Pathway Knowledge Base:
A Public Repository for Searching Biological Pathways

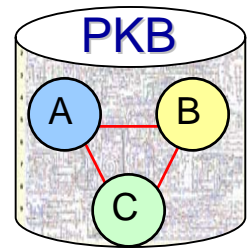***http://pkb.stanford.edu***

Nikesh Kotecha[1], Kyle Bruck[1], William Lu[1] and Nigam Shah[1]

*[1]Department of Biomedical Informatics, Stanford University*
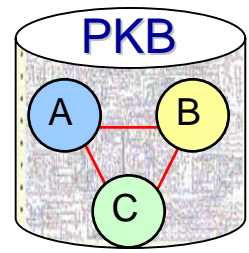
Email: pathwaykb@lists.stanford.edu

Outline

- The problem with pathways…

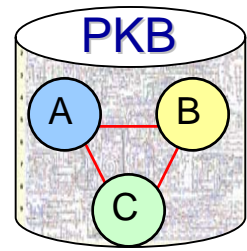- BioPAX and data exchange

- Pathway Knowledge Base
  - Querying over multiple data sources
  - Data import and query process

- Evaluation

- Conclusion & Next Steps
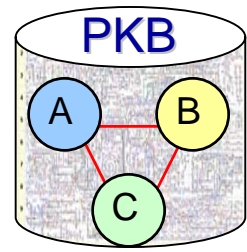
Pathways convey biological phenomena

