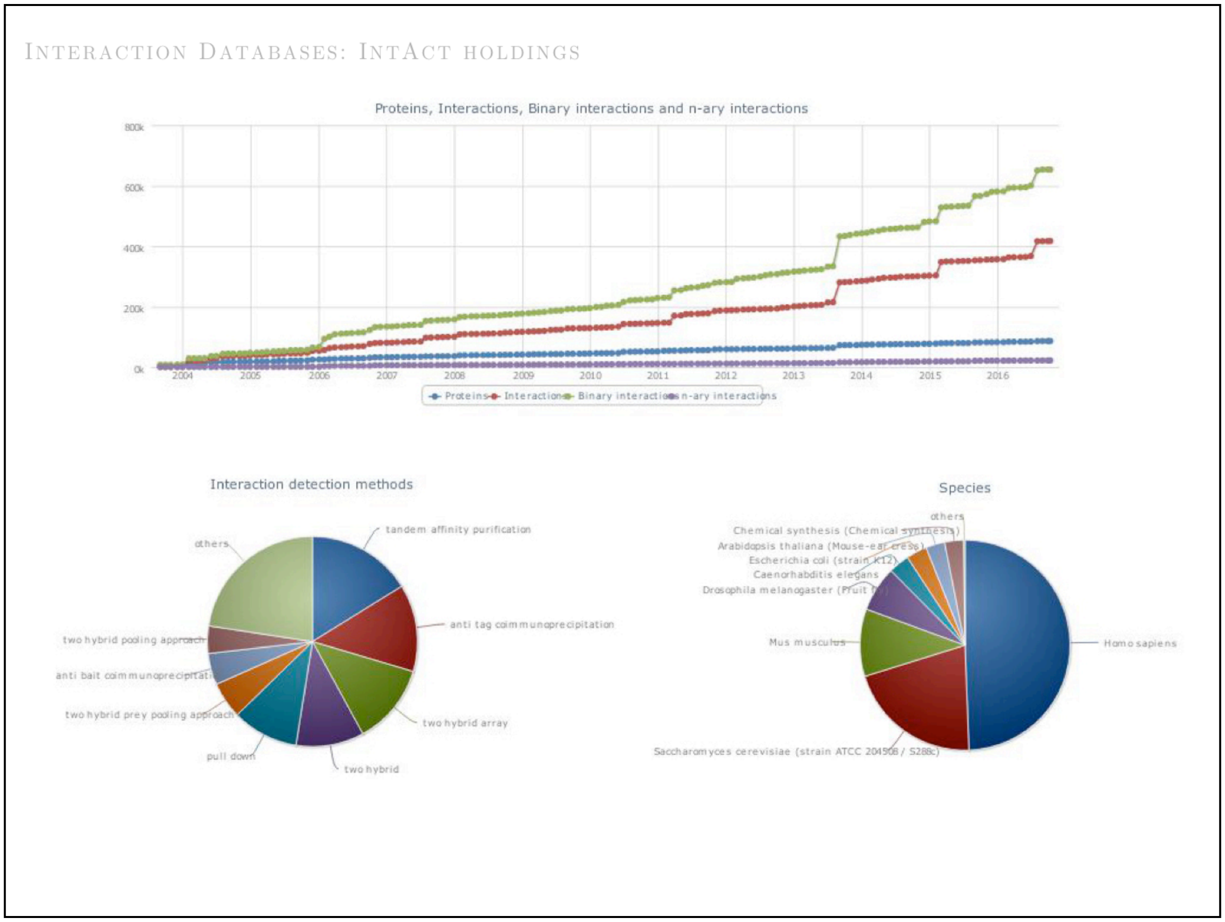
The **STRING** database integrates (functional) interaction information from multiple sources and provides organism-specific datasets for download.

The landscape of interaction databases has undergone very significant remodelling over the last decade. Currently, the databases of choice are IntAct, which spearheads the IMEx consortium of data-sharing member databases, and STRING, which integrates information into a very convenient format.

http://www.ebi.ac.uk/intact/

https://string-db.org/

Szklarczyk D *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39(Database issue): D561–D568.



