NPCDS 2003

- **Bioconductor for EDA and modeling in statistical genomics**
- VJ Carey <stvjc@channing.harvard.edu>
- http://www.bioconductor.org
- Founder Rob Gentleman, massive contributions from Jeff Gentry, Jianhua Zhang, Sandrine Dudoit, Jean Hee Hwa Yang, Rafael Irizarry, Laurent Gautier, Wolfgang Huber, Gordon Smyth and others
- errors and omissions in this talk are VC's responsibility

overview

- bioconductor: a set of software goals, methods, tools; a core set of developers, researchers, educators; a large and active user community (800 source downloads, 1100 windows downloads in June)

overview

- bioconductor: a set of software goals, methods, tools; a core set of developers, researchers, educators; a large and active user community (800 source downloads, 1100 windows downloads in June)
- talk

overview

- bioconductor: a set of software goals, methods, tools; a core set of developers, researchers, educators; a large and active user community (800 source downloads, 1100 windows downloads in June)
- talk
  - EDA: looking at high throughput bioinformatics data

overview

- bioconductor: a set of software goals, methods, tools; a core set of developers, researchers, educators; a large and active user community (800 source downloads, 1100 windows downloads in June)
- talk
  - EDA: looking at high throughput bioinformatics data
  - EDA? looking at and integrating biological metadata

overview

- bioconductor: a set of software goals, methods, tools; a core set of developers, researchers, educators; a large and active user community (800 source downloads, 1100 windows downloads in June)
- talk
  - EDA: looking at high throughput bioinformatics data
  - EDA? looking at and integrating biological metadata
  - Modeling: network visualization and inference

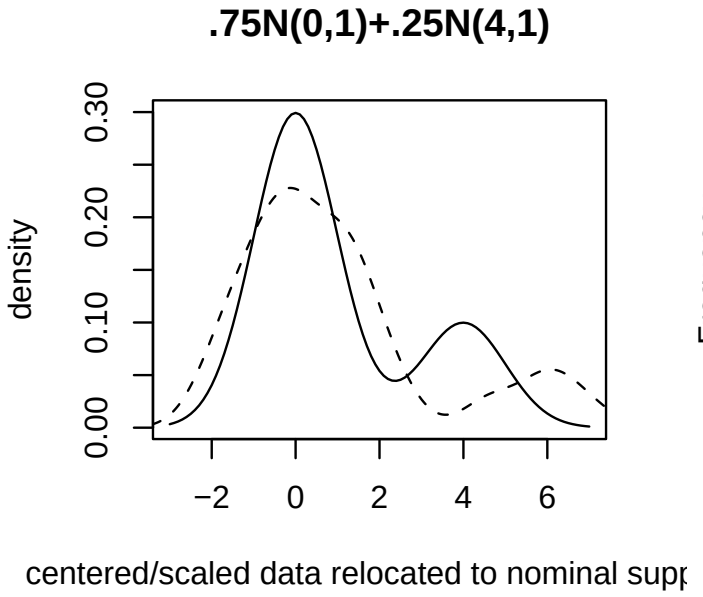
overview

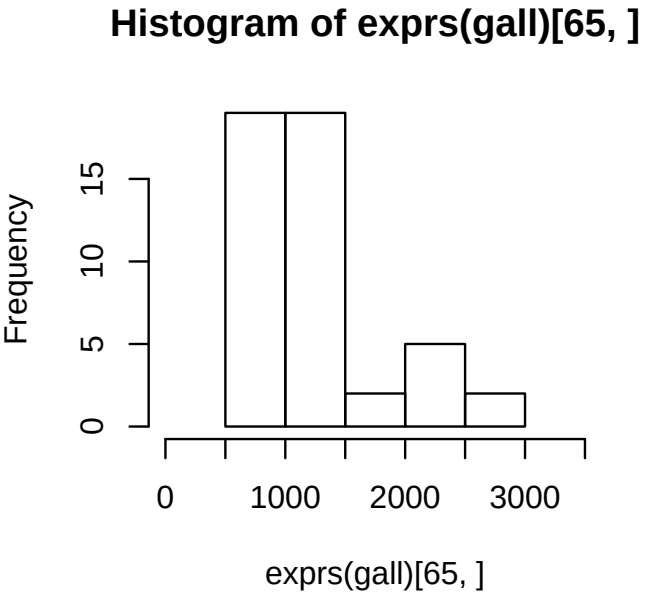
- bioconductor: a set of software goals, methods, tools; a core set of developers, researchers, educators; a large and active user community (800 source downloads, 1100 windows downloads in June)
- talk
  - EDA: looking at high throughput bioinformatics data
  - EDA? looking at and integrating biological metadata
  - Modeling: network visualization and inference
    - semantic web, logic programming roles?

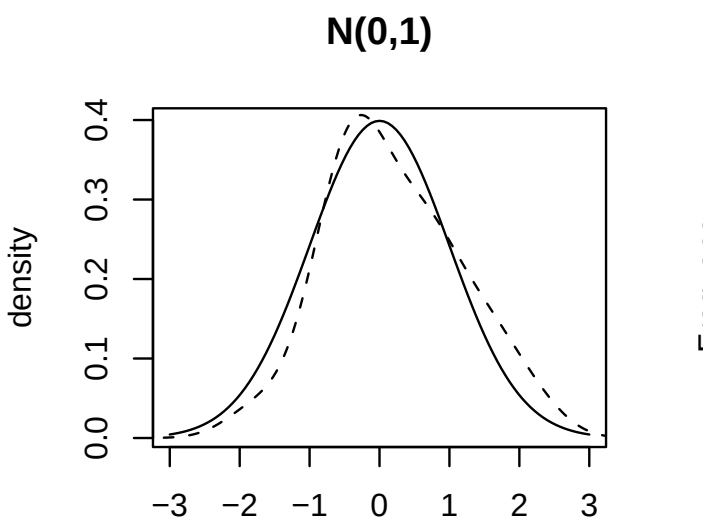
preview part 1

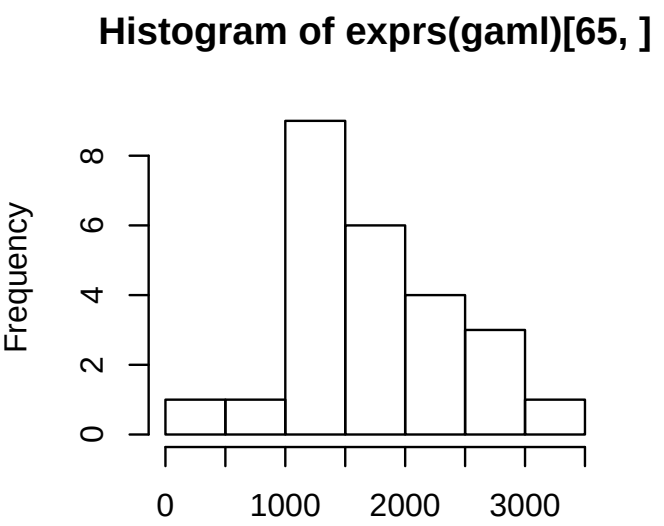
- looking at gene-specific expression distributions in high-throughput context
  - use merged golub data (47 ALL + 25 AML)
  - for illustration, use a severe filter:
  - retain genes with min expr $>$ 300, mad expr $>$ med mad expr leaving 540 genes
- objective: identify differentially expressed genes

