

Viruksien analysointi Markovin piilomalleilla

Tomas Laamanen

17.3.2022

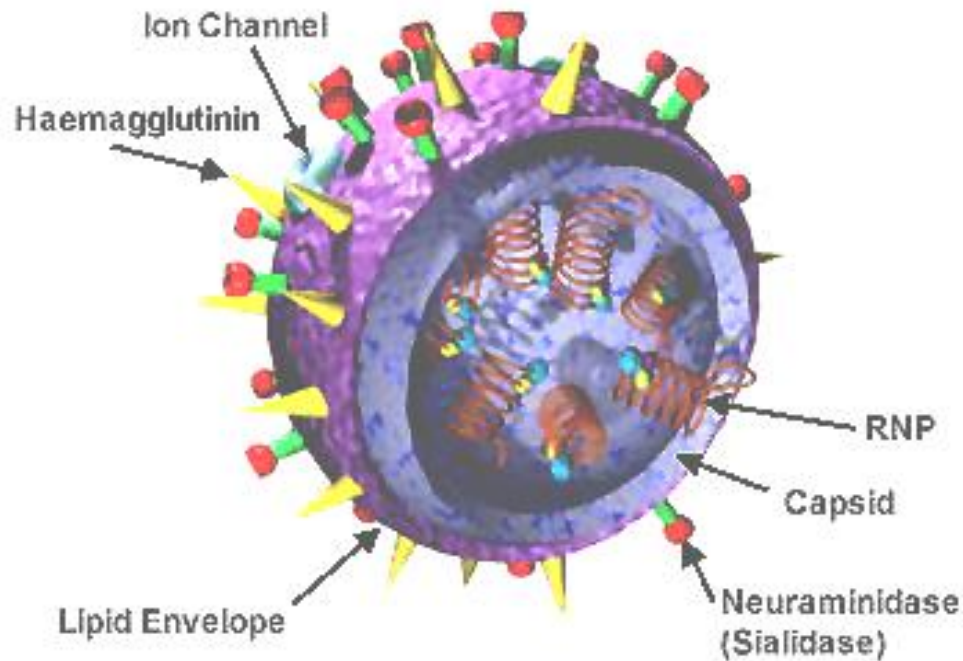
Vastuupettaja: Kai Virtanen

Ohjaaja: Kai Virtanen

Työn voi julkaista Aalto-yliopiston sivuilla. Kaikki muut oikeudet pidätetään.

Tausta 1/3

Virukset koostuvat geneettisestä materiaalista ja sen kuoresta:



Tausta 2/3

- Geneettinen materiaali (sekvenssi a, g, c ja t nukleotidejä) määrittelee virusten ominaisuudet
- Ymmärtämällä viruksien mutaatioita voidaan selvittää
 - minkälaisia oireita erilaiset infektiot aiheuttavat
 - miten virukset leviävät
- Stokastisia malleja käytetty aiemmin
 - virusten rakenteen analysointiin
 - virusten mutaatioiden ennustamiseen

Tausta 3/3

- Työssä tutkitaan kahta eri virusperhettä, jotta voidaan selvittää soveltuvatko sotkastiset mallit erilaisille viruksille
- Usean *Coronaviridae*- ja *Mitoviridae*-perheiden jäsenten genomit ollaan sekvenssoitu
- Näiden perheiden virusten rakenteet poikkeavat suuresti toisistaan

Markovin piilomallien terminologiaa

- Markovin piilomalleilla (Hidden Markov Model, HMM) analysoidaan emissioiden ketjuja
- Tässä tapauksessa emissiot ovat eri nukleotidejä
- Tilat ovat stokastisia arvoja, joissa ketju voi olla
- Mallin tilojen määrän neliö on yhtä suuri kuin mallin parametrien määrä

