

A Statistical Approach to Literature-based Microarray Annotation

Xin He

10/11/2006

Annotating a Gene List

- Understand the commonalities in a list of genes from:
 - Clustering by expression patterns
 - Differentially expressed genes
 - Genes sharing cis-regulatory elements
- How to automatically construct the annotations?

Gene Ontology-based Approach

- Each gene is annotated by a set of GO terms
- The importance of any term wrt the gene list is measured by the number of genes that are associated with this term
- Need to correct for the uneven distribution of GO terms: a hypergeometric test

Limitations of GO-based Approach

- GO annotations of all genes: may not be available
- Rapid growth of literature: constantly add new functions to existing genes
- Coverage is not even in all areas. E.g. ecology and behavior

Literature-based Approach

- Annotate via the analysis of text extracted from literature
- Advantages:
 - Not dependent on manually created data
 - Easy to keep up with the recent discoveries
 - Broad coverage
- Explorative tool: suggest new hypothesis

Literature-based Approach

- Extract abstracts for each gene
- **Idea:** If a word is overrepresented in the abstracts for the list, then it is likely to describe the common functions of the list
- A simple measure of significance: $Z\text{-score} = (\text{observed count} - \text{expected count under background distribution}) / \text{standard deviation}$

Motivations for the Current Work

- Drawbacks of the existing approach
 - False positives: overrepresented, but not common to the gene list
 - Genes that are well-studied will dominate the results
- Motivations for a new approach:
 - Need to capture overrepresentation of words
 - Favor words that are common to all or most genes
 - A unified way to solve both problems?

Ideas for the Statistical Model

- **Observation:** typically, some genes in the list are related to a given word, but the other genes are not (Few gene clusters are perfect!)
- **Assumption:** the count of a word in a document follows Poisson distribution
- **Idea:** the count of a word in one gene is either from a background distribution (if the gene is unrelated) or from a “signal” distribution (if the gene is related)

Poisson Mixture Model

Notations:

$d_1, d_2 \dots d_n$: the size of documents of genes 1 to n

$x_1, x_2 \dots x_n$: counts of the word in documents 1 to n

λ_0, λ : the rate of the background and signal Poisson distributions

θ : the mixing weight of the signal Poisson

Each count is generated from a mixture of the background and signal Poisson distributions:

$$P(x_i | d_i, \lambda_0, \theta, \lambda) = \theta P(x_i | \lambda d_i) + (1 - \theta) P(x_i | \lambda_0 d_i)$$

The probability of observing the data is thus:

$$\begin{aligned} P(\mathbf{x} | \mathbf{d}, \lambda_0, \theta, \lambda) &= \prod_{i=1}^n P(x_i | d_i, \lambda_0, \theta, \lambda) \\ &= \prod_{i=1}^n [\theta P(x_i | \lambda d_i) + (1 - \theta) P(x_i | \lambda_0 d_i)] \end{aligned}$$

Parameter Estimation

Maximum likelihood estimation of parameters:

$$(\hat{\theta}, \hat{\lambda}) = \arg \max_{0 \leq \theta \leq 1, \lambda > 0} P(\mathbf{x} | \mathbf{d}, \lambda_0, \theta, \lambda)$$

EM algorithm to maximize the likelihood function. The updating formula is given by:

$$\theta^{(t+1)} = \sum_{i=1}^n P(z_i = 1 | x_i, d_i, \lambda_0, \theta^{(t)}, \lambda^{(t)}) / n$$

$$\lambda^{(t+1)} = \sum_{i=1}^n P(z_i = 1 | x_i, d_i, \lambda_0, \theta^{(t)}, \lambda^{(t)}) x_i / \sum_{i=1}^n P(z_i = 1 | x_i, d_i, \lambda_0, \theta^{(t)}, \lambda^{(t)}) d_i$$

The posterior probability of missing label (z_i) is:

$$P(z_i = 1 | x_i, d_i, \lambda_0, \theta^{(t)}, \lambda^{(t)}) = \frac{\theta^{(t)} P(x_i | \lambda^{(t)} d_i)}{(\theta^{(t)} P(x_i | \lambda^{(t)} d_i) + (1 - \theta^{(t)}) P(x_i | \lambda_0 d_i))}$$

Evaluating the Statistical Significance

- The candidate words: those with a large θ (an estimate of the proportion of related genes)
- Need to assess the significance
 - E.g. a word from a distribution slightly different from the background. EM may estimate λ to be MLE and θ close to 1
- **Idea:** if the counts can be explained well by the background, then there is no need to use a mixture of two distributions. This word would be insignificant regardless of the estimated θ

Likelihood Ratio Test

Hypothesis testing:

$H_0 : \theta = 0$, the observed counts are generated from the background

$H_1 : \theta > 0$, the observed counts are generated from the mixture

Generalized Likelihood Ratio test Statistic (LRS):

$$\begin{aligned} T &= -2 \log \frac{P(\mathbf{x} | \mathbf{d}, \lambda_0)}{P(\mathbf{x} | \mathbf{d}, \lambda_0, \hat{\theta}, \hat{\lambda})} \\ &= -2 \sum_{i=1}^n \log \frac{P(x_i | \lambda_0 d_i)}{\hat{\theta} P(x_i | \hat{\lambda} d_i) + (1 - \hat{\theta}) P(x_i | \lambda_0 d_i)} \end{aligned}$$

Reject H_0 if T is greater than a certain threshold.

