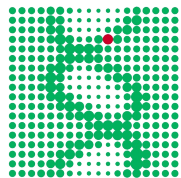




# ConsensusPathDB

a database for integrating  
human functional interactions



MPIMG

Atanas Kamburov



International  
Max Planck Research School  
for Computational Biology  
and Scientific Computing



# Introduction

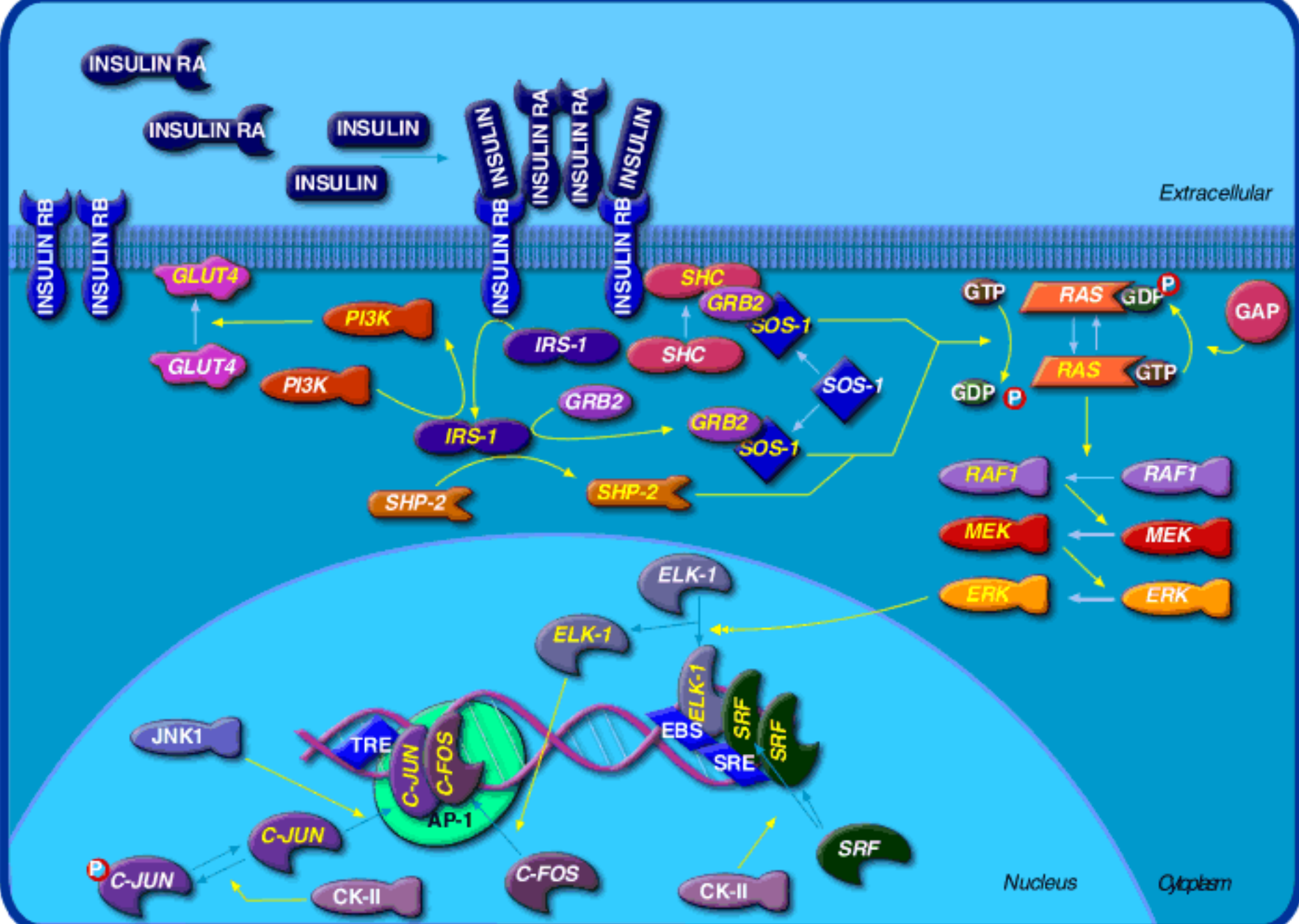
- Biological processes driven by **interacting** molecules
- Functional interactions:
  - protein-protein interactions
  - biochemical reactions
  - gene regulatory interactions
  - ...
- Different methods for interaction detection generate interaction knowledge → **interaction databases**



