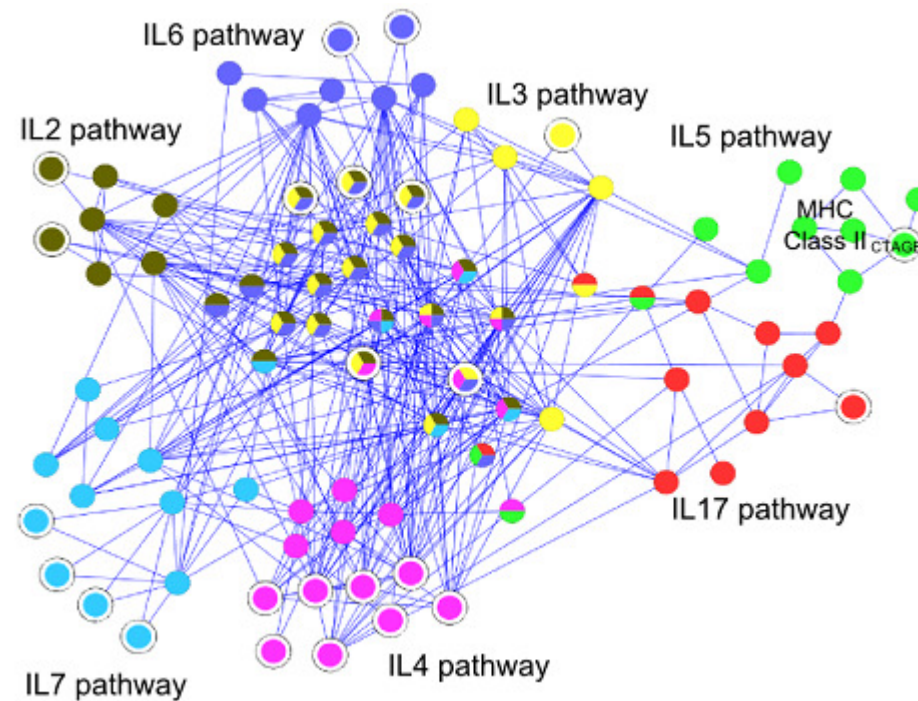


# PathExpand: Extending biological pathways using molecular interaction networks

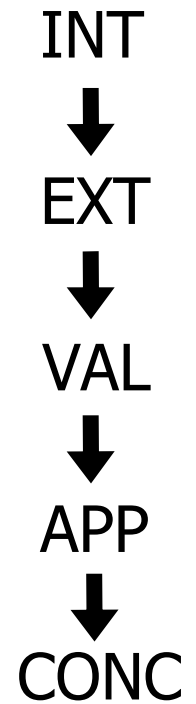


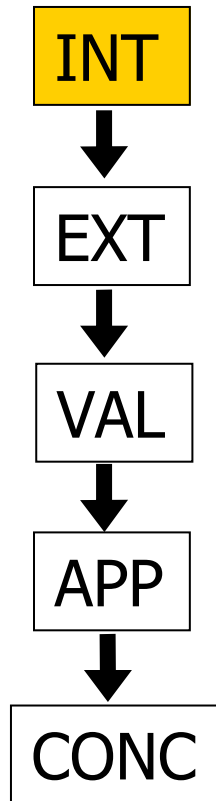
E. Glaab, A. Baudot, N. Krasnogor, A. Valencia



## Overview:

- Introduction / Motivation
- Pathway extension procedure and criteria
- Validation methods
- Biological application: Alzheimer and cancer pathways
- Conclusion



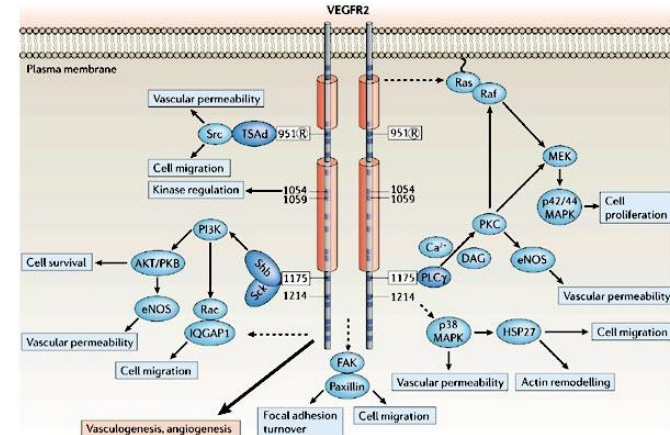


Introduction / Motivation:  
Why do we want to extend classical  
biological pathway definitions?

