Klyngeanalyse

ISTx1003 Statistisk læring og Data Science

Stefanie Muff, Institutt for matematiske fag

November 12, 2021

Plan for i dag

- Hva er klyngeanalyse
- Læringsmål, pensum og læringsressurser
- Avstandsmål
- K-gjennomsnitt ("K-means") klyngeanalyse
- Bruk av klyngeanalyse på et bilde (prosjektet fra i fjor)
- Hierarkisk klyngeanalyse
- Informasjon om prosjektet

Eksempel 1: Genaktivitet

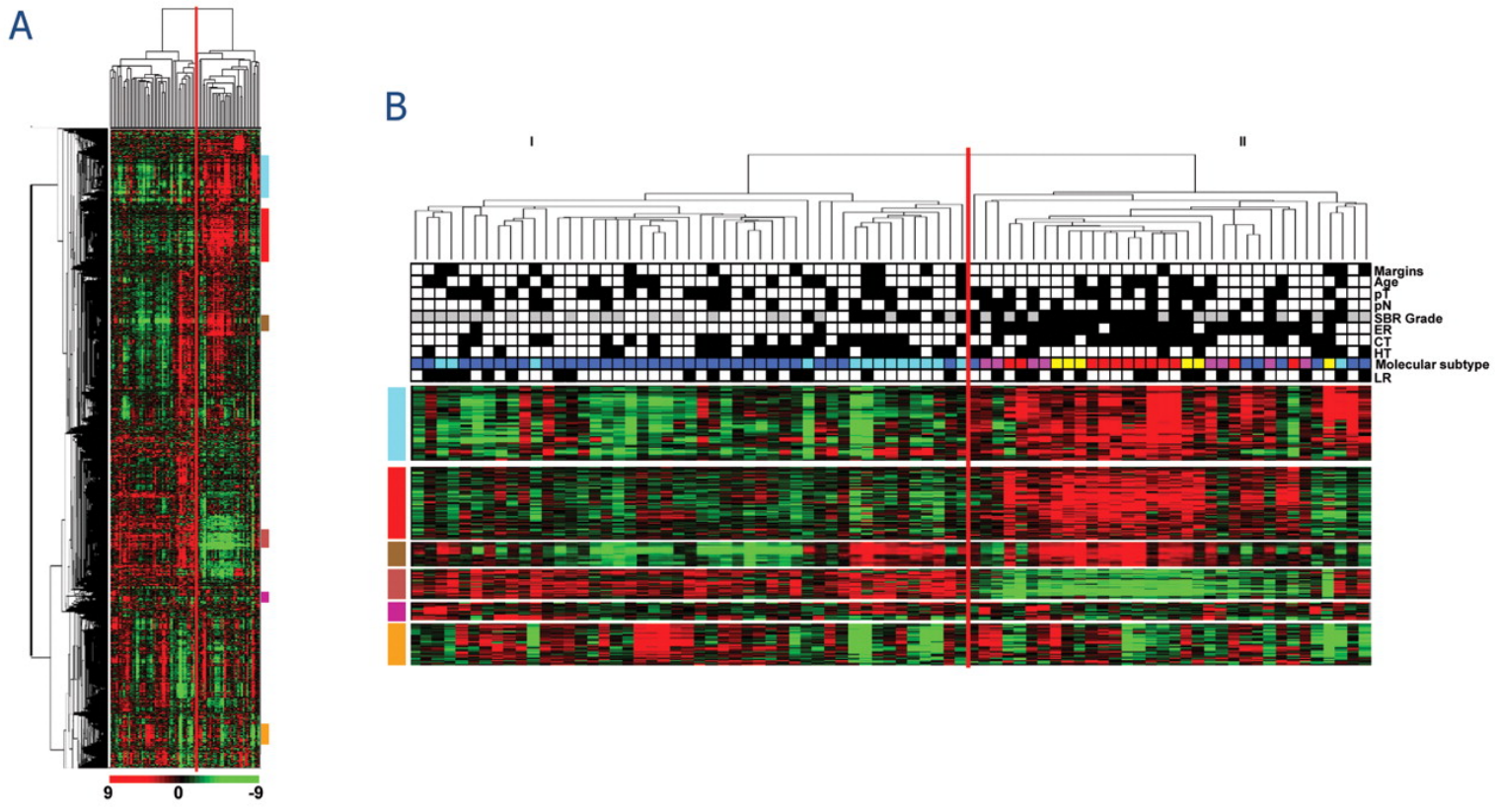
- $n = 81$ celleprøver fra kreftsvulster til ulike pasienter
- genaktivitet for $p = 12957$ gener

Spørsmål:

Hvilke celleprøver fra brystkreftpasienter ligner hverandre mest?

