Quantum for Life

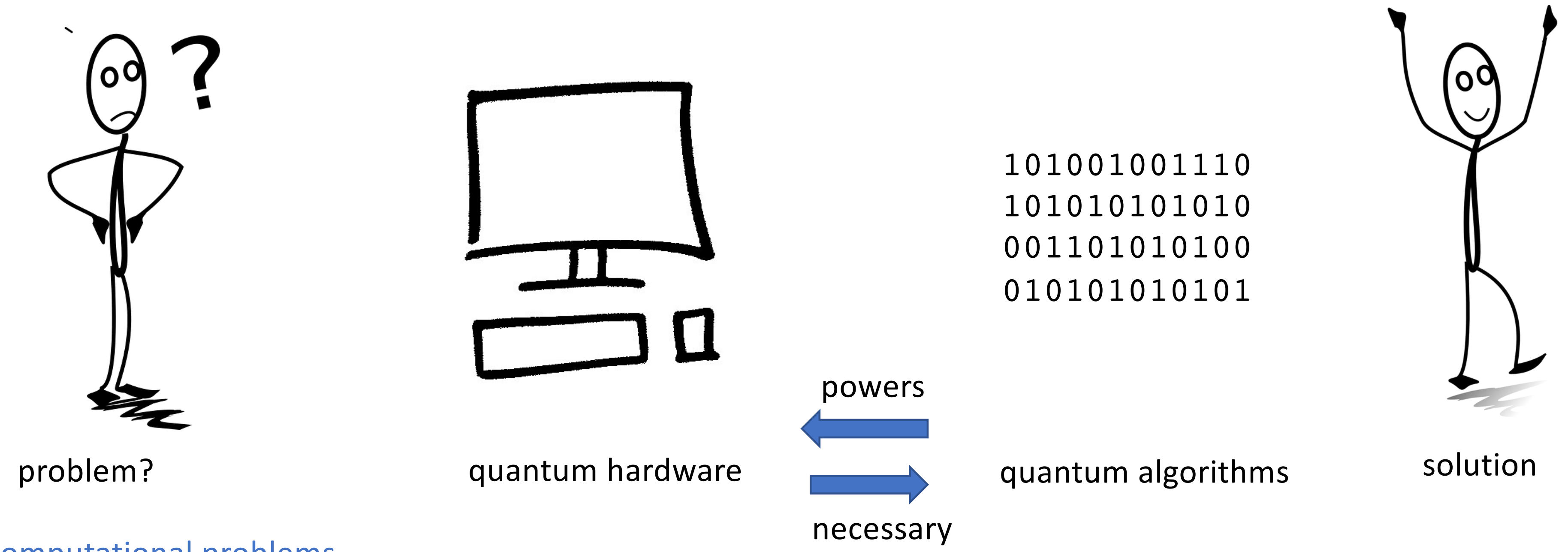
Funded by the NovoNordisk Foundation

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



101001001110
101010101010
001101010100
010101010101

problem?

quantum hardware

powers

necessary

quantum algorithms

solution

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

Jan Philip Solovej
(atoms)

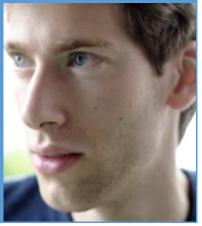
Laura Mancinska
(algorithms)

Albert Werner
(simulation)

QMATH Centre for the Mathematics of Quantum Theory

Ignacio Cirac

Pioneer in Theoretical Quantum Information Science
Director at Max-Planck-Institute