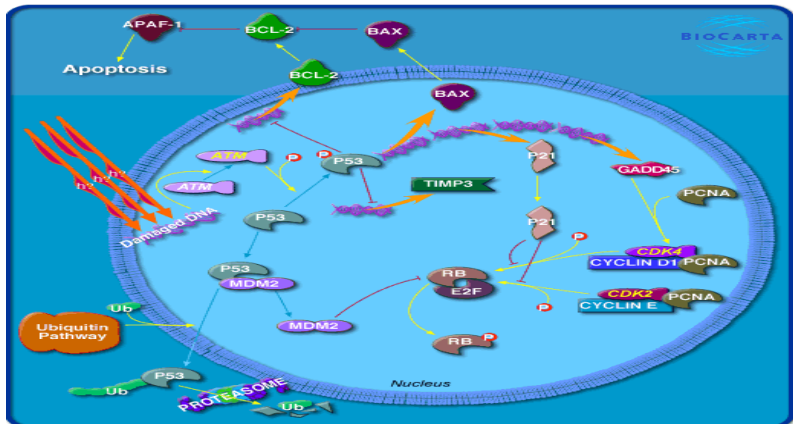


## Biological pathways and processes:

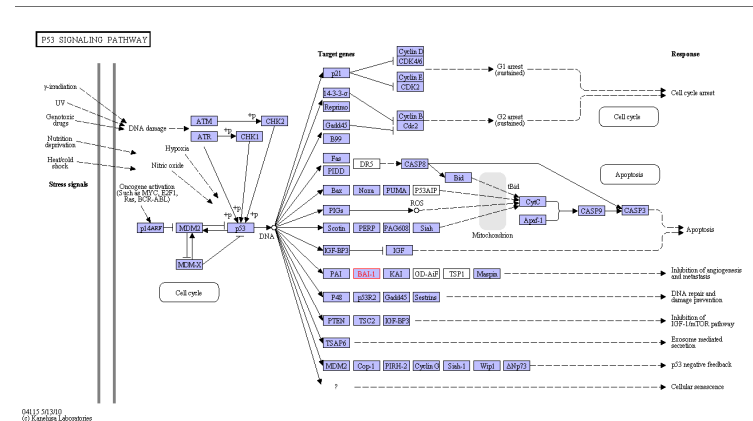
Rich sources of information, but partly subjective and inconsistent.



**Reactome (VEGF signalling)**



**BioCarta (p53 signalling)**



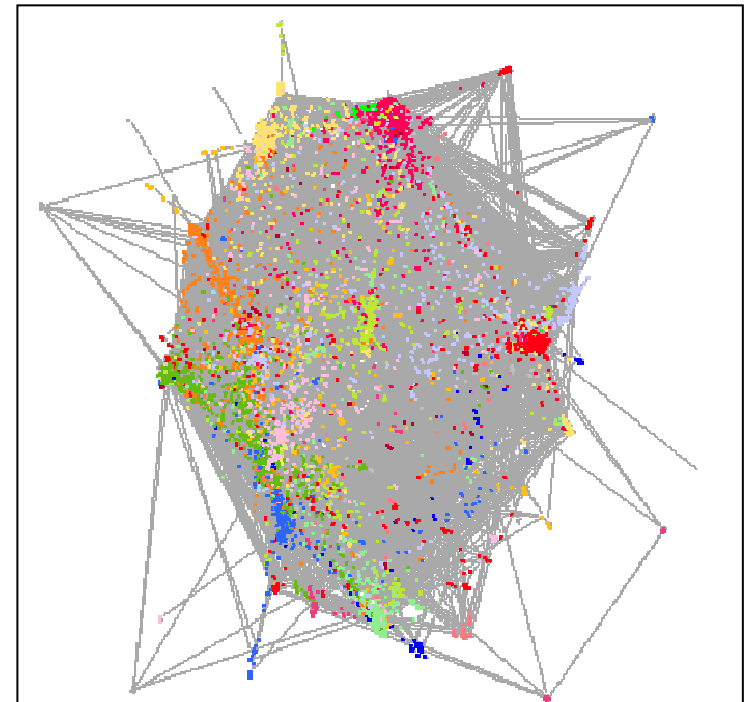
**KEGG (p53 signalling)**

## Include functional genomics data:

- protein-protein interactions
  - genetic interactions
  - gene co-expression
- large-scale, less biased

## Questions / Goals:

- Can we improve pathway definitions (compactness, connectivity, density)?
- How are pathways communicating (“cross-talk”)?

