Prediction of Operon in Prokaryotes

10/22/2008
Prediction of operons through integration of multi-source information


- Biological facts used:
  - Adjacent genes in an operon often have small distance, and sometimes small overlapping happens;
  - Genes in an operon often have similar functional category, e.g. COG or gene ontology (GO) annotations:
  - Genes in an operon tend to have similar phylogenetic profiles
The JPOP algorithm

For each adjacent pair of genes \((g_a, g_b)\) on the same strand of DNA, compute distance \(d_F(g_a, g_b)\) based on the biological feature \(F\), which potentially predict that \(g_a\) and \(g_b\) are within an operon (\(WO\) pair) or are on the boundary of two transcription units (\(TUB\) pair).

Define likelihood ratio

\[
LL(WO / d_F(g_a, g_b)) = \log \frac{P(d_F(g_a, g_b) / WO)}{P(d_F(g_a, g_b) / TUB)}
\]
The JPOP algorithm

Using intergenic distances

\[ d > 0 \quad d < 0 \]

\[ d_{\text{intergenic distance}} (g_a, g_b) = \text{base pairs between the closest ends of} \ g_a \text{ and } g_b \]
The JPOP algorithm

Phylogenetic profiles contain operon structure information

genes  Multiple microbe genomes

\[ d_{PP}(g_a, g_b) = L - (L - D_H \sqrt{E(p)}) \]  

Clustering
Prediction of operon structures
Phylogenetic profiles contain operon structure information
The JPOP algorithm

Using phylogenetic profile

$$d_{PP}(g_a, g_b) = L - (L - D_H \sqrt{E(p)});$$

$$E(p) = -p \log p - (1-p) \log(1-p).$$

Where $L$ is the length of the profile; $D$, the Hamming distance between the two profiles; $p$, the fraction of common 0’s in the two profiles.
The JPOP algorithm

Using COG functional categories

\[
P(d_{\text{COG}(g_a, g_b)}/\text{WO}) = \frac{\text{# WO pairs with the same second level of COG number}}{\text{# WO pairs}} = \%\text{WO pairs}
\]

\[
P(d_{\text{COG}(g_a, g_b)}/\text{TUB}) = \frac{\text{# TUB pairs with the same second level of COG number}}{\text{# TUB pairs}} = \%\text{TUB pairs}
\]

Frequency of adjacent pairs in COG functional categories and their log-likelihood scores

<table>
<thead>
<tr>
<th>COG functional categories</th>
<th>% WO pairs</th>
<th>% TUB pairs</th>
<th>Log-likelihoods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information storage and processing</td>
<td>0.143</td>
<td>0.074</td>
<td>0.942</td>
</tr>
<tr>
<td>Cellular processes</td>
<td>0.187</td>
<td>0.093</td>
<td>1.006</td>
</tr>
<tr>
<td>Metabolism</td>
<td>0.464</td>
<td>0.142</td>
<td>1.702</td>
</tr>
<tr>
<td>Different characterized categories</td>
<td>0.204</td>
<td>0.689</td>
<td>-1.75</td>
</tr>
</tbody>
</table>

Thus the log-likelihood is dependent on functional categories
The JPOP algorithm

The WO and TUB pairs in the training set are successfully separated in the likelihood space.
The JPOP algorithm

- Combine the three sources of information through a single layer neural network
- Train a perceptron using the *E. coli* data set
The JPOP algorithm

Summary of the prediction power of predictors and methods

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic distance</td>
<td>0.778</td>
<td>0.829</td>
<td>0.804</td>
</tr>
<tr>
<td>COG function</td>
<td>0.798</td>
<td>0.689</td>
<td>0.744</td>
</tr>
<tr>
<td>Phylogenetic profile</td>
<td>0.570</td>
<td>0.842</td>
<td>0.706</td>
</tr>
<tr>
<td>Joint prediction with equal weights</td>
<td>0.805</td>
<td>0.853</td>
<td>0.829</td>
</tr>
<tr>
<td>Joint prediction with a perceptron</td>
<td>0.824</td>
<td>0.851</td>
<td>0.838</td>
</tr>
<tr>
<td>Prediction by comparative analysis</td>
<td>0.829</td>
<td>0.721</td>
<td>0.775</td>
</tr>
</tbody>
</table>
Other operon prediction methods through information integration


The orf algorithm integrates several sources of information, including

1. Intergenic distance (ID);
2. Common functional annotation and;

Definitions

S pair: two adjacent genes are on the same strand.