edd: motivating example







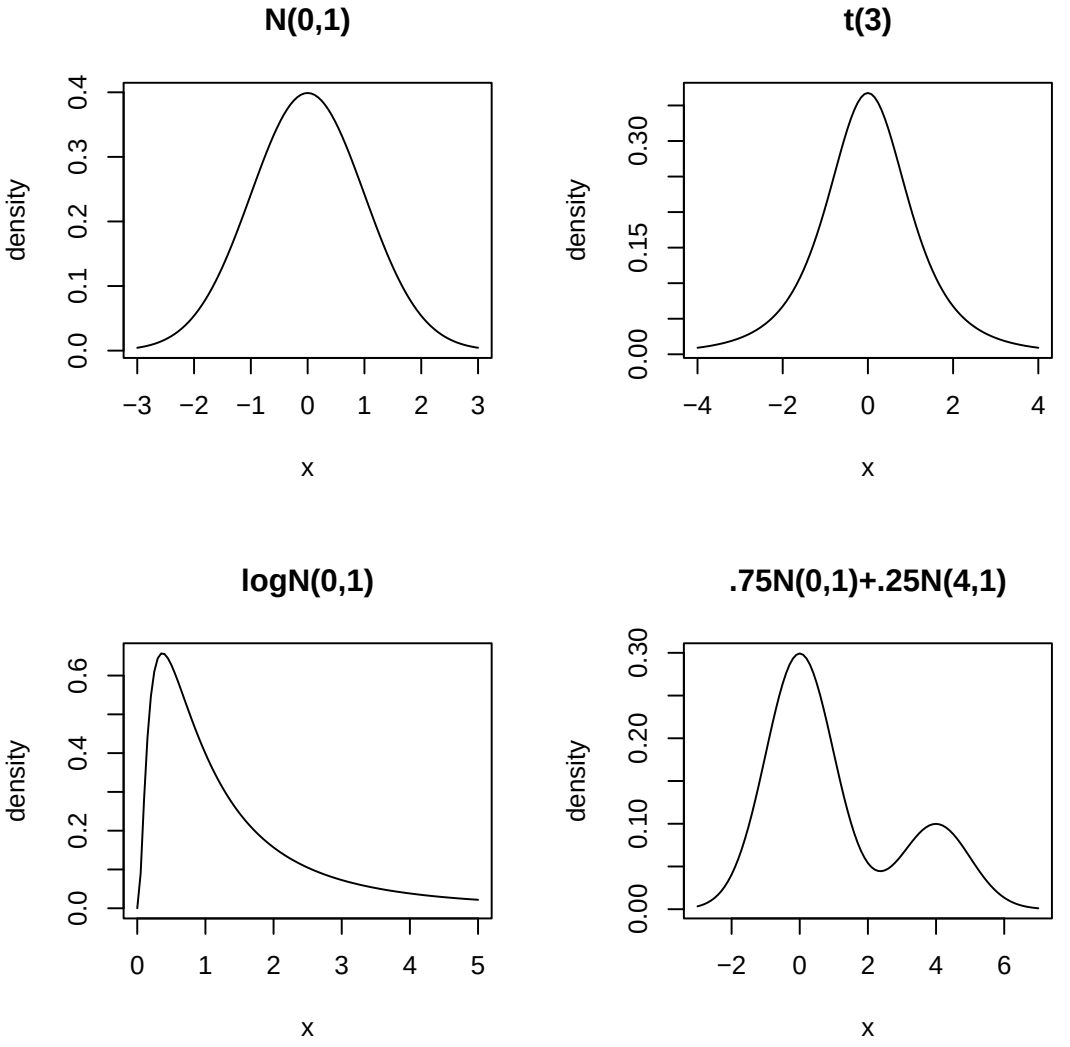


comments

- N-myc downstream regulated gene 1, enzyme and tumor suppressor activity, nuclear
- wilcoxon or t-statistic places it in the top 30 to 60 genes (location shift)
- shape of dist in ALL samples may be of interest in its own right
- right component exclusively B-cell tumors

catalog of dist. shapes

- 4 of 10 elements supplied with package



shape diversity within and between strata

- cell entries are counts of genes with ALL samples having dist. shape. labeled in row, and AML samples having dist. shape as labeled in column.

ALL/AML	Φ	t_3	$LN_{0,1}$	$U_{0,1}$	$\beta_{2,8}$	$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	$\frac{1}{4}\Phi + \frac{3}{4}\Phi$
Φ	28	13	2	3	10	7	3
t_3	40	35	7	5	15	19	7
$LN_{0,1}$	21	21	12	6	28	15	1
$\beta_{8,2}$	2	3	0	1	0	1	1
$\beta_{2,8}$	41	23	12	8	44	22	5
$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	21	12	6	1	10	2	0
$\frac{1}{4}\Phi + \frac{3}{4}\Phi_{4,1}$	1	1	0	0	0	0	1

