Comparative Genomics

Background & Strategy
March 30, 2016

What is the goal?

Develop a typing scheme for various Nontypeable *Haemophilus influenzae* (NTHi) strains
How specific are we getting?
How specific are we getting?
Typing of *H. influenzae*
The classification problem
The Classification Problem

Design a clear classification system which always correctly classifies objects as being part of or not part of a group.
Why is this problem hard?
Stated simply

Design a clear classification system which *always* correctly classifies objects as being part of or not part of a group.
Becomes complex, very quickly, and subjective...
Becomes complex, very quickly, and subjective...
Becomes complex, very quickly, and subjective…

Classification requires many-dimensional analysis
Distinguishing Key Elements

Cheung, 2005
Typing scheme: Which will happen?

Serotype A
  ├── Serotype B
  │    └── Serotype C
  │        ├── NT1
  │        │    ├── NT2
  │        │    │    └── NT3
  │        │    └── NT1
  │        └── NT2
  └── NT1

Serotype B
  ├── NT1
  │    └── NT2
  └── NT3

Serotype C
  └── NT3

Serotype A
  └── NT1

NT1

NT2
Global solutions
DNA-DNA Hybridization

- DNA from two species: A, a previously typed strain, and B, an unknown, are denatured and annealed.
  - Same species hybridize over >60% of their length
  - Closely related species hybridize over ~30-60% of their length
  - Evolutionarily unrelated (very distant) strains hybridize <20% of their length

- DDH is time consuming
  - Grows by $G^3$, for G genomes
  - Sample contamination is a significant problem

- DDH doesn’t differentiate very closely related species
  - *H. influenzae* and *H. haemolyticus* are ~62% identical by DDH
DNA-DNA Hybridization

1. Heat to separate strands.
2. Combine single strands of DNA.
3. Cool to allow renaturation of double-stranded DNA.
4. Determine degree of hybridization.

Complete hybridization: organisms identical
Partial hybridization: organisms related
No hybridization: organisms unrelated

Whitman et al, 2012
Average Nucleotide Identity

- Computational method to compare genomes
- Measures average % identity between two genomes
- Correlates strongly with DDH
- Likely same species: >70% DDH = >95% ANI
ANI - Method

1. Split genomes into 1020 nt fragments
2. Determine pairwise similarities
   a. Reciprocal best hits using BLAST (ANIb)
   b. MUM distance using nucmer (ANIim)
3. Remove low-scoring fragments
   a. Fragments with <70% ID across only a small portion of its length -- i.e. unique, non-shared blocks.
4. Average percent identities
Based on orthologs
MASH - Fast genome distance estimation

- Implementation of MinHash
  - An optimized set comparison algorithm
- Large sequence sets are converted to “sketches” (>5000x compression)
  - Equivalent to 1024nt blocks in an ANI
- Sketches can be compared to estimate similarity via MinHash
- Strong correlation between Mash Distance (D) and ANI
  - \( D = 1 - \text{ANI} \)
- Fast performance - 55,000 RefSeq genomes compared in 46 CPU hours
<table>
<thead>
<tr>
<th>$S_1$: CATGGACGACCAG</th>
<th>$S_2$: GCAGTACCGATCGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT</td>
<td>GTA</td>
</tr>
<tr>
<td>GAC</td>
<td>CGA</td>
</tr>
<tr>
<td>GAC</td>
<td>CGT</td>
</tr>
<tr>
<td>ATG</td>
<td>AGT</td>
</tr>
<tr>
<td>ACC</td>
<td>CCG</td>
</tr>
<tr>
<td>ACC</td>
<td>TCG</td>
</tr>
<tr>
<td>TGG</td>
<td>CAG</td>
</tr>
<tr>
<td>CCG</td>
<td>ACC</td>
</tr>
<tr>
<td>CCA</td>
<td>ATC</td>
</tr>
<tr>
<td>GGA</td>
<td>GCA</td>
</tr>
<tr>
<td>CGA</td>
<td>TAC</td>
</tr>
<tr>
<td>CAG</td>
<td>GAT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Gamma_1$</th>
<th>$\Gamma_2$</th>
<th>$\Gamma_3$</th>
<th>$\Gamma_4$</th>
<th>$\Gamma_1$</th>
<th>$\Gamma_2$</th>
<th>$\Gamma_3$</th>
<th>$\Gamma_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>14</td>
<td>57</td>
<td>36</td>
<td>CAT</td>
<td>GCA</td>
<td>36</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td>57</td>
<td>36</td>
<td>19</td>
<td>ATG</td>
<td>CAG</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>58</td>
<td>37</td>
<td>16</td>
<td>15</td>
<td>TGG</td>
<td>AGT</td>
<td>11</td>
<td>54</td>
</tr>
<tr>
<td>40</td>
<td>23</td>
<td>2</td>
<td>61</td>
<td>GGA</td>
<td>GTA</td>
<td>44</td>
<td>27</td>
</tr>
<tr>
<td>33</td>
<td>28</td>
<td>11</td>
<td>54</td>
<td>GAC</td>
<td>TAC</td>
<td>49</td>
<td>44</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>47</td>
<td>26</td>
<td>ACC</td>
<td>ACC</td>
<td>5</td>
<td>48</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>60</td>
<td>43</td>
<td>CCG</td>
<td>CCG</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
<td>50</td>
<td>45</td>
<td>CGA</td>
<td>CGA</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>33</td>
<td>28</td>
<td>11</td>
<td>54</td>
<td>GAC</td>
<td>GAT</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>47</td>
<td>26</td>
<td>ACC</td>
<td>ATC</td>
<td>13</td>
<td>56</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>62</td>
<td>41</td>
<td>CCA</td>
<td>TCG</td>
<td>54</td>
<td>33</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
<td>56</td>
<td>39</td>
<td>CAG</td>
<td>CGT</td>
<td>27</td>
<td>6</td>
</tr>
</tbody>
</table>