# Introduction

- Current interaction knowledge: ~10% of real human interactions\* interspersed in over 200 databases
- Main problems with interaction database compatibility:
  - focus: biochemical reactions **or** PPI **or** gene regulation **or** ...
  - incomplete overlap between similar databases
  - different data format and information detail
- Conclusion: interaction databases cannot represent biological truth completely and in a standard way.

\* Hart *et al.* 2006, *Genome Biol.* **7**: 120





# Introduction

- Biological processes contain biochemical reactions **and** PPI **and** gene regulation **and** ...
  - Interaction-based research: use several databases
  - ConsensusPathDB integrates:
    - protein interactions
    - metabolic reactions
    - signaling reactions
    - gene regulatory interactions
- heterogeneous interaction network in  
ConsensusPathDB is closer to biological reality



# ConsensusPathDB - content

- Integrated interaction data sources:

Database name	Version integrated	Web address
Reactome	27	<a href="http://reactome.org">http://reactome.org</a>
Kegg	49.0	<a href="http://kegg.org">http://kegg.org</a>
Humancyc	12.5	<a href="http://humancyc.org/">http://humancyc.org/</a>
Pid	2008_12_09	<a href="http://pid.nci.nih.gov">http://pid.nci.nih.gov</a>
Biocarta	2008_01_08	<a href="http://pid.nci.nih.gov">http://pid.nci.nih.gov</a>
Netpath	6.1.2009	<a href="http://www.netpath.org">http://www.netpath.org</a>
Intact	2008-12-12	<a href="http://www.ebi.ac.uk/intact/">http://www.ebi.ac.uk/intact/</a>
Dip	2008-01-13	<a href="http://dip.doe-mbi.ucla.edu">http://dip.doe-mbi.ucla.edu</a>
Mint	2008-05-19	<a href="http://mint.bio.uniroma2.it">http://mint.bio.uniroma2.it</a>
Hprd	I_090107	<a href="http://www.hprd.org">http://www.hprd.org</a>
Biogrid	2.0.48	<a href="http://www.thebiogrid.org/">http://www.thebiogrid.org/</a>
Spike	7.1.2009	<a href="http://www.cs.tau.ac.il/~spike/">http://www.cs.tau.ac.il/~spike/</a>

# ConsensusPathDB - content

- To avoid redundancy and assess the overlap of databases: **mapping algorithms**:
  - physical entities: compare identifiers → merge
  - interactions: compare relevant participants → similar
  - pathways not mapped: unclear boundaries

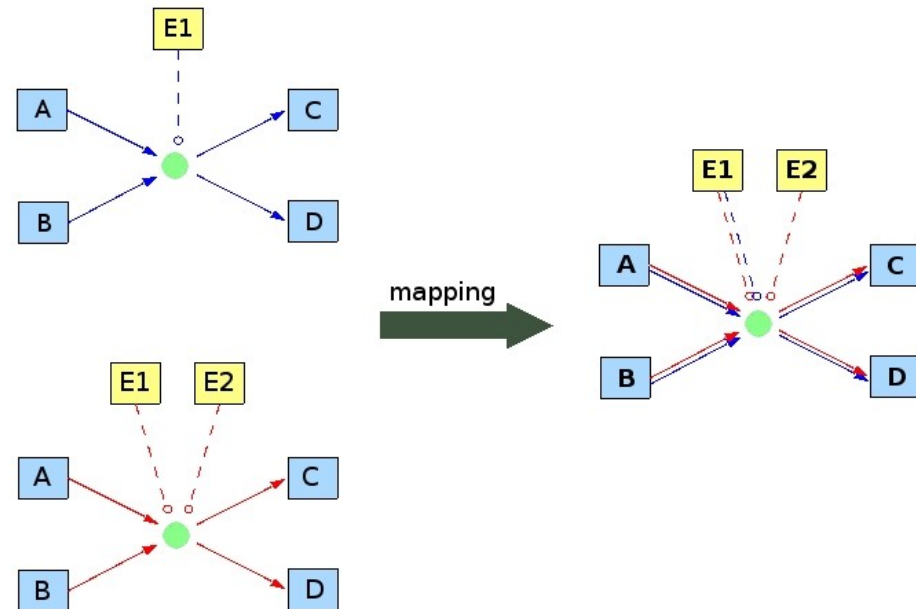
*DB X*

## • Example:

135851 interactions before integration

76167 interactions after integration

*DB Y*



# ConsensusPathDB - content

- Mapping results: source databases are complementary

	Reactome	Kegg	Humancyc	Pid	Biocarta	Netpath	Intact	Dip	Mint	Hprd	Biogrid	Spike
Reactome	4524	287	153	111	83	53	104	33	54	323	183	131
Kegg	287	1638	265	0	4	0	0	0	0	0	0	0
Humancyc	153	265	1427	0	3	0	1	1	1	6	2	2
Pid	111	0	0	3930	286	192	75	50	79	364	235	207
Biocarta	83	4	3	286	2218	124	54	36	44	145	106	173
Netpath	53	0	0	192	124	2340	305	224	398	1750	1057	636
Intact	104	0	1	75	54	305	8463	322	2828	3389	1596	4450
Dip	33	0	1	50	36	224	322	1216	395	824	597	448
Mint	54	0	1	79	44	398	2828	395	13174	7520	4382	6103
Hprd	323	0	6	364	145	1750	3389	824	7520	37947	16348	11891
Biogrid	183	0	2	235	106	1057	1596	597	4382	16348	26299	9447
Spike	131	0	2	207	173	636	4450	448	6103	11891	9447	22230



# Protein-protein interaction networks

- Many studies focus on PPI network topology to deduce biological knowledge (e.g. connectivity  $\leftrightarrow$  essentiality)
- How representative are DBs for the real PPI network?

# Protein-protein interaction networks

- Many studies focus on PPI network topology to deduce biological knowledge (e.g. connectivity  $\leftrightarrow$  essentiality)
- How representative are DBs for the real PPI network?

	IntAct	DIP	MINT	HPRD	BioGRID	SPIKE	Integrated
<b>Physical entities</b>	3,799	936	5,756	9,392	8,026	6,751	13,993
<b>Binary interactions</b>	51,626	1,139	13,509	40,716	23,734	21,880	132,735
<b>Average connectivity</b>	27.2	2.4	4.7	8.7	5.9	6.5	19.0
<b>Average clustering coefficient</b>	0.53	0.18	0.13	0.21	0.15	0.11	0.31
<b>Average path length</b>	4.2	6.0	4.6	4.2	4.7	4.2	3.9
<b>Diameter (average minimum distance between pairs of proteins)</b>	16	16	13	14	14	11	12
<b>Most frequent path length</b>	4 (32.7%)	5 (18.3%)	4 (35.3%)	4 (44.3%)	5 (34.5%)	4 (44.7%)	4 (46.3%)

Table generated with version 7 of ConsensusPathDB



# Protein-protein interaction networks

- Current separate PPI database contents are rarely representative of the real PPI network
- de Silva et al., 2006:  
“Present protein interaction network data sets include only interactions among subsets of the proteins in an organism. Previously this has been ignored, but in principle any global network analysis that only looks at partial data may be biased.”
- Data integration is a way towards meaningful network analyses



# ConsensusPathDB – public access

- ConsensusPathDB freely accessible through  
**<http://cpdb.molgen.mpg.de>**
- Web services access available shortly

home  
content information  
**search entities or pathways**  
shortest interaction paths  
over-representation analysis

**search shortest paths**

Version 7.06.07.0000  
**over-representation analysis**

**data upload**

Specify mapping criteria

Map and visualize interactions

select	Role	External links
<input type="checkbox"/>	<b>Interactions of Apoptosis regulator Bcl-X</b>	
<input type="checkbox"/>	Physical interaction of Apoptosis regulator Bcl-X and p53 protein	H
	<b>similar interactions</b>	<b>matching external information links</b>
	tp53-bcl2	●●● I
	tp53-bcl3	●●● I
	p53-bclx	●●● B
	Physical interaction of BCL2L1 and TP53	●●● B
	Protein-Protein interaction between BCL2L1 and TP53	●●● S
<input type="checkbox"/>	Protein-Protein interaction between BCL2L1 and BNIP3L	S
	<b>similar interactions</b>	<b>matching external information links</b>
	Physical interaction of bcl3_human and Apoptosis regulator Bcl-X	●●● H
	Physical interaction of BCL2L1 and BNIP3L	●●● B
<input type="checkbox"/>	bcl21-bcl-1	I
<input type="checkbox"/>	Interaction involving Apoptosis regulator Bcl-X	P
<input type="checkbox"/>	bcl3-bcl3-2	I
<input type="checkbox"/>	Protein-Protein interaction between BCL2L1 and BNIP3L	S

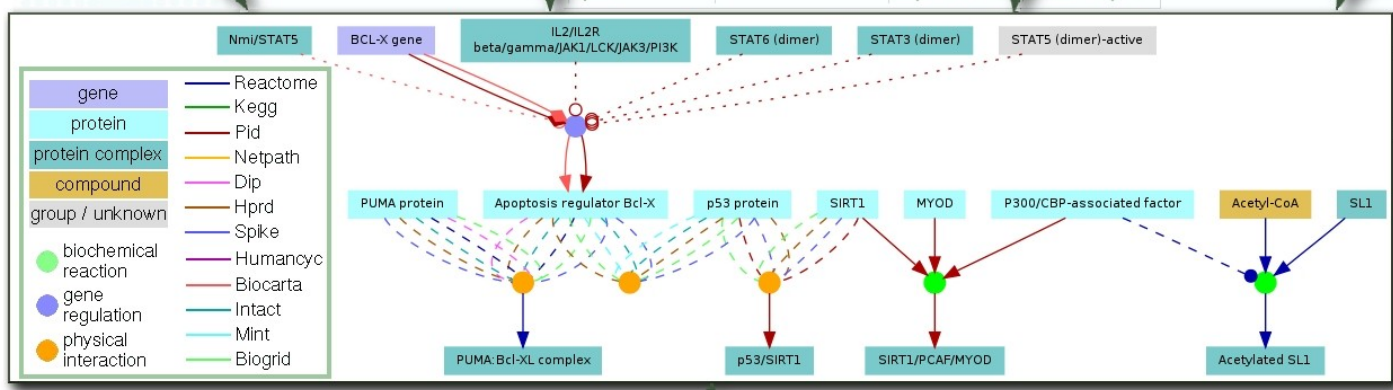


Enriched neighborhood-based sets (download)

set centers	radius	set size	candidates	p-value	set sources
CASP8 and FADD-like apoptosis regulator precursor	1	33	3 (4.5%)	8.12e-06	M D R P H B I S H
RIP	1	44	3 (4.5%)	2.61e-05	D B R P H B I M S
Caspase-10 precursor	1	48	3 (4.5%)	3.69e-05	M N R P H B I S S

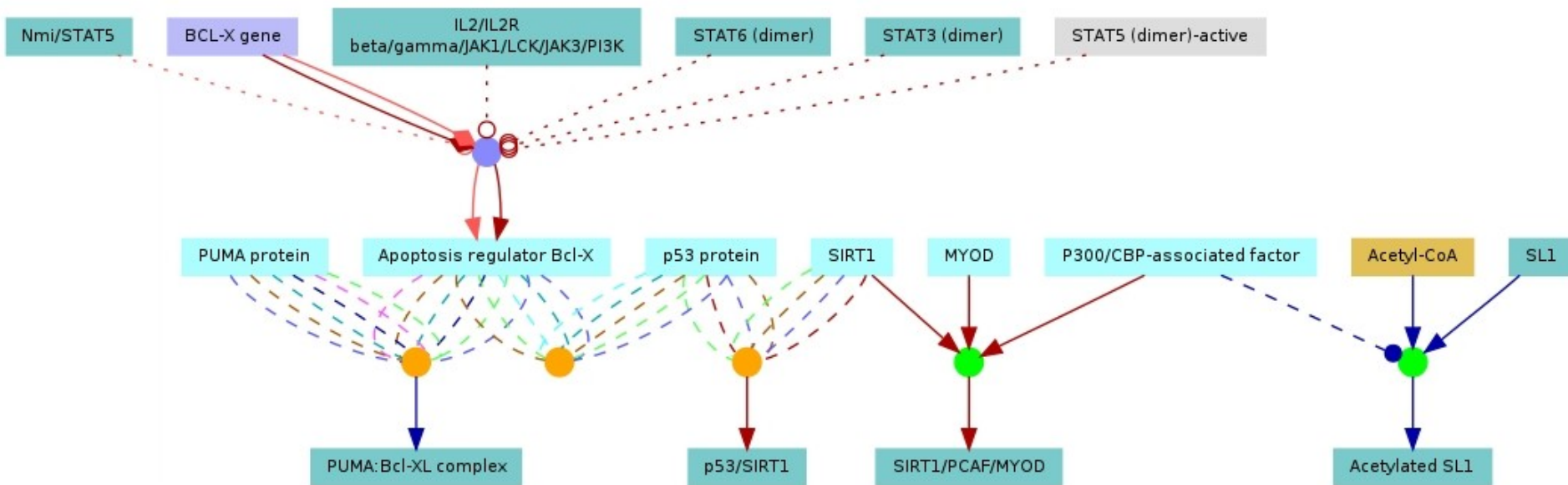
Enriched pathway-based sets (download)

pathway name	set size	candidates	p-value	pathway source
Eukaryotic Translation Initiation	117	5 (7.5%)	6.44e-06	Reactome
Cap dependent Translation Initiation	117	5 (7.5%)	6.44e-06	Reactome
Translation	124	5 (7.5%)	9e-06	Reactome
Insulin effects increased synthesis of γ-Glutose-5-Phosphate	2	1 (1.5%)	1.52e-05	Reactome
N-Glycan degradation - Homo sapiens (human)	16	2 (3.0%)	3.11e-05	KEGG
Formation of a pool of free 40S subunits	98	4 (6.0%)	4.09e-05	Reactome
3'-UTR-mediated translational regulation	110	4 (6.0%)	7.1e-05	Reactome
GTP hydrolysis and joining of the 60S ribosomal subunit	110	4 (6.0%)	7.1e-05	Reactome
L13a mediated translational silencing of Ceruloplasmin expression	110	4 (6.0%)	7.1e-05	Reactome

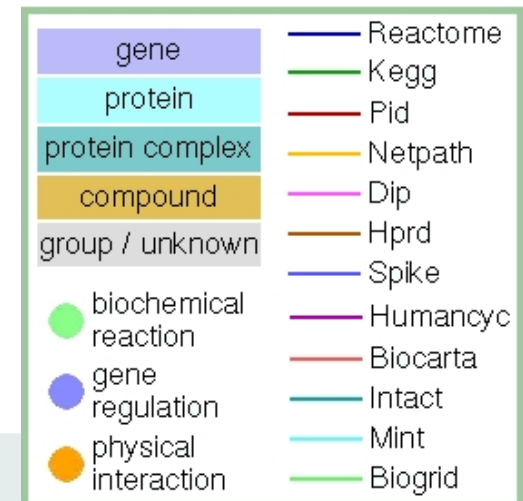


**data download**

# ConsensusPathDB – visualization



- Interaction network visualization
  - Bipartite network
  - Interaction sources encoded in edge colors





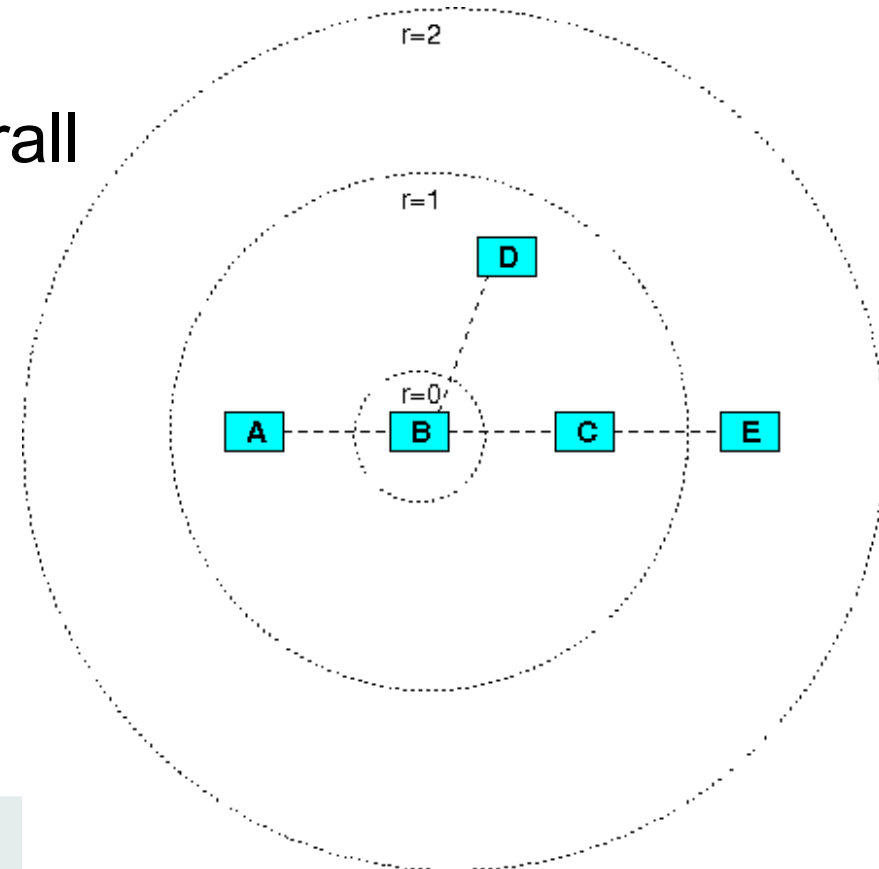
# ConsensusPathDB – ORA

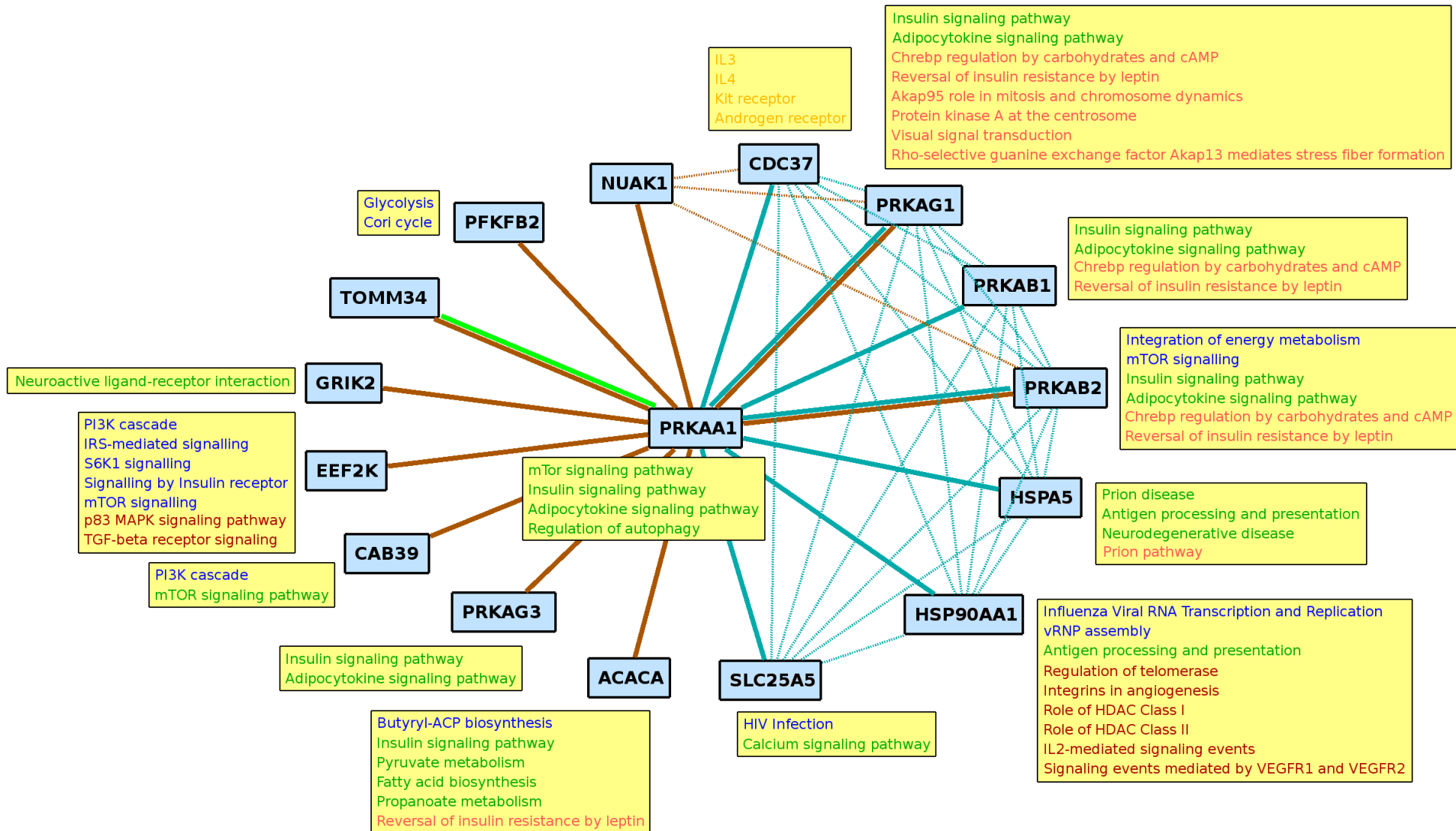
- Over-representation analysis: which **predefined** gene sets are statistically over-represented in an **input** gene list
- Predefined sets: pathways, GO categories, positional sets ...
- Input list: e.g. diff. expressed genes in cancer patients
- Method: P-value (e.g. hypergeometric test) for each predefined set reflects the significance of overlap between predefined set and input list, given a background “world”
  - points e.g. to dysregulated pathways, genomic aberrations, ...