Preview of part 2

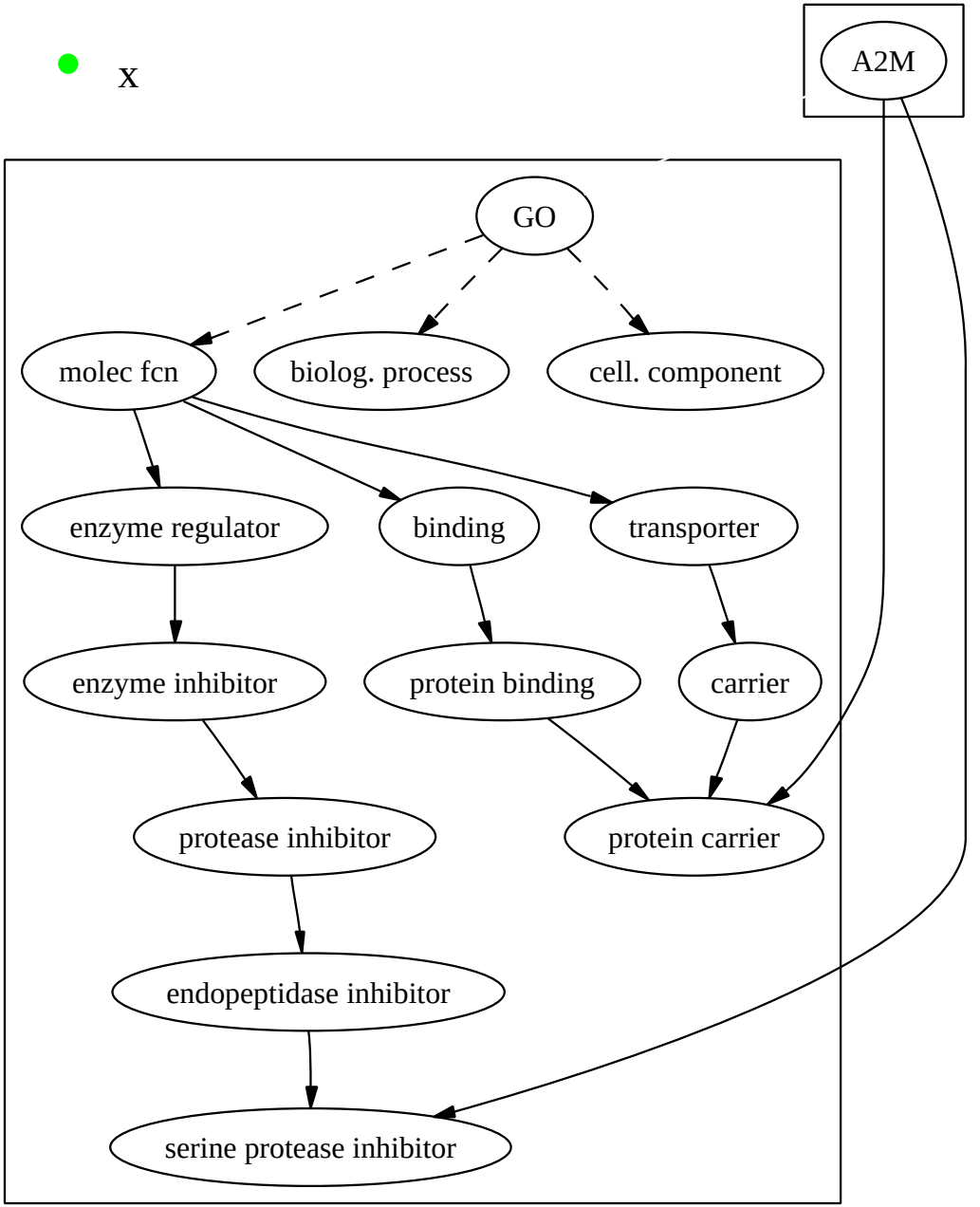
- curating images of biological metadata for use in a flexible statistical analysis environment
- using metadata in visualization and modeling

Looking at metadata

- 'metadata' here used to refer to any interpretive biological information pertinent to an analysis
- standard statistical metadata: variable names, labels, design documents, protocols, instruments
- biological metadata more complex
- should support strong 'dimension reduction', *a priori* specification of theory