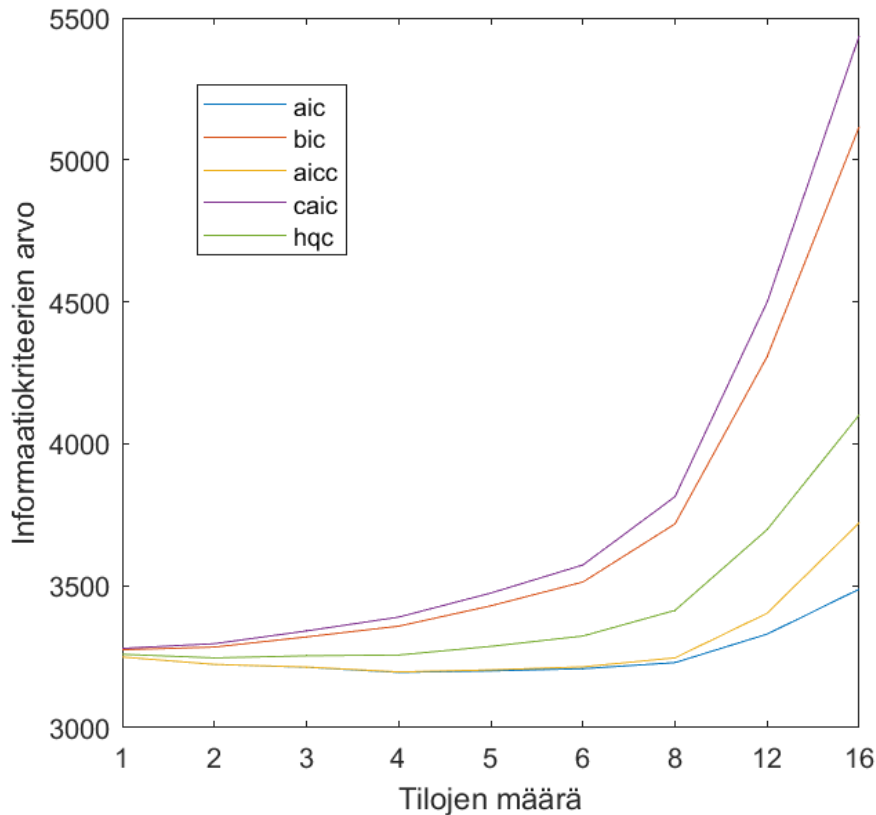
Tavoitteet

- Viruksia kuvaavien Markovin piilomallien tarkasteleminen
- Määritellä sellainen määrä paremètreja malleille, että ne
 - ovat optimaalisia informaatioteoreettisesti
 - ne pystyvät ennustamaan tulevia mutaatioita kaikkein tarkemmin
- Luoda kaksi mallia - yksi kummallekin virusperheelle
- Kuinka hyvin mallit ennustavat eri viruslajien olemassaolon todennäköisyyden

Työn rakenne

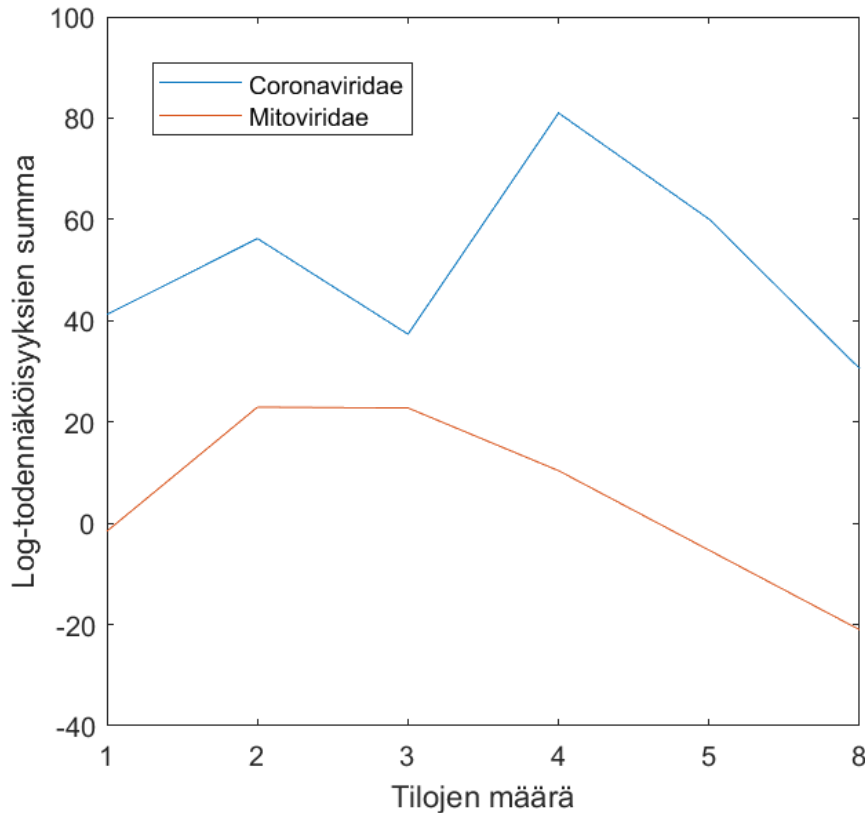
- Markovin mallien konstruointi
- Informaatiokriteereiden ja ristiinvalidoinnin käyttökelpoisuuden tarkastelu parametrien määrän valinnalle
- Mallit molemmille virusperheille ristiinvalidoinnin ja informaatiokriteerien perusteella valituilla parametrien määrillä