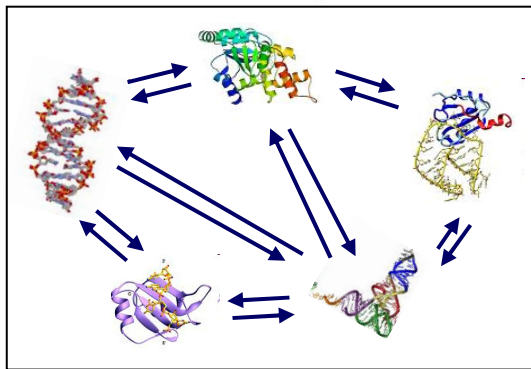


Human unweighted binary  
protein interaction network  
(9392 proteins, 38857  
interactions)

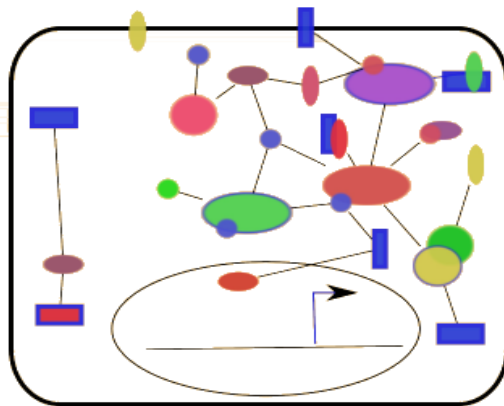


## Modelling and combining the data

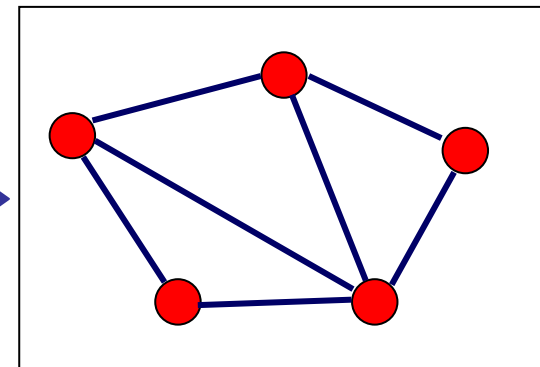
Molecular interactions:





Pathway diagrams:

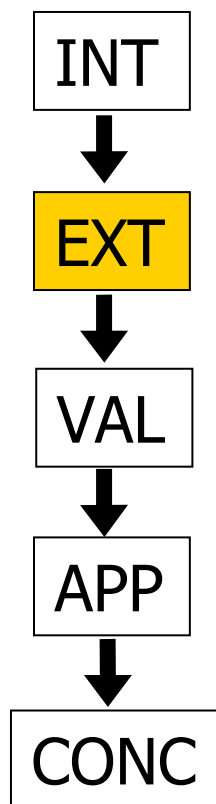


Model data as a graph  $G=(V,E)$ :  
 $V:=$  molecules;  $E:=$  interactions



Normalisation  
of gene/protein  
names

-  = annotated node
-  = unweighted edge



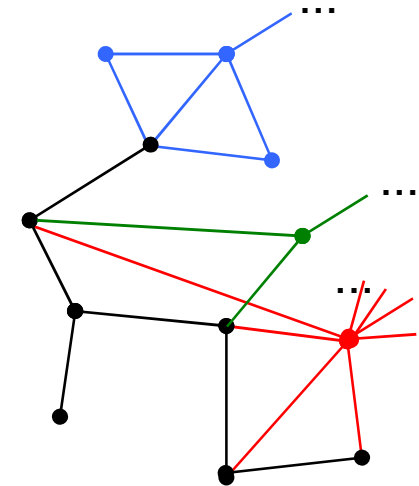
Pathway extension procedure and criteria:  
How do we recognize “good” pathway  
definitions and improve them?

## PathExpand – Idea:

Extend pathways by adding genes that are „strongly connected“ to the pathway-nodes or increase the pathway-“compactness“ in a PPI.

**Pathway extension criteria:** Add a node  $v$  to set  $P$  if:

- $v$  has a pathway-neighbour and  $\text{degree}(v) > 1$ ; and
- $\# \text{pathway-links}(v,p) / \# \text{outside-links}(v,p) > T_1$ ; or
- $\# \text{triangle-links}(v,p) / \# \text{possible\_triangles}(v,p) > T_2$ ; or
- $\# \text{pathway-links}(v,p) / \# \text{pathway-nodes}(p) > T_3$ ; and
- avg. shortest path distance in  $\{P,v\}$  smaller than in  $P$