O pair: two adjacent genes are in a operon.

X: a variable observed for two adjacent genes.
The orf algorithm

- Assumption 1: $|O| = 2|\text{directon}|$, so, on average an directon contains 2 operons, therefore, $|\text{SN}| = |\text{directon}|$, and,

$$P(O / S) = 1 - P(\overline{O} / S) = 1 - \frac{|\text{directon}|}{|S|}$$

- Assumption 2: the observed static $X$ has the same distribution for non operon gene pairs, $\overline{O}$, whether there are on the same strand or not.

$$P(X = x / \overline{O}) = P(X = x / \overline{S})$$
What is the probability for an S pair to be in an operon given information $X=x$?

$$P(O / S, X = x) = 1 - P(\overline{O} / S, X = x)$$

$$= 1 - \frac{P(X = x / \overline{O}, S)P(\overline{O}, S)}{P(S, X = x)}$$

$$= 1 - \frac{P(X = x / \overline{S})P(\overline{O}, S)}{P(S, X = x)}$$

$$= 1 - \frac{P(X = x / \overline{S})P(\overline{O} / S)P(S)}{P(X = x / S)P(S)}$$

$$= 1 - \frac{P(X = x / \overline{S})}{P(X = x / S)} P(\overline{O} / S)$$

$$= 1 - \frac{P(X = x / \overline{S})}{P(X = x / S)} \frac{|directon|}{|S|}$$
The orf algorithm

What is the probability for an $S$ pair to be in an operon given $ID = x$

$$P(O / S, ID = x) = 1 - \frac{P(ID = x / \bar{S})}{|directon|} \frac{\mid directon \mid}{P(ID = x / S)} \mid S \mid$$
The orf algorithm

What is the probability for an $S$ pair to be in an operon given the length of the longest common substring in their one line annotation, $CL = y$:

$$P(O / S, CL = y) = 1 - \frac{P(CL = y / \bar{S}) |direction|}{P(CL = x / S) |S|}$$

Define a informant genome: $\rho = (n-b)/n$,

$n$ is the total number of orthologous genes between the target and informant;

$b$ is the breakpoint distance between the two genomes, defined as the number of pairs of adjacent genes in the informant whose orthologs are not adjacent, or whose relative orientation is not preserved, in the target.

Keep only those informants with $\rho > \beta$. 
The orf algorithm

- Define a conserved gene cluster using orthologous relationship:

Target genome

Orthology by BDBH,
E value $< \tau = 10^{-5}$

Informatnt Genome

Distance $< \varphi = 4$
The orf algorithm

- Define a conserved gene cluster using homologous relationship:
  1. For each gene in a direction, find all homologous hits in the informant, with E value < $\tau = 10^{-5}$;
  2. Connect the two hits if they are in the same direction separated by fewer than < $\varphi = 4$ genes.
  3. Find all the chains for each direction by depth-first search starting from every nodes without an incoming edge.
  4. Find the a window $w_1$ of the direction in the target and a window $w_2$ in the informant, such that $w_1$ and $w_2$ share maximal number ($k$) of homologous genes.
The orf algorithm

- Define a conserved gene cluster using homologous relationship:

  - The target genome and the informant genome share \( n \) genes;
  - The window \( w_1 \) in target genome = \{13, 14, 15, 16, 17\};
  - The window \( w_2 \) in the informant genome = \{47, 48, 49, 50, 51\}, which are homologous to 4 genes in \( w_1 \), therefore number of matched genes \( k = 4 \).
The orf algorithm

- Define the statistical significant of a gene cluster:
  Assume the target and informant share $n$ orthologous genes. Given a direction $w_1$ of size $m$ in the target and a direction $w_2$ of $r$ genes in the informant, the probability $P_e(n,k,m,r)$ that $w_2$ contains exactly $k$ orthologous genes from $w_1$ in the same relative order is given by the hypergeometric distribution:

$$P_e(n,k,m,r) = \frac{\binom{m}{k} \binom{n-m}{r-k}}{\binom{n}{r} k!}$$

- The numerator counts the number of ways to divide the $r$ genes of $w_2$ into $k$ genes with matches in $w_1$ and $r - k$ non-cluster-associated genes not from $w_1$. The denominator counts the total number of ways to choose the $r$ genes in $w_2$, while $k!$ is the number of possible ways to order the $k$ cluster-associated genes of $w_2$. 

