

Statistical Method of Gene Set Annotation Based on Literature Information

Xin He

09/25/2007

Annotating a Gene List

- Goal: understand the functional themes from 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
- Other systems use more schemes (e.g. SwissProt keywords), but they all rely on manual annotations

Limitations of GO-based Approach

- GO annotations of all genes involve intense manual efforts
- Rapid growth of literature: constantly add new functions to existing genes
- Coverage is not even in all areas. E.g. ecology and behavior; medicine; anatomy and physiology; etc.

Literature-based Approach

- Annotate via the analysis of text extracted from literature, essentially the pattern of co-occurrence between terms and genes
- Explorative tool: suggest new hypothesis

Literature-based Approach

- GEISHA
 - Idea: measure the overrepresentation of a term in the documents of a gene list (Z-score)
 - Problem: some genes have much larger literature support
- TXTGate, MeSHer
 - Idea: each gene is associated with a weighted profile, consisting of MeSH terms
 - Problem: not weight term count (as long as there is one co-occurrence between a term and a gene, then they are associated)

Motivations for the Current Work

- 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
- **Assumption:** the count of a term in a document follows Poisson distribution
- **Idea:** the count of a term in one gene is either from a background distribution (if the gene is unrelated) or from a positive 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 positive Poisson distributions

θ : the mixing weight of the positive 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 terms: those with a large θ (an estimate of the proportion of related genes)
- Need to assess the significance
- **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.

Implementation

- Computational framework
 - Medline collections: yeast and fly
 - Indexing, retrieval, analysis by Indri 2.4
- Background distribution of terms
 - Organism-specific distribution
- Phrase construction: bigrams only
 - NSP (Ngram statistical package) for bigram selection: chi-square measure
- Procedure
 - Retrieval of articles about the given list of genes
 - Collection of counts of all words and significant bigrams
 - Statistical analysis by Poission mixture model
 - Presentation of results: (1) sorted by statistical significance; (2) remove concepts whose fractions (θ) are below a threshold

Solve the Gene Representation Bias Problem

- Gene List: Eisen K cluster (15 genes)
 - Mainly respiratory chain complex (13), one mitochondrial membrane pore (por1 or VDAC)
 - However, VDAC matches many more articles than other genes, thus dominates the results, eg. voltage-dependent channel
- Results:
 - GESHIA: voltage-dependent, outer member, pores, channel, succinate
 - StatAnnot: electron transport, respiratory, mitochondrial membrane, oxidoreductase, succinate dehydrogenase, ubiquinone, cytochrome

Agreement with GO-based Method

- Gene List: genes up-regulated by the manganese treatment (93 genes)

| GO Theme | Related Annotator terms |
|-----------------------|---|
| neurogenesis | axon guidance, growth cone, commissural axon, proneural gene |
| synaptic transmission | synaptic vesicle, neurotransmitter release, synaptic transmission, sodium channel |
| cytoskeletal protein | alpha tubulin, actin filament |
| cell communication | tight junction, heparan sulfate proteoglycan |

Discovering Novel Themes

- Gene List: genes up-regulated by the methoprene treatment (69 genes)

| Theme | Annotator terms |
|-----------------------|--|
| muscle | flight muscle, muscle myosin, nonmuscle myosin, light chain, myosin ii, thick filament, thin filament, striated muscle |
| synaptic transmission | neurotransmitter release, synaptic transmission, synaptic vesicle |
| signaling pathway | notch signal |

Conclusion from Experiments

- Solved the biased textual representation problem of the earlier literature-based method
- In general, the new method is able to cover a large proportion of terms from GO enrichment analysis
- Supplement with additional biological concepts, including many related genes
- May be particularly useful for studying aspects not focused in GO, such as medicine

Future Plan

- Phrases
 - N-gram, $N > 2$
 - Syntactic phrase construction
 - Removal of redundant concepts in the results
- Retrieval for genes
 - Gene name normalization and disambiguation
- Entity recognition
 - Gene names, chemicals, organisms, etc.