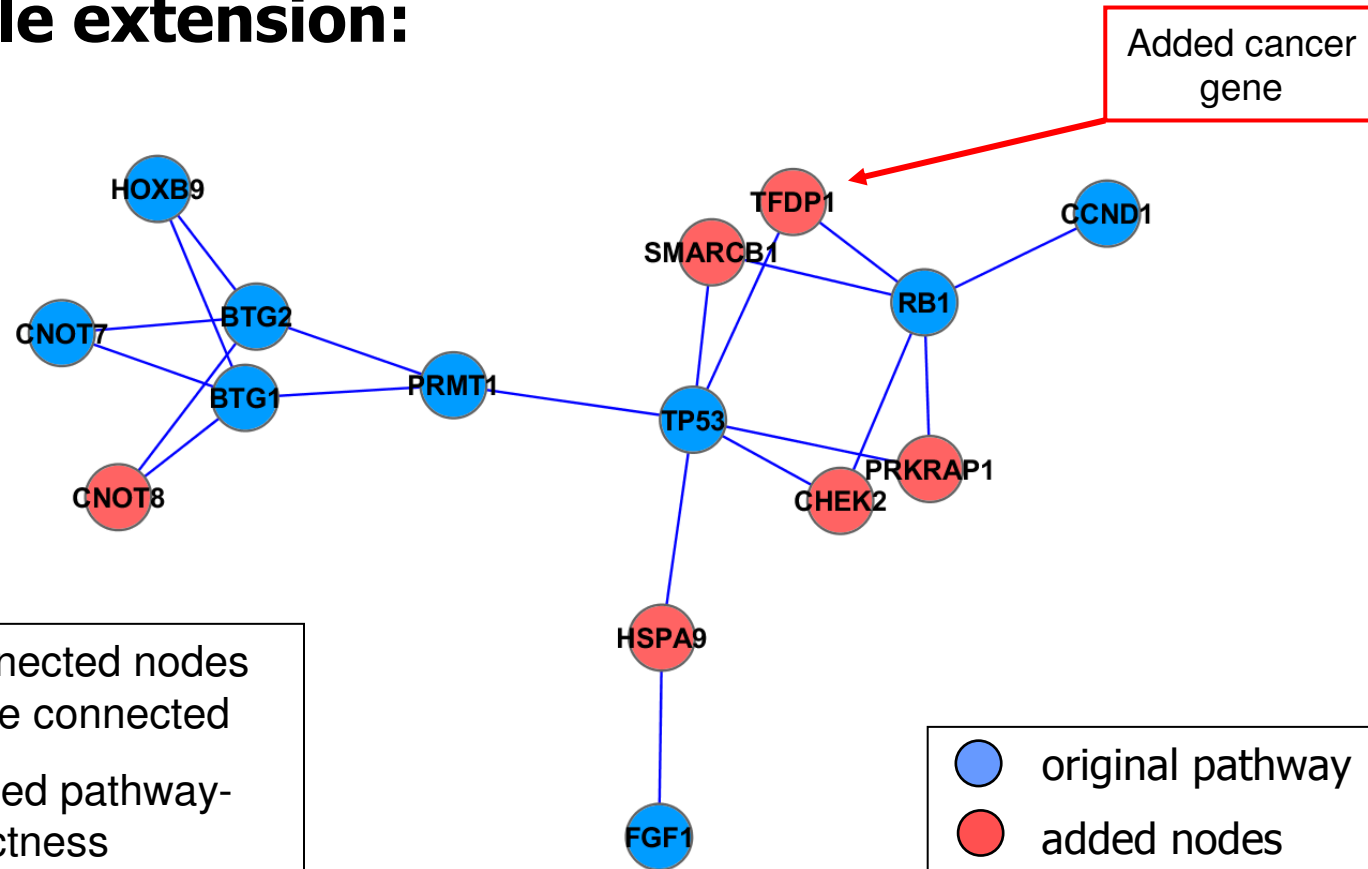
black = pathway-nodes

red blue green = nodes added based on different criteria

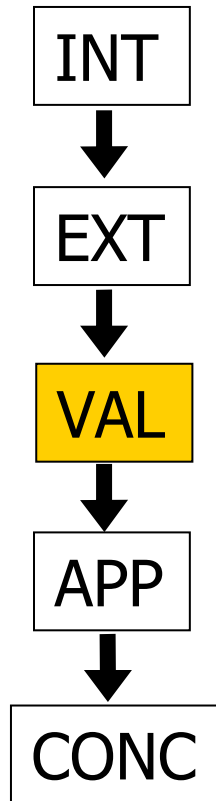




## Example extension:



Pathway: BioCarta “*BTG family proteins and cell cycle regulation*”



Validation:  
How to validate pathway extensions  
without a real “gold standard”?



## Cross-validation

Can randomly deleted genes in the original pathways be recovered by the expansion procedure?

