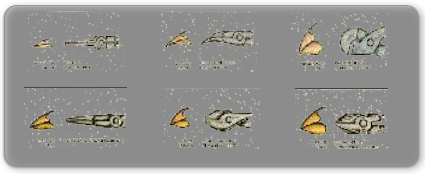
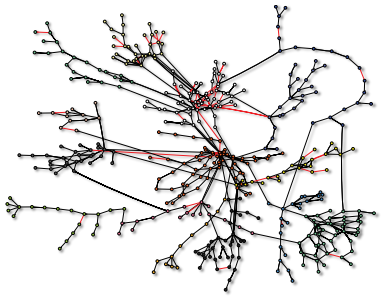
searching for functional divergence in genomes and metagenomes

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

morphology



metabolic networks



molecules

gene sequences

```
human
cow
rabbit
rat
opossum

GTG CTG TCT CCT GCC GAC AAG ACC AAC GTC AAG GCC GCC TGG GGC AAG GTT GGC GCG CAC
... ... ... G.C ... ... ... T.. ..T ... ... ... ... ... ... ... ... ... .GC A..
... ... ... ..C ..T ... ... ... ... A.. ... A.T ... ... .AA ... A.C ... AGC ...
... ..C ... G.A .AT ... ..A ... ... A.. ... AA. TG. ... ..G ... A.. ..T .GC ..T
... ..C ..G GA. ..T ... ... ..T C.. ..G ..A ... AT. ... ..T ... ..G ..A .GC ...

GCT GGC GAG TAT GGT GCG GAG GCC CTG GAG AGG ATG TTC CTG TCC TTC CCC ACC ACC AAG
... ..A .CT ... ..C ..A ... ..T ... ... ... ... ... ... AG. ... ... ... ... ...
.G. ... ... ... ..C ..C ... ... G.. ... ... ... ... T.. GG. ... ... ... ... ...
.G. ..T ..A ... ..C .A. ... ... ..A C.. ... ... ... GCT G.. ... ... ... ... ...
..C ..T .CC ..C .CA ..T ..A ..T ..T .CC ..A .CC ... ..C ... ... ... ..T ... ..A

ACC TAC TTC CCG CAC TTC GAC CTG AGC CAC GGC TCT GCC CAG GTT AAG GGC CAC GGC AAG
... ... ... ..C ... ... ... ... ... ... ..G ... ... ..C ... ... ... ... G..
... ... ... ..C ... ... ... T.C .C. ... ... ... .AG ... A.C ..A .C. ... ... ...
... ... ... T.T ... A.T ..T G.A ... .C. ... ... ... ... ..C ... .CT ... ... ...
..T ... ... ..C ... ... ... ... TC. .C. ... ..C ... ... A.C C.. ..T ..T ..T ...
```


gene sequences

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

introduction

gene sequences

large scale DNA sequencing project

organismal gene sequences → genome-scale inference → function / phenotype

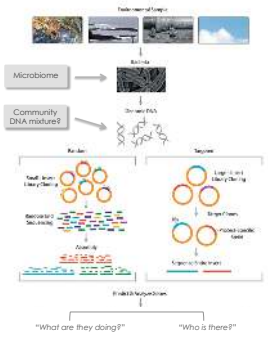
(part 2 of this talk)

environmental gene sequences → metagenomic inference → function / phenotype

(part 1 of this talk)

part 1: Searching for functional divergence metagenomics

metagenomics



network model

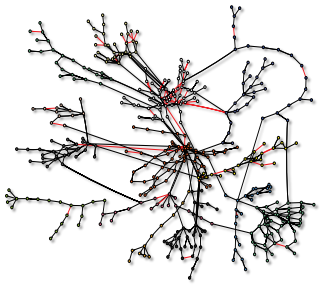
statistical modeling metabolism as a hierarchical mixture

samples

metabosystems

subnetworks

reactions



network model

metabolic reactions

samples

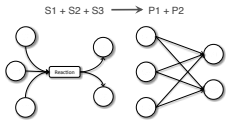
metabosystems

subnetworks

reactions

metagenomics:

- sequence → E.C. → reaction → substrate-product pairs
- intermediates are both substrate & product
- sequence reads → weighted abundance



network model

metabolic subsystems

samples

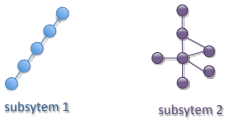
metabosystems

subnetworks

reactions

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

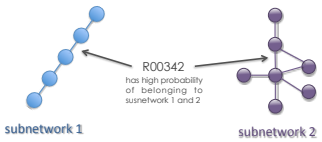
samples

metabosystems

subnetworks

reactions

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



network model

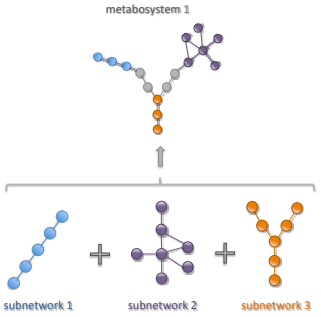
metabosystems

samples

metabosystems

subnetworks

reactions



network model

metabosystems: mixtures of subnetworks

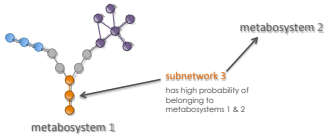
samples

metabosystems

subnetworks

reactions

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



network model

microbiome samples: mixtures of metabosystems

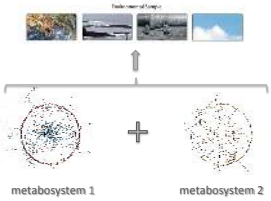
samples

metabosystems

subnetworks

reactions

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



network model

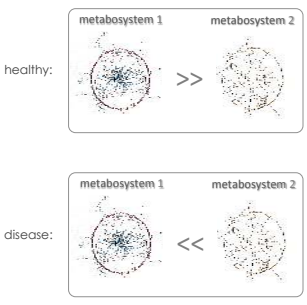
microbiome samples: mixtures of metabosystems

samples

metabosystems

subnetworks

reactions

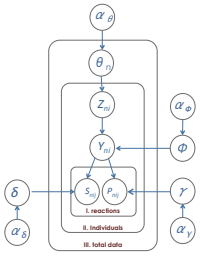


network model

analysis goals

1. Learn composition of metabosystems within samples

2. Discover discriminatory and core subnetworks



hypothetical mixture for $K = 3$

θ_j = (0.2, <0.01, 0.8)

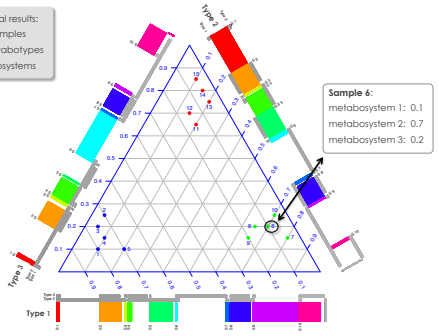
hypothetical mixture for $L = 10$

Φ_2 = (<0.01, <0.01, 0.2, <0.01, <0.01, 0.1, <0.01, <0.01, 0.7, <0.01)

network model

interpreting output

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

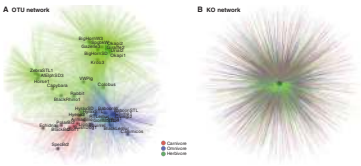


dataset 1

mammalian gut microbiomes

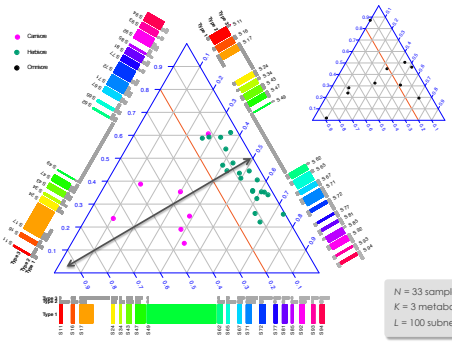
Muegge et al. (2011)

- *fecal samples: 33 mammals*
- *carnivore, herbivores, omnivores (K = 3)*
- *we obtained: 2,824 reactions*
  - *abundance: 11,744 to 161,626 counts / sample*



dataset 1

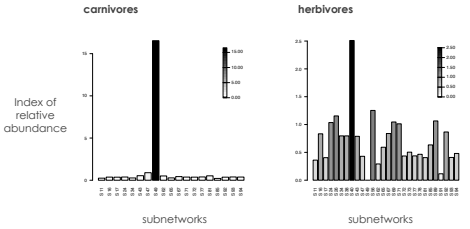
carnivores: high membership of metabosystem 1



