Kollekolle 2003

- Day 2, Session 3
- Annotation, ontologies, graph structures I
- VJ Carey <stvjc@channing.harvard.edu>
General annotation resources

> library(annotate)
> objects("package:annotate")

[1] "abstText"  "abstUrl"  "accessionToUID"
[4] "articleTitle"  "authors"  "buildChromLocation"
[10] "chromInfo"  "chromLengths"  "chromLocs"
[13] "chromNames"  "contents"  "dataSource"
[16] "fileName"  "findChr4LL"  "findNeighbors"
[19] "genbank"  "genelocator"  "geneSymbols"
[22] "getBoundary"  "getGO"  "getGOdesc"
[25] "getLL"  "getPMID"  "getPMInfo"
[28] "getQuery4LL"  "getQuery4UG"  "getQueryLink"
[31] "getSYMBOL"  "getTDRows"  "getValidChr"
[34] "installDataPackage"  "journal"  "ll.htmlpage"
[37] "locuslinkByID"  "locuslinkQuery"  "mainPage"
[40] "makeAnchor"  "nChrom"  "organism"
[43] "pageText"  "pageTitle"  "pm.abstGrep"
[46] "pm.getabst"  "pm.titles"  "pmAbst2HTML"
[49] "pmAbst2html2"  "pmid"  "pmidQuery"
[52] "probesToChrom"  "pubDate"  "pubmed"
[55] "pubMedAbst"  "sidePage"  "toFile"
[58] "topPage"  "UniGeneQuery"  "usedChromGenes"
Annotate package

- helps with nomenclature translations
- queries NCBI resources
- constructs hyperlinked web documents based on gene lists
- the PubMed-related utilities are of particular interest
Data packages

- two broad classes
  - platform-specific annotation (hu6800, hgu133abc... mgu74ab...)
  - synopses of curated databases (GO, KEGG, humanLLMappings)
- it is planned to pursue more genome-wide (as opposed to platform-specific) annotation
Working with a platform

> library(hu6800)
> print(hu6800())

Quality control information for   hu6800
Date built:   Tue Aug 26 12:58:02 2003
Number of probes: 7129
Probe number missmatch: None
Probe missmatch: None

Mappings found for probe based rda files:
  hu6800ACCNUM found 7092 of 7129
  hu6800CHRLOC found 6788 of 7129
  hu6800CHR found 7090 of 7129
  hu6800ENZYME found 1102 of 7129
  hu6800GENENAME found 7116 of 7129
  hu6800GO found 6376 of 7129
  hu6800GRIF found 4126 of 7129
  hu6800HGID found 6524 of 7129
  hu6800LOCUSID found 7129 of 7129
  hu6800MAP found 7082 of 7129
  hu6800PATH found 1480 of 7129
  hu6800PMID found 7048 of 7129
  hu6800SUMFUNC found 369 of 7129
  hu6800SYMBOL found 7116 of 7129
  hu6800UNIGENE found 7044 of 7129

Mappings found for non-probe based rda files:
  hu6800ENZYME2PROBE found 494
  hu6800GO2ALLPROBES found 3936
  hu6800GO2PROBE found 2874
  hu6800PATH2PROBE found 120
  hu6800PMID2PROBE found 37928
  NULL

{ p.5
Working with a platform

> print(objects("package:hu6800"))

[1] "hu6800"       "hu6800ACCNUM"     "hu6800CHR"
[4] "hu6800CHRLENGTHS" "hu6800CHRLOC"   "hu6800ENZYME"
[7] "hu6800ENZYME2PROBE" "hu6800GENENAME" "hu6800GO"
[10] "hu6800GO2ALLPROBES" "hu6800G02PROBE" "hu6800GRIF"
[13] "hu6800HGID"    "hu6800LOCUSID"   "hu6800MAP"
[16] "hu6800ORGANISM" "hu6800PATH"      "hu6800PATH2PROBE"
[19] "hu6800PMID"    "hu6800PMID2PROBE" "hu6800QC"
[22] "hu6800SUMFUNC" "hu6800SYMBOL"    "hu6800UNIGENE"
Quality assurance elements

• annotation is an evolving process

```r
> print(hu6800CHRLENGTHS)
```

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>246016287</td>
<td>243263343</td>
<td>199112551</td>
<td>191684869</td>
<td>180794049</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>135991854</td>
<td>134841696</td>
<td>133787053</td>
<td>131997912</td>
<td>112966179</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>81467123</td>
<td>76017630</td>
<td>63766217</td>
<td>63614468</td>
<td>46912768</td>
</tr>
</tbody>
</table>

• these Golden-path derived bounds on chromosome coordinates change from build to build

• want to guarantee that frank errors in CHRLOC elements are discovered
Application with pathway data

- customary formalization of Golub’s problem:
  - $X_{ig}, i = 1, \ldots, 38, g = 1, \ldots, G = 7129$ is the expression outcome
  - $P_{ij}, j = 1, \ldots, p$ is $j$th phenotypic feature on subject $i$

- case-control formulation: $P_{i1}$ is an indicator of AML (vs ALL); if distribution of $X_g$ depends upon value of $P_{.1}$, differential expression of gene $g$ is declared

- classification formulation: use $X_\cdot$ to predict value of $P_{.1}$
concrete pway example ct’d

- the $X_g$ may be organized in biologically interesting ways, e.g., subsets $W_k$ of $\{1, \ldots, G\}$, $k = 1, \ldots, K$ define memberships in $K$ genomic pathways

- new question: are the members of a given pathway differentially expressed?