→ 3-step cross-validation procedure:

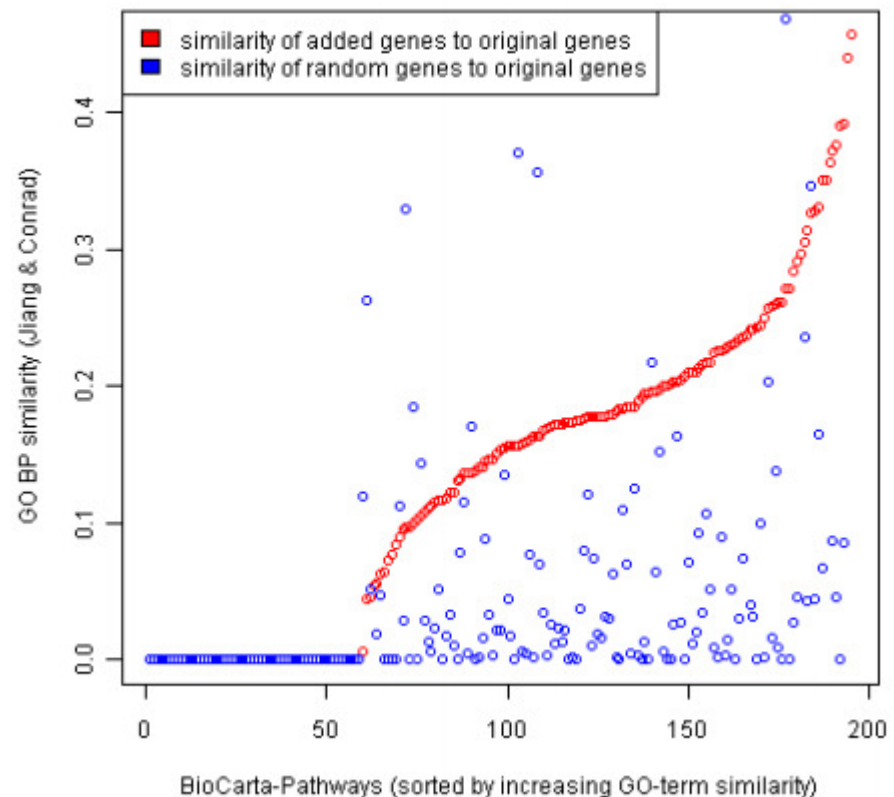
1. Randomly remove 10% of the pathway members (among proteins with at least one partner in the pathway)
2. Apply the proposed extension procedure as well as 100 random extensions (random sampling among candidates)
3. Estimate p-value-like significance scores:

$$\sum_{i \in P} \left( \frac{\sum_{i=1}^{100} I(\text{recovery\_random}_i \geq \text{recovery\_proposed})}{100} \right) / |P|$$

## Semantic similarity analysis (Gene Ontology)

- Quantify pairwise similarities between protein annotations using Jiang & Conrath's semantic similarity measure for GO-terms
- Compute avg. GO-term similarity between pathway-proteins and added proteins  
→ compare to random extension model

BioCarta - GO-term BP similarity between original pathway genes and added genes (connectivity-based and random)





## Extension statistics across all databases

Property	BioCarta	KEGG	Reactome
no. of used pathways	195	140	62
avg. pathway size	19	49	75
avg. size after expansion	24	61	85
total no. of added proteins	935	1745	622
no. of unique added proteins	280	623	409

Statistics on added proteins across 3 pathway databases:

- pathways increase to 113% - 126% of original size
- many proteins added to multiple pathways

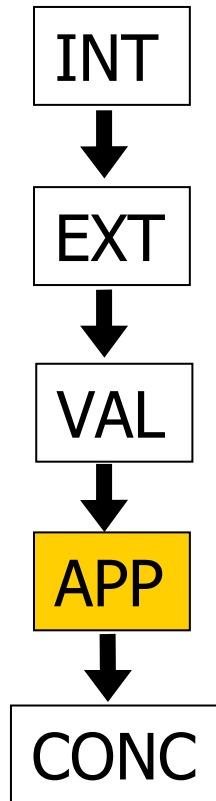


## Topological properties of added proteins

	<u>Protein set</u>	<u>Random set (mean)</u>	<u>Network (mean)</u>
Shortest path length	3.68	4.11 (0.03)	4.12 (0.94)
Node betweenness	21998	14545 (4751)	14669 (68893)
Degree	10.3	8.11 (0.94)	8.27 (16.2)
Clustering coefficient	0.34	0.11 (0.01)	0.11 (0.21)
Eigenvector centrality	0.04	0.01 (0.04)	0 (0.57)