Signalling pathway

STIMULUS

Genetic network

Metabolic pathway

Sundaraaj et. al, 2005
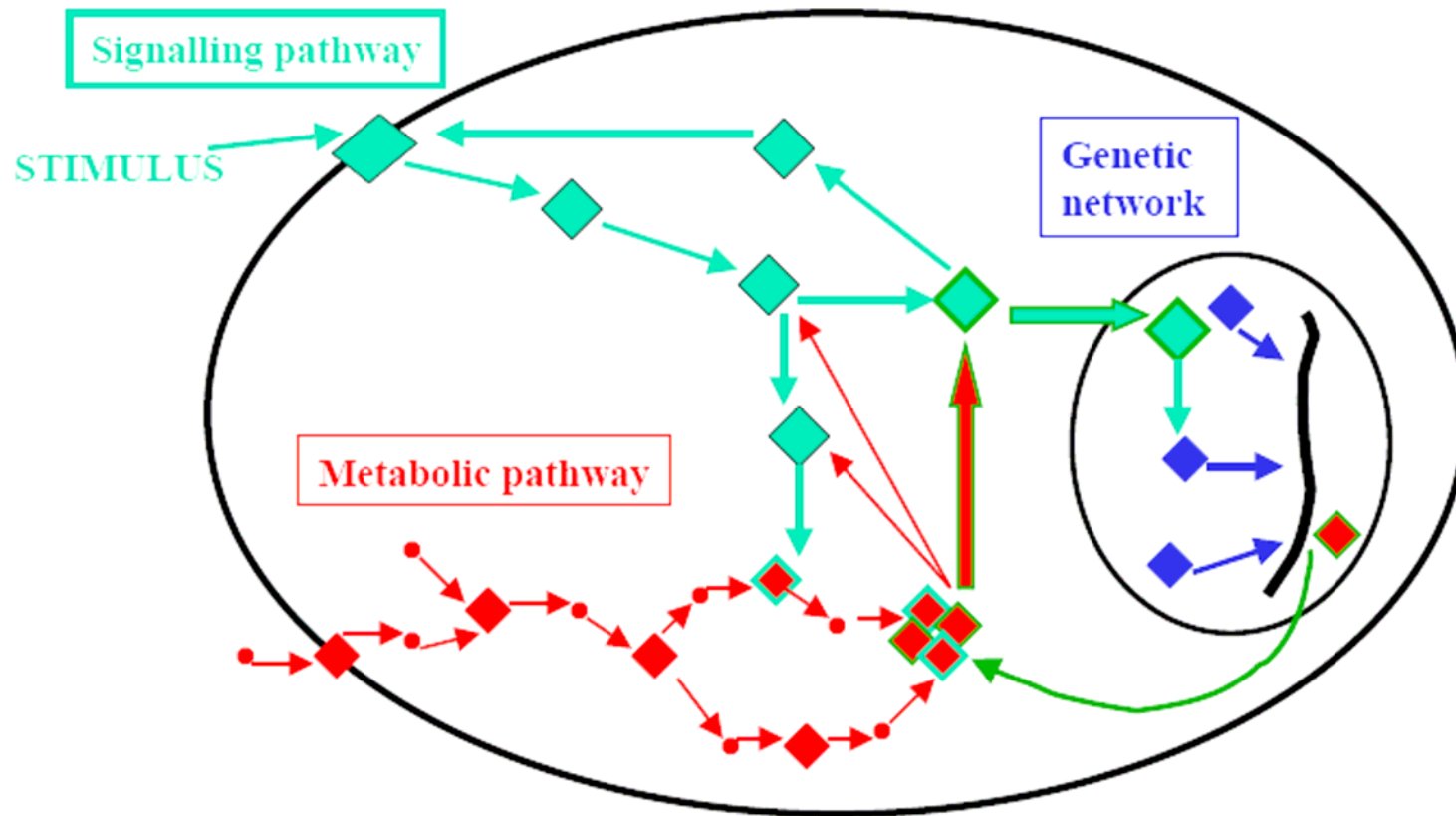
Users interpret non-standard pathway diagrams
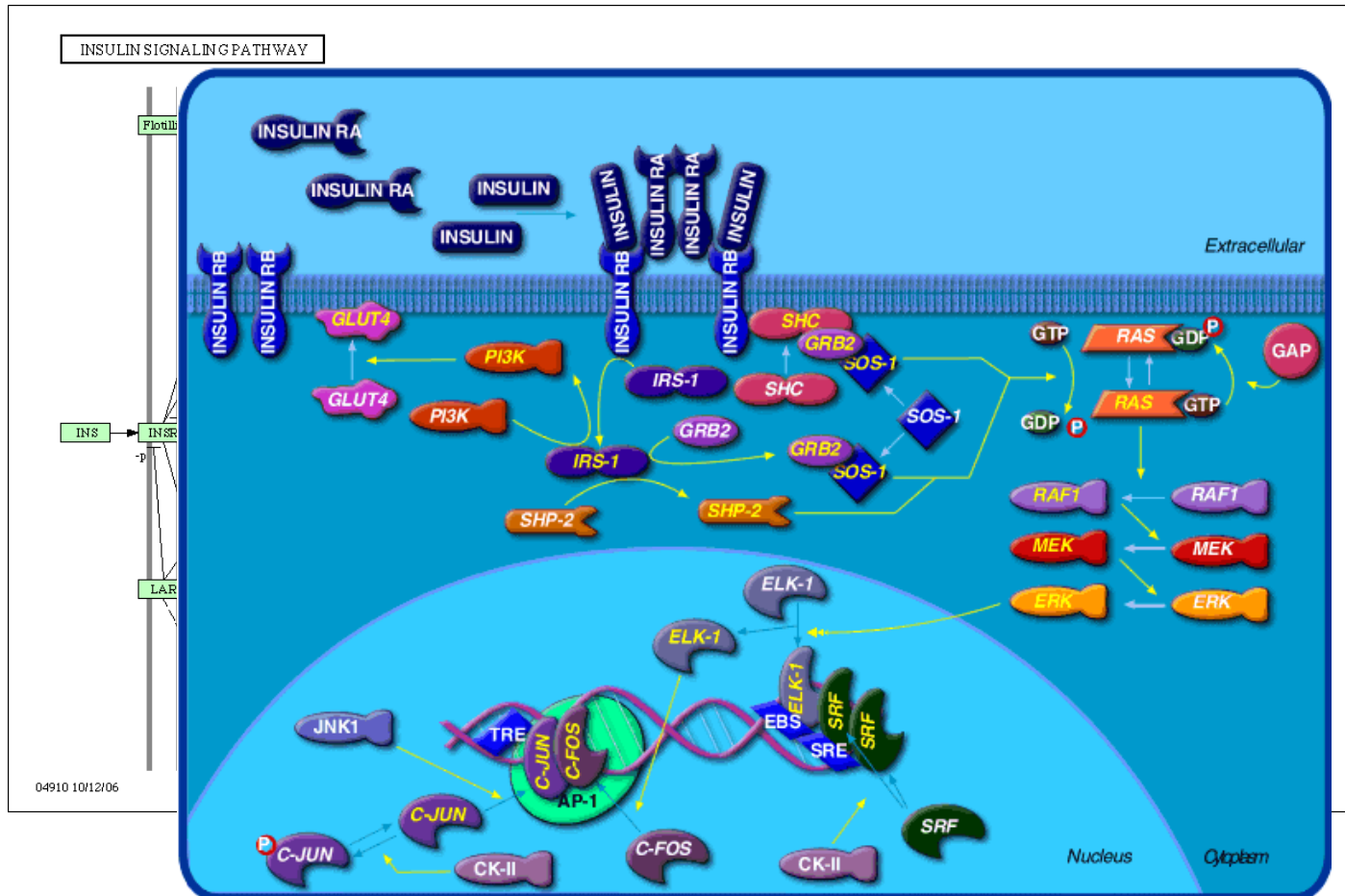
Insulin signaling pathways – KEGG & BioCarta
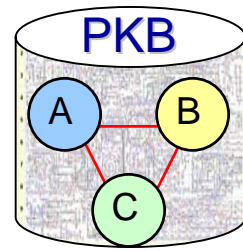
The problem with pathways…