Interaction Databases: IntAct holdings

IntAct statistics.

Growth of binary interactions is nearly exponential, albeit with a slow rate compared to eg. sequence databases. The most popular experiment is TAP-tag; by far the most interactions have been recorded for human proteins which alone account for as many interactions as all model-organisms combined.

cf. http://www.ebi.ac.uk/intact/about/statistics

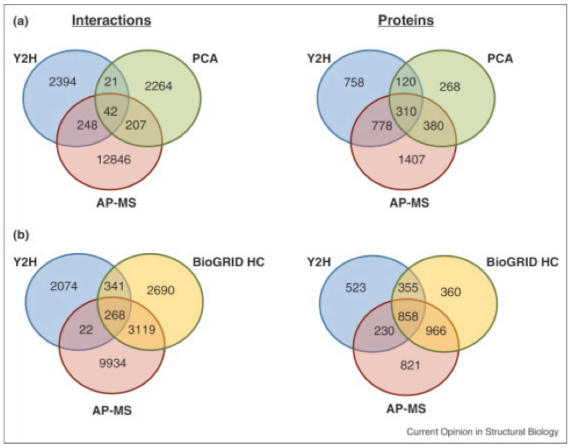
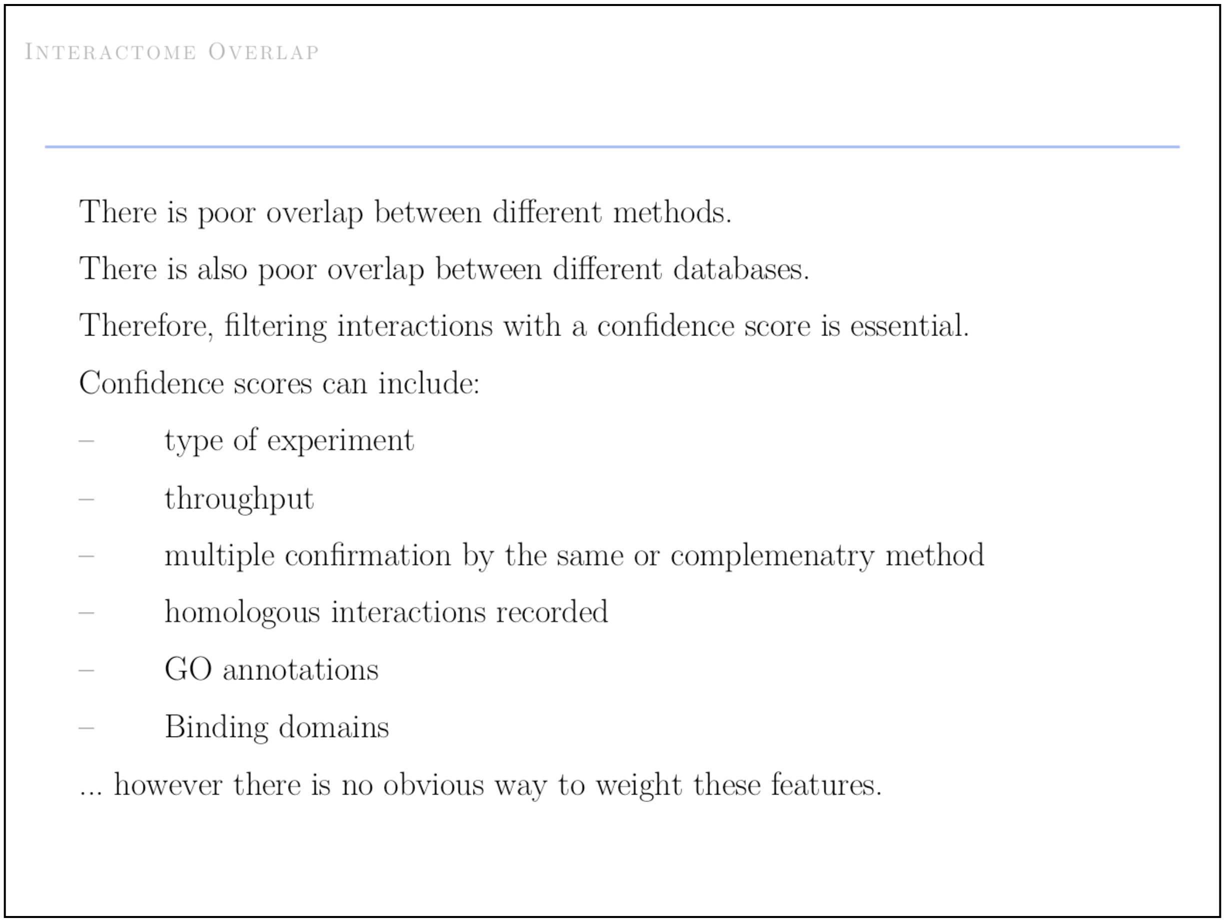