Network topological properties for proteins added to BioCarta pathways

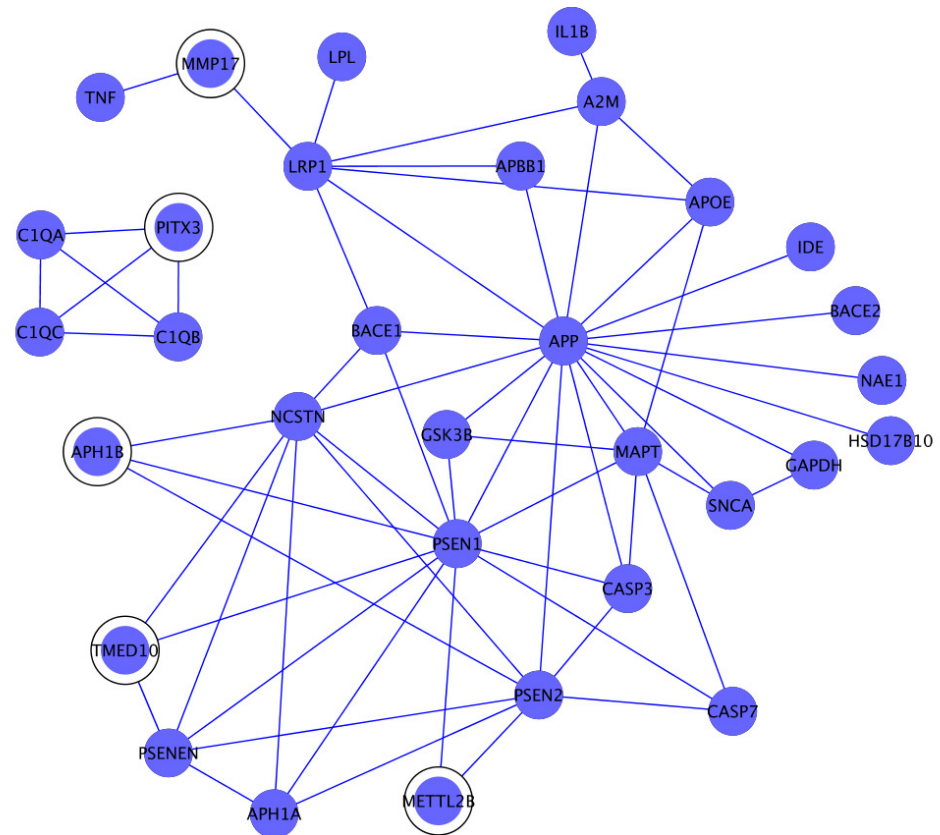
The **added proteins** are more central, more densely clustered and have shorter distances between them in comparison to **matched-size random proteins** and the **global network average**.



Biological application:  
Which insights do we gain when applying  
the approach to Alzheimer and cancer  
pathways?

## Application: Alzheimer disease pathway

- More than 20 proteins annotated in our PPI-network
- 5 proteins added by the extension process (circled)
- 3 known to be associated with the disease
- 2 novel candidates: METTL2B, TMED10

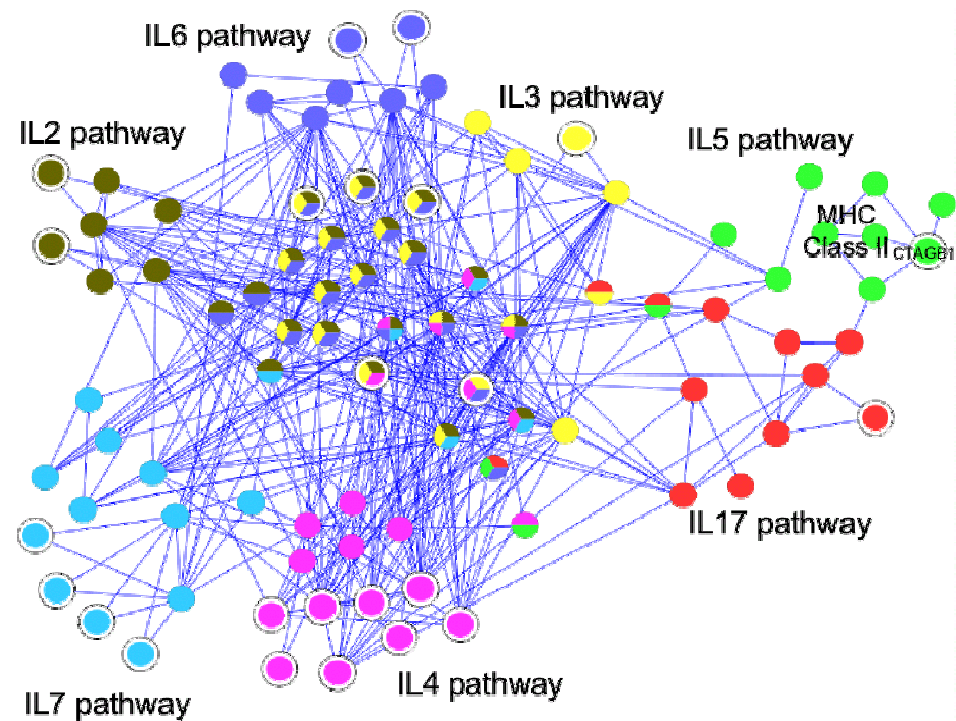


KEGG Alzheimer disease pathway  
mapped on human PPI-network



## Application: Interleukin signalling pathways

- Complex system of intracellular signalling cascades
- New putative pathway regulators identified
- New “crosstalk proteins” identified (associated with multiple pathways)



Two functions: pathway-regulation & pathway-communication?