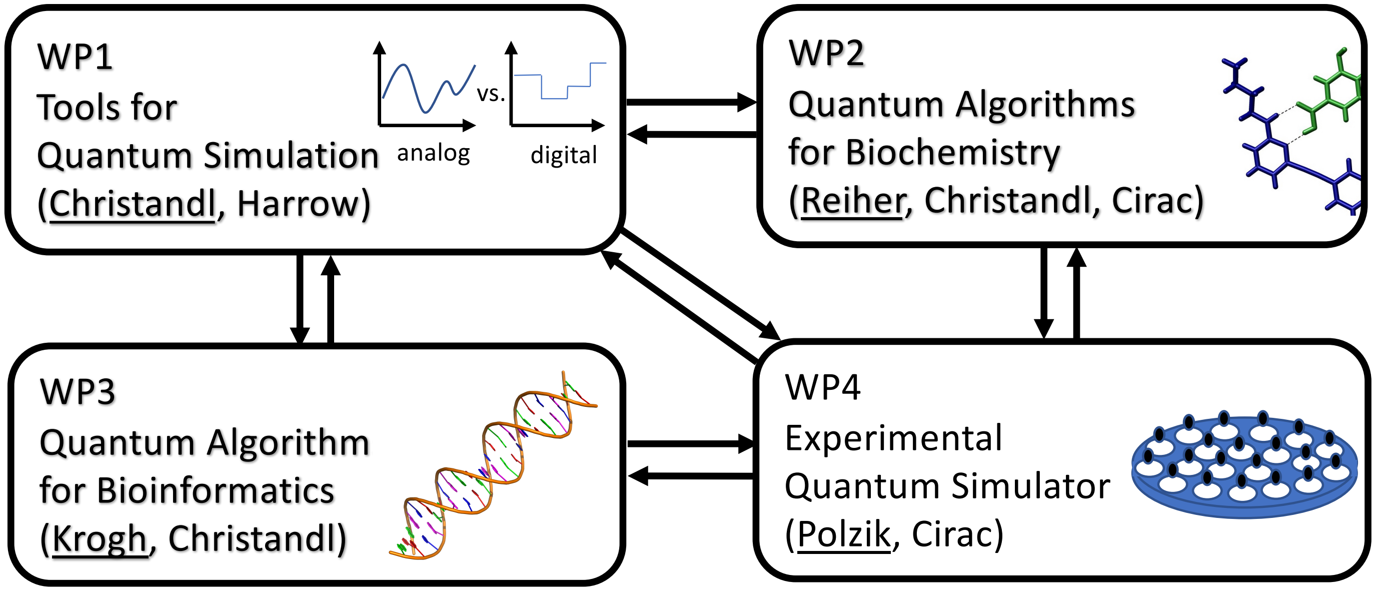


Aram Harrow

Leader in Quantum Algorithms
Associate Professor at MIT

- History of collaboration
- Distinct but overlapping expertise
- Looking forward to the project

Project Overview



WP1
Tools for
Quantum Simulation
(Christandl, Harrow)

WP2
Quantum Algorithms
for Biochemistry
(Reiher, Christandl, Cirac)

WP3
Quantum Algorithm
for Bioinformatics
(Krogh, Christandl)

WP4
Experimental
Quantum Simulator
(Polzik, Cirac)

- 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

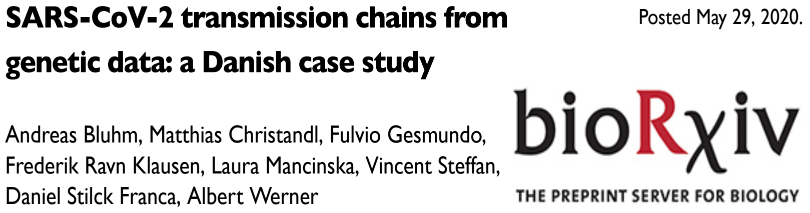
Frederik Ravn Klausen

Laura Mancinska

Vincent Steffan

Daniel Stilck Franca

Albert H. Werner



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

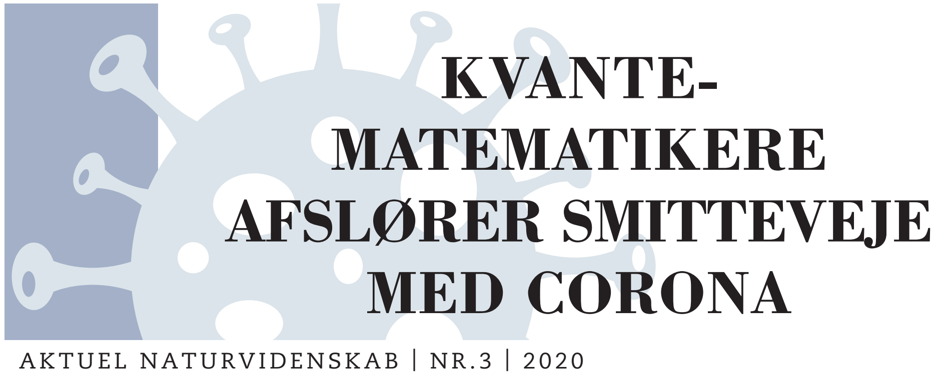
Posted May 29, 2020.

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

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

https://doi.org/10.1101/2020.05.29.123612



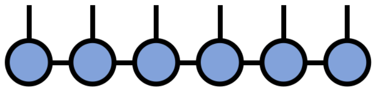
KVANTE-MATEMATIKERE AFSLØRER SMITTEVEJE MED CORONA

AKTUEL NATURVIDENSKAB | NR.3 | 2020

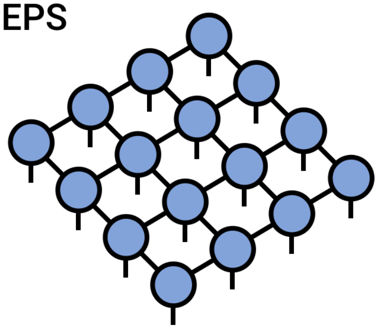
Pre Corona times

- TensorTalk started in the Fall (every Thursday lunch ever since)
- Tensor Networks (graphical models, hidden Markov models)
  - describe quantum matter

