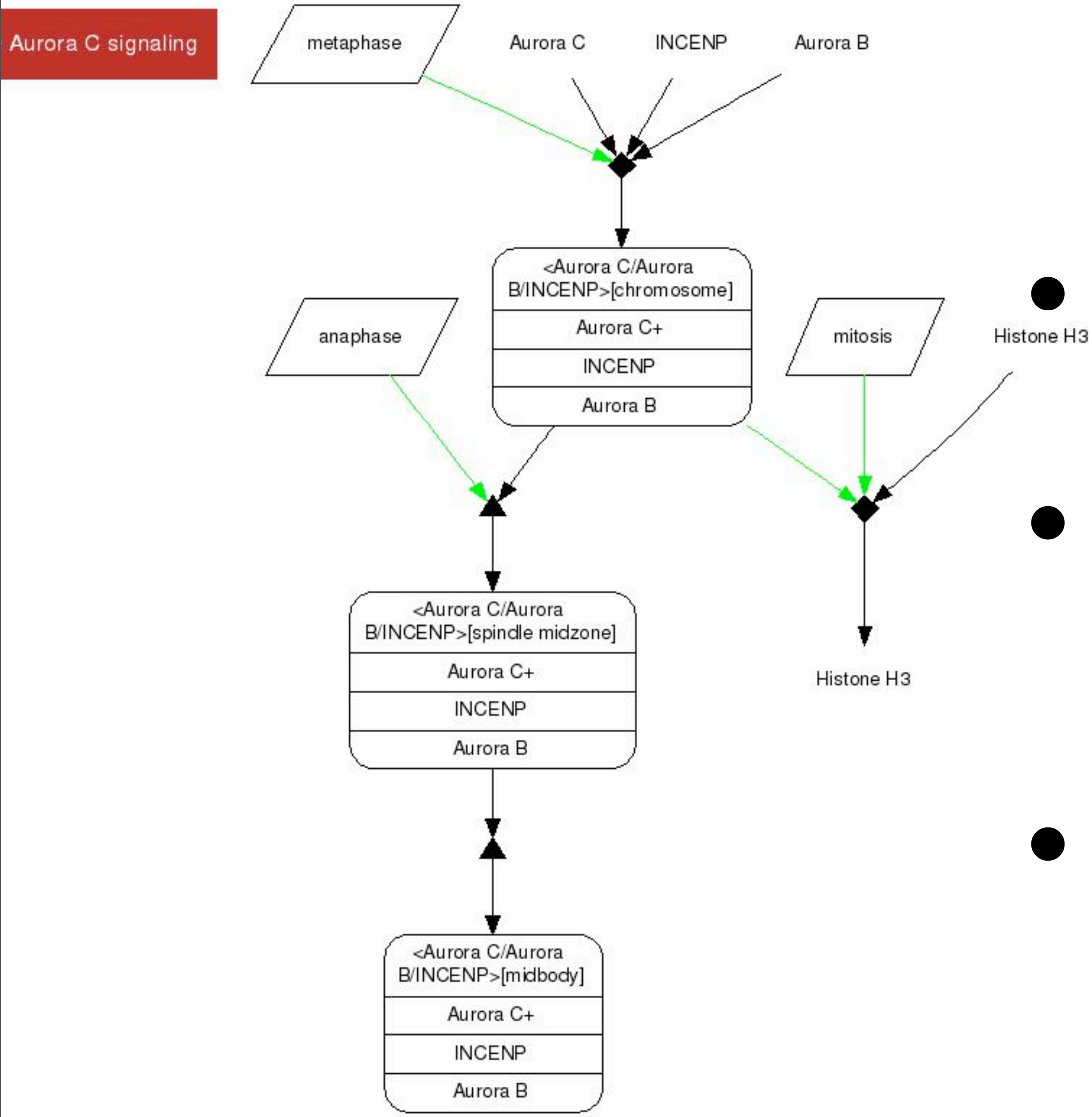
An attempt to use literature curated pathway DBs

Charles Vaske
Stuart Lab Meeting

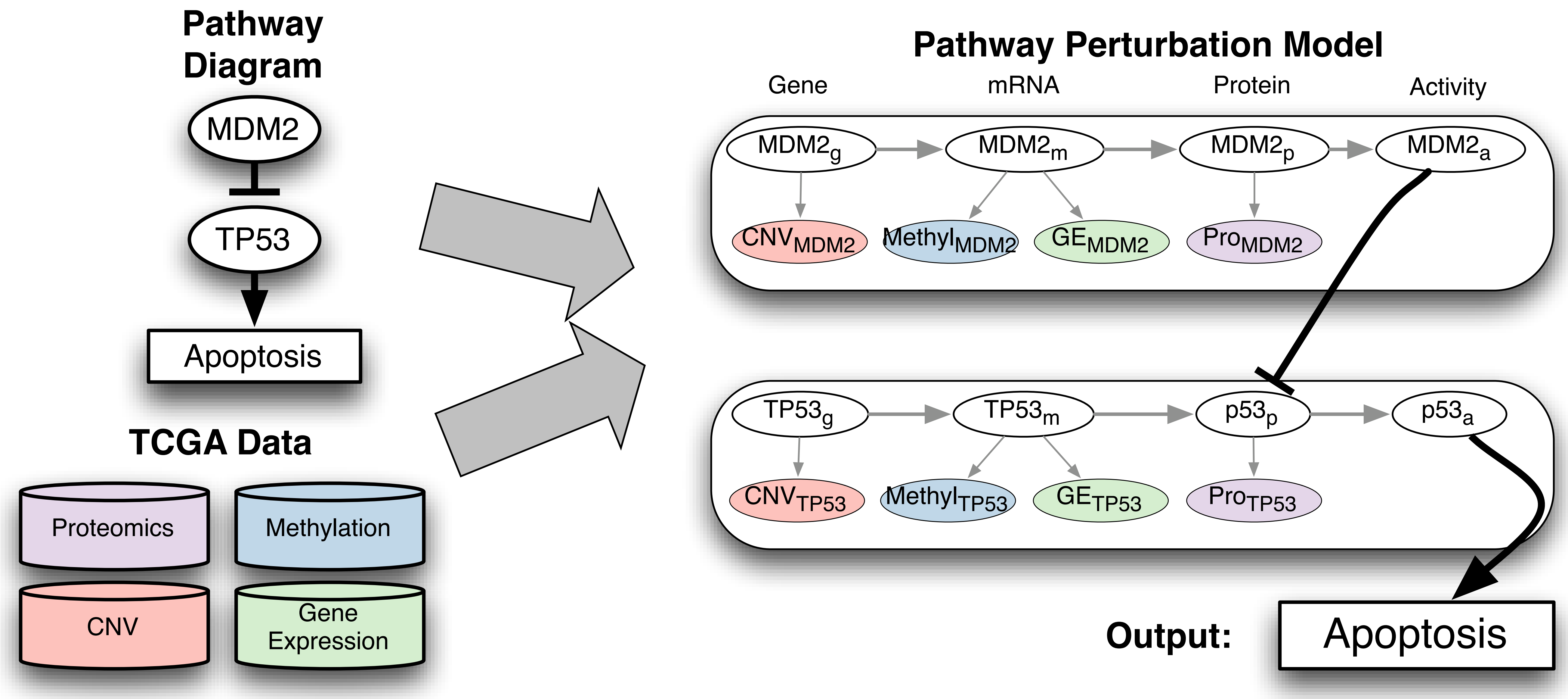
May 6th, 2009

Structured Pathways



- Lots of cancer research/ genes/data
- Subsequently, we know a lot about pathways active in cancer
- Can we use this structured knowledge?

