searching for functional divergence in genomes and metagenomes

Joseph P. Bielawski
Department of Biology
Department of Mathematics & Statistics
Dalhousie University

1. How do we efficiently extract information from large and complex datasets?
2. How to identify information (signals) most relevant to function?
gene sequences

large scale DNA sequencing project

organismal gene sequences

gene sequences

environmental gene sequences

metagenomic inference

metagenomic inference

function / phenotype

function / phenotype

(part 1 of this talk)

(part 2 of this talk)

13-01-14

part 1: Searching for functional divergence metagenomics

metagenomics

statistical modeling metabolism as a hierarchical mixture

samples

metabosystems

subnetworks

reactions

metabolic reactions

network model

metabolomics:

- sequence \( \rightarrow \) E.C. \( \rightarrow \) reaction \( \rightarrow \) substrate-product pairs
- intermediates are both substrate & product
- sequence reads \( \rightarrow \) weighted abundance

metabolic subsystems

subnetworks = subsets of reactions (S & P pairs)
- conditional dependence of reactions
- similar relative abundance (weights)
- could be functionally related or a pathway

samples

metabosystems

subnetworks

reactions
network model

subnetworks: mixtures of reactions

- reaction membership: ≥ one subnetwork
- reactions have mixing (posterior) probabilities for different subnetworks
- "soft" boundaries

samples

metabosystems

subnetworks

reactions


network model

metabosystems: mixtures of subnetworks

- subnetwork membership: ≥ one metabosystem
- subnetworks have mixing probabilities
- "soft" boundaries

samples

metabosystems

subnetworks

reactions


network model

microbiome samples: mixtures of metabosystems

- metabosystem membership: ≥ one sample
- metabosystems have mixing probabilities

samples

metabosystems

subnetworks

reactions


network model

analysis goals:

1. Learn composition of metabosystems within samples
2. Discover discriminatory and core subnetworks

Healthy: metabosystem 1 >> metabosystem 2

Disease: metabosystem 1 << metabosystem 2

Hypothetical mixture for K = 3
\[ \phi = (0.2, <0.01, 0.8) \]

Hypothetical mixture for L = 10
\[ \phi = (0.01, <0.01, 0.2, <0.01, <0.01, 0.1, <0.01, <0.01, 0.8, <0.01) \]
Hypothetical results:

- Carnivore
  - Type 2: S 11
  - Type 3: S 17

- Herbivore
  - Type 2: S 11
  - Type 3: S 17

Muegge et al. (2011)
- Fecal samples: 39 mammals
- Carnivores, herbivores, omnivores (K = 3)
- We obtained: 2,824 reactions
  - Abundance: 17,746 to 181,626 counts/sample

Carnivores: high membership of metabolosystem 1

Mammalian gut microbiomes

Discriminatory subnetworks

Discriminatory for carnivores

- Importation of certain extracellular saccharides
  - N-acetylmuramic acid
  - N-acetylgalactosamine
  - Fucose
  - Glucose
  - Mannose
- Good for exploiting alternate carbohydrate sources
  - Bacterial cell walls: N-acetylmuramic acid and N-acetylgalactosamine
  - Host secretions: fucosylated mucins

OTU network

Dataset 1

Counts

Counts
Human gut microbiomes

Gin et al. (2010)
- 124 adult humans
- 367.7 Gb of sequence data (4.5 Gb/sample)
- healthy, Crohn’s, ulcerative colitis (k = 3)
- we obtained: 3,433 reactions
- abundance: 5,948 to / 27040F counts/sample

discriminatory subnetworks for IBD
- Enriched for some PTS genes
  - suggests greater dependence on host-derived glycans for energy
    - mucus
    - shed-epithelial cells
  - suggests resistance to dietary changes
  - suggests close proximity to host cells
    - metabotype 2: greater impact on state of intestinal health
  - no apparent connection to cause of gut inflammation of IBD
  - other studies find enrichment of PTS genes in IBD patients

network model
- investigating community metabolic function
  - any environment (soil, ocean, biofilm reactor, ...)
  - at scale (distance, time, ...)
  - “soft” pathways, systems & metabosystems
  - metatranscriptomes of un-cultivable microbes
  - “superim” should improve power
  - number of reactions >> number of samples
  - depends on data quality & processing
Adapted from: http://legacy.camera.calit2.net/education/what-is-metagenomics

What are they doing?