Figure 2. Venn diagrams illustrating the overlap between major high confidence (HC) yeast PPI datasets derived from recent large scale studies and curated from the published literature. (a) Overlap between the following three datasets: the Y2H (Union) dataset ...

Shoshana J Wodak, James Vlasblom, Andrei L Turinsky, Shuye Pu
Protein–protein interaction networks: the puzzling riches
Current Opinion in Structural Biology, Volume 23, Issue 6, 2013, 941–953

An issue of great concern is the poor overlap between interactions in different PPI databases, even “high-confidence” (HC) interactions.

This is likely due to varying details of experimental conditions: interactions are heavily regulated and modulated through varying expression levels and post translational modification in response to stress, metabolic state, environmental signals etc., and they are limited to specifc cellular compartments, and fluctuate over time.



Interactome Overlap

There is poor overlap between different methods.

There is also poor overlap between different databases.

Therefore, filtering interactions with a confidence score is essential.

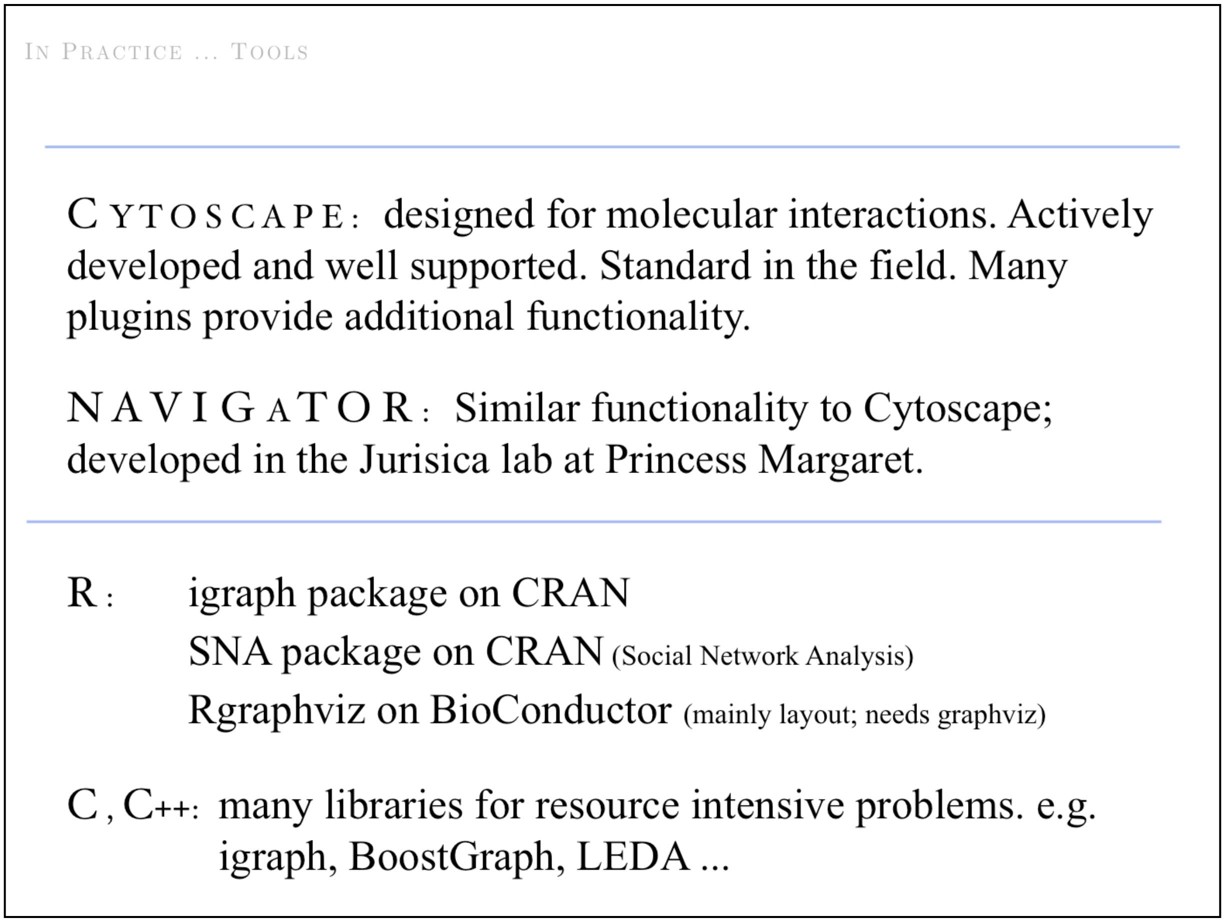
Confidence scores can include:

- type of experiment
- throughput
- multiple confirmation by the same or complemenatry method
- homologous interactions recorded
- GO annotations
- Binding domains

... however there is no obvious way to weight these features.

It is (nearly) never possible to define valid weights for orthogonal features and properties from first principles. Building valid, comprehensive confidence scores can be a good application of machine learning techniques. Training such computational tools requires both positive and negative examples. Gold-standard datasets exist for positive examples (albeit with a bias towards stable, time-independent interactions in well-defined complexes).

However, it is not trivial to define a good negative dataset: just becasue we have not observed an interaction does not mean it is not relevant or does not exist.



In Practice ... Tools

Cytoscape: designed for molecular interactions. Actively developed and well supported. Standard in the field. Many plugins provide additional functionality.

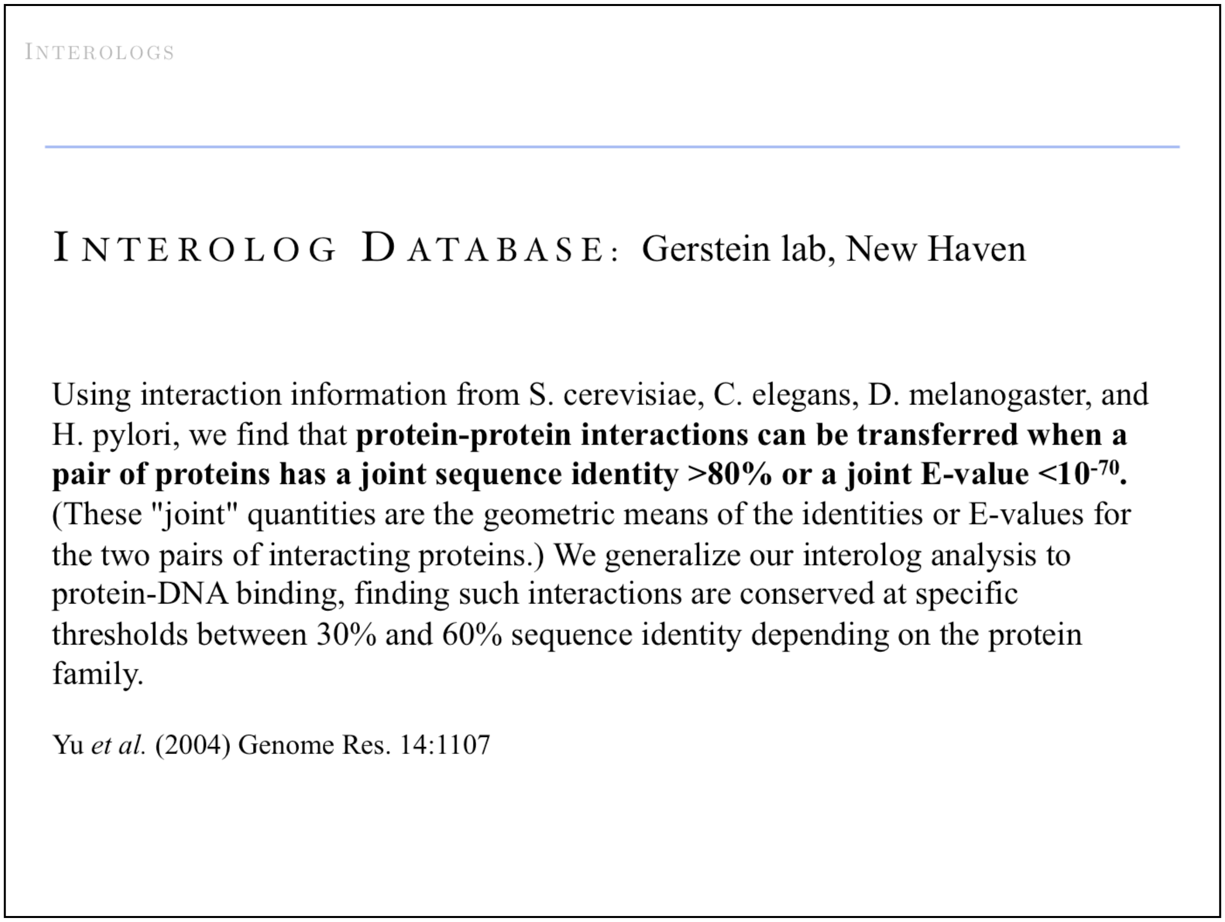
NAVIGaTOR: Similar functionality to Cytoscape; developed in the Jurisica lab at Princess Margaret.