# ConsensusPathDB – ORA

- ConsensusPathDB web interface offers ORA with:
  - pathway-based entity sets (PESTs)
  - neighborhood-based entity sets (NESTs)

- NESTs: subnetworks of overall  
FIN, defined with:
  - center
  - radius





Edge / pathway name colors



BioGrid



HPRD



KEGG



PID



IntAct



Reactome



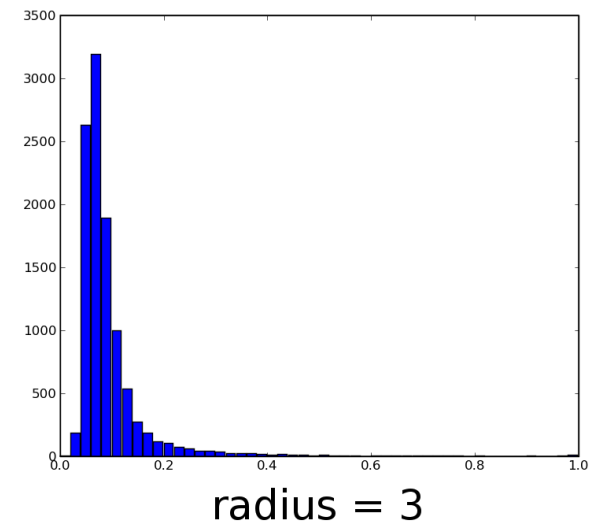
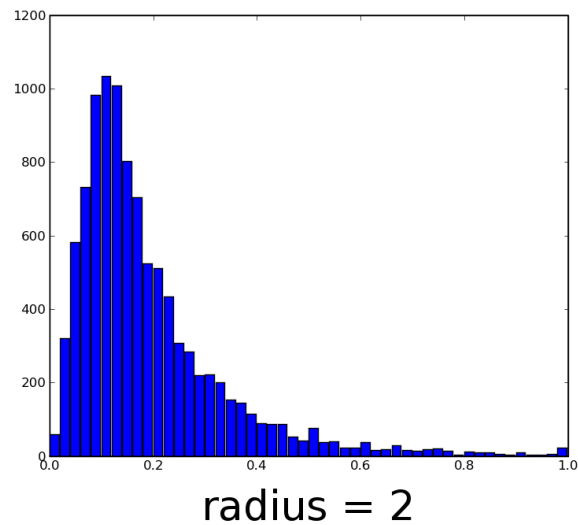
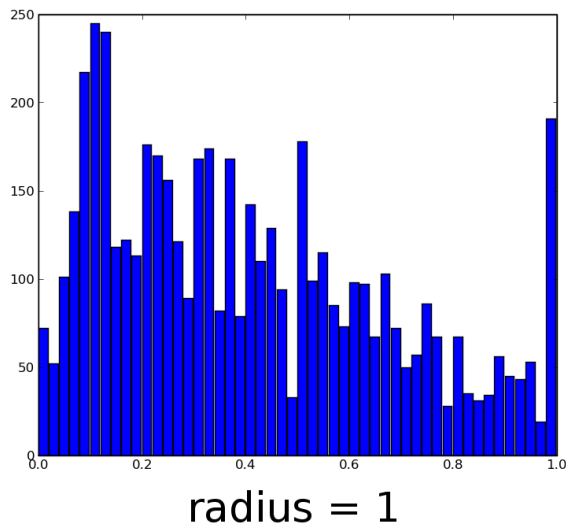
BioCarta



NetPath

# ConsensusPathDB – ORA

- NESTs go beyond single-database boundaries
- NESTs go beyond pathway boundaries
  - may represent crosstalks between pathways





# Summary and Outlook

- Data integration is important to get the full picture of cellular processes
- Network analysis based on single interaction resources may give misleading results
- ORA with NESTs may prove a useful approach, complementary to ORA with pathways
- Future plans:
  - more species
  - more resources
  - more interaction types
  - data curation
  - NESTs and disease



# Acknowledgments

Ralf Herwig

Hans Lehrach

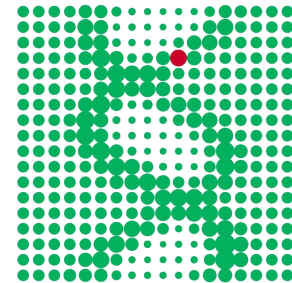
Christoph Wierling

Konstantin Pentchev

Thomas Meinel



International  
Max Planck Research School  
for Computational Biology  
and Scientific Computing



Thanks for your attention!