- **Too many sources to investigate**
  - >200 databases in the Pathway Resource List
    - http://www.pathguide.org

- **Pathway diagrams are the primary way for consuming information**
  - Non-standard
  - Incorporates a handful of genes and proteins
  - Difficult to compute over

- **"Pathway" representation is not easy to define**
  - Appropriate levels of abstraction
  - Standard naming systems for molecules
  - Appropriate numbers of components

BioPAX is an emerging standard for pathway representation

PKB

A B

C

Database Exchange Formats

Simulation Model Exchange Formats

BioPAX

PSI

Small Molecules (CML)

SBML, CellML

Molecular Interactions
Pro:Pro ⟷ All:All

Genetic Interactions

Biochemical Reactions

Rate Formulas

Metabolic Pathways
Low Detail ⟷ High Detail

Interaction Networks
Molecular ⟷ Non-molecular
Pro:Pro ⟷ TF:Gene ⟷ Genetic

Regulatory Pathways
Low Detail ⟷ High Detail

http://www.biopax.org

BioPAX allows for creation of a central resource

PKB
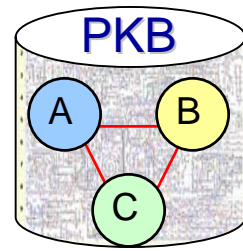
- ## BioPAX
  - ### Standardized representation of Pathway data in OWL
    - Provide method for exchange that promotes interoperability between known data sources.

| Development level	Scope of format	Sample data sources[*]
Level 1	Metabolic pathways	aMAZE, BioCyc, KEGG, PUMA2
Level 2	Level 1 plus molecular interactions	BIND, DIP, HPRD, IntAct, MINT
Level 3	Level 2 plus signaling pathways and gene regulation	CSNDB, INOH, PATIKA, Reactome, TRANSPATH
Level 4	Level 3 plus genetic interactions	FlyBase, MIPS
Future levels	Level 4 plus abstract associations	PubGene, GeneWays

[*]For a complete listing, see http://www.cbio.mskcc.org/prl.

