Quantum for Life
Funded by the NovoNordisk Foundation

Matthias Christandl
Centre for the Mathematics of Quantum Theory (QMATH)
Department of Mathematical Sciences
University of Copenhagen
Main goal

• Demonstrate the viability of quantum computing to the life sciences
• Payoff
  • Establishment of high-level research effort (near-term)
  • New computational tool for life sciences (medium-term)
  • Nucleus for Danish quantum life science community (medium-term)
  • Feedback to economy and society (long-term)
Quantum for Life

Computational problems in life science with potential quantum speedup:

- Biochemistry in drug discovery
- Big data in bioinformatics

Novel experimental quantum simulator

Novel quantum algorithms

Experimentally demonstrate use-cases
Quantum for Life Team

• Matthias Christandl
  • Quantum Information Theory
  • Professor at Department of Mathematics co-PI of QMATH Center, Copenhagen University (KU)

• Anders Krogh
  • Pioneer of machine learning in Bioinformatics
  • Professor at Department of Computer Science Head of Center for Health Data Science (KU)

• Eugene Polzik
  • Pioneer in Experimental Quantum Technologies (atoms and light)
  • Professor at Niels Bohr Institute Center Leader of QUANTOP Center

• Markus Reiher
  • Theoretical and Numerical Quantum Chemistry
  • Quantum Chemistry on Quantum Computer
  • Professor at ETH Zurich

• History of collaboration
• Distinct but overlapping expertise
• Looking forward to the project
In order for quantum computers to benefit life sciences
• Need to understand the current computational challenges
• Dive deep into bioinformatics
QMATH tracks the Danish Coronavirus from genomic data

Andreas Bluhm
Matthias Christandl
Fulvio Gesmundo
Frederik Ravn Klausen
Laura Mancinska
Vincent Steffan
Daniel Stilck Franca
Albert H. Werner

SARS-CoV-2 transmission chains from genetic data: a Danish case study

Andreas Bluhm, Matthias Christandl, Fulvio Gesmundo, Frederik Ravn Klausen, Laura Mancinska, Vincent Steffan, Daniel Stilck Franca, Albert Werner

https://doi.org/10.1101/2020.05.29.123612
Pre Corona times

• TensorTalk started in the Fall (every Thursday lunch ever since)
• Tensor Networks (graphical models, hidden Markov models)
  • describe quantum matter
  • phylogenetic trees (ancestry)
Corona times

• Quantum information group meetings on zoom
• International group
  • Brazil, Germany, Denmark, Latvia, Austria, ...
• Concerned about the Corona situation
  • Wanted to learn about the coronavirus
  • Wanted to use our abilities to make a small contribution to help
• Project background
  • Unprecedented amount of public genomic data about the coronavirus
  • Genome can be used to identify ancestry (phylogeny)
  • This connects to our tensor network experience
• Goal
  • Build phylogenetic tree of Danish coronavirus
  • Identify transmission chains relevant for Denmark
An unusual project

- Complex numbers
- Formulas and proofs
- Quantum Information Theory
- Not yet time critical 😊

- Real numbers
- Practical data analysis
- Bioinformatics
- Extremely time critical

- Office for live meeting
- Other interaction/email
- Pen and paper
- Normal work hours
- PI, Postdoc, PhD

- Zoom room
- Slack
- R-Studio
- don’t ask
- 8 members all newcomers
SARS-CoV-2 genome

• Genome contains the information needed to build an organism
  • String over the alphabet A, C, T, G (nucleotides)
  • Approx 30,000 nucleotides for SARS-CoV-2
    ...TACCCAAAAATGCTGTTGTTAATAATCCAGCATGTAC...
  • Three consecutive nucleotides (codon) encode an amino acid
    e.g. TAC=Tyrosine  CCC=Proline

• Genes are sections of genome that code for proteins
  • SARS-CoV-2 has 10 genes
Mutations

- As genome is copied, mistakes occur
  - Substitutions (most common)
    - Non-synonymous
    - Synonymous
  - Deletions, insertions etc.

