searching for functional divergence in genomes and metagenomes

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

molecules
gene sequences

1. How do we efficiently extract information from large and complex datasets?

2. How to identify information (signals) most relevant to function?
part 1: Searching for functional divergence
metagenomics
metagenomics

Adapted from: http://legacy.camera.calit2.net/education/what-is-metagenomics

Microbiome

Community DNA mixture?

Genomic DNA

Random

Small-insert Library Cloning

Random End Sequencing

Assembly

Predict & Analyze Genes

*What are they doing?*

*Who is there?*

network model

statistical modeling metabolism as a hierarchical mixture

**samples**

**metabosystems**

**subnetworks**

**reactions**
metabolic reactions

metagenomics:
• sequence $\rightarrow$ E.C. $\rightarrow$ reaction $\rightarrow$ substrate-product pairs
• intermediates are both substrate & product
• sequence reads $\rightarrow$ weighted abundance

subnetworks

metabosystems

samples

reactions

network model

metabolic subsystems

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

subnetworks: mixtures of reactions

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

samples

metabosystems

subnetworks

reactions

R00342 has high probability of belonging to subnetwork 1 and 2

network model

metabosystems

samples

metabosystems

subnetworks

reactions

subnetwork 1

subnetwork 2

subnetwork 3
network model

**metabosystems:** mixtures of subnetworks

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

network model

**microbiome samples:** mixtures of metabosystems

- metabosystem membership: ≥ one sample
- metabosysstems have mixing probabilities
microbiome samples: mixtures of metabosystems

samples ➔ metabosystems ➔ subnetworks ➔ reactions

healthy:

metabosystem 1 ➔ metabosystem 2

disease:

metabosystem 1 ➔ metabosystem 2

analysis goals

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

hypothesised mixture for $K = 3$

$\theta_i = (0.2, <0.01, 0.8)$

hypothesised mixture for $L = 10$

$\phi_j = (<0.01, <0.01, 0.2, <0.01, <0.01, <0.01, 0.1, <0.01, <0.01, 0.7, <0.01)$
Interpreting output

Hypothetical results:
N = 15 samples
K = 3 metabolotypes
L = 10 subsystems

Sample 6:
metabolotype 1: 0.1
metabolotype 2: 0.7
metabolotype 3: 0.2

Dataset 1

Mammalian gut microbiomes

Muegge et al. (2011)
- fecal samples: 33 mammals
- carnivore, herbivores, omnivores (K = 3)
- we obtained: 2,824 reactions
  - abundance: 11,744 to 161,626 counts/sample
dataset 1

carnivores: high membership of metabosystem 1

N = 33 samples
K = 3 metabosystems
L = 100 subnetworks

discriminatory subnetworks

Index of relative abundance

subnetworks

subnetworks
mammalian subnetworks

**Core:** correspond KEGG pathways

**Discriminatory:** do not correspond to KEGG pathways

herbivore: S40

carnivore: S49

discriminatory for carnivores

**subnetwork 49**

- importation of certain extracellular saccharides
  - N-acetylmuramic acid
  - N-acetylglucosamine
  - fucose
  - glucose
  - mannose

- good for exploiting alternate carbohydrate sources
  - bacterial cell walls: N-acetylmuramic acid and N-acetylglucosamine
  - hosts secretions: fucosylated mucins
Human gut microbiomes

Qin et al. (2010)

- 124 adult humans
  - 576.7 Gb of sequence data (4.5 Gb/sample)
  - healthy, Crohn’s, ulcerative colitis (K = 3)
  - we obtained: 3,433 reactions
  - abundance: 5,968 to 1,270,409 counts / sample

N = 124 samples
K = 3 metabosystems
L = 100 subnetworks
• Enriched for some PTS genes
  • suggests greater dependence on host-derived glycans for energy
    • mucin
    • shed-epithelial cells
  • suggests resistance to dietary changes
  • suggests close proximity to host cells
    • metabolotype 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
dataset 2

discriminatory for IBD

subnetwork 64
- enriched in ascorbate metabolism reactions
- metabolic profiling: IBD patients deficient in ascorbate
- ascorbate is an antioxidant
  - chronic IBD pathology related to excess reactive compounds
    - leads to oxidative stress within mucosa
    - IBD is associated with antioxidant levels and oxidative stress markers
  - absorption in gut depends on local gradient
  - metabotype 2 could interfere with human absorption

investigating community metabolic function

- any environment (soil, ocean, biofilm reactors, ...) or scale (distance, time, ...)
- “soft” pathways, systems & metabosystems
- metatranscriptomes of un-cultivatable microbes
- “supervision” 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?

let’s take a break: ~20 mins?
part 2: searching for functional divergence in genes and genomes

overview

1. introduction (to $\omega$ 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: $\omega$ ratio

Kimura (1968)

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

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

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

rate ratio | mode | example
--- | --- | ---
$\omega < 1$ | purifying (negative) selection | histones