Modeling clinical samples



Pathway Diagram

TCGA Data

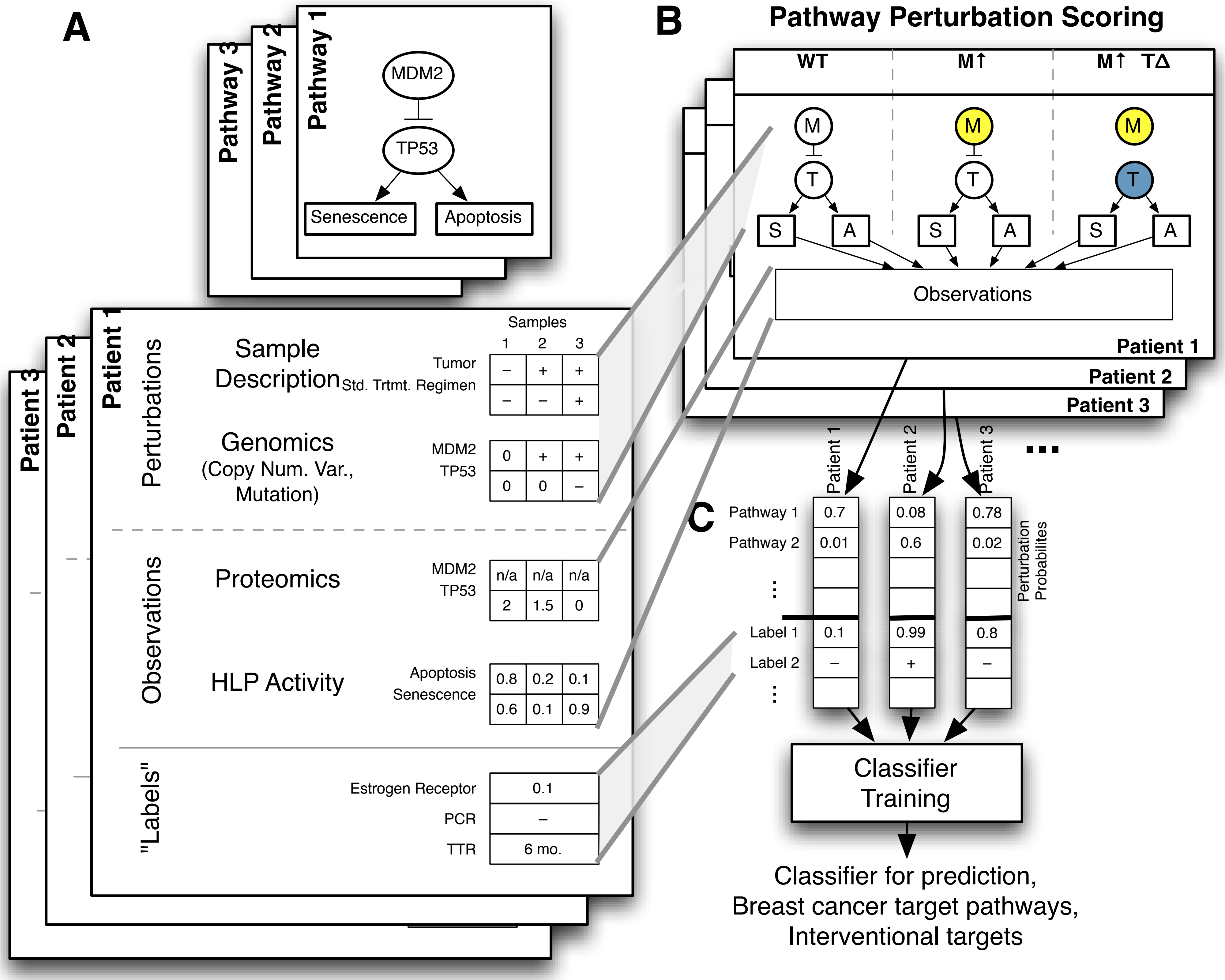
Proteomics
Methylation
CNV
Gene Expression

Pathway Perturbation Model

Gene
mRNA
Protein
Activity
MDM2_g
MDM2_m
MDM2_p
MDM2_a
CNV_{MDM2}
$\text{Methyl}_{\text{MDM2}}$
GE_{MDM2}
Pro_{MDM2}
TP53_g
TP53_m
p53_p
p53_a
CNV_{TP53}
$\text{Methyl}_{\text{TP53}}$
GE_{TP53}
Pro_{TP53}

Output: Apoptosis

Use in clinical samples



		Samples 1	2	3
Sample Description	Tumor	–	+	+
	Std. Trtmt. Regimen	–	–	+
Genomics (Copy Num. Var., Mutation)	MDM2	0	+	+
	TP53	0	0	–
Proteomics	MDM2	n/a	n/a	n/a
	TP53	2	1.5	0
HLP Activity	Apoptosis	0.8	0.2	0.1
	Senescence	0.6	0.1	0.9

Estrogen Receptor	0.1
PCR	–
TTR	6 mo.

	Patient 1	Patient 2	Patient 3
Pathway 1	0.7	0.08	0.78
Pathway 2	0.01	0.6	0.02
⋮			
Label 1	0.1	0.99	0.8
Label 2	–	+	–
⋮			

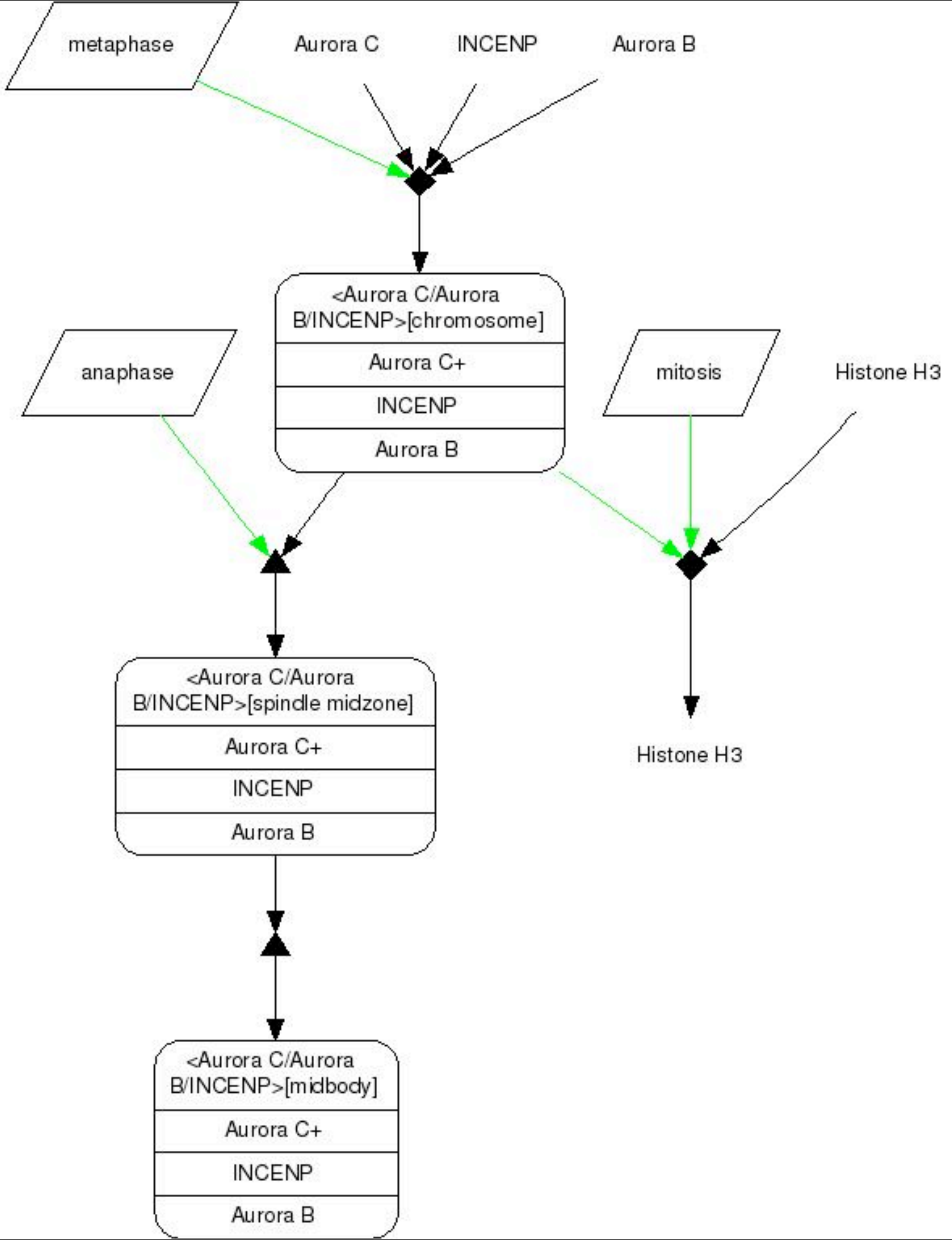
Classifier for prediction,
Breast cancer target pathways,
Interventional targets