- Notation: \(_{\text{REF}}\) \(_{\text{POS}}\) \(_{\text{ALT}}\) E.g. G6T
  - We use Wuhan-Hu-1/2019 as a reference

- As the virus spreads, mutations get passed on
Get the data

- All genomic data was downloaded from GISAID.org. As of May 26, there were more than 32,000 genomes of Coronavirus samples from all over the world are available. Of those, 742 are Danish.

- From each sample, we know the collection date and the submitting lab.

- For sequences from some countries, more metadata like travel history, gender, age, etc. is available.
Align the data

- Sequencing genomes is prone to errors, usually, sequences have missing parts in various places.
- By reintroducing gap characters, sequences can be ‘aligned’.
- We use the R-package DECIPHER for alignment.

After that, the data is ready to build a phylogenetic tree!
From sequences to phylogenetic trees

Idea: similar sequences are closely related

distance measure
e.g. Hamming distance = # of differences

distance matrix
\[ M_{i,j} = d(i, j) \]

Build tree by clustering
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix
• join closest sequences

Example:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix
• join closest sequences
• update distance matrix

Example:

<table>
<thead>
<tr>
<th></th>
<th>A/D</th>
<th>B</th>
<th>C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>0</td>
<td>4</td>
<td>6.5</td>
<td>9.5</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>6.5</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>9.5</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix
• join closest sequences
• update distance matrix

Example:

<table>
<thead>
<tr>
<th></th>
<th>A/D</th>
<th>B/C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>0</td>
<td>5.25</td>
<td>9.5</td>
</tr>
<tr>
<td>B/C</td>
<td>5.25</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>9.5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix
• join closest sequences
• update distance matrix
• repeat

Example:

<table>
<thead>
<tr>
<th></th>
<th>A/D</th>
<th>B/C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>0</td>
<td>5.25</td>
<td>9.5</td>
</tr>
<tr>
<td>B/C</td>
<td>5.25</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>9.5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix
• join closest sequences
• update distance matrix
• repeat

Example:

<table>
<thead>
<tr>
<th></th>
<th>A/D</th>
<th>B/C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>0</td>
<td>5.25</td>
<td>9.5</td>
</tr>
<tr>
<td>B/C</td>
<td>5.25</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>9.5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix
• join closest sequences
• update distance matrix
• repeat

Example:

<table>
<thead>
<tr>
<th></th>
<th>A/D</th>
<th>B/C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>0</td>
<td>5.25</td>
<td>9.5</td>
</tr>
<tr>
<td>B/C</td>
<td>5.25</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>9.5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean
• input: distance matrix
• join closest sequences
• update distance matrix
• repeat

Example:

<table>
<thead>
<tr>
<th></th>
<th>A/D</th>
<th>B/C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>0</td>
<td>5.25</td>
<td>9.5</td>
</tr>
<tr>
<td>B/C</td>
<td>5.25</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>9.5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

• input: distance matrix
• join closest sequences
• update distance matrix
• repeat

In practice, use more elaborate schemes, e.g. neighbor joining

Example:

<table>
<thead>
<tr>
<th></th>
<th>A/D</th>
<th>B/C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>0</td>
<td>5.25</td>
<td>9.5</td>
</tr>
<tr>
<td>B/C</td>
<td>5.25</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>9.5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
The Danish Tree – first observations

- 70% of the Danish sequences have 6 specific mutations compared with Wuhan.
- 100+ sequences with a mutation that primarily happened in Denmark.
- Many entries to Denmark and many chains of mutations inside Denmark
Danish transmission chain
Example: C1302T

Likely spread of the strain with haplotype C1302T from Denmark to other Northern European countries. Within Denmark, it also mutated further.

Understanding mutations is a fantastic tool to understand spread of the virus → inform mitigation strategies
Understand the computational problems → attack with quantum computers?
Summary

• Quantum for Life

problem?

quantum hardware

powers

necessary

quantum algorithms

solution

• Corona virus project