Tulokset 1/5



- Informaatiokriteerit pysyvät lähes vakiona pienillä tilojen määrillä
- Minimiarvo saavutetaan 2-4 parametrilla
- Eri värit ovat eri informaatiokriteerejä

Tulokset 2/5



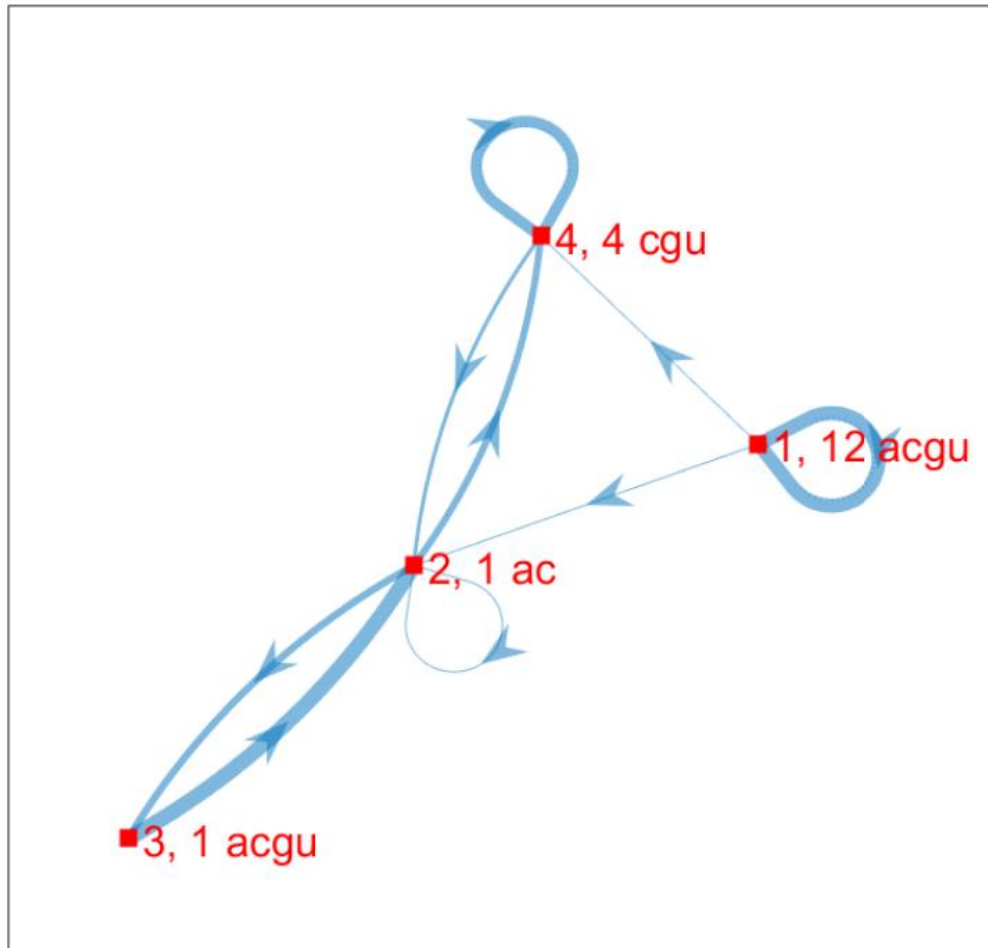
- Molemmille virusperheille laskettiin, kuinka hyvin muut näytteet selittävät yhden poistetun näytteen
- Tämä toistettiin eri määrille tiloja
- Korkein log-todennäköisyyksien summa saadaan kun tilojen määrä on 3-5

Tulokset 3/5

Seuraavien diojen kuvissa

- punaiset laatikot ovat tiloja
- nuolten paksuus on verrannollinen todennäköisyyteen siirtyä johonkin tilaan
- jos nuolta ei ole, siirtymistodennäköisyys on 0
- ensimmäinen numero on tilan nimi
- toinen numero on oletusmäärä askelia, jonka prosessi pysyy tilassa
- kirjaimet viittavat nukleotideihin