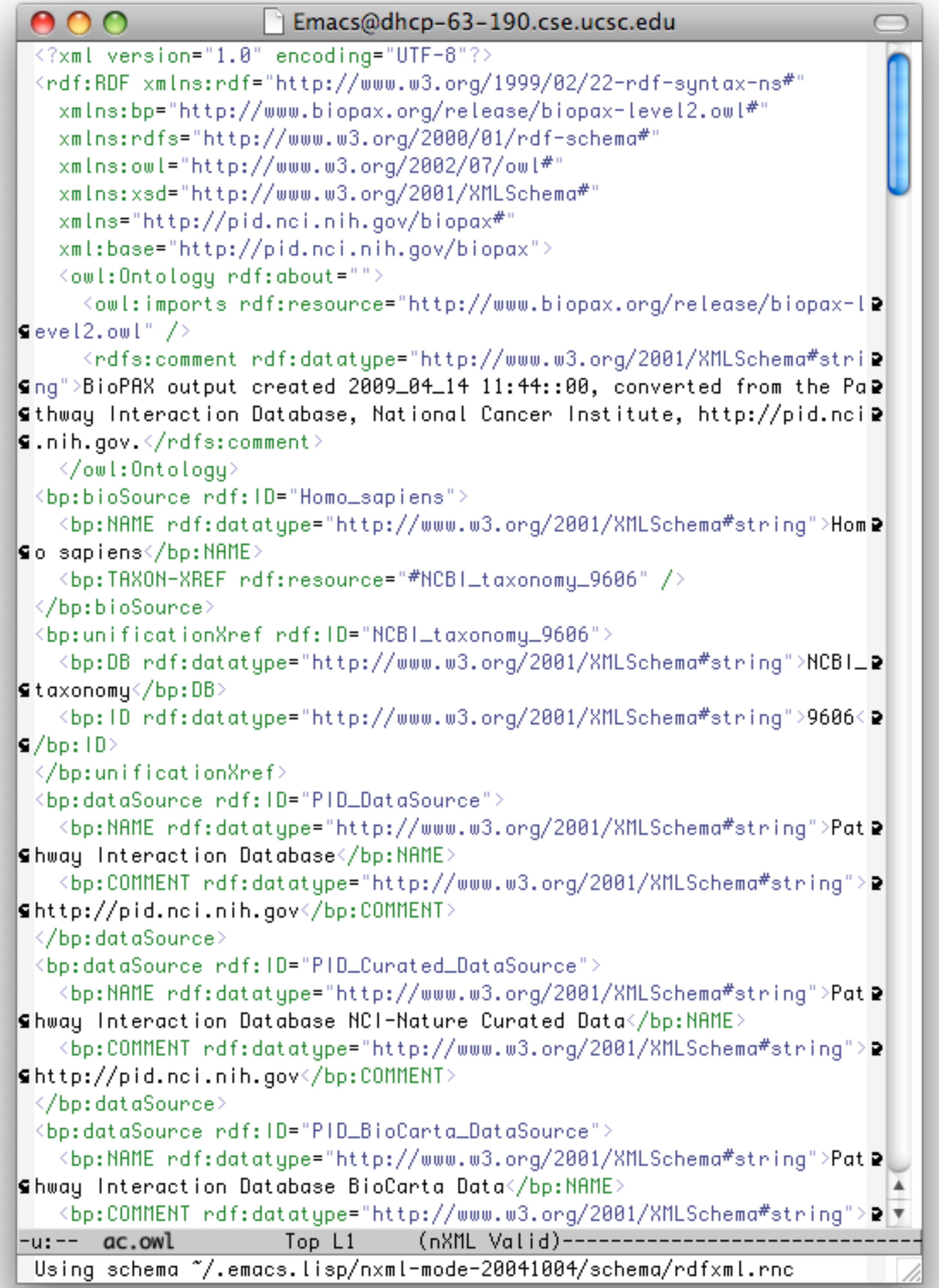
Outline

1. Get pathways (ugly, 50%-95% done)
2. Convert to graphical model
3. Add evidence from patient
4. Infer the value of hidden variables (i.e. Apoptosis, Chemotaxis)
5. Solve cancer (finally)

PID Aurora C signaling

- Proteins
- Complexes
- Abstract procceses
- Reactions/ modifications/ translocations
- Activation vs. participants





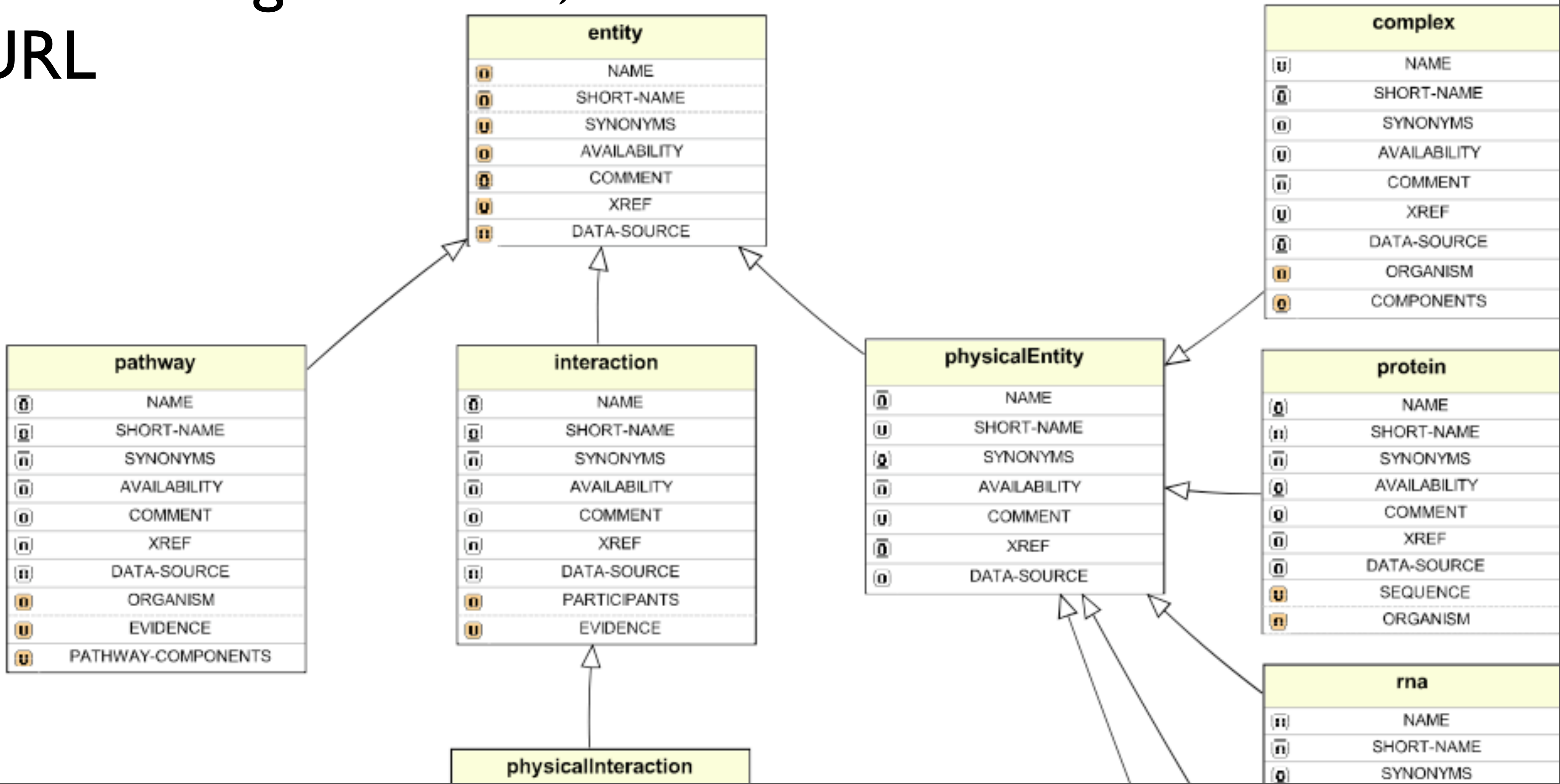
```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:bp="http://www.biopax.org/release/biopax-level2.owl#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns="http://pid.nci.nih.gov/biopax#"
  xml:base="http://pid.nci.nih.gov/biopax">
  <owl:Ontology rdf:about="">
    <owl:imports rdf:resource="http://www.biopax.org/release/biopax-l
evel2.owl" />
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#stri
ng">BioPAX output created 2009_04_14 11:44::00, converted from the Pa
thway Interaction Database, National Cancer Institute, http://pid.nci
.nih.gov.</rdfs:comment>
  </owl:Ontology>
<bp:bioSource rdf:ID="Homo_sapiens">
  <bp:NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Hom
o sapiens</bp:NAME>
  <bp:TAXON-XREF rdf:resource="#NCBI_taxonomy_9606" />
</bp:bioSource>
<bp:unificationXref rdf:ID="NCBI_taxonomy_9606">
  <bp:DB rdf:datatype="http://www.w3.org/2001/XMLSchema#string">NCBI_
taxonomy</bp:DB>
  <bp:ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">9606<
/bp:ID>
</bp:unificationXref>
<bp:dataSource rdf:ID="PID_DataSource">
  <bp:NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Pat
hway Interaction Database</bp:NAME>
  <bp:COMMENT rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
http://pid.nci.nih.gov</bp:COMMENT>
</bp:dataSource>
<bp:dataSource rdf:ID="PID_Curated_DataSource">
  <bp:NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Pat
hway Interaction Database NCI-Nature Curated Data</bp:NAME>
  <bp:COMMENT rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
http://pid.nci.nih.gov</bp:COMMENT>
</bp:dataSource>
<bp:dataSource rdf:ID="PID_BioCarta_DataSource">
  <bp:NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Pat
hway Interaction Database BioCarta Data</bp:NAME>
  <bp:COMMENT rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
-u:--  ac.owl          Top L1      (nXML Valid)--------------------------
 Using schema ~/.emacs.lisp/nxml-mode-20041004/schema/rdfxml.rnc
```

