What is computational biology?

David Welch

School of Computer Science, Centre for Computational Evolution, and Te Pūnaha Matatini
What is biology?

Biology is the study of life. This is a broad target!

• **individual organisms**
• **populations of organisms**
• **evolution of populations and species**
• **ecological systems** (interactions between diverse populations)
• **organism subsystems** (e.g. organs)
• **cell biology**
• **genetics**
• **metabolism**
• **ethology** (behaviour)
• **the origin of life** (how is life different from non-life, how did life emerge?)
• And more....!
Biological systems differ from eg physical ones

Biological systems are complex
• lots of interacting parts
• parts are different from each other (heterogeneity)
• the parts interact non-linearly
How can computers help?

• Modelling
• Simulation
• Statistical inference
• Data processing
• Visualisation
• Data collection
• Automation
• …
3 examples

• Boids: modelling, simulation, emergence
• Alignment and structural prediction: algorithms, gamification, heuristics
• Phylodynamics of COVID-19: inference, big data, visualisation
Boids: a simple model of flocking

Each bird is following a few simple rules and the complex pattern simply emerges from them.

Separation  
Alignment  
Cohesion

Images taken from Craig Reynold's website.  http://www.red3d.com/cwr/boids/
Central Dogma of Molecular Biology: Eukaryotic Mode
Alignments of Human Beta-Globin to Other Globins

Human beta-globin

Ring-tailed lemur beta-globin

Goldfish beta-globin

Bloodworm globin IV

Soybean leghemoglobin
How to align multiple sequences?

<table>
<thead>
<tr>
<th>Organism</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
<th>Sequence 4</th>
<th>Sequence 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_aminophilium</td>
<td>TGGACAGTT</td>
<td>TGGAAAGAT</td>
<td>ACTCCGTT</td>
<td>GGTACAGGAT</td>
<td></td>
</tr>
<tr>
<td>C_colinum</td>
<td>GGAAAA</td>
<td>GGAAAA</td>
<td>AGTAAAGTG</td>
<td>GGTATAGGAG</td>
<td></td>
</tr>
<tr>
<td>C_lentocellum</td>
<td>TGGATTGCT</td>
<td>TGGATTNAT</td>
<td>GATTAGT</td>
<td>GCGATAGGAT</td>
<td></td>
</tr>
<tr>
<td>C_botulinum_D</td>
<td>TTTACGTTAA</td>
<td>AAAATAACAA</td>
<td>GCCATAGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_novyi_A</td>
<td>TGGCA</td>
<td>GC</td>
<td>GCCATAGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_perfringens</td>
<td>TTTCGCA</td>
<td>TGGACAA</td>
<td>GGAACATCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_estertheticum</td>
<td>TTTCGCA</td>
<td>TGGGCA</td>
<td>GGAACATCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_gasovires</td>
<td>ATTCGCA</td>
<td>TGGGCA</td>
<td>GGAACATCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_perfringens</td>
<td>AAGTAGCA</td>
<td>TCATCA</td>
<td>TTCAACAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_cadaveris</td>
<td>TTTACTGCA</td>
<td>TGAAAAA</td>
<td>GGAACATCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_cellulosolvens</td>
<td>ATTCGCA</td>
<td>TGAGAG...</td>
<td>GGAACATCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_K21</td>
<td>TGATCAGGCA</td>
<td>ATCAAGGAT...</td>
<td>GCTTAGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_estertheticum</td>
<td>TTGATCGCA</td>
<td>TACCAAGGA...</td>
<td>TTTCTTTGGAATC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_botulinum_A</td>
<td>AAGATACGA</td>
<td>TACCAAGGA...</td>
<td>ATTGCTAGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_sporogenes</td>
<td>AAGATACGA</td>
<td>TACCAAGGA...</td>
<td>ATTGCTAGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_argentinense</td>
<td>AAGATACGA</td>
<td>TACCAAGGA...</td>
<td>ATTGCTAGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_subterminale</td>
<td>AAGATACGA</td>
<td>TACCAAGGA...</td>
<td>ATTGCTAGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_tetanomorphum</td>
<td>TTTTCGCA</td>
<td>TGAAAAACTA</td>
<td>ATCAAGGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_pasteurianum</td>
<td>AGTTTACCA</td>
<td>TGAGGCTTA</td>
<td>ATTAAAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_collagenovorans</td>
<td>TTGATCGCA</td>
<td>TGCTGAAAT...</td>
<td>ATTAAAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_histolyticum</td>
<td>TTTATGCA</td>
<td>TCTTGAAGA...</td>
<td>ATTAAAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_thermobutyracem</td>
<td>AGTTTACCA</td>
<td>TGAAATTTGAG</td>
<td>ATTAAAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_tetani</td>
<td>GTGGTCCGA</td>
<td>TGAAAATTTA</td>
<td>ATTAAAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_barkeri</td>
<td>GACATCGCA</td>
<td>TGAGTTGAA...</td>
<td>ACTCCGTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_thermocellum</td>
<td>GGGATCGTC</td>
<td>CTGTAT...</td>
<td>CAAAGGAGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_pep_fermentis</td>
<td>AGTCGGCA</td>
<td>TGGNGTTATC</td>
<td>ATCAAGAGA...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_innominatum</td>
<td>ACAGGAGCAG</td>
<td>TGCTGTTAAG...</td>
<td>CTTCAAGGCTGAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_ruminantium</td>
<td>AGTTTCCGCA</td>
<td>TGGGACCTTG</td>
<td>ATTAAGAGATG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GCTCTTAATGTAAGGTATCG
A phylogeny shows relationships between sampled species, populations, or individuals.
Phylogenetics builds trees from genomes by modelling mutation processes, typically ignores population altogether.

building a phylogenetic tree can tell us a lot about a virus: suggests local transmission chains, places sequences geographically, etc.

Phylodynamics accounts for the dynamics in the population of interest, incorporates information like

- number of cases
- Location and movement
- Sampling dates
Nextstrain.org provides an excellent global overview
220 NZ samples, 500 for rest of world, time increases towards leaves
A basic phylodynamic model

Two locations: World and NZ

Parameters

Migration rates
Transmission rates ("birth")
Removal rates ("death")
Sampling rates

Parameters can differ in different intervals
Major outbreaks in New Zealand

- Resthome (Christchurch)
- Wedding (Bluff)
- Hospitality (Matamata)
- Unknown (Auckland)
- Wedding (Wellington)
- Conference (Queenstown)

Dec 2019
Jan 2020
Feb 2020
Mar 2020
Apr 2020
May 2020

$R_W$ deme

$\mathcal{I}S$ deme

$t_1$: First reported case in $\mathcal{I}S$
$t_2$: Mobility decrease in $\mathcal{I}S$

Clade support:
- $p \geq 0.9$
- $0.5 \leq p < 0.9$
Hereford conference – TMRCA 25 Feb-2 Mar
When were cases introduced to New Zealand?
Estimating the reproduction number

Re in New Zealand

Interval 2
Mean: 1.32
n=85

Interval 3
Mean: 0.74
n=131
Using mobility data to define intervals
Challenges in computational biology

- More data being generated than we can handle...
- ...and data is noisy, often highly correlated
- High levels of cross disciplinary work required
- Models are easy to build and simulate but inference is hard
- Finding balance of simplicity and complexity is very hard
- How can we combine multiple sources of data?
Links

• Murmuration: https://www.youtube.com/watch?v=DmO4Ellgmd0
• Boids: https://www.youtube.com/watch?v=nbbd5uby0sY
• Protein folding animation: https://www.youtube.com/watch?v=jVOyaT56LEU
• Foldit: https://fold.it/portal/info/about
• Multiple alignment game: https://phylo.cs.mcgill.ca/play.php
• Nextstrain COVID: https://nextstrain.org/ncov/global