7.4.07
Flóvin Tór Nygaard Næs & Kristian Priisholm

# Retsgenetik - anvendelse af DNA-materiale i retssager

Studieretningsprojekt i biologi (A) og matematik (A)

## Emne:

Brug af genetisk materiale som grundlag for påvisning af statistisk sammenhæng mellem DNA fundet på et gerningssted og mistænktes DNA-profil.

En forholdsvis ny type bevismateriale i retssager er DNA-spor fundet på gerningssteder eller fra blodprøver fra mistænkte. Da selv en meget lille mængde DNA giver et meget stort mængde data, er denne type bevismateriale potentielt meget specifik og meget overbevisende. Dette emne omhandler muligheder og problemstillinger i den forbindelse.

Eleven skal gennemgå DNAs opbygning og funktion og forklare udvalgte mekanismer, som skaber de forskelle og ligheder som ses i arvematerialet og som danner grundlag for identifikation af personer ud fra DNA-profiler.

Matematisk set skal eleven arbejde med udvalgte aspekter af brugen af bayesiansk matematik (gennemgået nedenfor) i retssager. Fx ved ud fra egne eller udleverede cases at opstille Bayes' netværk, der angiver sandsynligheder for at en mistænkt er skyldig eller uskyldig bla. på baggrund af DNA-materiale. Eleven skal som grundlag gennemgå matematikken og tankegangen bag Bayes' netværk.

## Forudsætninger i matematik:

Kendskab til betingede sandsynligheder og Bayes formel. Det vil være en klar fordel, hvis eleven kan læse tekster på engelsk, da det vil give adgang til et meget bredere udvalg af litteratur.

## Faglige mål i matematik:

At sætte sig ind i, hvordan bayesianske matematik og netværk kan anvendes i forskellige sandsynlighedsteoretiske spørgsmål. Til dette formål inddrages computerprogrammer (f.eks. Hugin). Eleven skal blive fortrolig med begreber som betinget sandsynlighed og marginal sandsynlighed og kunne anvende formler som produktreglen, kædereglen og Bayes formel. Det er desuden et mål, at eleven opnår en kritisk sans for, hvornår de forskellige statistiske metoder kan bruges.

## Forudsætninger i biologi:

Kendskab til det menneskelige arvemateriales opbygning og funktion. Kendskab til DNA-sekvensering.

## Faglige mål i biologi:

Indsigt i forskellige biologiske mekanismer, der skaber forskellige og ligheder i arvemateriale i populationer så som kønnet formering, introns/exons, fitness, selektion, mutationer og genetisk drift. Indsigt i laboratoriemetoder til at påvise forskelle/ligheder i arvemateriale mellem individer, styrker og svagheder ved disse metoder og hvordan disse metoder kan bruges i retssystemet.

## Nærmere beskrivelse af emnet

### Matematikdelen

Her følger først en introduktion til bayesiansk sandsynlighedsregning, som bruges i projektet. Denne type sandsynlighedsregning adskiller sig væsentligt fra den sædvanlige, frekvensmæssige tilgang.

Frekvensmæssig tilgang:
Ofte bruger vi sandsynlighedsregning til at regne sandsynligheder ud på baggrund af *viden som vi allerede har*. Eksempelvis viser optællinger gennem mange år, at af alle nyfødte er 50.9 % piger. Med denne viden kan vi sige, at hændelsen A: "En tilfældig valgt nyfødt baby er en pige" forekommer med sandsynligheden P(A)=0,509.

Bayesiansk tilgang:
Hvis man ikke har adgang til data, som kan bruges som grundlag for en frekvensmæssige tilgang, kan forsøge at anvende den bayesianske tilgang, hvor man undersøger *sandsynligheder, givet visse betingelser*.
Det smarte ved den bayesianske tilgang er, at man gennem disse betingelser kan indarbejde sandsynligheder, som bygger på et objektivt grundlag (fx andelen af nyfødte piger) med sandsynligheder, som bygger på et mere subjektivt grundlag.
Eksempelvis er det umiddelbart svært at sige hvor sandsynligt udsagnet:
A: "FCK vinder det danske mesterskab i fodbold i 2010" er.