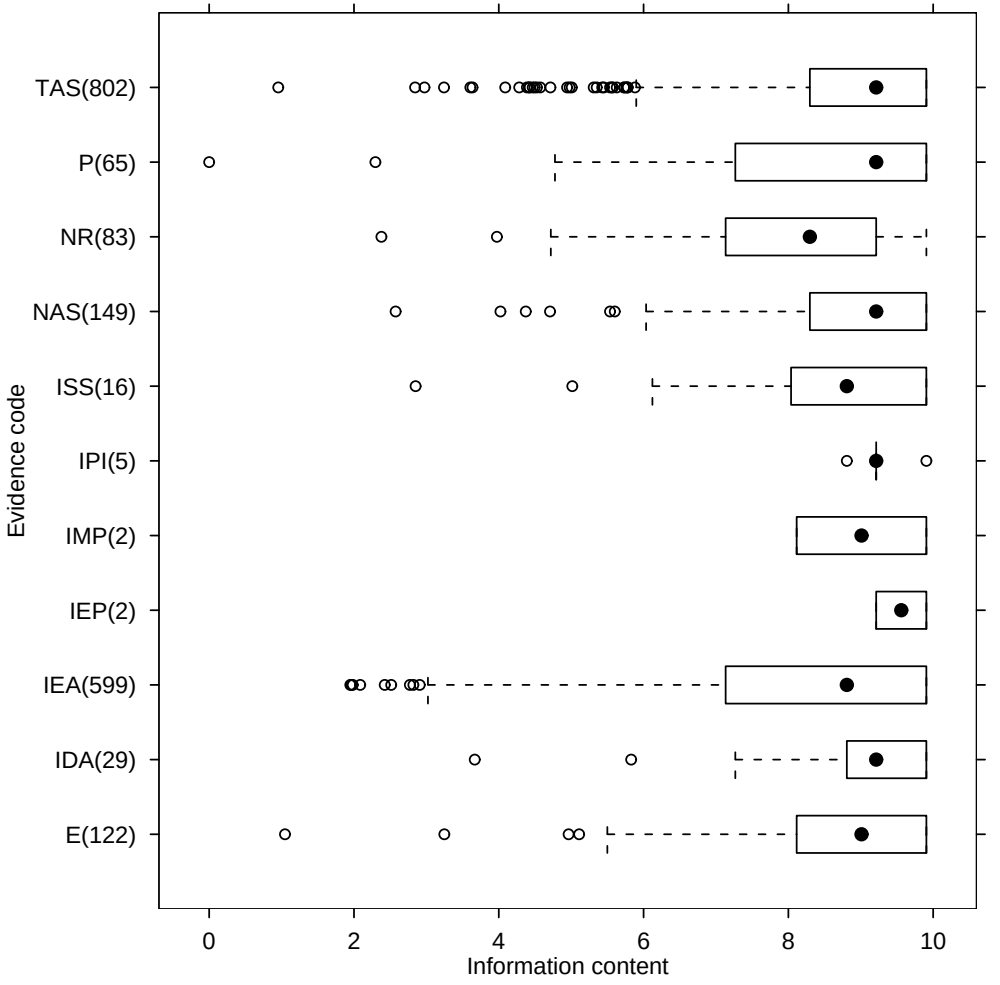
Gene ontology

- x

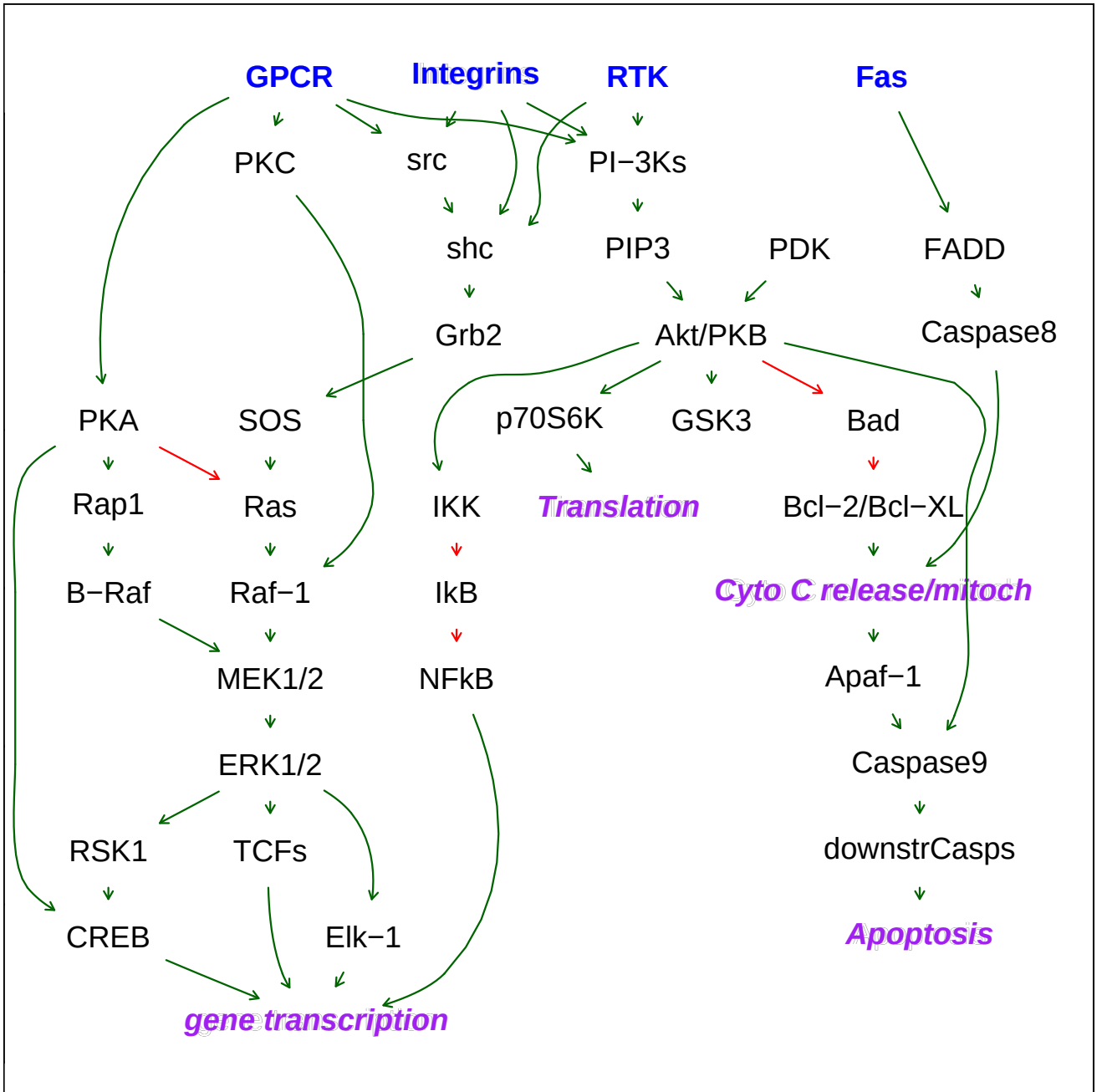


evidence code and info content

- GOA annotates from LocusLink to GO; here we consider human gene product annotations only
- evidence codes are TAS=traceable author statement, NR=not recorded, ...

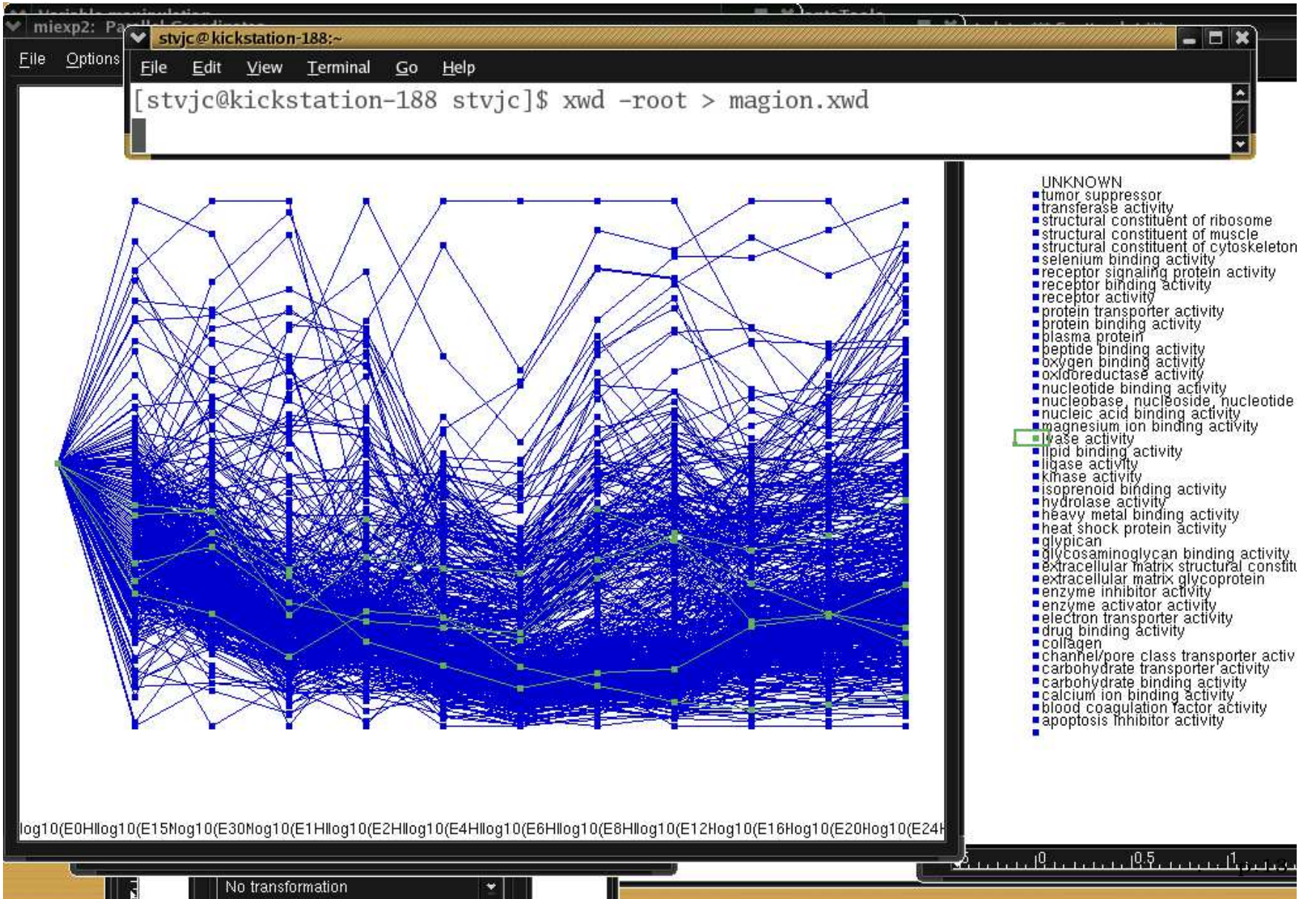


better (typed) graph example



binding metadata to dynamic graphics

- created using Rggobi with GO terms at right, Iyer517 at left



binding metadata to stat. modeling

```
testPway(golubTrain, "00130", "ALL.AML" )))
testPway(golubTrain, "00130", "ALL.AML", method="glmmPQL",
  family="gaussian" ))
```

- 00130 is ubiqinone biosynthesis, should use phrase
- address within-subject clustering in different ways
- returns ordinary R modeling object suitable for diagnostics, postprocessing, etc.