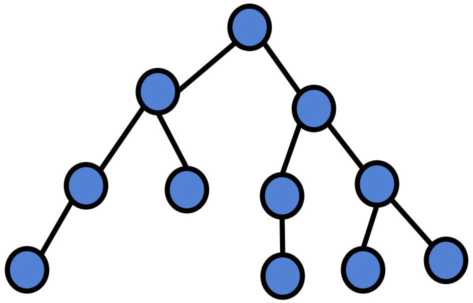
Matrix Product State / Tensor Train



PEPS

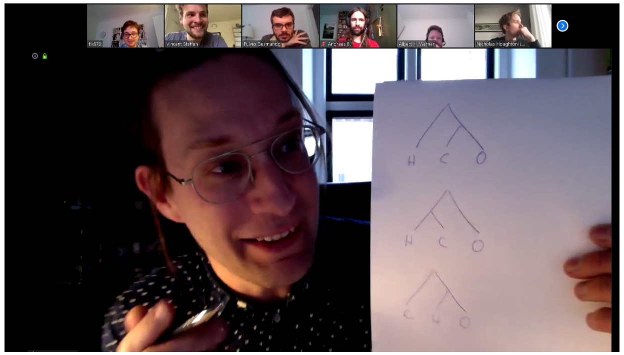


  - phylogenetic trees (ancestry)



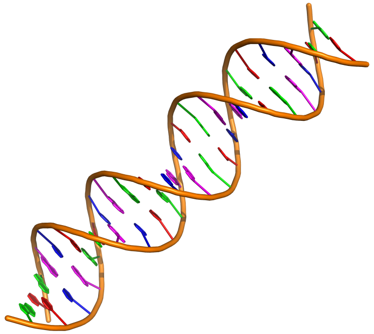
Corona times

- 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

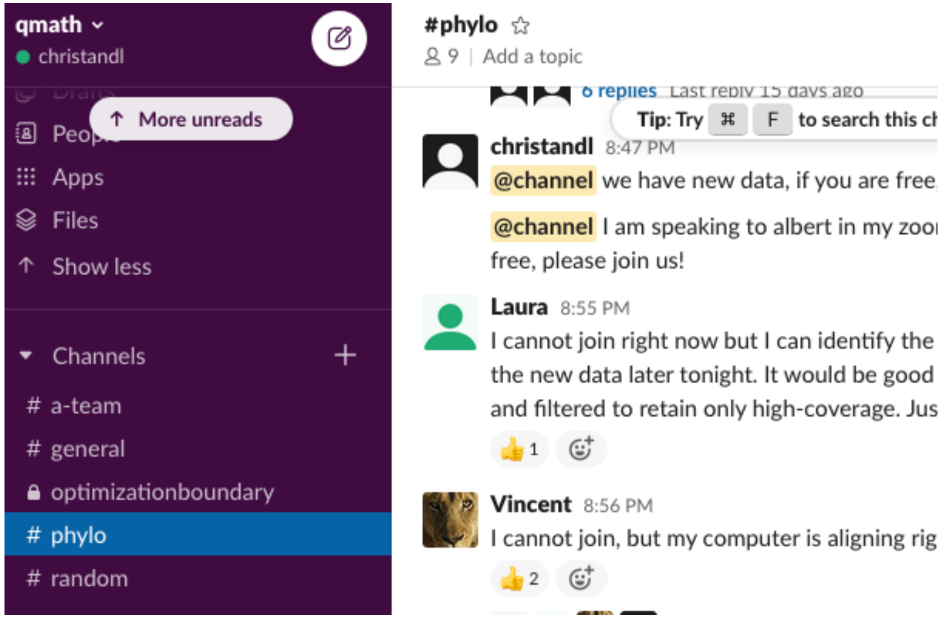


An unusual project

- Complex numbers → Real numbers
- Formulas and proofs → Practical data analysis
- Quantum Information Theory → Bioinformatics
- Not yet time critical ☺ → Extremely time critical



- Office for live meeting → Zoom room
- Other interaction/email → Slack
- Pen and paper → R-Studio
- Normal work hours → don't ask
- PI, Postdoc, PhD → 8 members all newcomers



christandl 8:47 PM

@channel we have new data, if you are free

@channel I am speaking to albert in my zoo
free, please join us!