Who is there?

metagenomic inference
joint use of both sources of info ...

let’s take a break: ~20 mins?

part 2: searching for functional divergence in genes and genomes

1. introduction (to ω ratio and codon models)
2. selected codon model
3. model based inference
4. gene-level example
5. genome-scale analyses

index of natural selection pressure: ω ratio

Kimura (1968)

$d_S$: number of synonymous substitutions per synonymous site ($K_S$)

$d_N$: number of nonsynonymous substitutions per nonsynonymous site ($K_A$)

the ratio: $d_N / d_S$; it measures selection at the protein level

Copyright © 2009 Paul O. Lewis
A model of sequence evolution: DNA

DNA state space

DNA transition matrix

The CODON matrix is huge

The Genetic Code

First 12 nucleotides at the 5' end of the 
rbcL gene in corn:

coding strand

template strand

mRNA

translation

DNA double helix

polypeptide

transcription

Codon models treat codons as the independent units, not individual nucleotide sites.

universal code: 61 sense codons

codon matrix: 61 × 61 = 3,721 matrix elements

Note: analysis is typically done by using an unrooted tree

\[ L_{0}(TTG, TTC) = \sum_{i} \pi_i P_{\text{freq}}(t_i) P_{\text{freq}}(t_i) \]

Note: analysis is typically done by using an unrooted tree
There are MANY different models we will look at three types of models:

1. selection varies over time
2. selection varies among sites
3. selection varies among sites AND over time

1. branch models: $\omega$ varies over time

2. site models: $\omega$ varies among sites

we made some progress ...
3. branch-site model: $\omega$ varies among sites & branches

$$P(x_i) = \sum_{\omega} P(x_i | \omega)$$

$\omega = 0.01$, $\omega = 0.90$, $\omega = 5.55$  

Foreground branches = $0.01$ or $0.90$  
Background branches = $(3.05 \times \omega)$

foreground = foreground branches only  
$\omega$ for background branches are from site-classes 1 and 2

analytical tasks

Task 1. parameter estimation (e.g., $\omega$)  
Task 2. hypothesis testing  
Task 3. make predictions (e.g., sites having $\omega > 1$)

maximum likelihood

• we let the data tell us what the value of a parameter (e.g., $\omega$) should be  
• the optimal value for a parameter is the value that maximizes the probability of observing the data

Task 1

maximize likelihood and probability are inverted

Task 1

maximum likelihood

Parameters: $t$ and $\omega$  
Gene: acetylcholine $\alpha$ receptor

Data

| Hypotheses | Data | $P(H)$  
|------------|------|---------|
| $H_1$: $P(H) = 1/4$ | Data 1: $1H$ & $1T$ | 0.375  
| $H_2$: $P(H) = 1/2$ | Data 2: $2H$ | 0.5  

$P(H)$ based on the data

Task 1

maximum likelihood

| Hypotheses | Data | $P(D|H)$  
|------------|------|---------|
| $H_1$: $P(H) = 1/4$ | Data 1: $1H$ & $1T$ | 0.0425  
| $H_2$: $P(H) = 1/2$ | Data 2: $2H$ | 0.25  

$P(D|H)$ based on the data

Data

| Hypotheses | Data | $P(D|H)$  
|------------|------|---------|
| $H_1$: $P(D|H) = \frac{1}{4}$ | Data 1: $1H$ & $1T$ | 0.0425  
| $H_2$: $P(D|H) = \frac{1}{2}$ | Data 2: $2H$ | 0.25  

$P(D|H)$ based on the data

Task 1

maximum likelihood
Task 2

How do we know that the estimate is significant?

Task 1. parameter estimation (e.g., $\omega$) ✔

Task 2. hypothesis testing ➙ LRT

Task 3. prediction / site identification

Task 2

Testing nested hypotheses by using the LRT

$\ell_0$ is the maximum log likelihood under $H_0$ with parameters $\theta_0$

and

$\ell_A$ is the maximum log likelihood under $H_A$ with parameters $\theta_A$

Test statistic: $2\Delta \ell = 2(\ell_0 - \ell_A)$

Degrees of freedom = difference in the number of parameters between the two models.

Task 2

LRT: Does selection pressure vary among sites?

$H_0$: uniform selective pressure among sites (M0)

$H_A$: variable selective pressure among sites (M3)

Compare $2\Delta \ell = 2(\ell_0 - \ell_A)$ with a $\chi^2$ distribution

$\hat{\omega} = 0.65$

$\hat{\omega} = 0.01$ $\hat{\omega} = 0.90$ $\hat{\omega} = 5.55$

Task 2

LRT: Have some sites evolved under positive selection?

$H_0$: variable selective pressure but NO positive selection (M1)

$H_A$: variable selective pressure with positive selection (M2)

Compare $2\Delta \ell = 2(\ell_1 - \ell_A)$ with a $\chi^2$ distribution

$\hat{\omega} = 0.5$ (M3)

$\hat{\omega} = 3.25$

Task 3

How do we identify the selected sites?