The orf algorithm

What is the probability for an S pair (A,B) to be in an operon given their respective homologs are from a gene cluster C(p) with \( p < e \)?

\[
P(O / S, C(p)) = P(O, F = 1 / S, C(p)) + P(O, F = 0 / S, C(p))
\]
\[
= P(F = 1 / S, C(p))P(O / S, F = 1) + P(F = 0 / S, C(p))P(O / S, F = 0)
\]
\[
= (1 - p) + P(O / S, F = 0)[1 - P(F = 1 / S, C(p))]
\]
\[
= (1 - p) + P(O / S)[1 - (1 - p)]
\]
\[
= (1 - p) + pP(O / S)
\]

\[
F = \begin{cases} 
1, & \text{if } p < e \\
0, & \text{if } p > e 
\end{cases}
\]
The orf algorithm

What is the probability for an \( S \) pair \((A,B)\) to be in an operon given three kinds of information:

\[
P(O / S, X_1 = x_1, \ldots, X_3 = x_3)
\]

\[
= \frac{P(X_1 = x_1, \ldots, X_3 = x_3 / S, O)P(O / S)}{P(X_1 = x_1, \ldots, X_3 = x_3 / S)}
\]

\[
= \prod_{i=1}^{3} \frac{P(X_i = x_i / S, O)P(O / S)}{P(X_1 = x_1, \ldots, X_3 = x_3 / S)}
\]

Using Bayes' rule, we have,

\[
P(X_i = x_i / S, O) = \frac{P(O / S, X_i = x_i)P(X_i = x_i / S)}{P(O / S)}
\]
The orf algorithm

What is the probability for an $S$ pair $(A,B)$ to be in an operon given three kinds of information:

$$P(O / S, X_1 = x_1, ..., X_3 = x_3)$$

$$\prod_{i=1}^{3} \left[ \frac{P(O / S, X_i = x_i)P(X_i = x_i / S)}{P(O / S)} \right] P(O / S)$$

$$= \prod_{i=1}^{3} P(O / S, X_i = x_i)P(X_i = x_i / S) \cdot \gamma.$$ 

$\gamma$ is an constant, and can be estimated from the fact that

$$P(O / S, X_1 = x_1, ..., X_3 = x_3) + P(\overline{O} / S, X_1 = x_1, ..., X_3 = x_3) = 1.$$
The orf algorithm

- Performance:
  True positive rate of 88% at 20% false positives.
Other operon prediction methods through information integration


- Use Bayesian networks to integrate different information:
  1. Codon usage statistics
  2. Intergenic distance;
  3. Gene expression profiles

- Identify over 78% of its operons at a 10% false positive rate.
Other operon prediction methods through information integration


- A graph theoretic algorithmic approach based on comparative genomics to identify clusters of conserved genes independent of intergenic distance and conservation of gene order.

- Then, intergenic distance distributions of operon pairs for any arbitrary prokaryotic genome can be inferred and used for operon predictions.

- For *E. coli*, the algorithm predicts 854 conserved adjacent pairs with a precision of 85%.
Other operon prediction methods through information integration


- Use a Bayesian hidden Markov model that incorporates
  1. Comparative genomic data with traditional predictors;
  2. Intergenic distances;

- It can be applied to essentially every gene in any sequenced bacterial genome.

- Does not depend on a specific relationship between the target genome and the comparative set.
Other operon prediction methods through information integration


- Apply a statistical model using logistic regression to predict operons in *Mycobacterium tuberculosis*. The model incorporates

  1. Intergenic distance;

  2. Correlation of gene expression calculated for adjacent gene pairs from over 474 microarray experiments with *Mycobacterium tuberculosis* RNA.