R: igraph package on CRAN
SNA package on CRAN (Social Network Analysis)
Rgraphviz on BioConductor (mainly layout; needs graphviz)

C, C++: many libraries for resource intensive problems. e.g. igraph, BoostGraph, LEDA ...

Tools to analyze PPI networks include the desktop programs Cytoscape (the most popular), and NAVIGaTOR, but an increasing number of PPI database providers offer embedded visualization tools on their Web pages.

The go-to R package is **igraph**.



INTEROLOGS

INTEROLOG DATABASE: Gerstein lab, New Haven

Using interaction information from S. cerevisiae, C. elegans, D. melanogaster, and H. pylori, we find that **protein-protein interactions can be transferred when a pair of proteins has a joint sequence identity >80% or a joint E-value <10^{-70}.** (These "joint" quantities are the geometric means of the identities or E-values for the two pairs of interacting proteins.) We generalize our interolog analysis to protein-DNA binding, finding such interactions are conserved at specific thresholds between 30% and 60% sequence identity depending on the protein family.

Yu *et al.* (2004) Genome Res. 14:1107

Other sources for interactions are predictions made for homologous proteins. The Gerstein Lab stiores predicted annotations for C. elegans, Drosophila, Arabidopsis, and Candida, but more modern methods are now available, e.g. the BIANE Interolog Prediction Server.