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

dataset 1

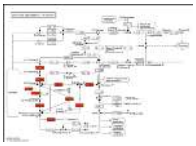
discriminatory subnetworks

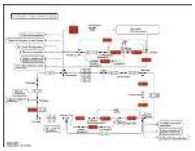


dataset 1

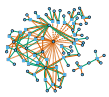
mammalian subnetworks

Core: correspond KEGG pathways

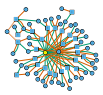




Discriminatory: do not correspond to KEGG pathways



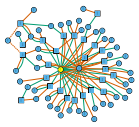
herbivore: S40



carnivore: S49

dataset 1

discriminatory for carnivores



subnetwork 49

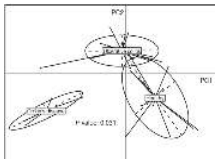
- importation of certain extracellular saccharides
  - N-acetylmuramic acid
  - N-acetylglucosamine
  - fucose
  - glucose
  - mannose
- good for exploiting alternate carbohydrate sources
  - bacterial cell walls: N-acetylmuramic acid and N-acetylglucosamine
  - hosts secretions: fucosylated mucins

dataset 2

Human gut microbiomes

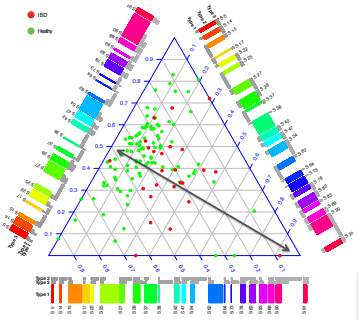
Qin et al. (2010)

- *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*



dataset 2

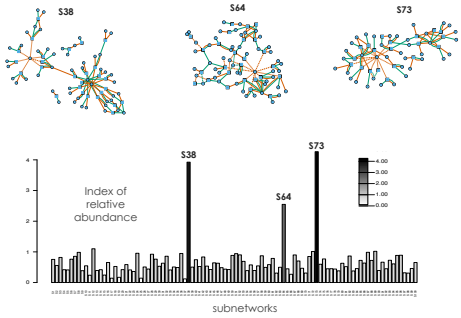
Human gut microbiomes



N = 124 samples
K = 3 metabosystems
L = 100 subnetworks

dataset 2

discriminatory subnetworks for IBD



dataset 2

discriminatory for IBD

subnetwork 38

- 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
    - 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

dataset 2

discriminatory for IBD

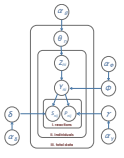
subnetwork 64

- 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

network model

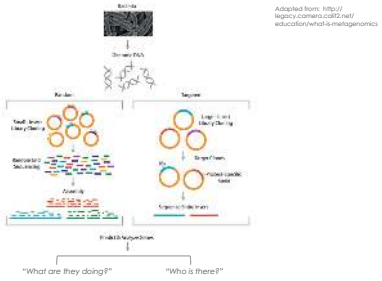
investigating community metabolic function

- 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



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

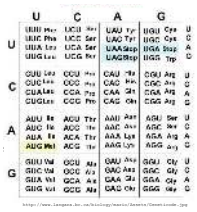
part 2: genes and genomes

overview

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

introduction

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)
- ω: the ratio d_N/d_S; it measures selection at the protein level

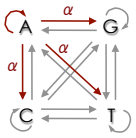
introduction

index of natural selection pressure: ω ratio

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



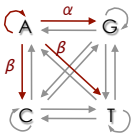
DNA transition matrix

$$\begin{array}{c|cccc} \text{from:} \backslash \text{to:} & A & C & G & T \\ A & -3\alpha & \alpha & \alpha & \alpha \\ C & \alpha & -3\alpha & \alpha & \alpha \\ G & \alpha & \alpha & -3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

* This is equivalent to most simple model for DNA (Jukes and Cantor, 1969).

a model of sequence evolution: DNA

DNA state space

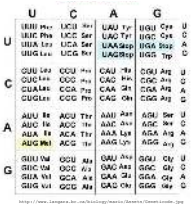


DNA transition matrix

$$\begin{array}{c|cccc} \text{from:} \backslash \text{to:} & A & C & G & T \\ A & * & \beta & \alpha & \beta \\ C & \beta & * & \beta & \alpha \\ G & \alpha & \beta & * & \beta \\ T & \beta & \alpha & \beta & * \end{array}$$

the CODON matrix is huge

universal code: 61 sense codons



codon matrix:

61 × 61 = 3,721 matrix elements

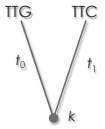
a model of sequence evolution: CODON

From codon below:	to codon below: TTT (Phe)	TTC (Phe)	TTA (Leu)	TTG (Leu)	CTT (Leu)	CTC (Leu)	⋯▸	GGG (Gly)
TTT (Phe)	——	$\kappa\pi_{TTC}$	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$	$\omega\kappa\pi_{TTT}$	0	⋯▸	0
TTC (Phe)	$\kappa\pi_{TTT}$	——	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$	0	$\omega\kappa\pi_{CTC}$	⋯▸	0
TTA (Leu)	$\omega\pi_{TTT}$	$\omega\pi_{TTC}$	——		0	0	⋯▸	0
TTG (Leu)	$\omega\pi_{TTT}$	$\omega\pi_{TTC}$	$\kappa\pi_{TTA}$	——	0	0	⋯▸	0
CTT (Leu)	$\omega\kappa\pi_{TTT}$	0	0	0	——	$\kappa\pi_{CTC}$	⋯▸	0
CTC (Leu)	0	$\omega\kappa\pi_{TTC}$	0	0	$\kappa\pi_{TTT}$	——	⋯▸	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	
GGG (Gly)	0	0	0	0	0	0	0	——

* This is equivalent to the codon model of Goldman and Yang (1994). Parameter ω is the ratio d_N/d_S. κ is the transition/transversion rate ratio, and π_i is the equilibrium frequency of the target codon (i).

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

compute the probability of codons



$$L_h(TTG,TTC) = \sum_k \pi_k p_{kTTG}(t_0)\, p_{kTTC}(t_1)$$

Note: analysis is typically done by using an unrooted tree

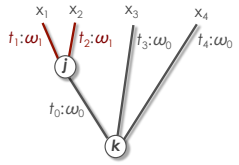
codon models

there are MANY different models

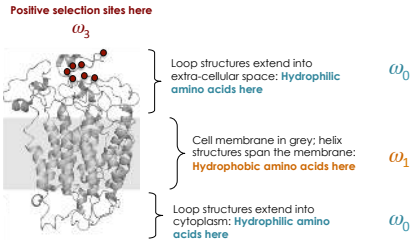
we will look at **three types of models**:

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

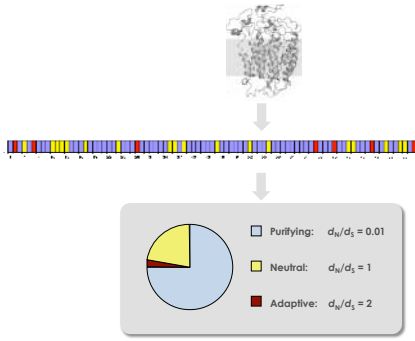
1. branch models: ω varies over time



2. site models: ω varies among sites

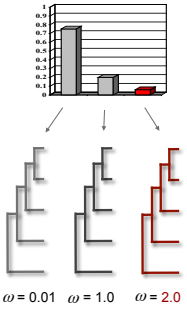


2. site models: ω varies among sites

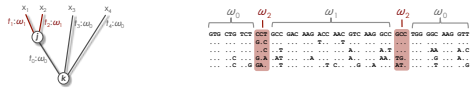


2. site models: ω varies among sites

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$



we made some progress ….



1. branch models
(ω varies among branches)

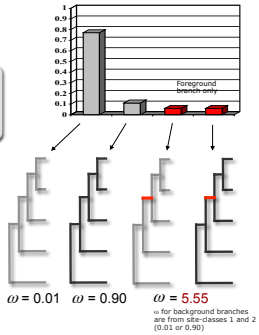
2. site models
(ω varies among sites)

3. branch-site models
(combines the features of above models)

codon models

3. branch-site model: ω varies among sites & branches

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$



model based inference

model based inference

codon models

analytical tasks

Task 1. parameter estimation (e.g., ω)

Task 2. hypothesis testing

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

Task 1

maximum likelihood

- we let the data tell us what the value of a parameter (e.g., ω) 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

Probabilities		Data	
		Data-1: 1H & 1T	Data-2: 2H
Hypotheses	H_1: P(H) = 1/4	0.375	0.0625
	H_2: P(H) = 1/2	0.5	0.25

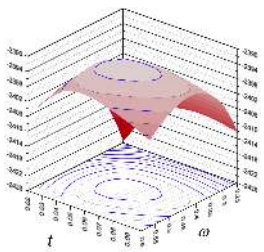
$$P(H \mid D) = \alpha P(D \mid H)$$

proportionality constant

Likelihoods		Data	
		Data-1: 1H & 1T	Data-2: 2H
Hypotheses	H_1: P(H) = 1/4	$\alpha_1 \times 0.375$	$\alpha_2 \times 0.0625$
	H_2: P(H) = 1/2	$\alpha_1 \times 0.5$	$\alpha_2 \times 0.25$

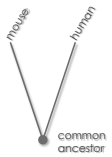
Task 1

maximum likelihood



Parameters: t and ω

Gene: acetylcholine α receptor



lnL = -2399

Task 2

How do we know that the estimate is significant?