Laura 8:55 PM

I cannot join right now but I can identify the
the new data later tonight. It would be good
and filtered to retain only high-coverage. Jus

👍 1

Vincent 8:56 PM

I cannot join, but my computer is aligning rig

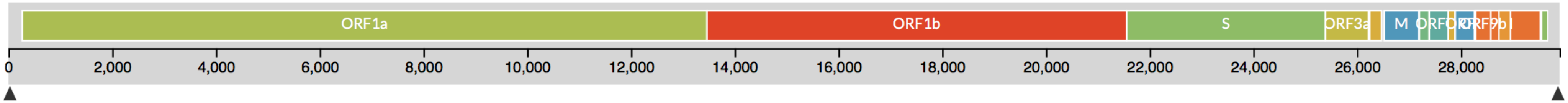
👍 2

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

    ...TACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCAC ...
  - Three consecutive nucleotides (codon) encode an amino acid
    e.g. TAC=Tyrosine CCC= Proline

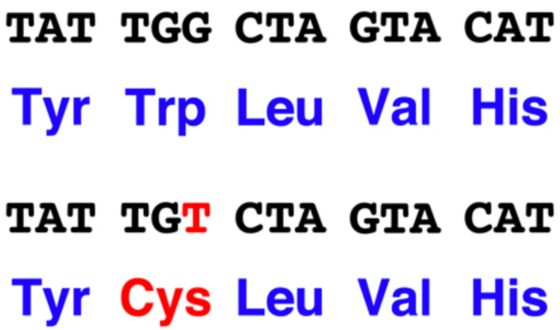
- Genes are sections of genome that code for proteins
  - SARS-CoV-2 has 10 genes



Mutations

- As genome is copied, mistakes occur
  - Substitutions (most common)
    - Non-synonymous
    - Synonymous
  - Deletions, insertions etc.
- Notation: `(_REF_)(_POS_)(_ALT_)` E.g. G6T
  - We use Wuhan-Hu-1/2019 as a reference

- As the virus spreads, mutations get passed on



Get the data

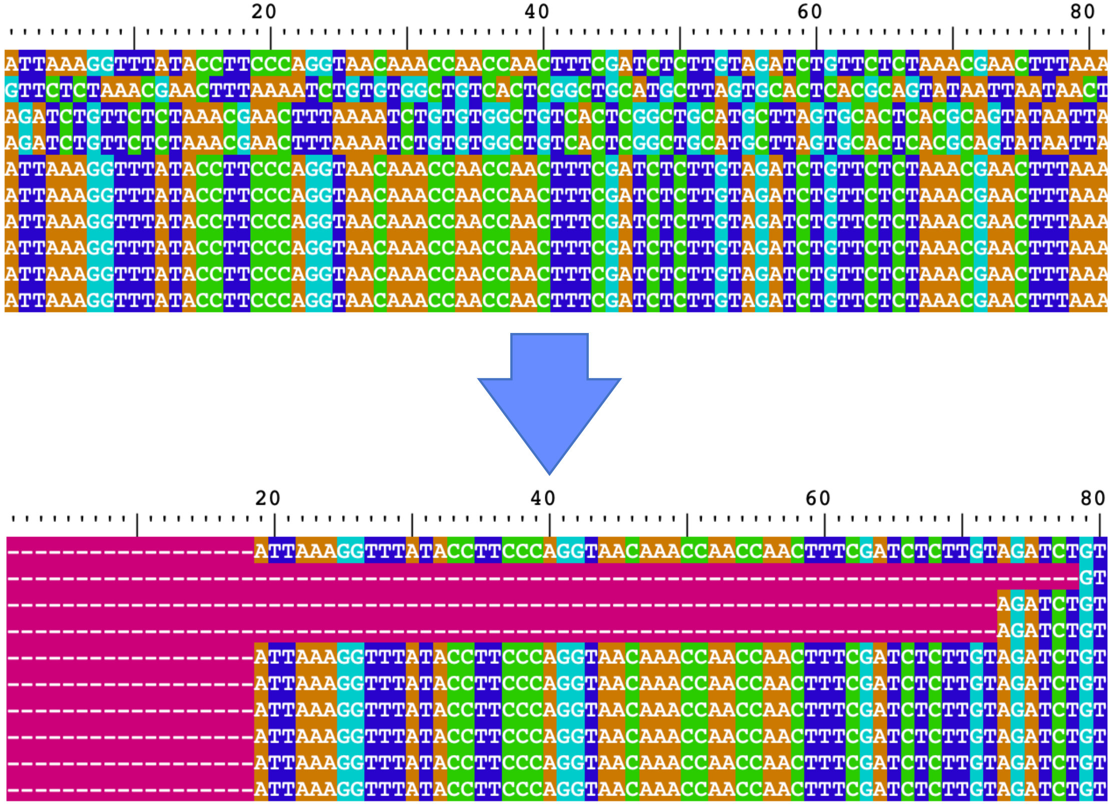
```
>hCoV-19/Denmark/ALAB-SSI-766/2020|EPI_ISL_452102|2020-03-19
```

- All genomic data was downloaded from GISAID.org. As of May 26, there were more than 32.000 genomes of Corona virus 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.

Source: GISAID.org

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.



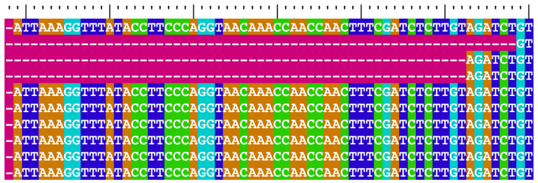
```
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAA
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACT
AGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTA
AGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTA
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAA
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAA
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAA
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAA
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAA
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAA
```

```
------------------ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
------------------------------------------------------------------------------GT
------------------------------------------------------------------------AGATCTGT
------------------------------------------------------------------------AGATCTGT
------------------ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
------------------ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
------------------ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
------------------ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
------------------ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
------------------ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
```

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

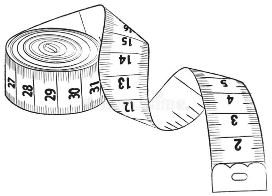
From sequences to phylogenetic trees

aligned sequences



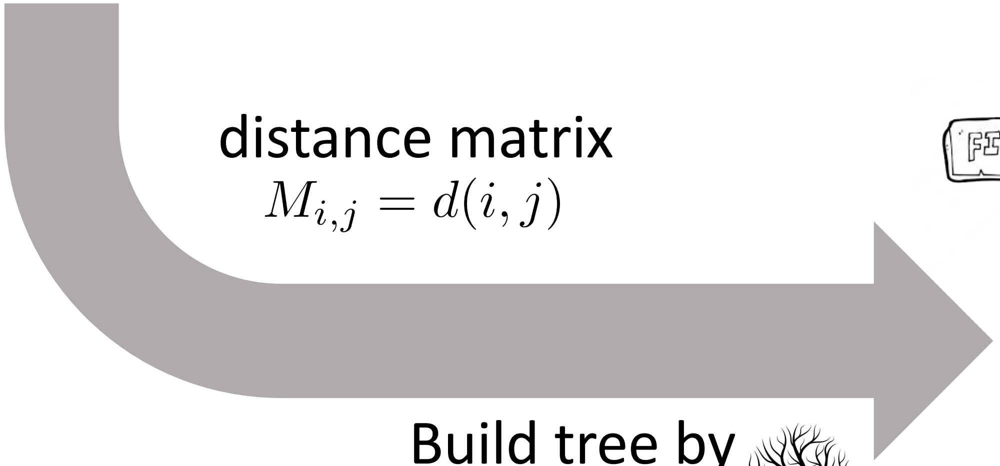
```
20                  40                  60                  80
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
---------------------------------------------------------------GT
-------------------------------------------------------AGATCTGT
-------------------------------------------------------AGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
```

Idea: similar sequences are closely related



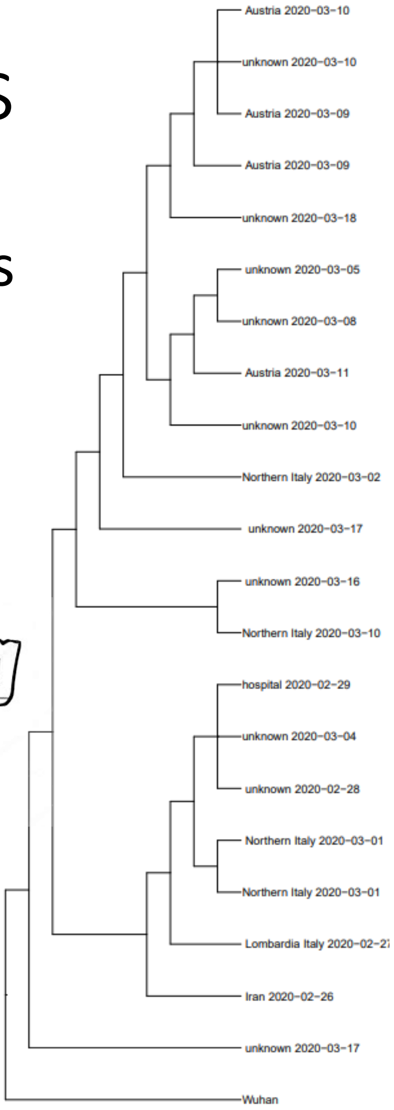
distance measure

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

Example:

	A	**B**	**C**	**D**	**E**
A	0	3	7	2	10
B	3	0	2	5	5
C	7	2	0	6	3
D	2	5	6	0	9
E	10	5	3	9	0

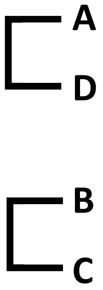
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

- input: distance matrix
- join closest sequences

Example:

	A	**B**	**C**	**D**	**E**
A	0	3	7	2	10
B	3	0	2	5	5
C	7	2	0	6	3
D	2	5	6	0	9
E	10	5	3	9	0

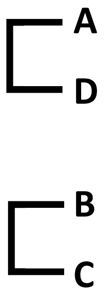


Building a tree: UPGMA

unweighted pair group method with arithmetic mean

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

Example:



	A/D	**B**	**C**	**E**
A/D	0	4	6.5	9.5
B	4	0	2	5
C	6.5	2	0	3
E	9.5	5	3	0

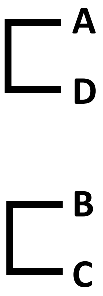
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

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

Example:

	A/D	**B/C**	**E**
A/D	0	5.25	9.5
B/C	5.25	0	4
E	9.5	4	0



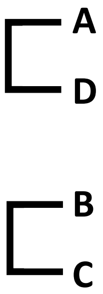
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

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

Example:

	A/D	**B/C**	**E**
A/D	0	5.25	9.5
B/C	5.25	0	4
E	9.5	4	0



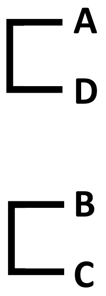
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

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

Example:

	A/D	**B/C**	**E**
A/D	0	5.25	9.5
B/C	5.25	0	4
E	9.5	4	0



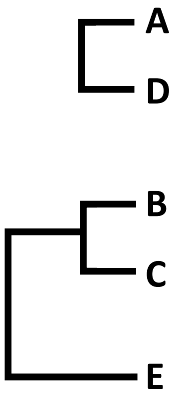
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

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

Example:

	A/D	**B/C**	**E**
A/D	0	5.25	9.5
B/C	5.25	0	4
E	9.5	4	0



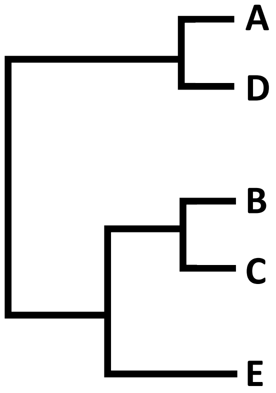
Building a tree: UPGMA

unweighted pair group method with arithmetic mean

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

Example:

	A/D	**B/C**	**E**
A/D	0	5.25	9.5
B/C	5.25	0	4
E	9.5	4	0



Building a tree: UPGMA

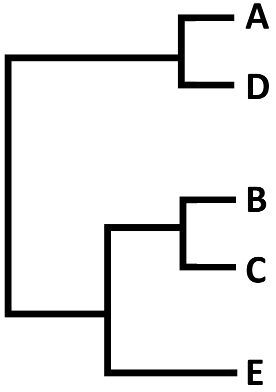
unweighted pair group method with arithmetic mean

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

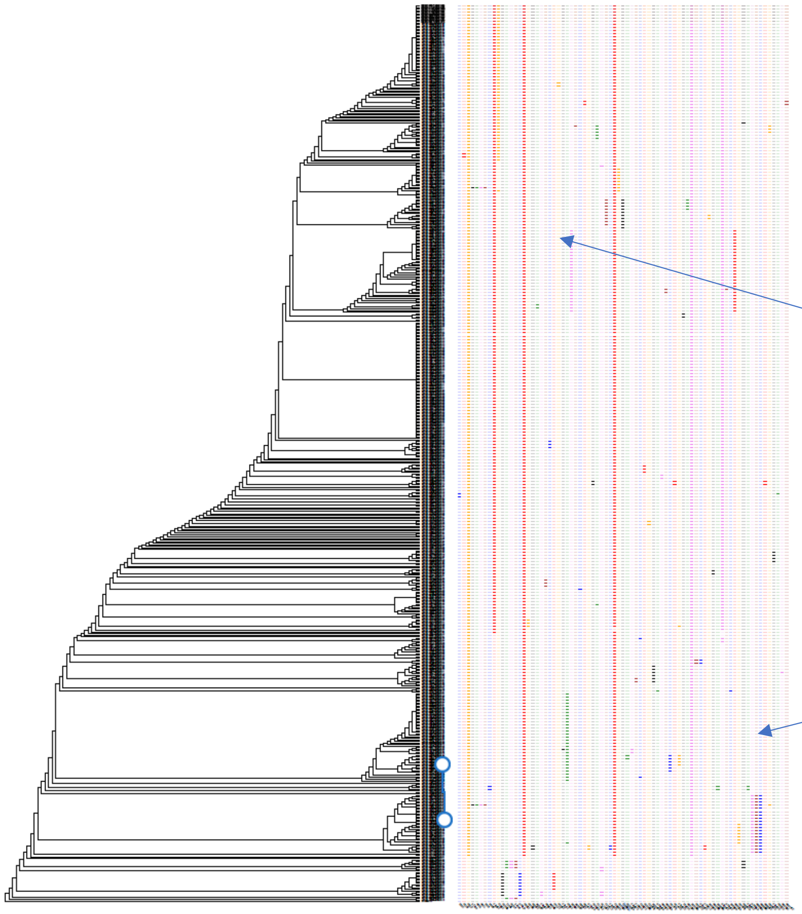
In practice, use more elaborate schemes, e.g. neighbor joining

Example:

	A/D	**B/C**	**E**
A/D	0	5.25	9.5
B/C	5.25	0	4
E	9.5	4	0

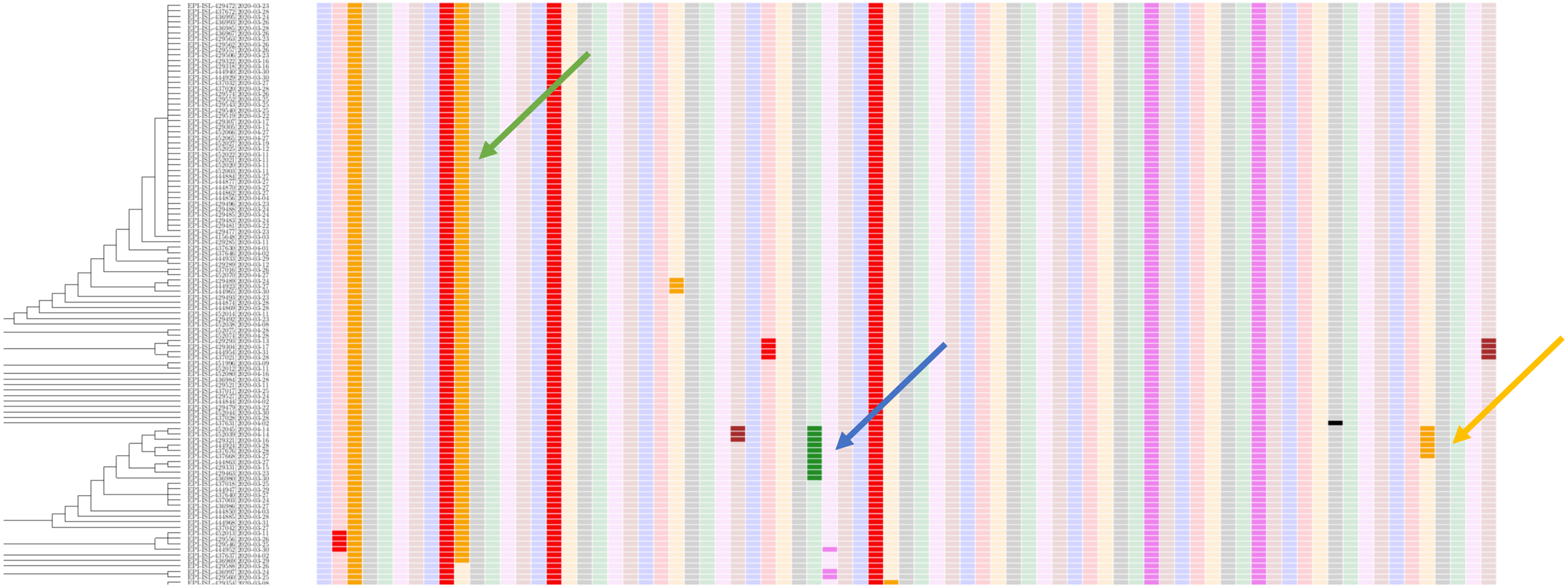


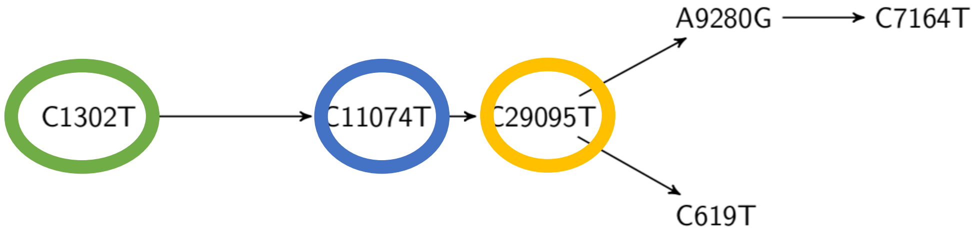
The Danish Tree – first observations



- 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



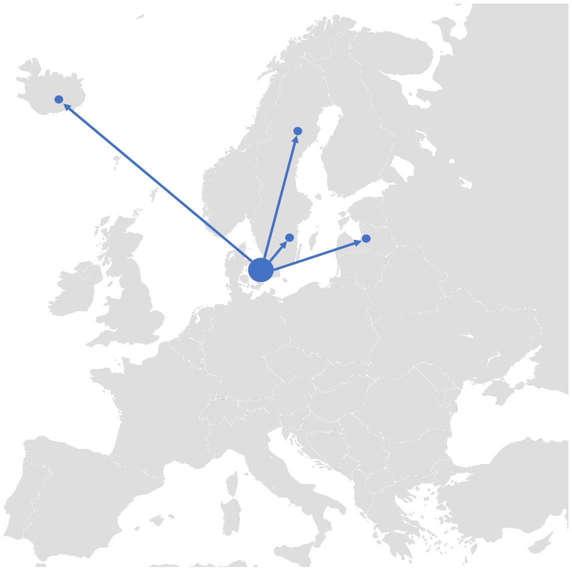


A9280G → C7164T

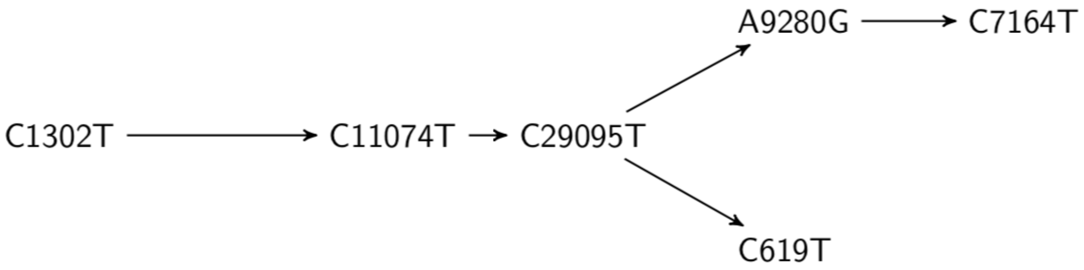
C1302T → C11074T → C29095T

C619T

Example: C1302T



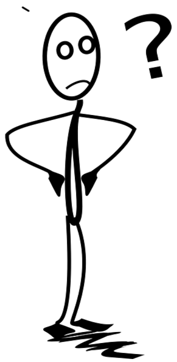
Likely spread of the strain with haplotype C1302T from Denmark to other Northern European countries. Within Denmark, it also mutated further.

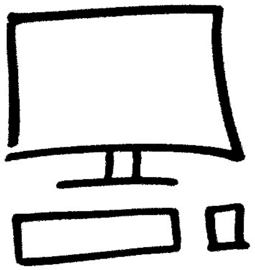


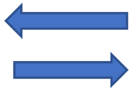
Understanding mutations is a fantastic tool to understand spread of the virus
→ inform mitigation strategies
Understand the computational problems
→ attack with quantum computers?

Summary

- Quantum for Life


problem?

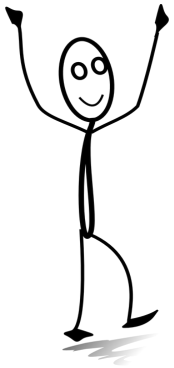

quantum hardware


powers

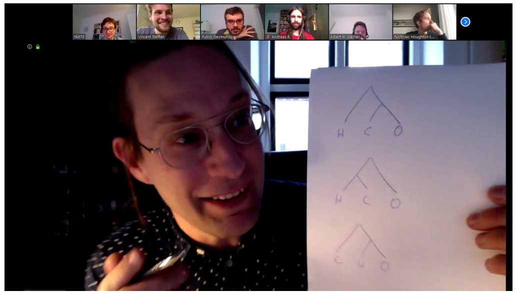
necessary

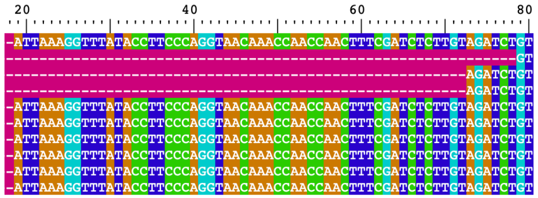
101001001110
101010101010
001101010100
010101010101

quantum algorithms

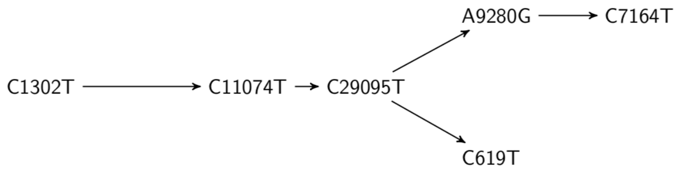

solution

- Corona virus project





```
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-------------------------------------------------------------GT
-------------------------------------------------------AGATCTGT
-------------------------------------------------------AGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT
```


C1302T → C11074T → C29095T

C29095T → A9280G → C7164T

C29095T → C619T