http://interolog.gersteinlab.org/

http://sbi.imim.es/web/index.php/research/servers/bips

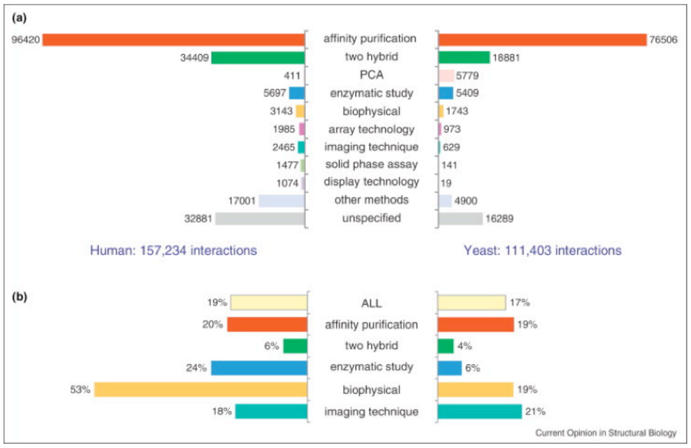
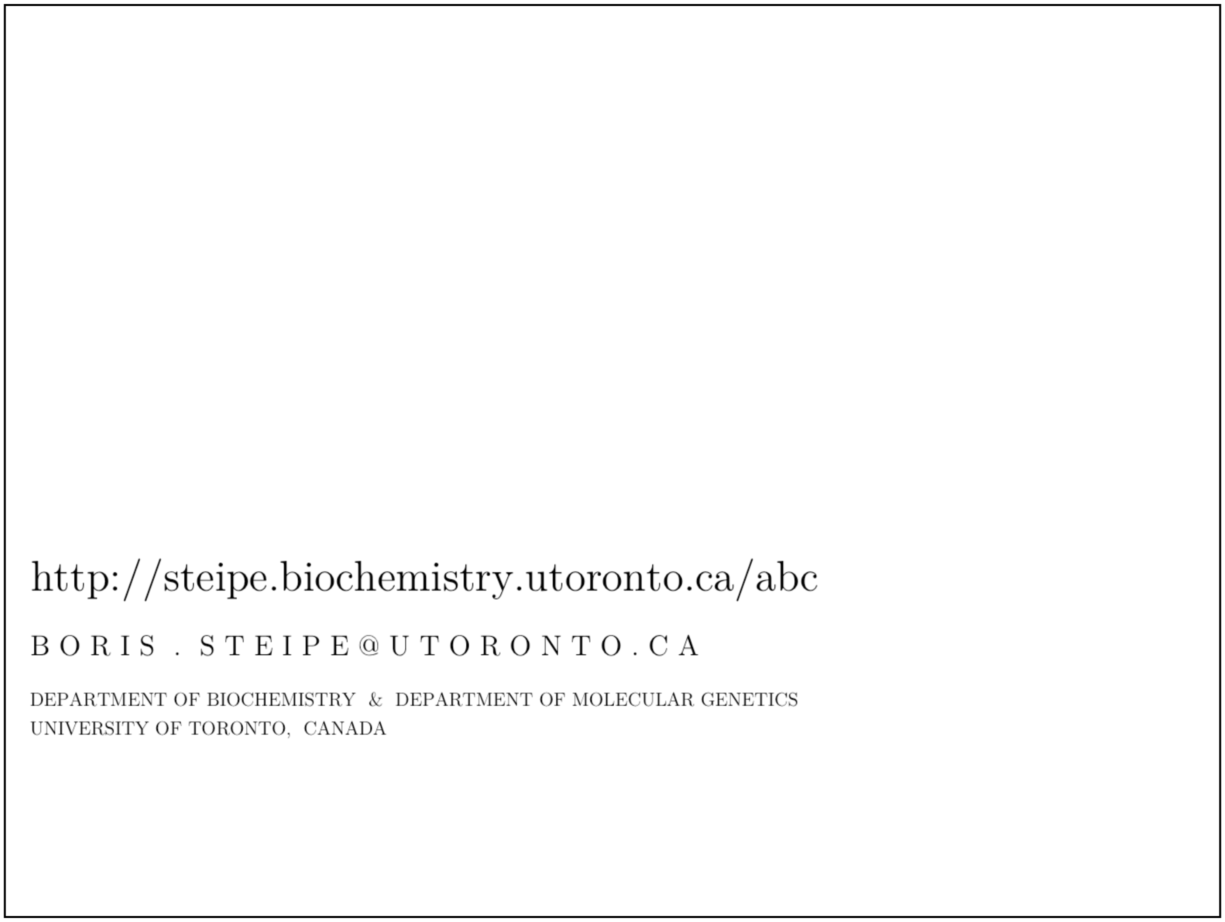


Figure 4. Overview of the PPI landscape for yeast and human consolidated from 14 public databases by the iRefIndex/iRefWeb resource. (a) Statistics on the number of interactions in human (left) and yeast (right) identified by different detection methods in hig...

Shoshana J Wodak, James Vlasblom, Andrei L Turinsky, Shuye Pu
Protein–protein interaction networks: the puzzling riches
Current Opinion in Structural Biology, Volume 23, Issue 6, 2013, 941–953

The proportion of experimental methods applied to different organisms is reoughly similar.



http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA