Detection and disambiguation of geospatial locations for phylogeography

D. Weissenbacher1, T. Tahsin1, R. Beard1,2, M. Figaro1,2, R. Rivera1, M. Scotch1,2 and G. Gonzalez1

Email: davy.weissenbacher@asu.edu

1 Department of Biomedical Informatics, ASU
2 Center for Environmental Security, Biodesign Institute, ASU
Phylogeography: virus mutation/migration

Bioinformatics 31 (12), 2015
Knowledge-driven geospatial location resolution for phylogeographic models of virus migration

Supported by NIAID - NIH Award N°R56AI102559
Resolving toponyms in publications

Pubmed Article PMID 22279582

a H1N1 virus was isolated in 2009 from a child hospitalized in Nanjing, China

Problem:
80% of name of places (toponyms) NOT precise enough for phylogeography

Organism = H1N1
Host = Homo Sapiens
Country = “China” + Nanjing

GenBank
H3N2 swine influenza virus strain that was isolated in Quebec.

GenBank Record

<table>
<thead>
<tr>
<th>PMID</th>
<th>18032</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>Canada</td>
</tr>
<tr>
<td>Strain</td>
<td>A/swine/Quebec/4001/2005</td>
</tr>
</tbody>
</table>
Detection/Disambiguation methods

- **Detection**: Dictionary-based + Rules-based approach
  - GeoNames (10 millions names)
  - Rules (Manually defined): {Not in stop list; NP starting with capital letter preceded by “in”}

- **Disambiguation heuristics**:
  - Population: choose most populated places
  - Distance: choose candidates in the smallest area
  - Knowledge: rules based on GenBank metadata and document information
Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Detection only</strong></td>
<td>.60</td>
<td>.90</td>
<td>.72</td>
</tr>
<tr>
<td><strong>Disambiguation only</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>.75</td>
<td>1.0</td>
<td>.86</td>
</tr>
<tr>
<td>Distance</td>
<td>.62</td>
<td>1.0</td>
<td>.77</td>
</tr>
<tr>
<td>Knowledge</td>
<td><strong>.89</strong></td>
<td>1.0</td>
<td><strong>.94</strong></td>
</tr>
<tr>
<td><strong>Detection + disambiguation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>.49</td>
<td>.89</td>
<td>.63</td>
</tr>
<tr>
<td>Distance</td>
<td>.41</td>
<td>.87</td>
<td>.56</td>
</tr>
<tr>
<td>Knowledge</td>
<td><strong>.55</strong></td>
<td><strong>.90</strong></td>
<td><strong>.68</strong></td>
</tr>
</tbody>
</table>

**Gold Standard**

[diego.asu.edu/downloads](http://diego.asu.edu/downloads)

60 Pubmed articles linked to 5730 GenBank records

1881 toponyms occurrences

**Inter-annotators Agreement**

Detection: 97P 98R

Disambiguation: 96P 98R