Phylogenomics

Gene History, Genome History and Organismal Phylogeny (genealogy of cellular life)

"Universal" Unrooted Phylogenetic Tree

MB&B 452 Genomics and Bioinformatics
Lecturer: Patrick O’Donoghue

Overview

1. Molecular Phylogenetics
Basis of Molecular Phylogenetics
   Homologous proteins, nucleic acids
   Sequence or structure based alignment
   Vertical Gene Transfer - Canonical Phylogenetic Pattern
   Horizontal Gene Transfer

2. Inferring (Computing) Phylogeny
   Algorithmic methods
   Optimization methods

3. Phylogenomics
   What is it?
   Three applications: tree of life
      pathogen evolution
      gene function prediction
   Advantages & Disadvantages of Phylogenomic Approach
   How Genome Evolution, HGT relate to cellular character.
Sequence Alignment
the basis of molecular phylogeny

Homologous protein or nucleic acid sequences from different organisms can be aligned.

Sequence similarity is assumed to be proportional to evolutionary distances.

Screen shot from the sequence alignment editor Cinema 5
http://aig.cs.man.ac.uk/research/utopia/cinema/cinema.php
Universal Phylogenetic Tree (UPT) based on the ribosome (rRNA).

The Last Universal Common Ancestor (LUCA) is represented by the base of the UPT.


Forterre, Fondation des Treilles 1996.
Orthologous relationships can reveal organismal phylogeny.

Ancient paralogous relationships indicate gene history that extends earlier than LUCA, i.e., prior to the origin of species.

Recent gene duplications can result in paralogs, which are also uninformative regarding organismal phylogeny.

Organismal phylogeny is a subset of gene history, determining which part of the genetic record tells of organismal relationships is a challenge.
A number of gene phylogenies, e.g., universal components of translation, transcription, protein secretory pathway (SecY), are congruent with rRNA.

2/3 of aaRSs show at least basal canonical pattern.

Canonical pattern recurs in aaRS phylogenies, HGT patterns are unique.

Although the phylogenetic distribution is limited for the circled genes, we can infer that these gene must have been extant prior to & in LUCA.

The Relationship Between Sequence & Structure

Sequence identity > 20%

The sequence degrades rapidly. Sequence identity signal < 10%

Structural superposition of AlaRS & AspRS.

Sequence id = 0.055, $Q_{H}=0.48$

Protein Homology in Structure and Sequence

3 homologous structures;
2 closely related (A,B),
1 more distant (C).

Overlap of protein backbones.

A: E---GARDFLV-PYRHE----------PGLFYALPQS
B: -E--GARDYLV-PSRVH---------KGKFYALPQS
C: ---DMWDTFWLT-GE--GFRLEGPLGEEVEGRLLLRTTH
Evolutionary History in Sequence & Structure
aspartyl-tRNA synthetase

Congruence justifies using structure to trace back evolutionary events beyond the reach of sequence phylogeny.

Evolution of Structure and Function in AspRS

i) class II

ii) subclass IIB
   - anticodon binding (ACB) domain

iii) LUCA AspRS

iv) bacterial AspRS

v) E. coli AspRS

Multiseq 2.0 in VMD 1.8.5

http://www.scs.uiuc.edu/~schulten/multiseq
Tree representations

These 2 trees are equivalent, line lengths represent the evolutionary distance.

The trees indicate that A, B & C have evolved at equal rates since their divergence from a common ancestor, i.e., a constant molecular clock is assumed.

Not all genes (or organisms) evolve at the same rate.

A & B share the most recent common ancestry, but B has evolved with a faster “clock” than A.

This is one reason that organisms can share a recent common ancestor, but have more distantly related genes than expected. (HGT, and loss of orthologs are other reasons.)

Adapted From Gary Olsen’s notes on classification and phylogeny: http://www.bact.wisc.edu/Bact303/phylogeny
Algorithms & Programs

2. Inferring Phylogeny

Algorithmic or Clustering Methods
Add sequences to a tree according to similarity relationships. Produces one tree.

```
    D
   / 
  C   B
 / 
A
```

Optimality Methods
Heuristic search through the space of possible trees. One tree is optimal according to:

- **Parsimony** (fewest changes)
- **Maximum Likelihood** (most probable tree)

Each alignment position is considered independently on a test tree. Branch topologies and lengths are sampled until the best tree is found.

Adapted from Gary Olsen at http://geta.life.uiuc.edu/~gary/
Algorithms & Programs

Methods That Impose a Molecular Clock

Clustering Methods
- UPGMA
- WPGMA
- Single-linkage
- Complete-linkage

Objective Criterion-Based Methods
- Least-squares distance (e.g., KITSCH)
- Maximum likelihood (e.g., dnamlk)

Methods That Do Not Impose a Molecular Clock

Neighbor-joining

Other Considerations

Substitution cost matrix. (for distance & parsimony)
- Cost of replacing one nucleic acid (or amino acid) for another.

Evolutionary models (for likelihood).
- Invariant positions, evolutionary rate heterogeneity among positions, can estimate rates of change from one base (or amino acid) to another from the alignment data.