Asymptotic Distribution of LRS

- It is well known that the distribution of LRS converges to chi-square, with degree of freedom equal to the difference between the number of free parameters of null and alternative hypothesis
- However, this does not apply in mixture models because the regularity condition is violated
- Analytically difficult: relies on simulation
- In practice: a LRS cutoff is empirically determined by inspecting the words. An open problem

Experimental Validation

- Test data sets
 - DNA replication cluster (J) in the paper “Cluster analysis and display of genome-wide expression patterns”, Eisen et al, PNAS’98
 - Pelle system: a set of genes involved in Drosophila pattern formation
- Procedure
 - Extract abstracts for each gene
 - Apply the EM estimation and compute LRS for each word
 - Output words whose LRS is greater than some threshold and sort the words by the estimated mixing weight (θ)

CDC54	DNA replication	MCM initiator complex
MCM3	DNA replication	MCM initiator complex
MCM2	DNA replication	MCM initiator complex
CDC47	DNA replication	MCM initiator complex
DBF2	Cell cycle	Late mitosis, protein kinas

1. minichromosome maintenance

minichromosome (0.800001), maintenance (0.8), chromatin (0.8)

2. DNA synthesis

nuclear (1), prereplicative (0.800703), replicative (0.800432), initiation (0.800004), replication (0.8), origins (0.8), origin (0.8), helicase (0.8), dna (0.8), licensing (0.799969), forks (0.778143), prereplication (0.647716), orc (0.8), orc2 (0.8103)

3. cell cycle

mitosis (1), g2 (1), g1 (1), cyclin (1), cycle (1), cdks (1), checkpoint (0.992086), phase (0.80003), mitotic (0.800001)

4. other genes not in the given list

dbf4 (1), cdc7 (1), cdc28 (1), mcm7p (0.860795), mcms (0.802664), mcm5 (0.800872), rcs (0.80046), cdc21 (0.80029), cdc45p (0.800021), cdc46 (0.800013), mcm7 (0.800005), rc (0.8), pre (0.8), mcm4 (0.8), cdc6 (0.8), cdc45 (0.8), mcDC21 (0.708466)

5. biological, not specific, but related

yeast (1), saccharomyces (1), cerevisiae (1), budding (1), complex (1), fission (0.812616), nucleus (0.799965), sv40 (0.820404)

6. noninformative words

we (1), the (1), that (1), show (1), protein (1), is (1), effects (1), cell (1), 2 (1), 0 (1)

spatzle	an extracellular ligand for toll
toll	a transmembrane receptor
pelle	a serine threonine protein kinase
tube	unknown function associated to membrane
cactus	inhibitor of dorsal
dorsal	transcription factor

1. embryonic development: dorsal-ventral axis formation

polarity (1), patterning (1), embryo (1), dorsoventral (1), dorsal (1), larval (0.999987), axis (0.999976), ventral (0.999683), larvae (0.995213), embryos (0.991556), dv (0.886486), gradient (0.701978)

2. defense response

drosomycin (1), immunity (0.989891), immune (0.83402), defense (0.642896), host (0.502721)

3. other genes not in the input list

ikappab (1), kappab (0.999999), nf (0.999648), gal4 (0.846292), easter (0.833718), dif (0.690716), hopscotch (0.682867), kra (0.669438), rel (0.666697), myd88 (0.66669), sog (0.522902)

4. biological, not specific, but related

drosophila (1), melanogaster (0.99999), zygotic (0.999996), lamellocytes (0.862279), innate (0.834148), nuclear (0.820233), receptor (0.757764), import (0.666706), hemocyte (0.527793)

5. noninformative words

were (1), we (1), was (1), the (1), that (1), show (1), is (1), here (1), genes (1), effects (1), 0 (1), function (1), protein (0.998548)

Future Plan

- Web-based system
- Automatically determine the threshold: via the simulation of LRS distribution
- Include phrases: choose candidate phrases via hypothesis testing
- Sentence selection: convert significant words into a language model and do retrieval
- Customize the background distribution
- Extensions: use the representative words for other purposes such as gene clustering