modeling results

```
Call:
geese(formula = EXPR ~ ptype + GFAC,
  id = ID, data = df, corstr = "exchangeable")

 Mean Link:                  identity
 Variance to Mean Relation: gaussian

                   estimate san.se   wald        p
(Intercept)           -68.2   52.6   1.68 1.95e-01
ptypeAML             -149.2   68.1   4.79 2.86e-02
GFACL00634.s.at       313.8   47.5  43.72 3.79e-11
...
Correlation Model:
 Correlation Structure:     exchangeable
 Correlation Link:          identity

 Estimated Correlation Parameters:
      estimate san.se wald       p
alpha     0.18 0.0562 10.3 0.00135
```

preview of part 3

- inference on network structure
- a ‘toy’ problem to expose the required infrastructure

‘Disconnected’ facts: B. Zupan+, *Bioinfo*, 19:383 (2003)

Table 1. Experimental data on *Dictyostelium* aggregation

Exp. No.	Genotype	Aggregation [−, ±, +, ++]
1	wild-type	+
2	*yakA*$^-$	−
3	*pufA*$^-$	++
4	*pkaR*$^-$	++
5	*pkaC*$^-$	−
6	*acaA*$^-$	−
7	*regA*$^-$	++
8	*acaA*$^+$	++
9	*pkaC*$^+$	++
10	*pkaC*$^-$, *regA*$^-$	−
11	*yakA*$^-$, *pufA*$^-$	++
12	*yakA*$^-$, *pkaR*$^-$	+
13	*yakA*$^-$, *pkaC*$^-$	−
14	*pkaC*$^-$, *yakA*$^+$	−
15	*yakA*$^-$, *pkaC*$^+$	++

Network resolved by GenePath

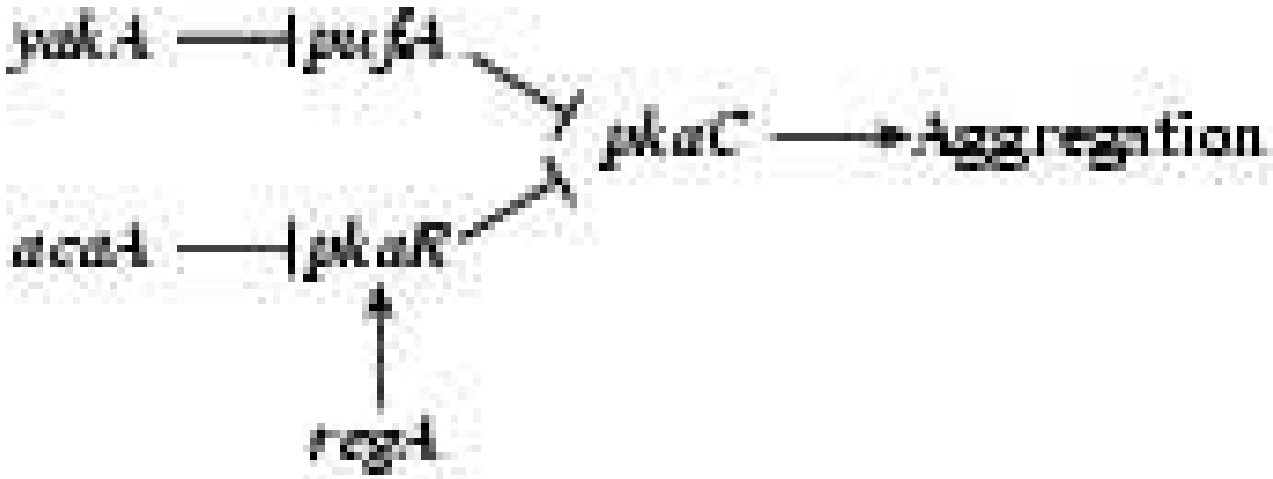


Fig. 2. A regulatory network for *Dictyostelium* aggregation. T
network was derived by GenePath from the data shown in Table
See text for detail.

issues for network inference

- various types of interaction
- behaviors of agents dependent upon context
- evidence measures important to drive interpretation
- problems and technologies similar to those related to semantic web development

looking at microarray data

- a fundamental dogma of EDA: $look\ at\ the\ data$
- problems with microarrays (post processed)
  - visualization too hard: 10K gene-specific distributions
  - visualization too easy: free iconic visualizations with heat maps etc.
  - N.B.: enthusiasm for/adoption of software in scientific research should be linked to transparency of method, documentation with working test cases and results

details of edd

- focus on distributional shapes for gene-specific expression
- reference catalog of shapes: F_c, $c = 1, \ldots, C$ a modest number
  - F_c has median 0 and mad 1
  - transformed from a substantively interesting parametric family (gaussian, t_3, log-normal, various mixtures)
  - map $\hat{F}_g$, expression distribution of each (transformed) gene g, to catalog element achieving $\min_{c \leq C} ||\hat{F}_g - F_c||$

shape diversity within and between strata

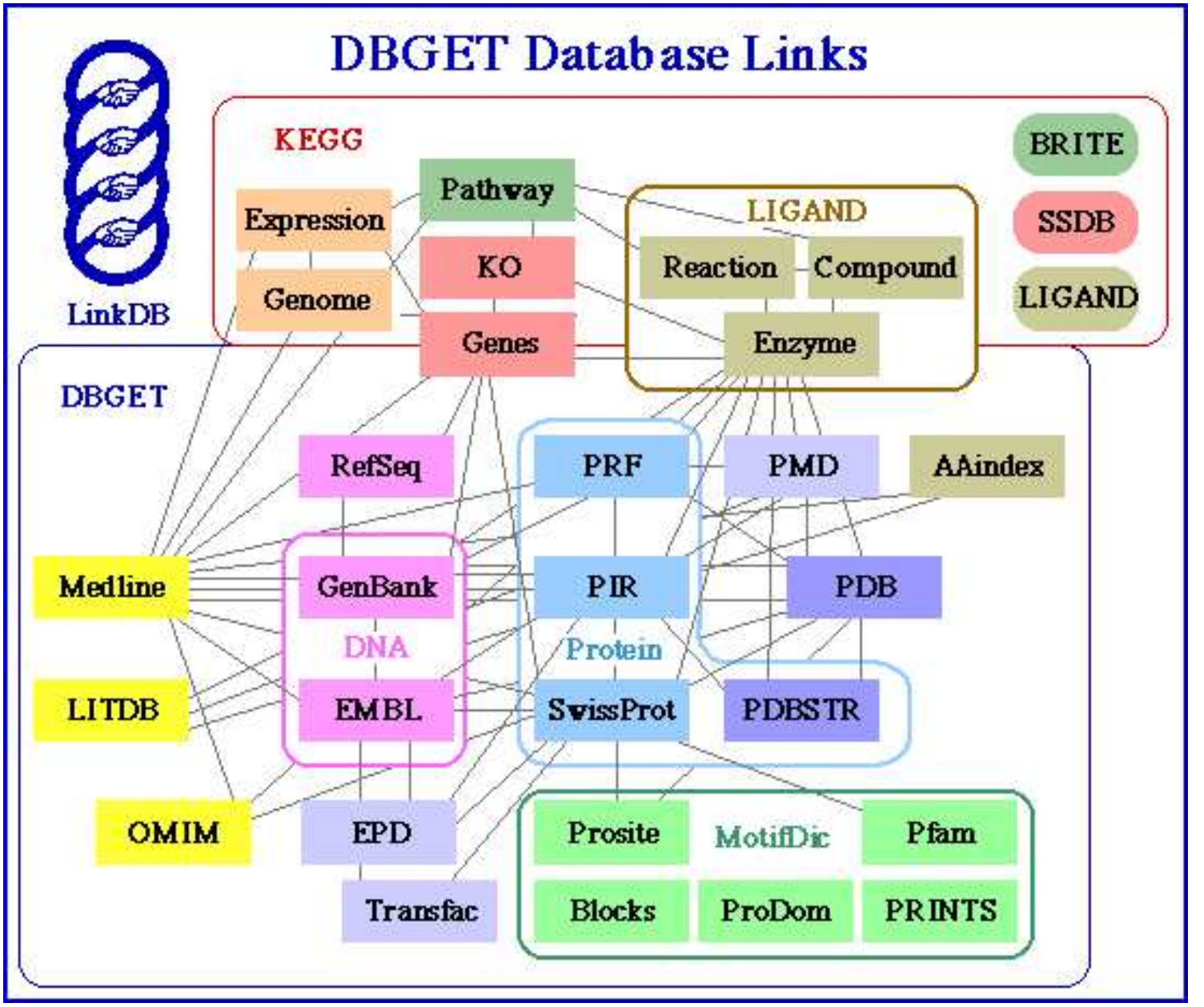
- cell entries are counts of genes with ALL samples having dist. shape. labeled in row, and AML samples having dist. shape as labeled in column.