At FCK har vundet mesterskabet et antal gange tidligere i historien kan ikke direkte bruges til at forudsige P(A) modsat tilfældet med andelen af nyfødte piger. En person kan selvfølgelig på baggrund af sin viden om FCK og dansk fodbold give sin helt personlige vurdering om chancerne. At sandsynligheden for A afhænger af personens viden, lad os kalde denne $K_1$, kan skrives som $P(A|K_1)$. En anden person kan dog sidde inde med en helt anden viden $K_2$, som har en radikalt anden indflydelse på sandsynligheden for A.
F.eks. kan denne anden person vide, at der faktisk er stor risiko for at FCK i 2009 kommer i økonomisk krise, som medfører at de må sælge deres bedste spillere, hvormed sandsynligheden for hændelsen A ville reduceres drastisk.
Det er netop sådanne subjektive vurderinger af en situation, som kan bruges f.eks. af jurymedlemmer, når de skal vurdere sandsynligheder for at visse hændelser er indtruffet.

Centralt i den bayesianske tilgang står den såkaldte Bayes' formel: $P(A|B) = P(B|A)*P(A)/P(B)$. (Se eksempelvis Michael Sørensen noter s. 25 for et bevis for formlen). Her regnes P(B) ofte som den marginaliserede sandsynlighed $\sum P(B|A_i)P(A_i)$

**Eksempel:**
(inspireret fra siden http://www.dcs.qmw.ac.uk/%7Enorman/BBNs/BBNs.htm)
Antag at sandsynligheden for hændelsen A: "S-togene er forsinkede" er P(A) = 0,1. Derfor er $P(\neg A)$ = 0,9.
Lader nu Martin og Norman være to arbejdskollegaer og betragter hændelserne
B: "Martin møder for sent på arbejde."
C: "Norman møder for sent på arbejde."

Norman kører normalt med tog på arbejde, så han vil være mere påvirket af en eventuel forsinkelse hos DSB end Martin, som dog også forsinkes af togforsinkelser, da der i så fald vil være flere biler

på vejene. Dette kommer til udtryk i følgende betingede sandsynligheder. Vi antager, at der uden togforsinkelser er 50 % chance for at Martin, som har lidt svært ved at komme op af sengen om morgenen, møder for sent:
$P(B|\neg A) = 0,5$ og $P(\neg B|\neg A) = 0,5$
Hvis der er forsinkelser antager vi at chancen for at Martin kommer for sent forøges til 60 %:
$P(B|A) = 0,6$ og $P(\neg B|A) = 0,4$.
Norman er derimod mere morgenmenneske end Martin, så chancen for at Norman møder for sent uden nogle togforsinkelser er blot 10 %:
$P(C|\neg A) = 0,1$ og $P(\neg C|\neg A) = 0,9$
Da Norman er en trofast bruger af DSB's togsystem, så rammes han hårdere af eventuelle forsinkelser end Martin. Vi antager, at chancen for at Norman møder for sent på arbejde en dag, hvor togene er forsinkede er 80 %:
$P(C|A) = 0,8$ og $P(\neg C|A) = 0,2$.

Den marginale sandsynlighed, som vi omtalte før, kan nu udregnes, og den fortæller os, hvad den samlede sandsynlighed er for at hhv. Martin og Norman kommer for sent:
$P(B) = P(B|A) * P(A) + P(B|\neg A) * P(\neg A) = 0,6*0,1 + 0,5*0,9 = 0,06 + 0,45 = 0,51$
$P(C) = P(C|A) * P(A) + P(C|\neg A) * P(\neg A) = 0,8*0,1 + 0,1*0,9 = 0,08 + 0,09 = 0,17$.