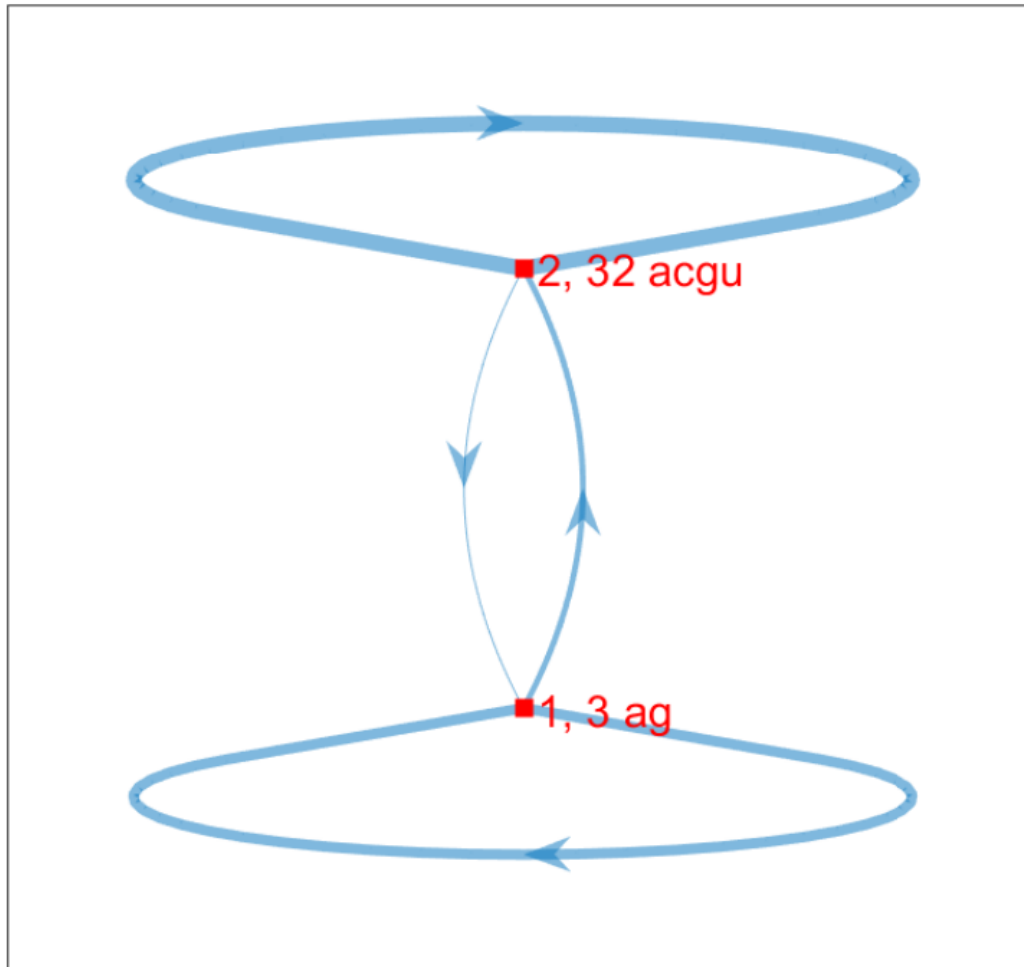
Tulokset 4/5 *Coronaviridae*-perhe



- Suuri todennäköisyys pysyä tiloissa 1 ja 4
- Vaihtelua tilojen 2 ja 3 välillä
- Eri nukleotidit yleisiä sekvenssin alussa ja lopussa

Tulokset 5/5 *Mitoviridae*-perhe

- Vain kaksi tilaa
- Tila 1 tarkoittaa vahvaa a ja g toistoa



Johtopäätökset

- HMM on tehokas tapa ennustaa minkälaisia uusia viruslajeja saattaa syntyä tiettyihin virusperheisiin
- Mahdollistaa virusten sukulaisuuden etäisyyden määrittämisen laskemalla kuinka suuri on viruslajin nukleotidin sekvenssin todennäköisyys, jos HMM on koulutettu toisella viruslajilla
- *Mitoviridae*-perheen jäsenten sekvenssit pystytään selittämään merkittävästi yksinkertaisemmalla mallilla kuin *Coronaviridae*-perheen
 - Nähtävissä tilojen määrässä

Oleellisia lähdeviitteitä

- Byung-Jun, Yoon: "Hidden Markov Models and their Applications in Biological Sequence Analysis." Current Genomics, vol. 10, no. 6, 2009, pp. 402-415.
- Pierce, Benjamin: Genetics: A Conceptual Approach. 5th ed., United States of America, Freeman, 2013.
- Weinan, E, Tiejun, Li, and Vanden-Eijnden, Eric: Applied Stochastic Analysis. United States of America, American Mathematical Society, 2019.