ALL/AML	Φ	t_3	$LN_{0,1}$	$U_{0,1}$	$\beta_{2,8}$	$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	$\frac{1}{4}\Phi + \frac{3}{4}\Phi$
Φ	28	13	2	3	10	7	3
t_3	40	35	7	5	15	19	7
$LN_{0,1}$	21	21	12	6	28	15	1
$\beta_{8,2}$	2	3	0	1	0	1	1
$\beta_{2,8}$	41	23	12	8	44	22	5
$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	21	12	6	1	10	2	0
$\frac{1}{4}\Phi + \frac{3}{4}\Phi_{4,1}$	1	1	0	0	0	0	1

upshots

- diversity of distributional shapes clearly evident
- test sensitivity typically enhanced by tailoring to distributional shape
- discovery of multimodal dists can proceed in other ways – e.g., projection pursuit indices
- ideal functional for scale-invariant visualization/grouping: p-p plot? (Holmgren JASA ’95; Handcock/Morris Relative Dist’n monograph)
- mining the shapes (consistency of component occupancy across multimodal genes) of interest

metadata explosion



metadata explosion upshots

- volume
- redundancies, inconsistencies
- dynamic: version tracking essential
- authorities will be needed to certify/qualify assertions

Formalisms for annotation system

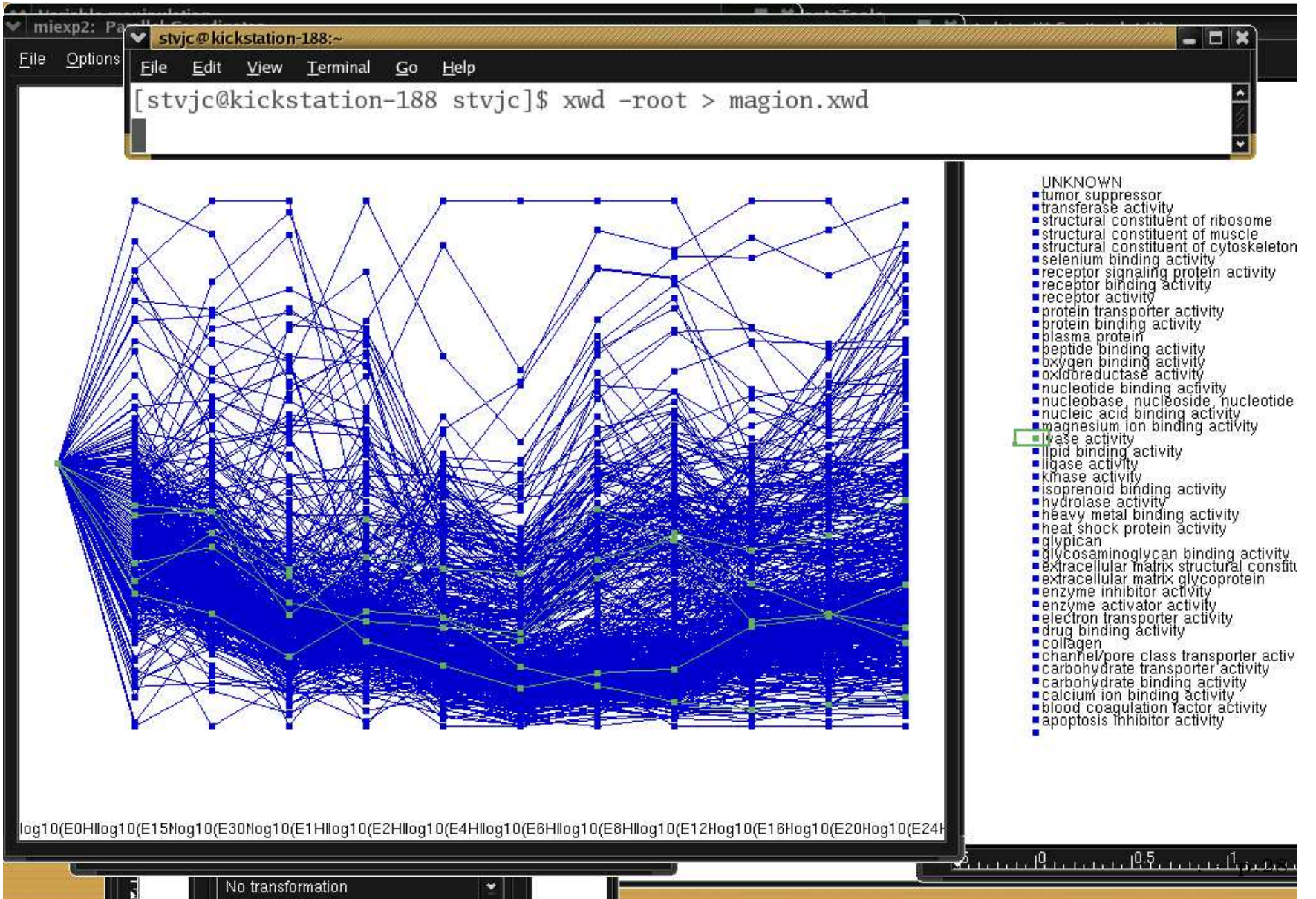
- GO is a DAG
- GOA (annotation project run by EBI) defines associations between various objects (genes, proteins, etc.) and GO terms
- object:term mapping matrix like a ‘corpus’
- Lord PW 2003 Bioinfo: information content of term c is $-\log p(c)$ [$p(\cdot)$ defined relative to corpus, e.g., $p(root) = 1$]
- semantic similarity of c_1, c_2 is the information content of their most informative subsumer (shared ancestral concept)

upshots

- GO is quite useful for interpretation of concepts (e.g., seeing what molecular function descriptions subsume others can help a newcomer deal with terminology)
- construct validity examined by PW Lord (sequence similarity correlated with semantic similarity)
- annotation from genes/proteins to GO is complex, involves evidence measures, potential circularity if a database of interest was used to infer an association

binding metadata to analysis

- created using Rggobi with GO terms at right, Iyer517 at left



issues in binding metadata to analysis

- graphical structures require layout algorithms
- many popular layout systems are commercial; AT+T graphviz a nice exception, but R support on Windows is currently infeasible
- with layout, need to render in the analysis environment and provide a navigation model – Rggobi requires sophistication

part 3: network inference/ontology/semWeb

- inference on gene interactions at a premium
- hunch: technologies that are emerging to understand and improve WWW can play a favorable role in facilitating inference on general network structures
- T. Berners-Lee May 2001 Scientific American background on semantic web; also www.w3.org