## Using extended pathways for functional enrichment analysis

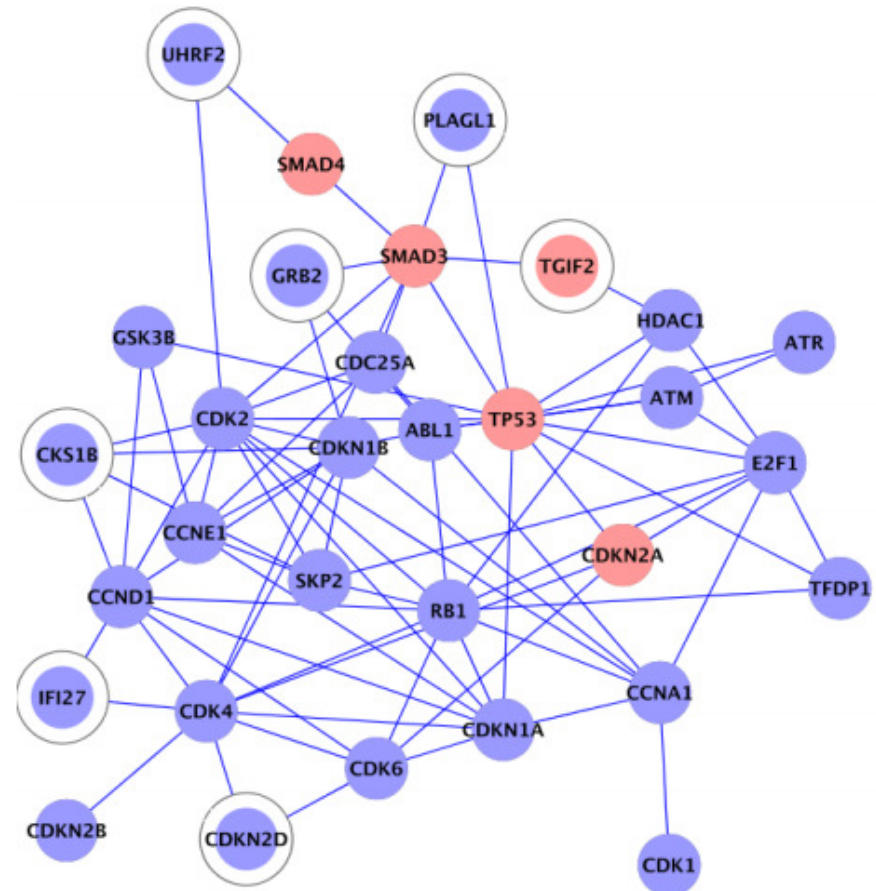
- classical approach:  
Test enrichment of experimentally derived gene sets in cellular pathway members (one-sided Fisher exact test)
- idea: replace original pathways by extended versions

Example: Enrichment analysis for pancreatic mutated genes

Cellular Process database	Cellular process	Pathway size	Number of pathway mutated genes	Number of mutated genes among added proteins	Mutated genes among added proteins
Biocarta	Agrin Postsynaptic Differentiation	38	5	2	PGM5, PLEKHG2
Kegg	Fc epsilon RI signaling pathway	112	10	5	DOCK2,MAPKBP1, DUSP19,ATF2,RASGRP3
Kegg	ErbB signaling pathway	190	13	7	VPS13A,MAPKBP1,NEK8, LIG3,DUSP19,AFF2,GLTSCR1