- what would be involved in testing $\beta = 0$ in

$$EX_{ig} = \alpha + P_i \beta + \gamma_g + f(P_{i,-1})$$

for $g$ in a given $W_k$?
concrete pway example ct’d

- provided we have a well-designed data structure
  - binding of expression array output to gene metadata
  - binding of gene metadata to pathway classification
  - binding of phenotype data to expression array output

the following suffices

testPway <- function(eset, # bioC
    pwname, # KEGG
    phenoFac, # investigator
    chipType="hu6800", # manufacturer
    method=c("gee","glmmPQL")[1], ...) {
    [12 lines of R code]
}
picking a pathway

> print(exprs(golubTrain)[1200, 1:5, drop = FALSE])

L04490_at -10 175 635 -154 352

> pname <- row.names(exprs(golubTrain))[1200]
> print(get(pname, env = hu6800SYMBOL))

[1] "NDUFA9"

> print(get(pname, env = hu6800CHR))

[1] "12"

> print(get(pname, env = hu6800CHRLOC))

 12
4637543

> print(pway <- get(pname, env = hu6800PATH))

[1] "00130"
metadata 3

> print(get("00130", env = KEGGPATHID2NAME))

```

00130
"Ubiquinone biosynthesis"

> print(get("00130", env = hu6800PATH2PROBE))
```

```
[1] "L10413_at"       "L00634_s_at"       "L00635_at"
[4] "Y10807_s_at"    "U94586_at"       "U53468_at"
[7] "HG3141-HT3317_f_at" "L04490_at"    "M33374_at"
[10] "X61100_rna1_at" "U65579_at"       "M22538_at"
[13] "X99728_at"      "U34343_at"       "Y08200_at"
[16] "X98001_at"
```

- various attributes of probe; could store in big database table
- we have tailored to platform/annotation source with hashed environments, use programming to create the interface
modeling calls

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

- 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

<table>
<thead>
<tr>
<th>estimate</th>
<th>san.se</th>
<th>wald</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-68.2</td>
<td>52.6</td>
<td>1.68</td>
</tr>
<tr>
<td>ptypeAML</td>
<td>-149.2</td>
<td>68.1</td>
<td>4.79</td>
</tr>
<tr>
<td>GFACL00634.s.at</td>
<td>313.8</td>
<td>47.5</td>
<td>43.72</td>
</tr>
</tbody>
</table>

... Correlation Model:
Correlation Structure: exchangeable
Correlation Link: identity

Estimated Correlation Parameters:
<table>
<thead>
<tr>
<th>estimate</th>
<th>san.se</th>
<th>wald</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>0.18</td>
<td>0.0562</td>
<td>10.3</td>
</tr>
</tbody>
</table>
modeling results

using

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

Random effects:
  Formula: ~1 | ID
       (Intercept) Residual
       StdDev:  236    504

Variance function:
  Structure: fixed weights
  Formula: ~invwt

Fixed effects: EXPR ~ ptype + GFAC

<table>
<thead>
<tr>
<th>Search</th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-68</td>
<td>95.8</td>
<td>555</td>
<td>-0.71</td>
<td>0.4765</td>
</tr>
<tr>
<td>ptypeAML</td>
<td>-149</td>
<td>97.1</td>
<td>36</td>
<td>-1.54</td>
<td>0.1331</td>
</tr>
<tr>
<td>GFACL00634.s.at</td>
<td>314</td>
<td>117.2</td>
<td>555</td>
<td>2.68</td>
<td>0.0077</td>
</tr>
</tbody>
</table>
| ...
Upshot

• the interface is a bit crude (could help with isolation of pathways or combinations thereof, introduce a formula element)

• key notions are
  • restructuring expression data to conform to statistical modeling input requirements goes on reliably behind the scenes owing to fixed APIs
  • the modeling procedure has a ‘drop-in’ status
  • estimation and testing structures are passed out for generic diagnosis
Ontologies

• in computer science, ontology has come to refer to a structured vocabulary
• structure often governed by is-a or part-of relationships among term denotations
• bioinformatics introduction to the concept: GO = Gene Ontology (www.geneontology.org)
GO excerpt

molec fcn  biolog. process  cell. component
enzyme regulator  binding  transporter
enzyme inhibitor  protein binding  carrier
protease inhibitor  protein carrier
dendopeptidase inhibitor
serine protease inhibitor

A2M
More generally

- need to deal with structures like
adjacency matrix for ontology A

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
incidence matrix (object:term) $W$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Z</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- accumulated boolean exponentiation of $A$ yields the accessibility matrix $C$;
- boolean product $CW$ yields the coverage matrix
### Coverage matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Z</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- useful for computing information content of terms relative to a corpus ($- \log p(c)$, where $p(c)$ is ratio of number of times term $c$ is used divided by maximum number of times any term is used)
- ontoTools package supports these manipulations with sparse matrices
Association evidence codes

GOA is an EBI-based project that associates terms with genes. Evidence codes are TAS (traceable author statement), P (predicted), NR (not recorded), NAS (non-traceable author statement), ISS (inferred from structural similarity), IPI (inferred from protein interaction), IPI (inferred from mutant phenotype), IEP (inferred from expression pattern), IEA (inferred from electronic annotation), IDA (inferred from direct assay), E (experimental).
Evidence and information
Other ontologies

- **OBO** ([obo.sourceforge.net](http://obo.sourceforge.net))
- a customized ontology for knockout experiment databases
Summary

- BioC provides annotation resources readily accessible to statistical analysis environment
- facilitates classical hypothesis testing with high-throughput resources (testPway), provided a prospective hypothesis is of interest
- Gene Ontology one of a growing number of structured terminology aids; GO is used by analysts to regulate confidence that gene semantics are understood
- ontoTools supports reasoning on information content and semantic similarity of terms (relative to an association of terms and objects)
- ontology design and processing promises to be a source of improved efficiency of data use and sharing (OBO, Semantic web...)