Semantic web big picture

- current WWW searches are string-based
- document metadata can specify document semantics (Carber is proper name of a person vs. one who eats carbs before athletics)
- if metadata is properly regimented, logical analysis can derive relationships among documents and their contents that are not discoverable in any one document
- regimentation schemes: RDF (resource description framework), OWL (web ontology language)

RDF

- all assertions have the form subject-predicate-object
- predicate regarded as directed arc between nodes subject and object
- construction and interpretation of graphs built from assertions proceed by strict rules
- serialization to XML is specifi ed, simplifying programmatic manipulation
- possible motivation: instead of forming complex typed graphs from the start, create only minimal elements and use rules of composition and inference to derive more complex structures

‘Disconnected’ facts: B. Zupan+, *Bioinfo*, 19:383 (2003)

Table 1. Experimental data on *Dictyostelium* aggregation

Exp. No.	Genotype	Aggregation [−, ±, +, ++]
1	wild-type	+
2	*yakA*$^-$	−
3	*pufA*$^-$	++
4	*pkaR*$^-$	++
5	*pkaC*$^-$	−
6	*acaA*$^-$	−
7	*regA*$^-$	++
8	*acaA*$^+$	++
9	*pkaC*$^+$	++
10	*pkaC*$^-$, *regA*$^-$	−
11	*yakA*$^-$, *pufA*$^-$	++
12	*yakA*$^-$, *pkaR*$^-$	+
13	*yakA*$^-$, *pkaC*$^-$	−
14	*pkaC*$^-$, *yakA*$^+$	−
15	*yakA*$^-$, *pkaC*$^+$	++

Network resolved by GenePath

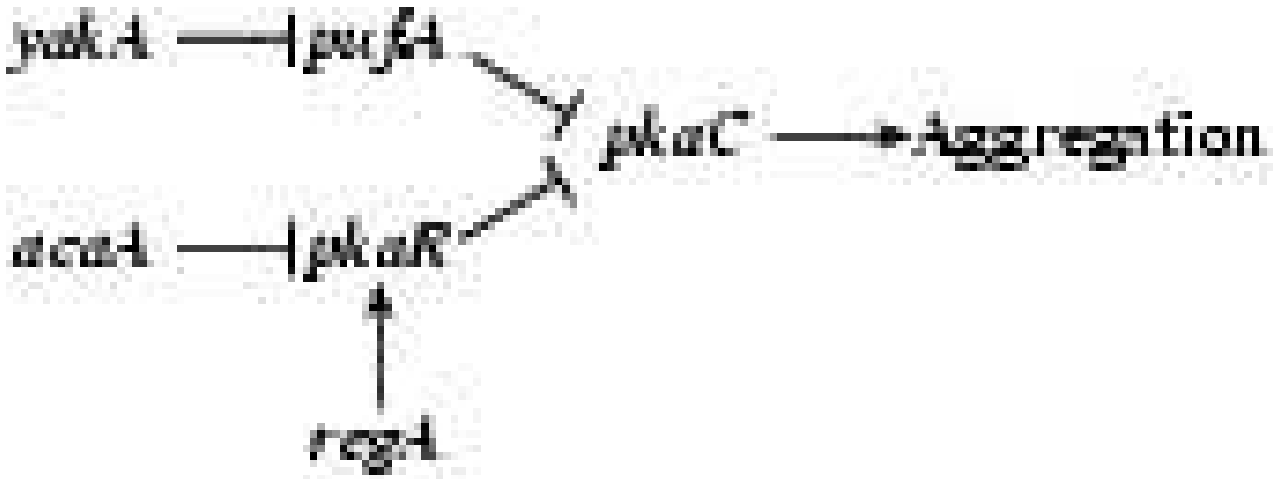


Fig. 2. A regulatory network for *Dictyostelium* aggregation. T
network was derived by GenePath from the data shown in Table
See text for detail.