BioPAX

- Based on OWL
  Web Ontology Lang.
- Based on RDF
  Resource Desc. Format
- Not human-readable
- Must use tools!
- I love to complain about it

BioPAX

- Three levels (versions), people only use level 2 (I think)
- Defines "things" which have various properties, including a "class"
- Each "thing" is a URI, which looks like a URL


entity
NAME
SHORT-NAME
SYNONYMS
AVAILABILITY
COMMENT
XREF
DATA-SOURCE

complex
NAME
SHORT-NAME
SYNONYMS
AVAILABILITY
COMMENT
XREF
DATA-SOURCE
ORGANISM
COMPONENTS

pathway
NAME
SHORT-NAME
SYNONYMS
AVAILABILITY
COMMENT
XREF
DATA-SOURCE
ORGANISM
EVIDENCE
PATHWAY-COMPONENTS

interaction
NAME
SHORT-NAME
SYNONYMS
AVAILABILITY
COMMENT
XREF
DATA-SOURCE
PARTICIPANTS
EVIDENCE

physicalEntity
NAME
SHORT-NAME
SYNONYMS
AVAILABILITY
COMMENT
XREF
DATA-SOURCE

protein
NAME
SHORT-NAME
SYNONYMS
AVAILABILITY
COMMENT
XREF
DATA-SOURCE
SEQUENCE
ORGANISM

rna
NAME
SHORT-NAME
SYNONYMS

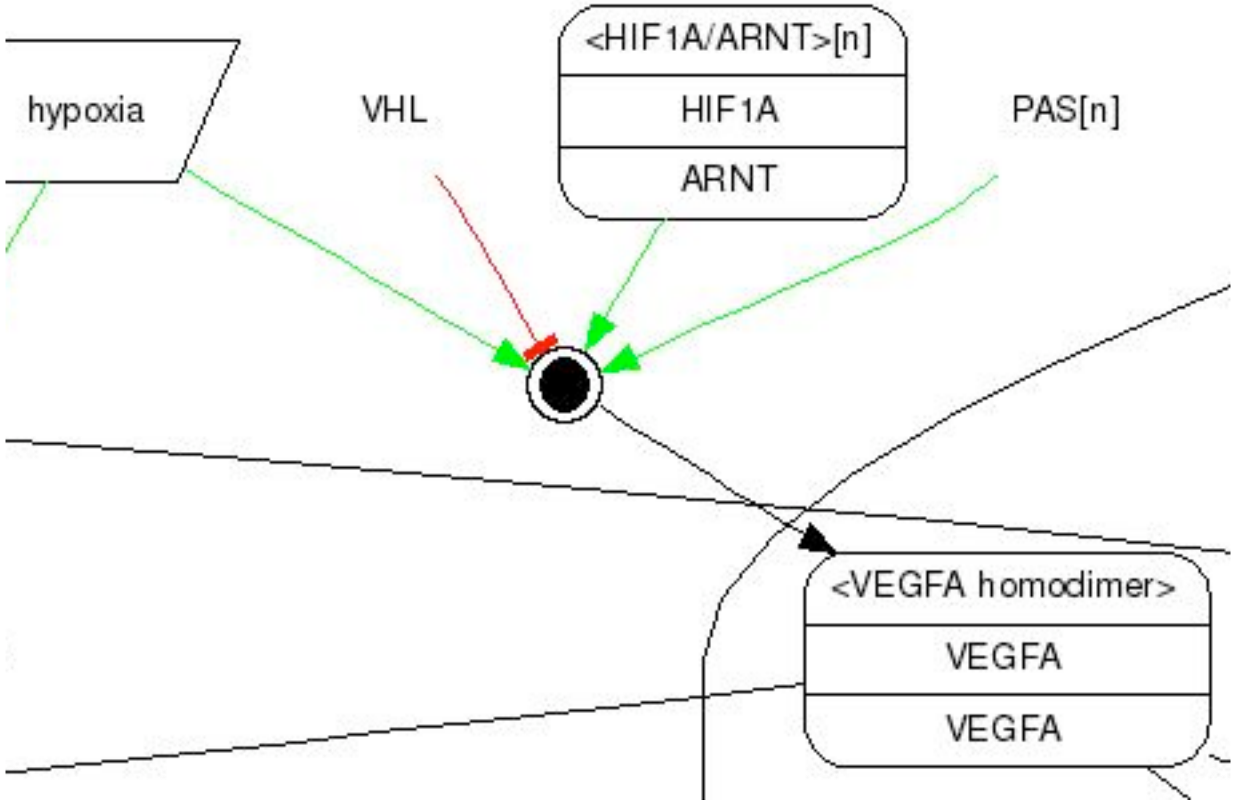
physicalInteraction

RDF/OWL/BioPAX Tools

- **Protege**: from Stanford, designed more for creating a BioPAX more than looking at “data” in that “format”
- **SPARQL/roqet**: sort of like SQL for RDF. Don’t use XML tools, you may miss things due to variations in serializations.

Caveat

- All this dense typing and formating is *extremely* expressive
- However, the amount of expression impedes programmatic understanding
- Test, test, test

