Understanding microbes through the lens of comparative genomics... a biased perspective

Aaron Darling
A/Prof. ithree institute - UTS
Molecular Evolution of Bacteria

- Bacteria reproduce clonally
A brief history of recombination

Joshua Lederberg
Discovered conjugative gene transfer in bacteria
Won Nobel Prize in 1958
Early work on phage transduction at UW Madison
Genomics revolutionizes our understanding of bacterial evolution

Perna *et al* 2001 *Nature*, Welch *et al* 2002 *PNAS*

Three genomes, same species, nearly identical 16S rRNA, only 40% genes in common

**What is a bacterial species??**
Need to compare genomes automatically! How??

Positional homology:
- Mauve, M-GCAT, MUMmer, Mugsy, TBA, others

Glocal:
- Enredo/PECAN, Cactus, SuperMap/Multi-LAGAN, eulerAlign, Repeatoire

Darling et al 2004, 2006, 2010
Sequence alignment

Most sequence alignment utilizes dynamic programming

**Key assumption:** each site evolves independently of neighbors (i.i.d.)

![Sequence alignment diagram]

**Substitution matrix:**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

Gap open: -2
Gap extend: -2
Scaling up to genomes: anchored alignment

Calculate only parts of the DP matrix that are very likely to be part of the optimal path

Each diagonal ‘band’ is a high-scoring local alignment of the sequences.

The highest scoring chain of local alignments become anchors

Multiple alignment: some algorithms anchor in $N$ dimensions, others progressively build from 2...$N$ dimensions
The architecture of Mauve

- 69k lines of C++ code. Organized into two development libraries plus refactored Clustal-W and MUSCLE into libraries.
- Uses: Boost, Subversion, Doxygen, valgrind
- Cross-platform autotools and Viz Studio builds
- Snapshot server runs VMs with 4 OSes to complete nightly builds
- Automated accuracy testing using cluster computing (SGE + perl + R)
- Java GUI is 31k lines of code
- 5 developers, **25 years effort** under basic COCOMO model
Mauve 1.0 could align genomes in 2004... but soon there were too many!!

Cost of genome sequencing

US$ per million nucleotides

- Introduction of 454 technology
- Illumina/Solexa/SOLiD
With many genomes, subsets become critical

Mauve 1.0 only aligned “core genome”
Many “almost core” regions conserved in some, but not all genomes
Main innovations in progressiveMauve

• Tolerate mismatches in initial alignment anchors
• Sum of pairs breakpoint score:

\[ \text{Score}(z) = \sum_{k=1}^{\mid z \mid} Q(z_i+k, z_j+k) \]

  – Score(z) is the sum of substitution scores for all sites in genome \( i \) and \( j \) that are aligned in block \( z \)
  – Penalize blocks containing repeats (not shown)
• Define a genome-pair-specific breakpoint penalty \( \beta_{ij} \)
• Add & remove blocks to OPTIMIZE:

\[ -\beta \| Z \| + \sum_{z \in Z} \text{Score}(z) \]

• Apply a HMM to determine gain/loss regions

Darling et al 2010 PLoS ONE
Alignment accuracy based on simulation studies

Blue asterisk indicates rates predicted among Enterobacteriaceae
50 large gene gain/loss events means total flux of 2Mbp in the 1Mbp genomes
10k small events means total flux of 3.56Mbp in the 1Mbp genomes
The Alignathon: more alignment accuracy metrics

Data simulated to mimic four primate genomes
Some variation in precision & recall among methods
For some applications, precision more important than recall & vice versa

Earl et al 2014, Genome Research
Alignments are hard to get right...

A: gap wander
B, C: gap attraction
D: gap annihilation

Source: Lunter et al 2008 Genome Research

Prakash and Tompa, 2007 Genome Biology: “We identify 9.7% (21 Mbp) of the human chromosome 1 alignment as suspiciously aligned.” in the UCSC genome browser
Great...so do we learn any new biology?

Some case studies:

- Genome rearrangement in *Yersinia*
- Gene gain and loss in *E. coli* and *Shigella*
- Homologous recombination in *E. coli*
Genome alignment of nine Yersinia
What inversions happened on this branch?

Use Bayesian MCMC to sample all possible inversion histories according to their probability.
Maximum a posteriori estimate of *Y. pestis* Antiqua inversion history
Maximum a posteriori estimate of *Y. pestis* Antiqua inversion history.
Rearrangement hotspots in *Yersinia*

- Few rearrangements surrounding *dif* site
- Could disorient FtsK binding sites (KOPS)
Great...so do we learn any new biology?