**We use BioPAX level 2 to integrate Kegg, BioCyc and Reactome**

Pathway Knowledge Base (PKB)
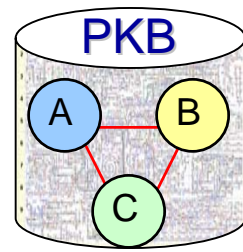
- ## Infrastructure to integrate pathway information from multiple sources into a central resource
  - BioPAX and Oracle RDF model

- ## Methods to query pathways in these databases
  - SQL/SPARQL

- ## A web interface that allows users and other programs to query and access these pathways
  - JSP/AJAX
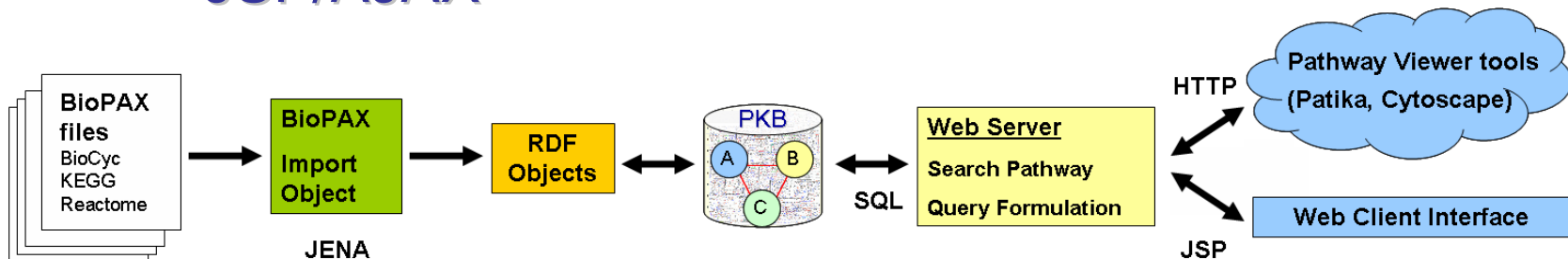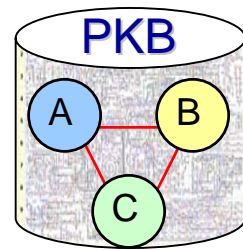
Search pathways across data sources and species

PKB

A B C

Pathway Knowledge Base
a public repository for searching biological pathways

(A) Search for phospholipase pathways: returns three pathways from two data sources and two species.

Name: lipases pathway
Species: human
Source: biocyc

Show Details>>

Name: Phospholipase-mediated signalling
Species: human
Source: reactome

Show Details>>

Name: lipases pathway
Species: yeast
Source: biocyc

Show Details>>

(B) Search for protein bcl1 which returns the human bcl1 oncogene.

Name: lipases pathway   Show results from   Show results in

Name: UniProt:P24385 G1/S-specific cyclin D1 (PRAD1 oncogene) (BCL-1 oncogene)
Short Name:
Species: human
Source: reactome
Comment: FUNCTION: Essential for the control of the cell cycle at the G1/S (start) transition.
SUBUNIT: Interacts with the CDK4 and CDK6 protein kinases to form a serine/threonine
kinase holoenzyme complex. The cyclin subunit imparts substrate specificity to the complex.
DISEASE: Involved in B-lymphocytic malignancy (particularly mantle-cell lymphoma (MCL)) by a
chromosomal translocation t(11;14)(q13;q32) that involves CCND1 and immunoglobulin gene
regions (BCL1 oncogene). Activation of CCND1 may be oncogenic by directly altering
progression through the cell cycle. DISEASE: Involved in a subset of parathyroid adenomas
by a chromosomal translocation t(11;11)(q13;p15) that involves CCND1 and the parathyroid
hormone (PTH) enhancer (PRAD1 oncogene). SIMILARITY: Belongs to the cyclin family. Cyclin
D subfamily. DATABASE: NAME=Atlas Genet. Cytogenet. Oncol. Haematol.;
WWW=\"http://www.infobiogen.fr/services/chromcancer/Genes/BCL1.html\".

Source: reactome
Comment: FUNCTION: Essential for the control of the cell cycle at the G1/S (start) transition.
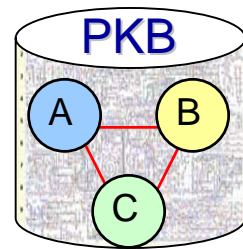
Visualize and navigate pathway results

Comment: FUNCTION: Essential for the control of the cell cycle at the G1/S (start) transition. SUBUNIT: Interacts with the CDK4 and CDK6 protein kinases to form a serine/threonine kinase holoenzyme complex. The cyclin subunit imparts substrate specificity to the complex. DISEASE: Involved in B-lymphocytic malignancy (particularly mantle-cell lymphoma (MCL)) by a chromosomal translocation t(11;14)(q13;q32) that involves CCND1 and immunoglobulin gene.