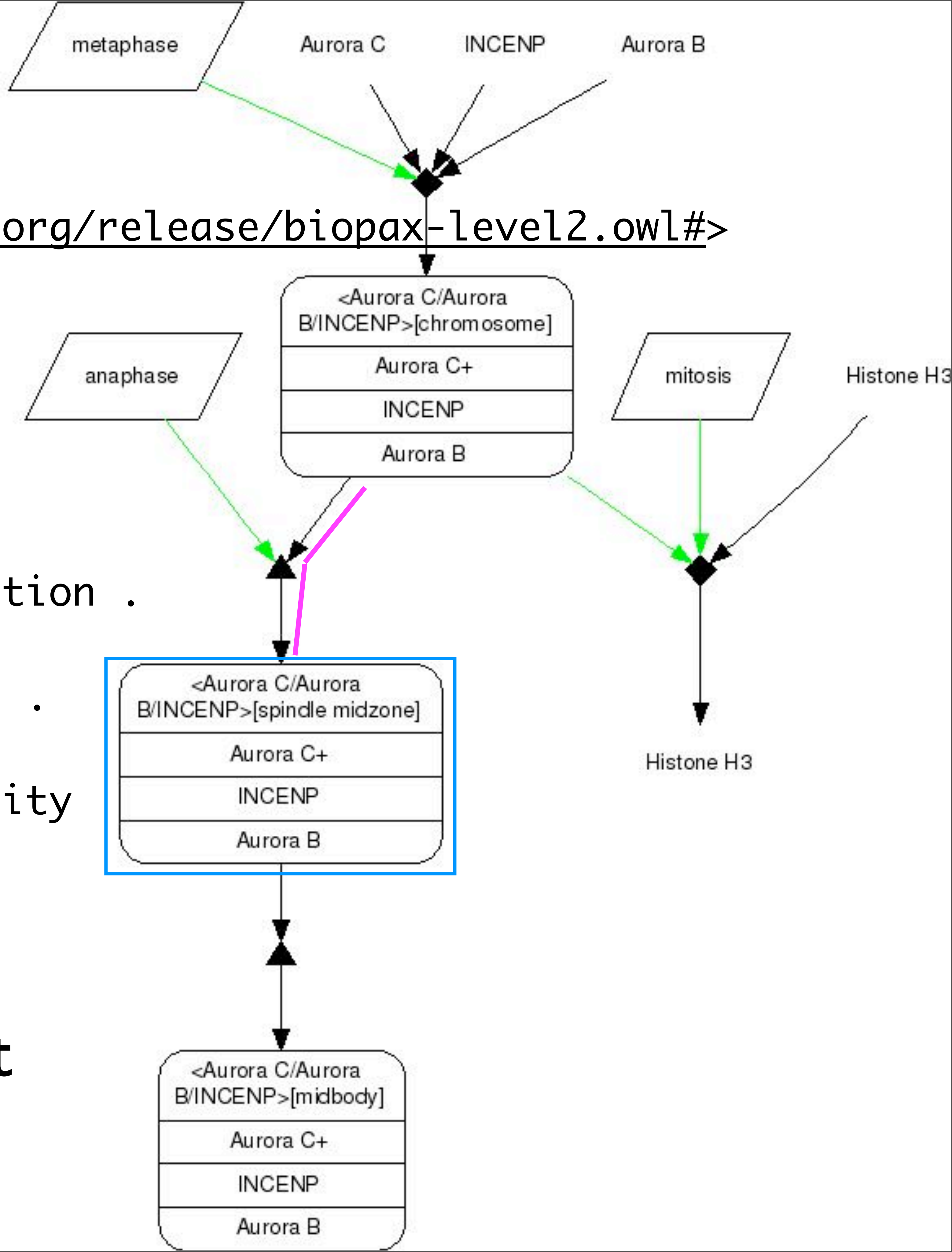


This shows the "transcription" of a complex. The meaning is obvious to a human, but befuddling to my naive scripts.

PID Aurora C signaling

```
PREFIX bp: <http://www.biopax.org/release/biopax-level2.owl#>

SELECT
  ?goname
  ?entity
  ?activation
WHERE {
  ?mod bp:CONTROL-TYPE ?activation .
  ?mod bp:NAME ?goname .
  ?mod bp:CONTROLLED ?reaction .
  ?reaction bp:RIGHT ?pep .
  ?pep bp:PHYSICAL-ENTITY ?entity
}
```



Example query: find abstract processes that are parents

Parsing

- Started by finding the proper queries to extract interactions, names, parts of complexes...

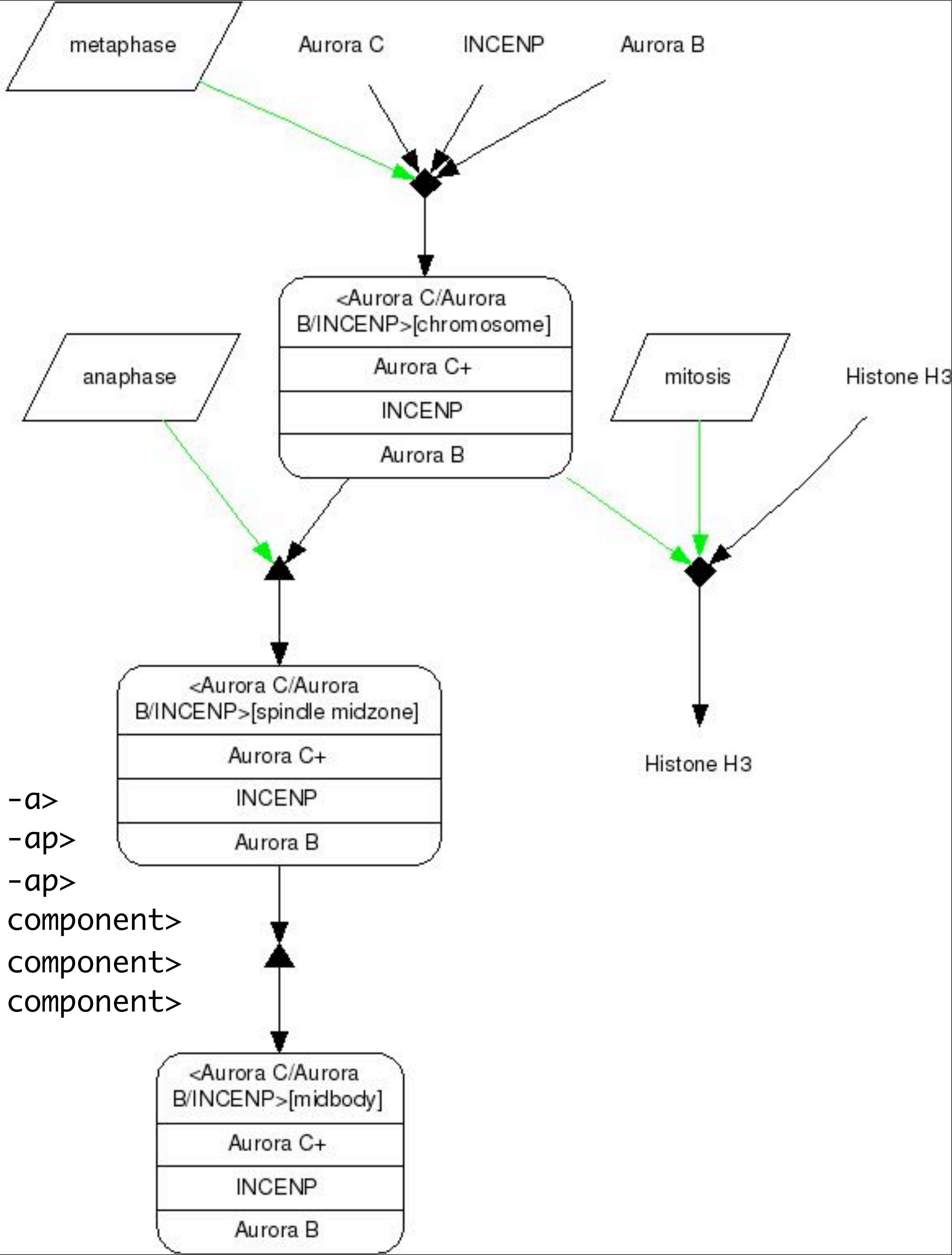
- Want a simple tab-delimited format:

```
abstract          metaphase
abstract          mitosis
complex           AurC/AurB/INCENP
protein           H3F3A
protein           AuroraB
protein           AuroraC
protein           INCENP
AurC/AurB/INCENP  H3F3A             -a>
mitosis           H3F3A             -ap>
metaphase         AurC/AurB/INCENP  -ap>
AuroraB           AurC/AurB/INCENP  component>
INCENP            AurC/AurB/INCENP  component>
AuroraC           AurC/AurB/INCENP  component>
```

Entity Definitions

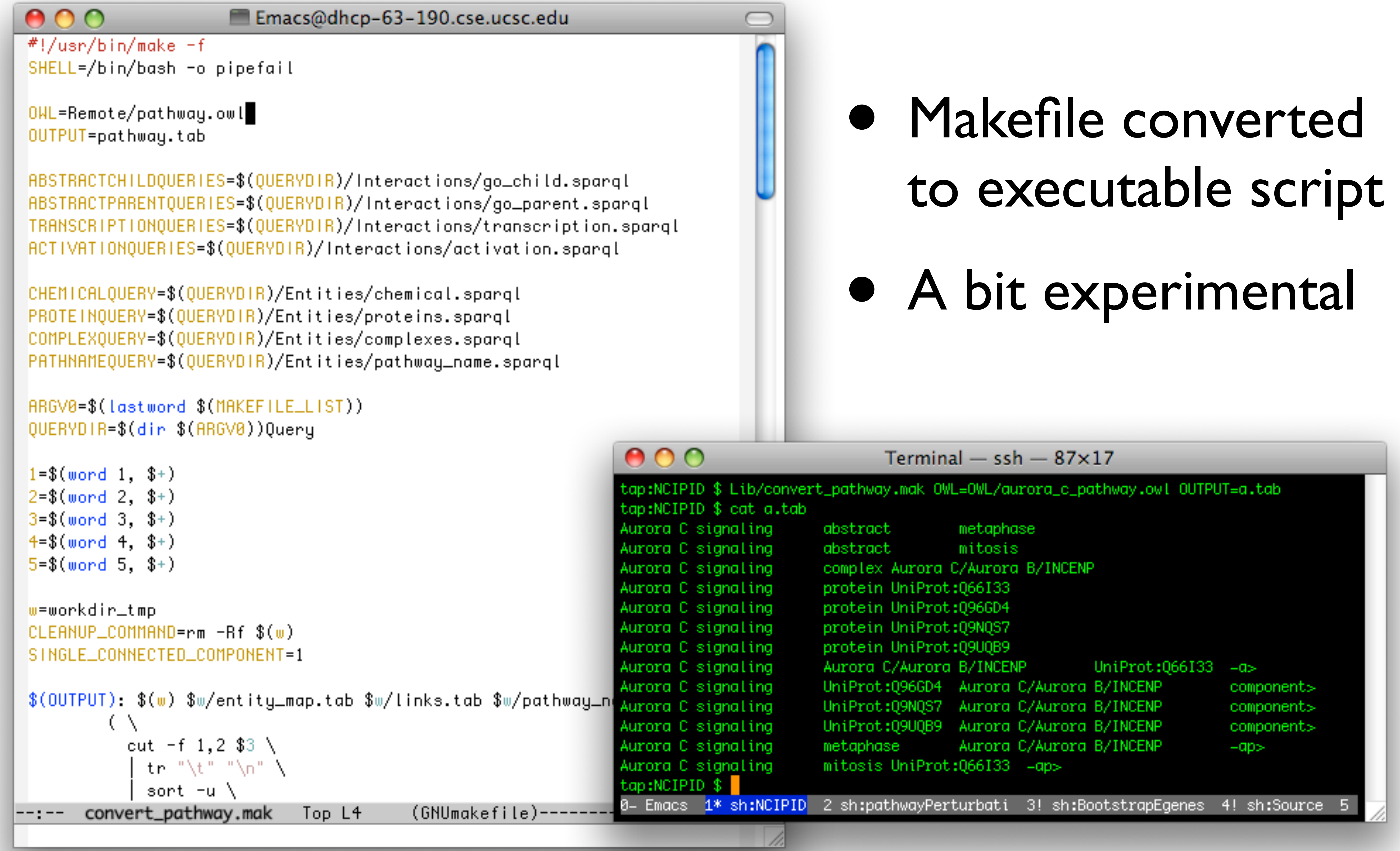
Entity Interactions

PID Aurora C signaling



abstract	metaphase	
abstract	mitosis	
complex	AurC/AurB/INCENP	
protein	H3F3A	
protein	AuroraB	
protein	AuroraC	
protein	INCENP	
AurC/AurB/INCENP	H3F3A	-a>
mitosis	H3F3A	-ap>
metaphase	AurC/AurB/INCENP	-ap>
AuroraB	AurC/AurB/INCENP	component>
INCENP	AurC/AurB/INCENP	component>
AuroraC	AurC/AurB/INCENP	component>

My hopeful monster



```
#!/usr/bin/make -f
SHELL=/bin/bash -o pipefail

OWL=Remote/pathway.owl
OUTPUT=pathway.tab

ABSTRACTCHILDQUERIES=$(QUERYDIR)/Interactions/go_child.sparql
ABSTRACTPARENTQUERIES=$(QUERYDIR)/Interactions/go_parent.sparql
TRANSCRIPTIONQUERIES=$(QUERYDIR)/Interactions/transcription.sparql
ACTIVATIONQUERIES=$(QUERYDIR)/Interactions/activation.sparql

CHEMICALQUERY=$(QUERYDIR)/Entities/chemical.sparql
PROTEINQUERY=$(QUERYDIR)/Entities/proteins.sparql
COMPLEXQUERY=$(QUERYDIR)/Entities/complexes.sparql
PATHNAMEQUERY=$(QUERYDIR)/Entities/pathway_name.sparql

ARGV0=$(lastword $(MAKEFILE_LIST))
QUERYDIR=$(dir $(ARGV0))Query

1=$(word 1, $+)
2=$(word 2, $+)
3=$(word 3, $+)
4=$(word 4, $+)
5=$(word 5, $+)

w=workdir_tmp
CLEANUP_COMMAND=rm -Rf $(w)
SINGLE_CONNECTED_COMPONENT=1

$(OUTPUT): $(w) $w/entity_map.tab $w/links.tab $w/pathway_n
        ( \
          cut -f 1,2 $3 \
          | tr "\t" "\n" \
          | sort -u \
--:--  convert_pathway.mak   Top L4    (GNUmakefile)--------
```

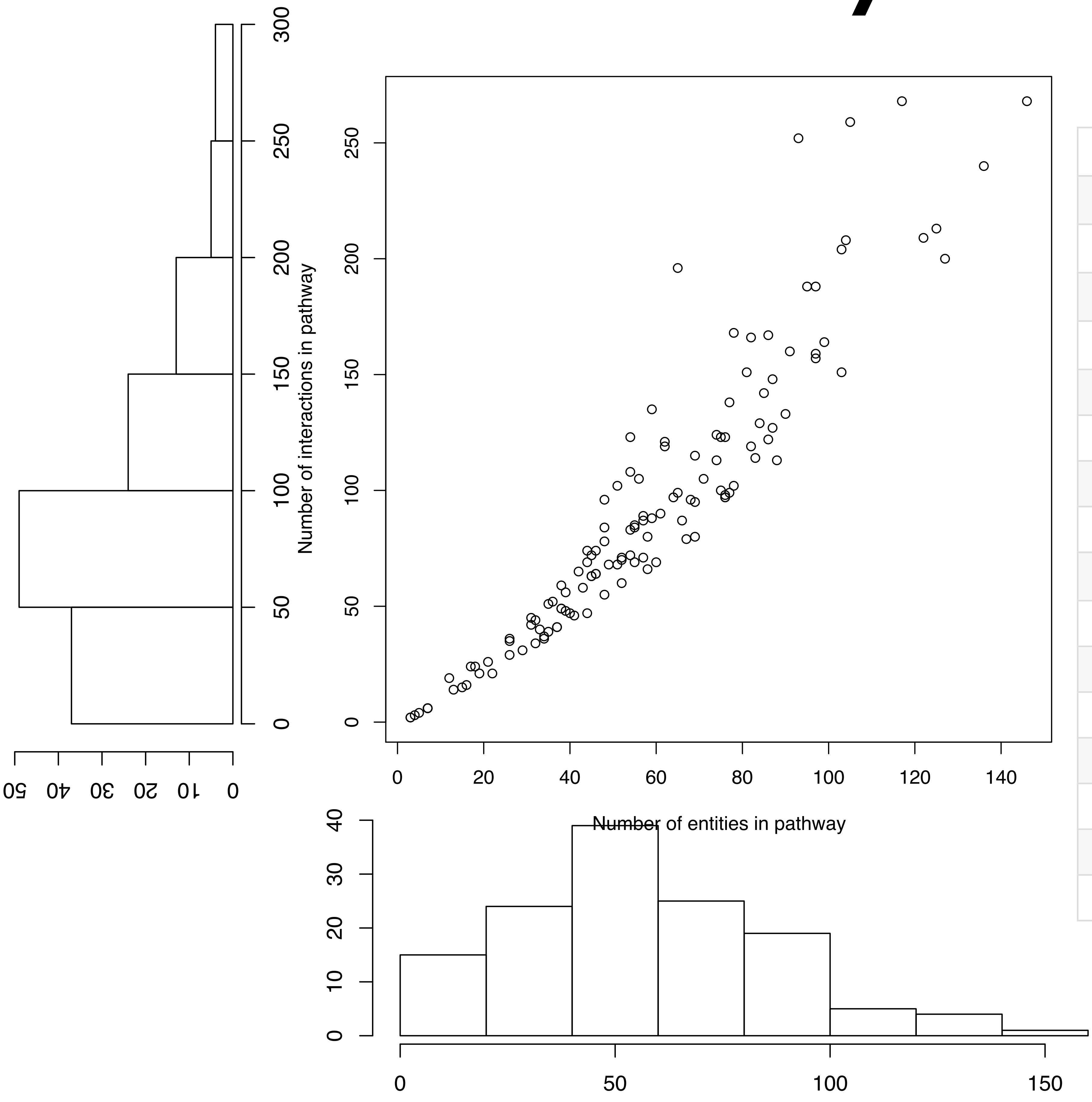
```
tap:NCIPID $ Lib/convert_pathway.mak OWL=OWL/aurora_c_pathway.owl OUTPUT=a.tab
tap:NCIPID $ cat a.tab
Aurora C signaling      abstract        metaphase
Aurora C signaling      abstract        mitosis
Aurora C signaling      complex Aurora C/Aurora B/INCENP
Aurora C signaling      protein UniProt:Q66I33
Aurora C signaling      protein UniProt:Q96GD4
Aurora C signaling      protein UniProt:Q9NQS7
Aurora C signaling      protein UniProt:Q9UQB9
Aurora C signaling      Aurora C/Aurora B/INCENP        UniProt:Q66I33  -a>
Aurora C signaling      UniProt:Q96GD4  Aurora C/Aurora B/INCENP        component>
Aurora C signaling      UniProt:Q9NQS7  Aurora C/Aurora B/INCENP        component>
Aurora C signaling      UniProt:Q9UQB9  Aurora C/Aurora B/INCENP        component>
Aurora C signaling      metaphase       Aurora C/Aurora B/INCENP        -ap>
Aurora C signaling      mitosis UniProt:Q66I33  -ap>
tap:NCIPID $ 
0- Emacs  1* sh:NCIPID  2 sh:pathwayPerturbati  3! sh:BootstrapEgenes  4! sh:Source  5
```

- Makefile converted to executable script
- A bit experimental

$MAPDIR/Data/Pathways

- Early, molten stage, but useful
- Human/NCIPID has NCI pathways
- Human/KEGG has early KEGG attempts

Pathway stats



132	**Pathways**
3766	**Unique Entities**
7569	**Entity instances**
10182	**Unique Interactions**
	Entity Breakdown
1742	protein
1638	complex
296	abstract
90	chemical
	Interaction Breakdown
2619	-a> (activation)
278	-a| (inhibition)
874	-ap> (abstract)
103	-ap|
528	-t> (transcription)
104	-t|
5676	component>

Outline

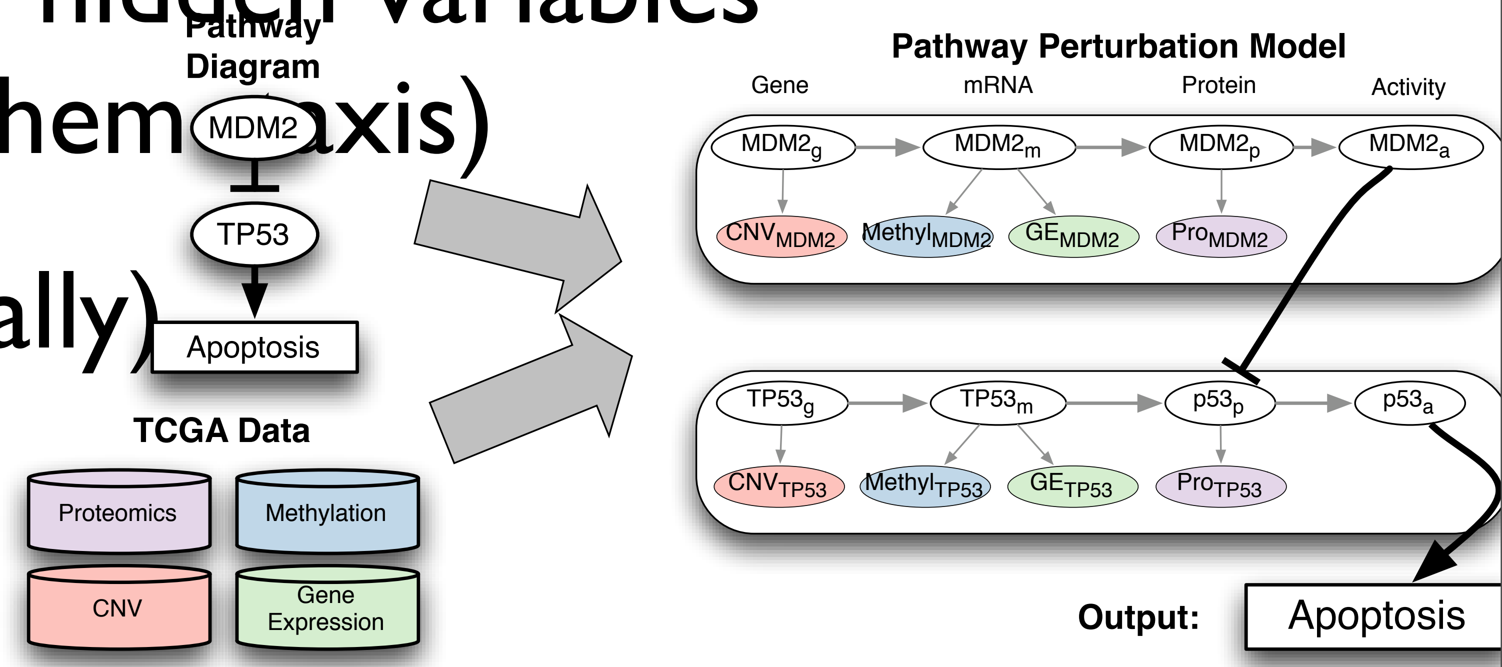
1. ~~Get pathways (ugly, 50%-95% done)~~