$\omega = 1$ | Neutral Evolution | pseudogenes

$\omega > 1$ | Diversifying (positive) selection | MHC, Lysin
a model of sequence evolution: DNA

DNA state space

DNA transition matrix

* This is equivalent to most simple model for DNA (Jukes and Cantor, 1969).
a model of sequence evolution: CODON

<table>
<thead>
<tr>
<th>From codon below:</th>
<th>TTT (Phe)</th>
<th>TIC (Phe)</th>
<th>TTA (Leu)</th>
<th>TIG (Leu)</th>
<th>CTT (Leu)</th>
<th>CTC (Leu)</th>
<th>GGG (Gly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT (Phe)</td>
<td>---</td>
<td>aTTC</td>
<td>aTIA</td>
<td>aTG</td>
<td>aTTT</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TIC (Phe)</td>
<td>kTTT</td>
<td>---</td>
<td>aTTA</td>
<td>aTGG</td>
<td>0</td>
<td>aTCT</td>
<td>0</td>
</tr>
<tr>
<td>TTA (Leu)</td>
<td>aTTT</td>
<td>aTTA</td>
<td>---</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>0</td>
</tr>
<tr>
<td>TIG (Leu)</td>
<td>aTTT</td>
<td>aTTA</td>
<td>kTTA</td>
<td>---</td>
<td>0</td>
<td>---</td>
<td>0</td>
</tr>
<tr>
<td>CTT (Leu)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>kTCT</td>
<td>---</td>
<td>0</td>
</tr>
<tr>
<td>CTC (Leu)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>---</td>
<td>0</td>
</tr>
<tr>
<td>GGG (Gly)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>---</td>
</tr>
</tbody>
</table>

* This is equivalent to the codon model of Goldman and Yang (1994). Parameter $\omega$ is the ratio $d_{v}/d_{s}$, $\kappa$ is the transition/transversion rate ratio, and $\pi_j$ is the equilibrium frequency of the target codon $j$. 

\[ P(t) = \{p_j(t)\} = e^{Qt} \]
compute the probability of codons

\[ L_k(TTG, TTC) = \sum_k \pi_k p_{TTG}(t_0) p_{TTC}(t_1) \]

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

Positive selection sites here

\[ \omega_3 \]

Loop structures extend into extra-cellular space: Hydrophilic amino acids here

Cell membrane in grey; helix structures span the membrane: Hydrophobic amino acids here

Loop structures extend into cytoplasm: Hydrophilic amino acids here

\( \omega_0 \)

\( \omega_1 \)

\( \omega_0 \)

Purifying: \( d_{\text{u}}/d_s = 0.01 \)
Neutral: \( d_{\text{u}}/d_s = 1 \)
Adaptive: \( d_{\text{u}}/d_s = 2 \)
2. site models: $\omega$ varies among sites

$$P(x_h) = \sum_{i=0}^{K-1} p_i P(x_h | \omega_i)$$

-we made some progress ....

1. branch models
   ($\omega$ varies among branches)

2. site models
   ($\omega$ varies among sites)

3. branch-site models
   (combines the features of above models)
3. branch-site model: \( \omega \) varies among sites & branches

\[
P(x_h) = \sum_{i=0}^{K-1} p_i P(x_h | \omega_i)
\]

\( \omega = 0.01 \)  \( \omega = 0.90 \)  \( \omega = 5.55 \)

\( \omega \) for background branches are from site-classes 1 and 2 (0.01 or 0.90)

model based inference
analytical tasks

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

Task 2. hypothesis testing

Task 3. make predictions (e.g., sites having $\omega > 1$)

• 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
maximize likelihood and probability are inverted

<table>
<thead>
<tr>
<th>Probabilities</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypotheses</td>
<td>Data-1: 1H &amp; 1T</td>
</tr>
<tr>
<td>H₁: P(H) = 1/4</td>
<td>0.375</td>
</tr>
<tr>
<td>H₂: P(H) = 1/2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[ P(H | D) = \alpha P(D | H) \]

proportionality constant

<table>
<thead>
<tr>
<th>Likelihoods</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypotheses</td>
<td>Data-1: 1H &amp; 1T</td>
</tr>
<tr>
<td>H₁: P(H) = 1/4</td>
<td>(\alpha_1 \times 0.375)</td>
</tr>
<tr>
<td>H₂: P(H) = 1/2</td>
<td>(\alpha_1 \times 0.5)</td>
</tr>
</tbody>
</table>

Task 1

maximum likelihood

Parameters: t and \(\omega\)
Gene: acetylcholine \(\alpha\) receptor

\(\ln L = -2399\)
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 $\theta_0$

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

Test statistic $= 2\Delta \ell = 2(\ell_0(\theta_0) - \ell_A(\theta_A))$

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 \( 2\Delta l = 2(l_1 - l_0) \) with a \( \chi^2 \) distribution

\( H_0: \) Model 0  
\( \hat{\omega} = 0.65 \)