Task 1. parameter estimation (e.g., ω) ✔

Task 2. hypothesis testing ⬅ **LRT**

Task 3. prediction / site identification

Task 2

testing nested hypotheses by using the LRT

ℓ_0 is the maximum log likelihood under H_0 with parameters θ_0

and

ℓ_A is the maximum log likelihood under H_A with parameters θ_A

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

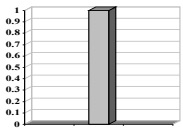
Degrees of freedom = difference in the number of parameters between the two models

Task 2

LRT: Does selection pressure vary among sites?

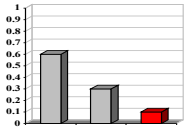
H_0: uniform selective pressure among sites (M0)
H_A: variable selective pressure among sites (M3)
Compare $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution

H_0: Model 0



$\hat{\omega}$ = 0.65

H_A: Model 3



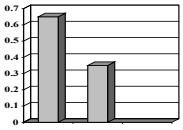
$\hat{\omega}$ = 0.01 $\hat{\omega}$ = 0.90 $\hat{\omega}$ = 5.55

Task 2

LRT: Have some sites evolved under positive selection?

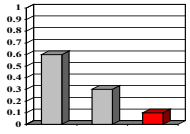
H_0: variable selective pressure but NO positive selection (M1)
H_A: variable selective pressure with positive selection (M2)
Compare $2\Delta l = 2(l_1 - l_A)$ with a χ^2 distribution

H_0: Model 1a



$\hat{\omega}$ = 0.5 (ω = 1)

H_A: Model 2a



$\hat{\omega}$ = 0.5 (ω = 1) $\hat{\omega}$ = 3.25

Task 3

How do we identify the selected sites?

Task 1. parameter estimation (e.g., ω) ✔

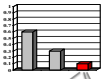
Task 2. hypothesis testing ✔

Task 3. prediction / site identification ⬅ **Bayes' rule**

Task 3

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: ω_0 = .03, 85% of codon sites

site class 1: ω_1 = .40, 10% of codon sites

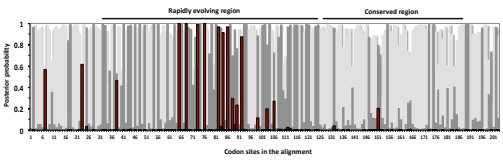
? → **site class 2: ω_2 = 14, 05% of codon sites**

Bayes' rule:

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

Task 3

Bayes' rule for identifying selected sites



Site class 0: ω_0 = .03 (strong purifying selection)

Site class 1: ω_1 = .40 (weak purifying selection)

Site class 2: ω_2 = 5 (positive selection)

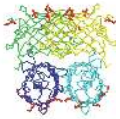
NOTE: The posterior probability should NOT be interpreted as a "P-value"; it can be interpreted as a measure of relative support

gene-level example

gene-level example

colour diversity of coral pigments

Red/blue colour morphs of the great star coal *Montastraea cavernosa*



- Is color diversity tuned by natural selection?
- Is there a relationship between colour and endosymbiotic algae?

gene-level example

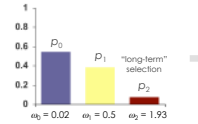
green fluorescent proteins (GFPs) in corals

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?

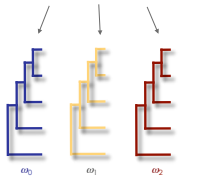
gene-level example

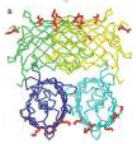
signal 1: long term (diversifying) selection



Bayes' rule:

$$P(\omega_2 \mid \mathbf{x}_h) = \frac{p_2 P(x_h \mid \omega_2)}{P(x_h)} = \frac{p_2 P(x_h \mid \omega_2)}{\sum_{i=0}^{2} p_i P(x_h \mid \omega_i)}$$