Programs


**PAUP** [http://paup.csit.fsu.edu/](http://paup.csit.fsu.edu/)
  [http://paup.csit.fsu.edu/paupfaq/faq.html](http://paup.csit.fsu.edu/paupfaq/faq.html)

**PHYML** [http://atgc.lirmm.fr/phyml/](http://atgc.lirmm.fr/phyml/)

Adapted from Gary Olsen at [http://geta.life.uiuc.edu/~gary/](http://geta.life.uiuc.edu/~gary/)
Phylogenomics

What is Phylogenomics?

Inference of phylogeny, for some group of taxa, based on the comparative analysis of some property of genomes or gene clusters/groups.


Only 135 citations in pubmed for “phylogenomic*”.

What is being compared?

Genome Identity
Gene Presence/Absence
“Genome Conservation”
Gene expression

Why phylogenomics?

Hypothesis: Additional information from genome sequences should help resolve evolutionary histories.

i. Tree of life
ii. Tracking pathogenic lineages across a population. Identify infection source & virulence genes.
iii. Gene annotation, protein function prediction.
Tree of Life 1

Phylogeny determined by protein domain content

Song Yang*, Russell F. Doolittle*, and Philip E. Bourne**

Departments of *Chemistry and Biochemistry and **Pharmacology and San Diego Supercomputer Center, University of California at San Diego, La Jolla, CA 92037

Contributed by Russell F. Doolittle, November 26, 2004

Protein Domains and Superfamilies

PDB: 1pkp

Presence/Absence Matrix

<table>
<thead>
<tr>
<th>Protein Domain Superfamilies</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
<th>M</th>
</tr>
</thead>
</table>

Genome 1 = [0 0 0 1 1 0 1 ... 1]
Genome 2 = [0 0 0 1 1 0 0 ... 1]
...
Genome N = [1 1 1 0 1 1 1 ... 0]

Evolutionary Distance

\[ D = A' / (A' + AB) \]

A' is number of superfamily domains.
AB is the number of shared superfamily domains.

http://scop.mrc-lmb.cam.ac.uk/scop/
Genome Conservation weights the average sequence similarity with the number of homologs between two genomes.

Genome conservation produces a tree that is mainly congruent with rRNA.

Blue branches - vertical inheritance
Red branches - horizontal gene transfer

Identify Virulence Factors
tcdA & tcdB encode proteins that synthesize toxin A & B, respectively.

Apparent deletion or highly divergent sequences at the end of tcdB is specific to the hypervirulent (HY) strains.

Microarray experiment tests if DNA from other strains hybridizes (yellow) or not (blue) to C. difficile 630.

Host Source: human, mouse, bovine, swine, equine

Geographic Spread of Pathogen Lineages
“The 20 [hypervirulent] strains were from diverse locations in the United States, Canada, and the United Kingdom, confirming their transcontinental spread.”

Microarrays can be used to “type” strains.


“proteins that function together in a pathway or structural complex are likely to evolve in a correlated fashion”

“Myxococcus xanthus ORFs blasted against genomes.

Blast bit scores

Similarity between Blast hit lists.

“Mythe products of genes with similar evolutionary histories cluster together in mountains, and where local height is proportional to the density of proteins within an area.”

Function Prediction

Phylogenetic clustering of all genes in the *Myxococcus xanthus* genome.

“Mountains” include genes with similar evolutionary histories.

“M. xanthus (δ-proteo) can exist as both a single-species biofilm and a free-living cell.

The biofilm is a self-organizing predatory swarm that has many of the characteristics of a multicellular organism.”

<table>
<thead>
<tr>
<th>Putative annotation</th>
<th>Swarm expansion</th>
<th>Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MXAN1095 TonB system transport protein</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>MXAN1324 TPR domain protein</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MXAN0346 TolB protein, putative</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MXAN0275 Twitching mobility protein</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Genes with the same evolutionary history produce a similar phenotype when disrupted by plasmid insertion.

In summary, 12 of 15 ORFs targeted for disruption on the basis of phylogenomic proximity to known motility proteins produced obvious defects in swarm expansion and/or aggregation when inactivated.

Assessing Phylogenomics

Advantages

The complete information in the genome sequence can be used for constructing the tree of life, monitoring pathogen evolution, locating virulence genes and predicting gene function.

When rRNA is nearly identical, whole genome comparisons provide a more sensitive measure of relationships.

Disadvantages

Not all genes have the same history.
Incongruent gene histories result from various sources:
1. loss of close orthologs
2. horizontal gene transfer
3. lineage specific evolutionary rate acceleration.

Data Selection
“selecting only data that contain minimal nonphylogenetic signals takes full advantage of phylogenomics and markedly reduces incongruence.” (Jeffroy et al. 2006 Trends in Genetics v22)
Remove genes with different histories.

Explicit representation of Horizontal, Vertical Gene Transfer.
1/3 of Methanosarcina mazei’s genome is of bacterial origin (red).

(Black) Core energetic (methanogenesis), information processing genes, lipid membranes, and gene order are all characteristic of its archaeal relatives.

A genome may acquire a large fraction of foreign genes, without fundamentally transforming the core cellular subsystems or the cell itself into another type.
Adaptive characters, though of paramount importance to the being, are of hardly any importance in classification.

We have to discover the lines of descent by the most permanent characters, however slight their vital importance may be.

It may even be given as a general rule, that the less any part of the organisation is concerned with special habits, the more important it becomes for classification.