Kan vi finne ukjente klynger (av celleprøver) i dataene?

Dette kan hjelpe for å forutsi sannsynlighet for en tilbakefall.



$$X = p \times n = \text{gener} \times \text{prøver} \ .$$

Finn ut mer: https://cgp.iiarjournals.org/content/8/4/199

Eksempel 2: Proteininteraksjonsnettwerk

Kan vi finne klynger med relatert funksjon?

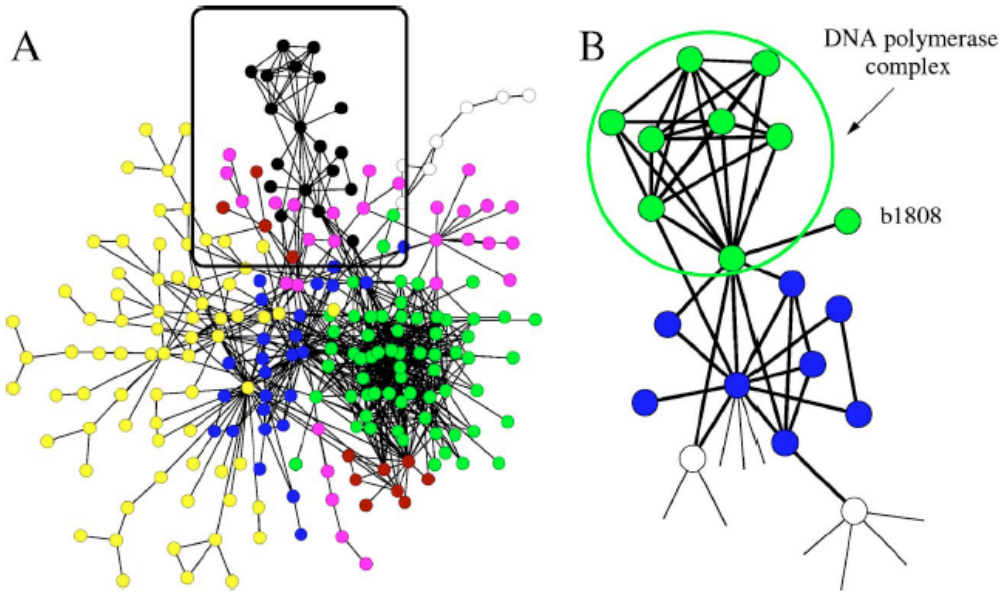


FIG. 4. (Color online) (a) Largest connected component of the PPI of *E. coli*. The colors represent the clusterization found by optimizing modularity. (b) *LQ* clusterization of the black *Q* cluster. The green circle contains proteins belonging to the DNA polymerase complex. The unknown protein b1808 is assigned to this complex according to *LQ* while the complete *Q* cluster is heterogeneous.

Læringsmål

- forstå hvorfor det er interessant å gjøre klyngeanalyse
- kjenne igjen situasjoner der klyngeanalyse vil være en aktuell metode å bruke
- kjenne begrepene avstandsmål, koblingstype, dendrogram
- forstå algoritmen for å utføre K-gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse
- forstå hvordan klyngeanalyse utføres i Python
- kunne besvare oppgave 3 av prosjektoppgaven på en god måte!

Læringsressurser

Tema Klyngeanalyse:

- **Kompendium**: Klyngeanalyse (pdf og html, by Mette Langaas)
- **Korte videoer**: (by Mette Langaas)
  - Klyngeanalyse (8:43 min)
  - Hierarkisk klyngeanalyse (11:26 min)
  - K-gjennomsnitt-klyngeanalyse (8:38 min)
- Denne forelesningen
- **Disse slides** med notater

https://wiki.math.ntnu.no/istx1003/2021h/start

Klyngeanalyse – hva er det?

- Mål:
  - tilordne en ny observasjon til en av flere *kjente* klasser
  - lage en klassifikasjonsregel
  - estimere sannsynligheten for at en ny observasjon tilhører de ulike klassene

For hver av de uavhengige observasjonene $i = 1, \ldots, n$ har vi

- Forklaringsvariabler $(x_{1i}, x_{2i}, \ldots, x_{pi})$
- En kategorisk responsvariablel y_i.

Eksempler