2. Convert to graphical model

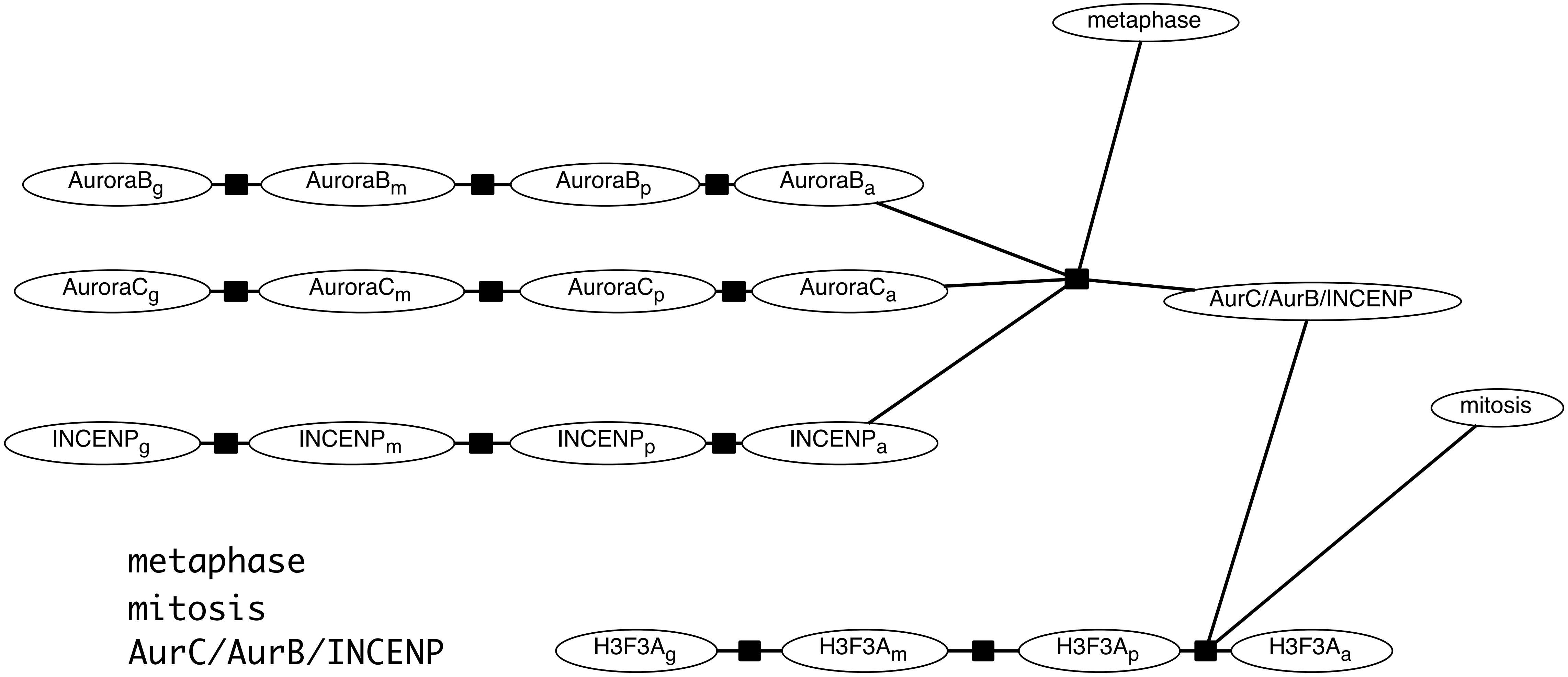
3. Add evidence from patient

4. Infer the value of hidden variables (i.e. Apoptosis, Chemotaxis)

5. Solve cancer (finally)

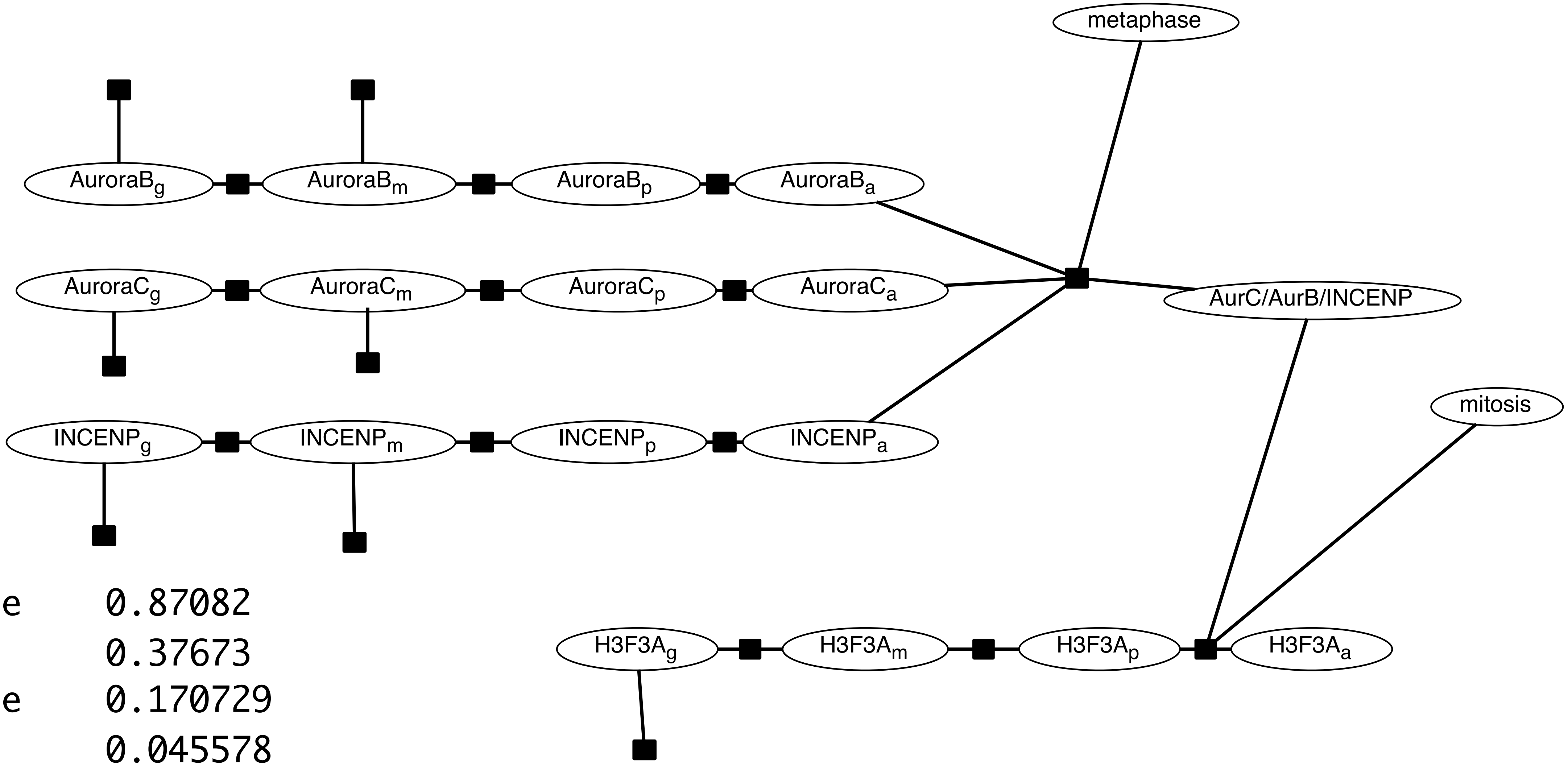


Aurora C Factor Graph



```
abstract         metaphase
abstract         mitosis
complex          AurC/AurB/INCENP
protein          H3F3A
protein          AuroraB
protein          AuroraC
protein          INCENP
AurC/AurB/INCENP H3F3A            -a>
mitosis          H3F3A            -ap>
metaphase        AurC/AurB/INCENP -ap>
AuroraB          AurC/AurB/INCENP component>
INCENP           AurC/AurB/INCENP component>
AuroraC          AurC/AurB/INCENP component>
```

Aurora C Evidence



AuroraB	genome	0.87082
AuroraB	mRNA	0.37673
AuroraC	genome	0.170729
AuroraC	mRNA	0.045578
INCENP	genome	-0.082277
INCENP	mRNA	-0.060272
H3F3A	genome	-0.411328

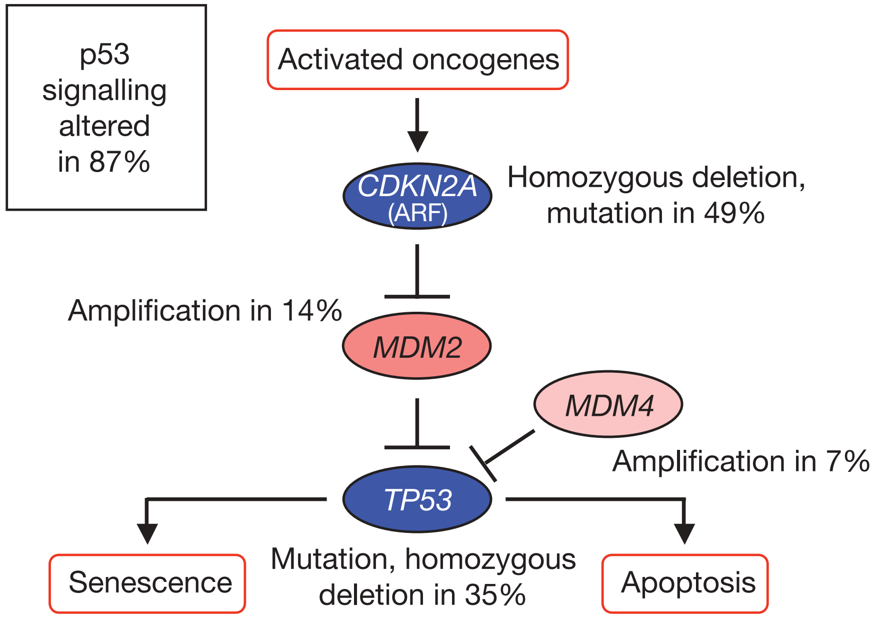
- Data points are signed, log p-values
- Right now, I discretize into up/down/same at 0.05 level
- Therefore, many patients look “identical” on hidden variables

Aurora C Inference

- using the package libDAI, which implements many approximate inference algorithms (and exact)
- Using exact at the moment
- 128 patients, 132 pathways ~ 2 hours

Prelim. Pathway results

- 2 data sets
  - Glioblastoma 224 samples
  - Ovarian Cancer 128 samples
- Still working out kinks in pipeline
- Not satisfied with data treatment


p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Senescence
Mutation, homozygous deletion in 35%
Apoptosis