(C) Neighborhood search for pathways containing bcl1 returns five pathways

Pathways:

- reactome Ubiquitin dependent degradation of Cyclin D1
- reactome Cyclin D Cdk4 6 mediated phosphorylation of Rb and dissociation of Rb from the
- reactome Formation of Cyclin D Cdk4 6 complexes
- reactome Phosphorylation of Cyclin D Cdk4 6 complexes
- reactome Translocation of Cyclin D Cdk4 6 complexes from the cytoplasm to the nucleus

(D) Tree view of the reactions in the first pathway in the result

Ubiquitin-dependent degradation of Cyclin D1

Relocalization of nuclearly localized Cyclin D1 to the cytoplasm
Proteasome mediated degradation of Cyclin D1
Ubiquitination of Cyclin D1
Phosphorylation of Cyclin D1 at T286 by glycogen synthase kin
Relocalization of nuclearly localized phospho-(T286):cyclin D1:C
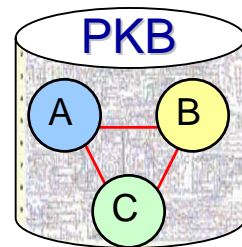
Custom queries enabled via advanced search

Copyright (c) 2006 by PKB.

http://pkb.stanford.edu

Under the Hood…

Architecture

PKB

Pathway Viewer tools

Patika, Cytoscape etc.

HTTP

Web Server

Search Pathway

BioPAX output

JSP

Web Client Interface

SQL

PKB

BioPAX files
Reactome
BioCyc

BioPAX

Import Object

JENA

RDF Objects

select distinct t,x,y FROM

(

Each species has its own RDF model

**Query across species via UNION ALL**

Determine data source via URI:

http://pkb.stanford.edu/biopax#biocyc_biochemicalReaction12408630 1HUMANRecruitment of elongation factors to form HIV-1 elongation complex

**http://pkb.stanford.edu/biopax#reactome_**Recruitment_of_elongation _factors_to_form_HIV_1_elongation_complex

sel...
:bi...
nu...
ase...
UN...
sel...
:bi...
nu...
ase...
UN...
sel...
:bi...
null,SDO_RDF_Alias(SDO_RDF_Alias('',http://www.biopax.org/rele
ase/biopax-level2.owl#')),null))
)ORDER BY x

Adding new species and data sources
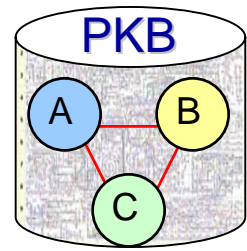
- **Add new species**
  - Create TABLE for storing rdf data
  - Create an RDF_MODEL for that species
  - Insert data as N-Triples

- **Add new data source**
  - Obtain BioPAX pathway data
  - Convert BioPAX OWL files into the triples format using JENA
  - Update all unique identifiers to have a custom URI* (http://pkb.stanford.edu/biopax#datasourcename)
  - Upgrade biopax identifiers from level 1 to level 2*
  - Insert data into table
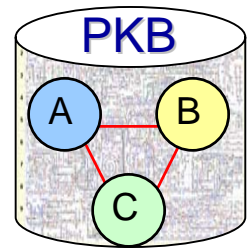
Evaluation
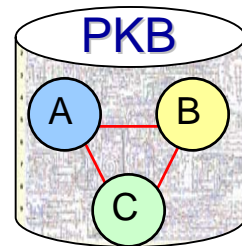
PKB Query Performance

PKB

A *B* *C*

Summary of data available in PKB

	Human	E. coli	Yeast	Total
# of pathways	1072	515	606	2193
# of reactions	3550	1966	1794	7310
# of triples	525148	367454	238877	1,131,479

Example query times

Query	Time (s)
Search for pathway titles containing phospholipase	1.3
Find all proteins with bcl1 in their name or synonym	16.1
Find all pathways with protein bcl1	35.3
List all entities to the left or right of hydroxyanthranilate	39.7