Some case studies:

• Genome rearrangement in *Yersinia*

• Gene gain and loss in *E. coli* and *Shigella*

• Homologous recombination in *E. coli*
Inferring rate changes in gain&loss

Phylogenetic tree from ClonalFrame on core genome
Bayesian mapping of gains, losses onto tree
Compound Poisson process

BLUE: rate of gene loss
RED: rate of gene gain
Great...so do we learn any new biology?

Some case studies:

• Genome rearrangement in *Yersinia*

• Gene gain and loss in *E. coli* and *Shigella*

• Homologous recombination in *E. coli*
Bayesian inference of homologous recombination

Given a genome alignment,
infer:
- species phylogeny
- local phylogenies
- boundaries of recombination events

Approach:
- Bayesian phylogenetic model
- Recombination are edges connecting points on species tree
- Each recombination affects a local region of the genome
- Inference via MCMC sampler ClonalOrigin

The “weak ARG” model:
Species tree in black
Recombination in red

Didelot et al 2010 Genetics, Didelot et al 2012 BMC Genomics
Inferring recombination in *E. coli*

Mauve Genome Alignment of 27 finished *E. coli* genomes
- 765 alignment blocks
- 3.22 Million alignment columns in *core genome only*
- Interrupted by gene content differences, rearrangements

+ and * are in our study

Darling, Mau, Perna 2010
Global spatial pattern of recombination

E. coli K12 red bar terminus, black bar origin
Y-axis = Average number of recombination events per site

O antigen proteins
Membrane-localizing proteins
Mystery hotspot
Mucho differential gene content
Integron??
Ribosomal protein cluster

Didelot, Meric, Falush, Darling 2012. BMC Genomics
Asymmetric gene flow across clonal lineages

Heat colors give Observed/Expected recombination events
Blue=fewer than expected, Red=more

E. coli ATCC8739
E. coli HS
E. coli BL21/DE3
E. coli B REL606
E. coli BW2952
E. coli K-12 DH10B
E. coli K-12 MG1655
E. coli K-12 W3110
E. coli IAI
E. coli SE11
E. coli 55989
E. coli O103:H2
E. coli E24377A
E. coli O111:H
E. coli O26:H11
E. coli O157:H7 EC4115
E. coli O157:H7 TW14359
E. coli O157:H7 EDL933
E. coli O157:H7 Sakai
E. coli O157:H7 Sakai
E. coli O55:H7
E. coli APEC O1
E. coli UT18927
E. coli S8822
E. coli CFT073
E. coli ED1A
E. coli 536
E. coli E2348/69

Didelot, Meric, Falush, Darling 2012. BMC Genomics
### Functional enrichment of recombination

Global over- and under-representation in the class of highly recombinogenic genes:

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Exp</th>
<th>Chi²</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>13</td>
<td>22.6</td>
<td>4.1</td>
<td>Protein related information transfer</td>
</tr>
<tr>
<td>Over</td>
<td>7</td>
<td>3.23</td>
<td>4.4</td>
<td>Metabolism of other compounds</td>
</tr>
<tr>
<td>Under</td>
<td>1</td>
<td>6.33</td>
<td>4.5</td>
<td>Energy production/transport</td>
</tr>
<tr>
<td>Under</td>
<td>26</td>
<td>40.0</td>
<td>4.9</td>
<td>Function Unknown</td>
</tr>
<tr>
<td>Over</td>
<td>1</td>
<td>0.12</td>
<td>6.2</td>
<td>Accessory Factors Involved in Transport</td>
</tr>
<tr>
<td>Over</td>
<td>51</td>
<td>30.7</td>
<td>13.5</td>
<td>Building block biosynthesis</td>
</tr>
<tr>
<td>Over</td>
<td>20</td>
<td>7.70</td>
<td>19.6</td>
<td>DNA related information transfer</td>
</tr>
</tbody>
</table>

Categories with Chi² values < 4 not shown

Darling, unpublished
Thanks to...

Nicole T. Perna  
Univ. of Wisconsin

Bob Mau  
Univ. of Wisconsin

Mark Ragan  
Univ. of Queensland

Daniel Falush  
Max Plank Institut

Xavier Didelot  
Imperial College

Daniel Lawson  
Univ. of Bristol

Erick Matsen  
FHCRC

Todd Treangen  
BNBI, Johns Hopkins

Jonathan Eisen  
UC Davis, JGI

Guillaume Jospin

Jenna Lang

Dongying Wu

Students and other staff involved in this work:
- Anna Rissman, Andrew Tritt, Paul Infield-Harm, Andres Esteban-Marcos, Tiffanie Nelson, Eric Lowe, John Zhang, David Larsen, A/Prof Marc Facciotti, A/Prof Rachel Whitaker

Open source software:
- Mauve  http://gel.ahabs.wisc.edu/mauve
- ClonalOrigin  http://code.google.com/clonalorigin
- GenoPlast  http://www.xavierdidelot.xtreemhost.com/genoplast.htm