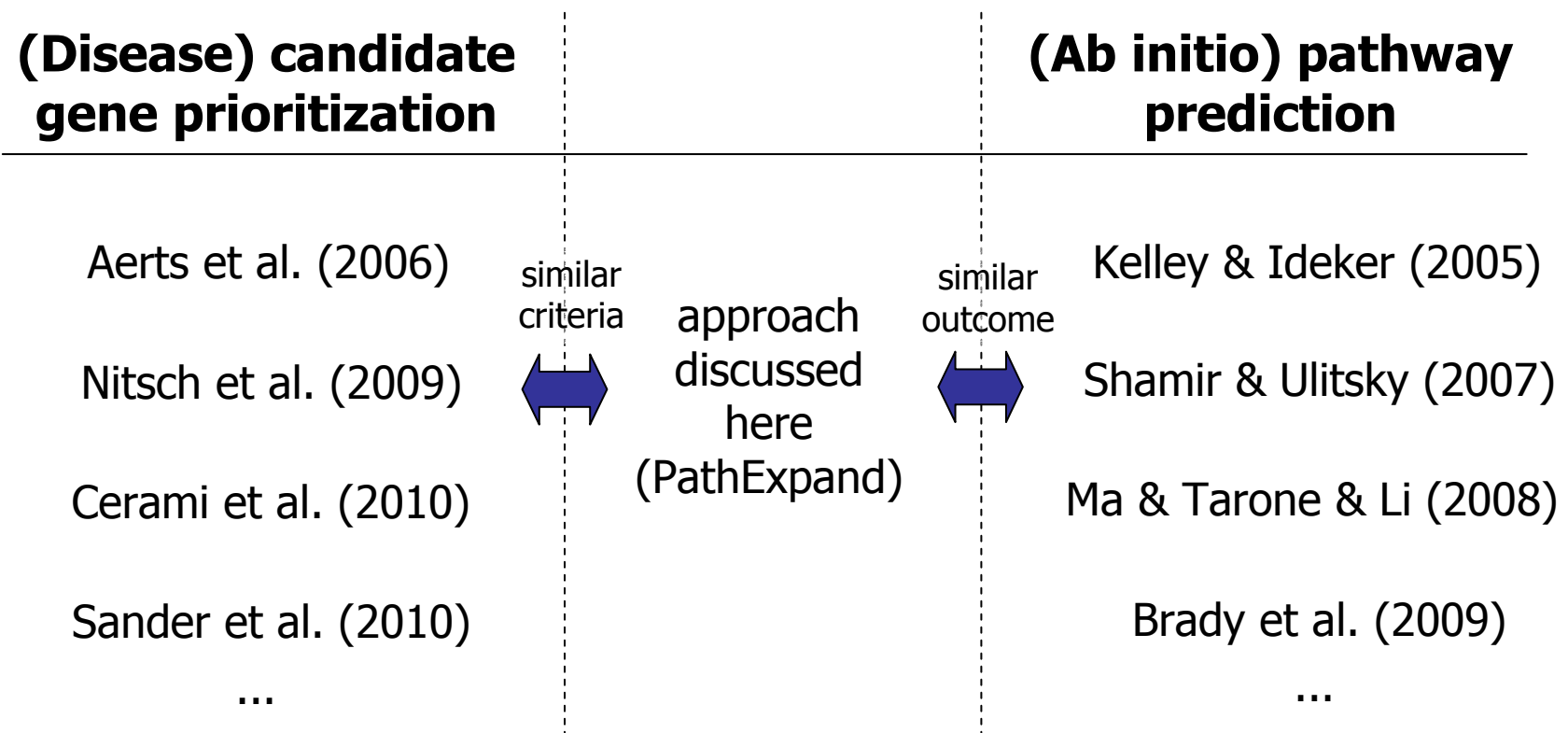
## Pancreatic mutated genes in expanded pathways

- “Cell cycle G1/S check point process” - extension procedure adds 7 proteins
- 6 of the added proteins are involved in cell cycle regulation
- the 7<sup>th</sup> (TGIF2) is known to be mutated in pancreatic cancer
- points to functional role of added proteins





## Relation to other network analysis methods





## Conclusion & Summary

- The method integrates two sources of information, extending **canonical pathways** using large-scale **protein interaction data**
- Three **validation** methods: cross-validation, GO-term semantic similarity and enrichment analysis
- Extended pathways have advantages in terms of network-compactness and can provide new insights on **pathway regulators**, the **cross-talk** between pathways and gene set **functional enrichment**

## References

1. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *Extending pathways and processes using molecular interaction networks to analyse cancer genome data*, BMC Bioinformatics, 11(1), 597, 2010
2. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *TopoGSA: network topological gene set analysis*, Bioinformatics, 26(9):1271-1272, 2010
3. E. Glaab, J. M. Garibaldi and N. Krasnogor. *ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization*, BMC Bioinformatics, 10:358, 2009
4. E. Glaab, J. M. Garibaldi, N. Krasnogor. *Learning pathway-based decision rules to classify microarray cancer samples*, German Conference on Bioinformatics 2010, Lecture Notes in Informatics (LNI), 173, 123-134
5. H. O. Habashy, D. G. Powe, E. Glaab, G. Ball, I. Spiteri, N. Krasnogor, J. M. Garibaldi, E. A. Rakha, A. R. Green, C. Caldas, I. O. Ellis. *REMG (Ras-like, oestrogen-regulated, growth-inhibitor) expression in breast cancer: a marker of ER-positive luminal-like subtype*, *Breast Cancer Research and Treatment*, (Epub ahead of print)
6. E. Glaab, J. M. Garibaldi and N. Krasnogor. *VRMLGen: An R-package for 3D Data Visualization on the Web*, Journal of Statistical Software, 36(8), 1-18, 2010