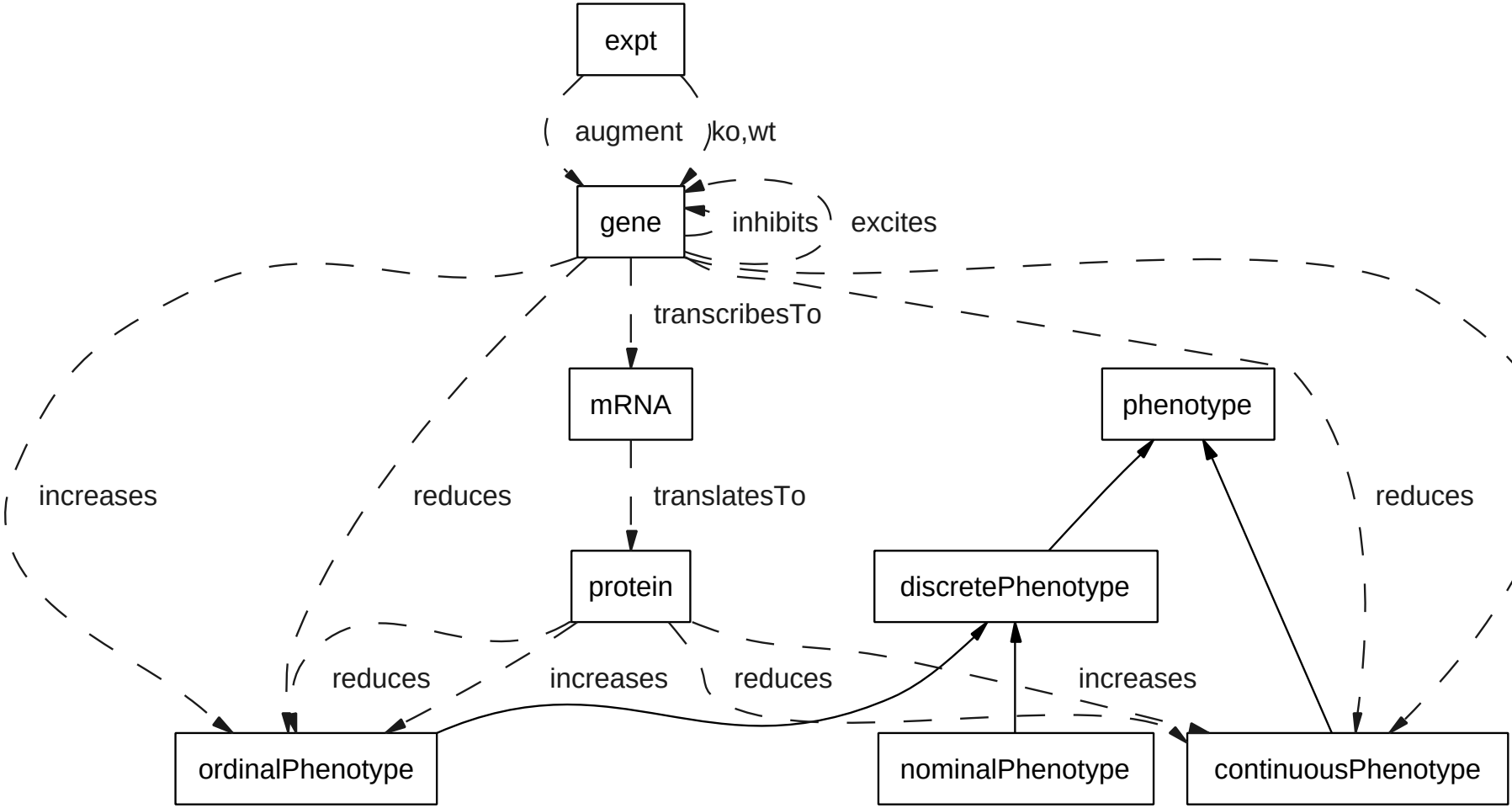
Comments

- GenePath web service uses Prolog back end
- modest number of rules to map from knockout-phenotype (biological process) observations to gene:gene relations
  - $influence$: is a gene involved with a certain process or not?
  - $parallelism$: do two genes act independently on a certain process?
  - $epistasis$: can gene involvements be ordered linearly? (G1 epistatic to G2 if G1 is 'closer' to the process)
- input data structures straightforward; standards? biopathways.org, BIOPAX, etc.

objectives

- standard representation of experiment database with qualifications of results (standard errors...)
- conventional and standardly expressed rules for interpretation of groups of experiments (influence, epistasis, etc.)
- transparent and portable inference tool

database representation standard: ontology



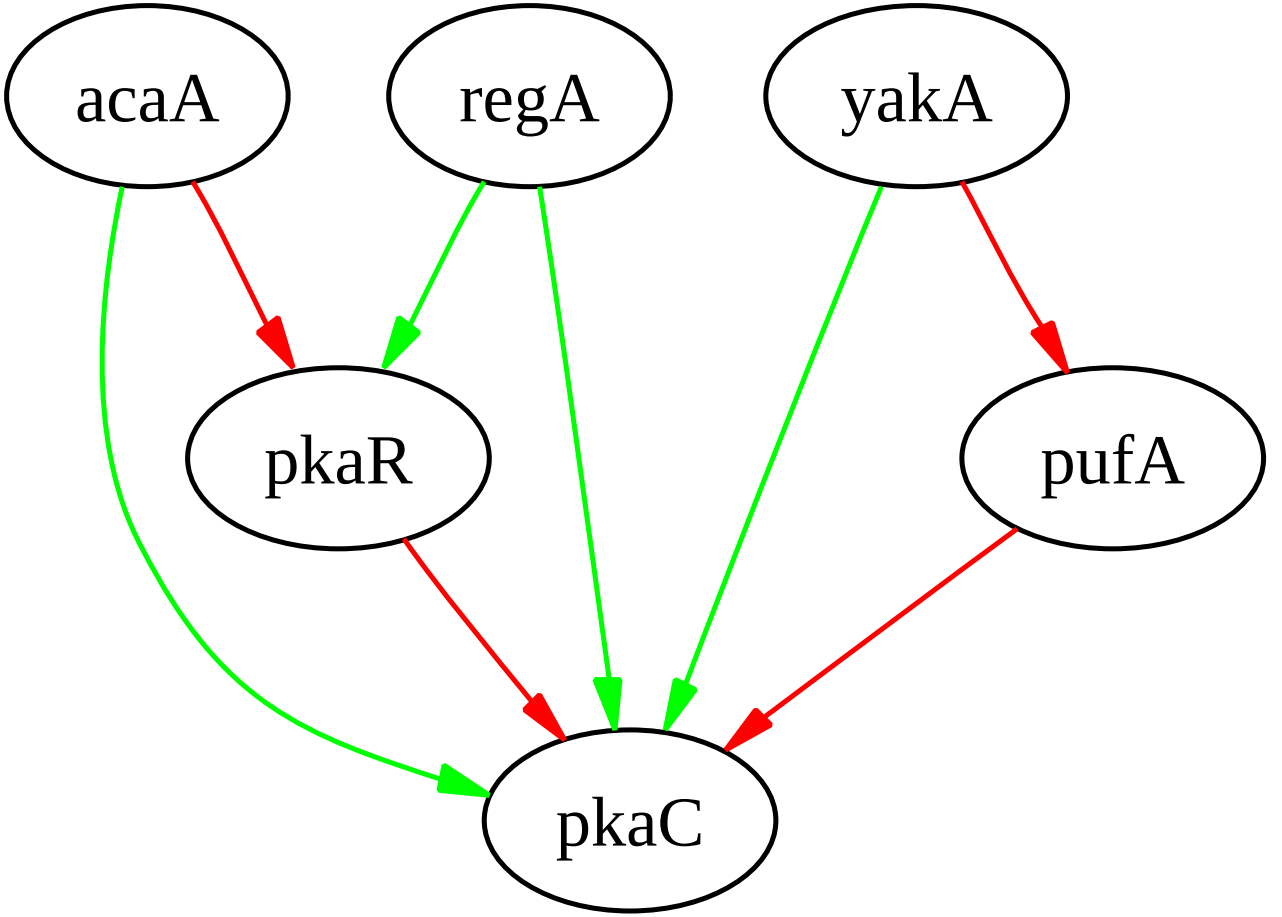
- also needs ‘outcome’ property with domain `expt` and range `phenotype`

standard rule for G 'increasing' P

```
increases( G, P ) :-
 gene(G),
 expt(F), expt(E), F \= E,
 ;((wt(E,G), ko(F,G)), (augment(E,G), wt(F,G))),
 outcome(E,X), outcome(F,Y),
 ordinalPhenotype(X), ordinalPhenotype(Y),
 X > Y .
```

- here expressed in prolog
- can be expressed in RDF via Notation 3
- CWM RDF processor can infer the network structure (with additional rules)

CWM/R(XML/graph)/dot result



comments

- qualitatively similar to genepath result
- possible advantages: explicit database schema and semantics, transparent expression of rules, inferences serialized to RDF/XML
- needs propagation of measures of evidence for individual experiment assertions

Conclusions

- EDA at the microarray expression level is feasible
- considerable diversity in expression distribution shapes should be acknowledged; tests should adapt
- managing access to and analysis with metadata is a significant undertaking; analysts cannot escape the complexity of the information and its evolution
- get in touch with the semantic web initiative; logic programming attack on network inference only scratches the surface