Task 1. parameter estimation (e.g., $\omega$) ✔

Task 2. hypothesis testing ✔

Task 3. prediction / site identification ➙ Bayes’ rule

Task 3

Which sites have $\omega > 1$?

Bayes’ rule:

sites 4, 12 & 13

structure:

sites are in contact

model:

5% have $\omega > 1$
**Task 3**

Bayes’ rule for identifying selected sites

- Site class 0: $\omega_0 = 0.03$, 85% of codon sites
- Site class 1: $\omega_1 = 0.40$, 10% of codon sites
- Site class 2: $\omega_2 = 5$, 0.5% of codon sites

**Bayes’ rule:**

$$P(\omega_k | x_h) = \frac{P(x_h | \omega_k) P(\omega_k)}{P(x_h)}$$

NOTE: The posterior probability should NOT be interpreted as a "$P$-value"; it can be interpreted as a measure of relative support.

---

**gene-level example**

colour diversity of coral pigments

---

**gene-level example**

green fluorescent proteins (GFPs) in corals

questions we have:

1. What is the intensity of selection on coral GFPs?
2. Have there been episodes of positive selection during the evolution of colour diversity?
3. Are some sites in GFPs positively selected?
4. Which sites?
5. What happens to the colour when the amino acid is changes at these sites?

---
Bacteria were engineered to express the extant and ancestral GFP-like proteins. These bacteria were then cultured in a pattern that corresponded to the GFP-LIKE gene tree just for fun.

Baye's rule:

\[
\text{Prior} \times \text{Likelihood} = \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{	ext{Evidence}}
\]

Baye's rule was applied using the above model. The model's expected value was calculated for each gene. The model was then scored to find a match with the data from the 2D protein cartography.

Foreground branch signal 2: episodic selection

• gram positive
• no capsule
• non-sporing forming
• low G+C%
• facultative anaerobes
• rod shaped
• opportunistic pathogens

• genus contains pathogenic and non-pathogenic species
• lineages of L. monocytogenes differ in (i) population structure, (ii) ability to respond to stress, and (iii) pathogenicity
• L. monocytogenes is serious food-borne pathogen; hence, several genomes

**lineage analysis**

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Species</th>
<th>Pathogenicity</th>
<th>Env.</th>
<th>%GC</th>
<th>ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lm-L2 1/2a</td>
<td>L. monocytogenes</td>
<td>lower</td>
<td>higher</td>
<td>38%</td>
<td>2973</td>
</tr>
<tr>
<td>Lm-L2 1/2a</td>
<td>L. monocytogenes</td>
<td>lower</td>
<td>higher</td>
<td>37%</td>
<td>2971</td>
</tr>
<tr>
<td>Lm-L1 4b</td>
<td>L. monocytogenes</td>
<td>higher</td>
<td>lower</td>
<td>38%</td>
<td>3024</td>
</tr>
<tr>
<td>Lm-L1 4b</td>
<td>L. monocytogenes</td>
<td>higher</td>
<td>lower</td>
<td>37%</td>
<td>3199</td>
</tr>
<tr>
<td>Li</td>
<td>L. innocua</td>
<td>None</td>
<td>None</td>
<td>higher</td>
<td>37%</td>
</tr>
</tbody>
</table>

L. monocytogenes: L. monocytogenes, L. monocytogenes, L. monocytogenes, L. monocytogenes, L. monocytogenes, L. innocua

L. monocytogenes

L. monocytogenes

L. monocytogenes

L. monocytogenes

L. monocytogenes

L. innocua

**core genome analysis**

• 1905 "core" genes
• concatenated genes
• 82% of single core genes

mosaic genome:
• >1952 recombination (HR)
• loci w/ "extended gene pools"
genome-scale analyses
branch models: divergent selection pressure

null (1 parameter): 2 parameter alternatives: 3 parameter alternative:

alternatives (2 parameters):
H0
H1
H2
H3

outgroup(s)

Niche-2 Pathogens

patterns of divergent selection

outgroup(s)

Niche-2 Pathogens

genome-scale analyses
patterns of divergent selection

3 modules, and 3 KEGG pathways, were significantly associated with a particular pattern of selection pressure (H1, H2, or H3).

Odds of divergent selection when gene expression is altered during invasive disease

<table>
<thead>
<tr>
<th>genes</th>
<th>odds of divergent selection</th>
<th>altered vs. unaltered expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth &amp; survival</td>
<td>other</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td></td>
<td>odds ratio</td>
</tr>
</tbody>
</table>

For genes critical to intracellular growth:
The odds of divergent selection pressure was significantly higher if a gene had altered expression during intracellular growth.

genome-scale analyses
divergent selection & gene expression

The odds of divergent selection pressure was significantly higher if a gene had altered expression during intracellular growth.

"We are drowning in information and starving for knowledge."
—Rutherford D. Roger