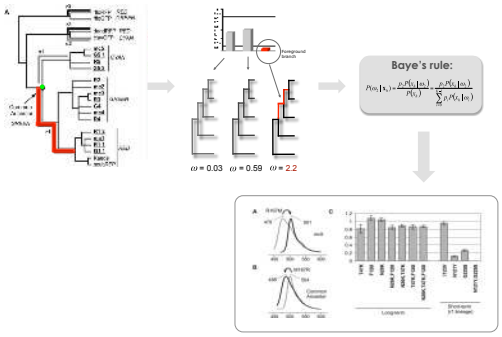




sites in red correspond to the protein-binding region of ***non-colored*** *homologs of these GFP proteins*

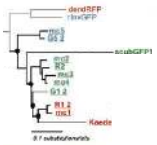
gene-level example

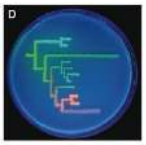
signal 2: episodic selection



gene-level example

just for fun ....





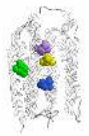
Bacteria were engineered to express the extant and ancestral GFP-like proteins. These bacteria were then cultured in a pattern that corresponded to the GFP-LIKE gene tree

genome-scale analysis

genome-scale analyses

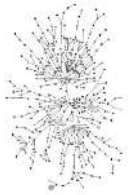
still about genotype-phenotype

gene analysis



3D protein cartography

genomic analysis

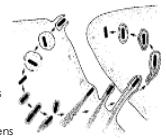


cartography of metabolic network

genome-scale analyses

bacterial genus *Listeria*

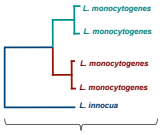
- gram positive
- no capsule
- non-spore forming
- low G+C%
- facultative anaerobes
- rod shaped
- opportunistic pathogens



- 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 serious food-borne pathogen; hence, several genomes

genome-scale analyses

core genome analysis



Lineage	Serotype	Pathogenicity	Envir. S.R.	%GC3	ORFs
Lm-L2	1/2a	lower	higher	38%	2973
Lm-L2	1/2a	lower	higher	37%	2971
Lm-L1	4b	higher	lower	38%	3024
Lm-L1	4b	higher	lower	37%	3199
Li	None	None	higher	37%	3065

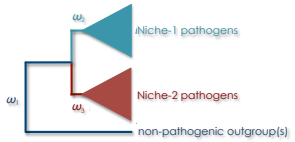
"core genome phylogeny"

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

mosaic genome:
- > 18% recombination (HR)
- loci w/ "extended gene pools"

genome-scale analyses

branch models: divergent selection pressure



null (1 parameter):

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

2 parameter alternatives:

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

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

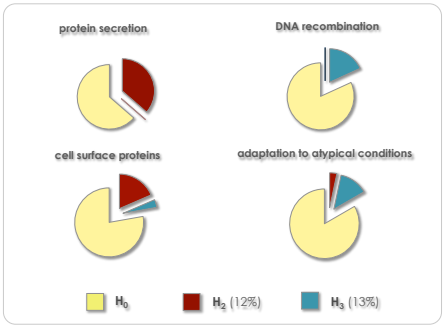
$H_3: \omega_1 = \omega_3 \neq \omega_2$

3 parameter alternative:

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

genome-scale analyses

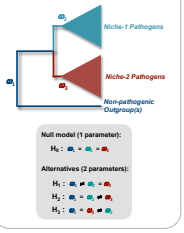
patterns of divergent selection



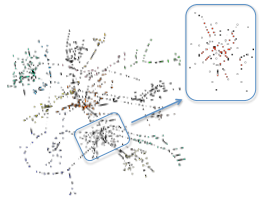
genome-scale analyses

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).

genome-scale analyses

divergent selection & gene expression

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

	odds of divergent selection		altered vs. unaltered expression	
genes	altered expression	unaltered expression	odds ratio	*P*-value
growth & survival*	0.506	0.293	1.72	0.017
other	0.320	0.392	0.82	0.1098

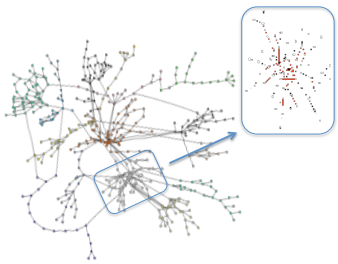
* genes in ListiList categories considered critical to intracellular growth & survival

For genes critical to intracellular growth:

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

genome-scale analyses

genotype - phenotype



adaptive phenotypes are a function of **networks of genes**

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

— Rutherford D. Roger