Queries run on a Dell PowerEdge 2800 w/ 4 GB RAM and Oracle 10.2.0.2
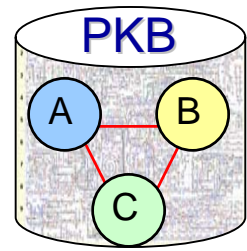
Certain queries are easier with relational structures

PKB

A   B

C
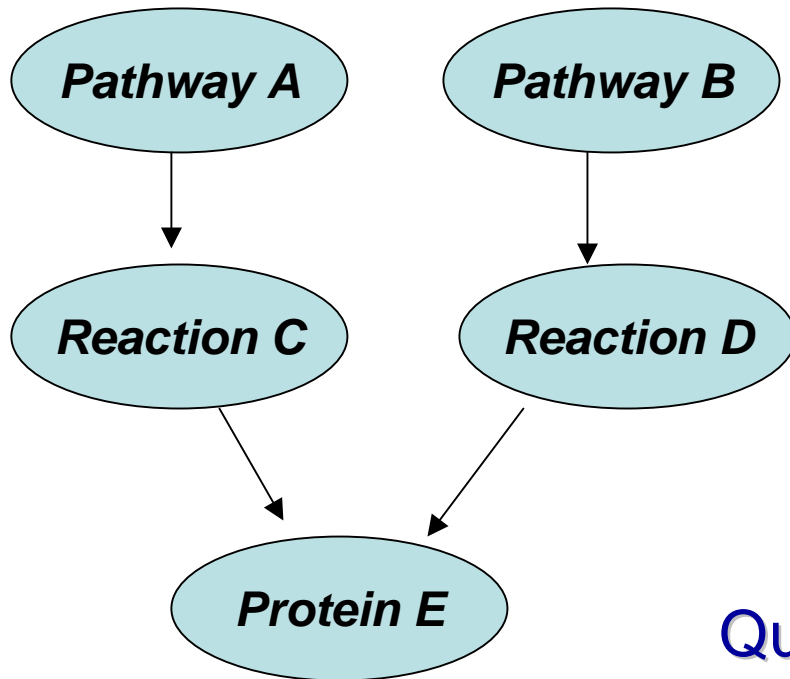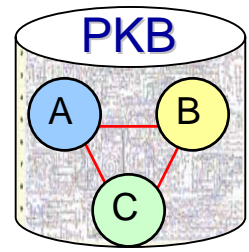
Pathway	Reaction	Protein
A	C	E
B	D	E

Relational Structure

Pathway A   **Pathway B**

Reaction C   **Reaction D**

Protein E

Directed Graph

Query:
Given protein E, find all pathways

Pathway Knowledge Base

- Benefits
  - Centralized resource for querying pathway information across data sources and species

  - Demonstrates use of BioPAX and RDF for exchange and integration of pathway data

- Limitations
  - Storing data *only* as RDF graph structures limits query performance

  - Import requires manipulation of namespaces

  - Querying options and methods are based on SPARQL but do not currently support all the features specified in SPARQL
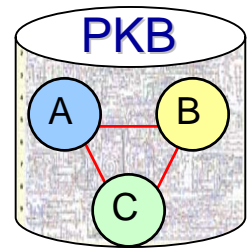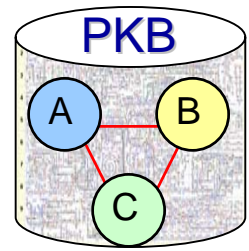
Next Steps in PKB?
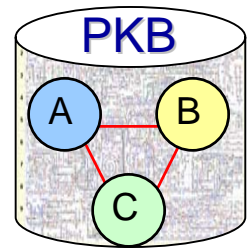
- Pathway differences between data sources
  - How does the galactose metabolism differ b/w ecocyc and reactome

- Pathway merging
  - Merge information from the same pathways from different data sources

- Pathway consistency checking
  - BioPAX rules for testing pathway models

Acknowledgements

Thanks. Any Questions?

http://pkb.stanford.edu

pathwaykb@lists.stanford.edu