\( H_A: \) Model 3  
\( \hat{\omega} = 0.01 \)  
\( \hat{\omega} = 0.90 \)  
\( \hat{\omega} = 5.55 \)

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 \( 2\Delta l = 2(l_1 - l_0) \) with a \( \chi^2 \) distribution

\( H_0: \) Model 1a  
\( \hat{\omega} = 0.5 \)  
(\( \omega = 1 \))

\( H_A: \) Model 2a  
\( \hat{\omega} = 0.5 \)  
(\( \omega = 1 \))  
\( \hat{\omega} = 3.25 \)
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

Which sites have $\omega > 1$?

model:
5% have $\omega > 1$

Bayes’ rule:
site 4, 12 & 13

structure:
sites are in contact
Task 3

Bayes’ rule for identifying selected sites

- Site class 0: $\omega_0 = .03$, 85% of codon sites
- Site class 1: $\omega_1 = .40$, 10% of codon sites
- Site class 2: $\omega_2 = 14$, 05% of codon sites

Bayes’ rule:

$$P(\omega_2 \mid x_h) = \frac{p_2 P(x_h \mid \omega_2)}{P(x_h)}$$

NOTE: The posterior probability should NOT be interpreted as a "P-value"; it can be interpreted as a measure of relative support.
Is color diversity tuned by natural selection?
Is there a relationship between color and endosymbiotic algae?
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 changes at these sites?

Bayes’ rule:

\[
P(\omega_{2} \mid \mathbf{x}) = \frac{p_{2}P(x_{2} \mid \omega_{2})}{P(x_{2})} = \frac{p_{2}P(x_{2} \mid \omega_{2})}{\sum_{i} p_{i}P(x_{2} \mid \omega_{i})}
\]

*signal 1: long term (diversifying) selection*

sites in red correspond to the protein-binding region of non-colored homologs of these GFP proteins
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.
genome-scale analysis

still about genotype-phenotype

gene analysis

3D protein cartography

genomic analysis

cartography of metabolic network
bacterial genus *Listeria*

- gram positive
- no capsule
- non-spore 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 a serious food-borne pathogen; hence, several genomes

---

**Genome-scale analyses**

**Core genome analysis**

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Serotype</th>
<th>Pathogenicity</th>
<th>Envir. S.R.</th>
<th>%GC3</th>
<th>ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lm-L2</td>
<td>1/2a</td>
<td>lower</td>
<td>higher</td>
<td>38%</td>
<td>2973</td>
</tr>
<tr>
<td>Lm-L2</td>
<td>1/2a</td>
<td>lower</td>
<td>higher</td>
<td>37%</td>
<td>2971</td>
</tr>
<tr>
<td>Lm-L1</td>
<td>4b</td>
<td>higher</td>
<td>lower</td>
<td>38%</td>
<td>3024</td>
</tr>
<tr>
<td>Lm-L1</td>
<td>4b</td>
<td>higher</td>
<td>lower</td>
<td>37%</td>
<td>3199</td>
</tr>
<tr>
<td>Li</td>
<td>None</td>
<td>None</td>
<td>higher</td>
<td>37%</td>
<td>3065</td>
</tr>
</tbody>
</table>

“core genome phylogeny”

- 1905 “core” genes
- concatenated genes
- 82% of single core genes

mosaic genome:
- > 18% recombination [HR]
- loci w/ “extended gene pools”
null (1 parameter):

$H_0: \omega_1 = \omega_2 = \omega_3$

2 parameter alternatives:

$H_1: \omega_1 \neq \omega_2 = \omega_3$

$H_2: \omega_1 = \omega_2 \neq \omega_3$

$H_3: \omega_1 = \omega_2 \neq \omega_3$

3 parameter alternative:

$H_4: \omega_1 \neq \omega_2 \neq \omega_3$

patterns of divergent selection

protein secretion

DNA recombination

cell surface proteins

adaptation to atypical conditions

$H_0$

$H_2 (12\%)$

$H_3 (13\%)$
genome-scale analyses

patterns of divergent selection

A: Gene-level data analysis

B: Phenotype-level data analysis

3 modules, and 3 KEGG pathways, were significantly associated with a particular pattern of selection pressure ($H_1$, $H_2$, or $H_3$).

divergent selection & gene expression

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

<table>
<thead>
<tr>
<th>genes</th>
<th>altered expression</th>
<th>unaltered expression</th>
<th>odds ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth &amp; survival*</td>
<td>0.506</td>
<td>0.293</td>
<td>1.72</td>
<td>0.017</td>
</tr>
<tr>
<td>other</td>
<td>0.320</td>
<td>0.392</td>
<td>0.82</td>
<td>0.1098</td>
</tr>
</tbody>
</table>

* genes in ListiList categories considered critical to intracellular growth & survival

For genes critical to intracellular growth:

The odds of divergent selection pressure was significantly higher if a gene had altered expression during intracellular growth.
adaptive phenotypes are a function of **networks of genes**

"We are drowning in information and starving for knowledge."

— Rutherford D. Roger