min-mers
min-mers

\[ [5, 1, 2, 15] \]
Sketch \((S_1)\)

\[ [5, 1, 6, 6] \]
Sketch \((S_2)\)

\[ J(S_1, S_2) \approx \frac{2}{4} = 0.5 \]

\(S_1:\) CATGGACCGACCAG

\(S_2:\) GCAG\underline{TACCG}ATCGT
MASH correlates strongly with ANI

Ondov et al, 2015
\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|} \]
Marker based solutions
What we observe can be divided into:

What we see
What we observe can be divided into:

- **What we see**

- **Signal**

- **Noise**
**Multilocus sequence typing (MLST)**

- Described in 1998 as a portable method for characterizing bacteria from nucleic sequences (from gels)

- Uses 450-500bp fragments of common bacterial housekeeping genes

- The sequences of each fragment are compared with all the previously identified sequences (alleles) at that locus and, thereby, are assigned allele numbers at each of the seven loci.

- MLST is now widely used for molecular epidemiology as it allows strains studied by different groups to be compared and MLST schemes have been developed for ~20 bacteria (mostly pathogens)
7 housekeeping genes

**Strain A**
- **abcZ**: G
- **adk**: ACTG
- **araE**: A
- **gdh**: C
- **pdhC**: CT
- **pgm**: TG
- **fumC**: TGA

**Strain B**
- **abcZ**: C
- **adk**: CCAG
- **araE**: G
- **gdh**: T
- **pdhC**: AG
- **pgm**: TG
- **fumC**: AGA

**Strain C**
- **abcZ**: G
- **adk**: ACTA
- **araE**: G
- **gdh**: T
- **pdhC**: CT
- **pgm**: GA
- **fumC**: TGT
Select 7 housekeeping genes in h.influenzae

BLAST search genomes for housekeeping genes, compares sequences and assign allele numbers and sequence types (ST)

Conduct a phylogenetic analysis based on differences between allelic profiles or sequence types to infer genetic relatedness between isolates

Sequence samples and assemble genomes
What if?
In our case

H. influenzae
Type X

~2200 genes

Signal
In our case...  

What we see

H. influenzae Type X

~2200 genes

Signal
Do we group by present regions?

Do we separate by absent regions?

What we see

Signal

H. influenzae

~2200 genes
Where do we start?
Where do we start?
Antibody typing

Capsule protein
Capsule related genes

Satola et al., 2003
Other interesting categories

Virulence factors
- Fimbriae proteins
- Other membrane proteins

Metabolism proteins
Clustering and self organizing maps
Clustering and self organizing maps
Clustering and self organizing maps
Clustering and self organizing maps

Whole genome
Clustering and self organizing maps

Whole genome

Selected markers

Zamani et al, 2013
Where do we finish?

A serotype-like distinguishable genome; among the *H. influenzae* species characterized by a common set of physiologically relevant markers.

With a clear experimental validation protocol (e.g. PCR, antibodies)
Experimental protocol?

Forward Primer

Reverse Primer
And finally!
Conclusion: Flow Chart

1. MASH
2. ANI
3. HMM-SOM
4. MLST

Intermediate:
- Global information
- Locally based information

Final:
- Consensus
- Phylogenetic trees
References