En af pointerne ved den bayesianske fremgangsmåde er, at man kan ændre på sandsynlighederne, alt efter hvor mange oplysninger er tilgængelige. F.eks. kan man se på hvilken indflydelse det har på de øvrige sandsynligheder, hvis vi ved at Norman er kommet for sent:
Vha. Bayes formel bliver sandsynligheden for at der er togforsinkelse nu ændret til at være
$P(A|C) = P(C|A) * P(A)/P(C) = 0,8 * 0,1/0,17 = 0,47$.

Vi kan altså konkludere, at det faktum, at Norman møder for sent på arbejde øger sandsynligheden for at der er forsinkelser hos DSB fra at være 0,1 til at være 0,47. Man kan derfor sige, at viden om, at en hændelses sandsynlighed ændres medfører, at der forplanter sig en ændring af de andre sandsynligheder i nettet.

Man kan tilføje flere nye parametre til systemet, hvis der er behov for dette. Dette vil selvfølgelig have indflydelse på sandsynlighederne gennem hele systemet. For eksempel kunne vi tilføje hændelsen D: "Martin sover over sig", som selvfølgelig vil have betydning for P(B).
Lad os antage, at $P(D)=0,4$ og $P(\neg D)=0,6$. et spørgsmål man nu kunne stille sig var, at hvis man vidste, at Martin kom for sent, hvad er sandsynligheden så for at Martin har sovet over sig og hvad er sandsynligheden for, at der har været en togforsinkelse? Sådanne spørgsmål er det oplagt at regne ud vha. Bayes formel, men udregningerne bliver hurtigt lidt besværlige. Eksempelvis er sandsynligheden for at der er en togforsinkelse, når det er givet at Martin kommer for sent:
$P(A|B) = P(B|A) * P(A)/P(B) = (0,8 * 0,4 + 0,6 * 0,6) * 0,1/0,51 = 0,133$ og den er altså vokset en smule, da $P(A) = 0,1$ før vi fik oplysningen om at Martin var for sent på den.

Jo flere parametrene er, desto mere kompliceret bliver situationen også, og derfor er det en stor fordel at tegne de forskellige hændelser ind i et kausalt netværk, hvor pile angiver, hvilke hændelser er afhængige af hinanden. Der findes udmærkede og lettilgængelige computerprogrammer (eks. Hugin, se links), hvor man kan beskrive sit netværk og tildele de forskellige hændelser nogle sandsynligheder. Derefter kan man så lade programmet regne på, hvad der sker, hvis man får flere oplysninger, dvs. hvilken indflydelse det f.eks. har på resten af systemet, hvis vi ved, at Martin møder for sent på arbejdet, jvf. eksemplet ovenfor.

**Bayesiansk matematik i retssalen**
Vi vil nu også give et specifikt eksempel på, hvordan den bayesianske tilgang kan bruges i retssalen. Specielt vil vi se på det tilfælde, hvor DNA-materiale fra den formodede gerningsmand er det eneste konkrete bevismateriale i sagen, såkaldt "cold hit probability" (beskrevet i de to artikler af Devlin, som der er links til).

I grove træk kan anvendelse af Bayes formel i retssager kan beskrives på følgende vis:
Vi definerer hændelserne
S: Den anklagede person er skyldig
D: Den anklagedes DNA-profil matcher det DNA, som blev fundet på gerningsstedet.

Hvis man nu antager at man har fundet en matchende DNA-profil på gerningsstedet, så kan man regne på sandsynligheden for at den anklagede er skyldig:
$P(S|D) = P(D|S) * P(S)/P(D)$.

På højre side er leddet $P(D|S) = 1$, da man netop har fundet en matchende DNA-profil. P(S) er sandsynligheden for at den anklagede er skyldig, når man ser bort fra DNA-bevismaterialet. Hvis der ikke er andet materiale at gå ud fra, kan man her anvende fakta som, hvor mange der bor i lokalområdet, sandsynligheden for at vidneudsagnene er rigtigt osv., men det er klart, at her bliver bevisførelsen og udregningerne mere uklare.
Leddet P(D) beskriver sandsynligheden for at der findes et match mellem den anklagedes DNA-profil og DNA-materialet på gerningsstedet. Her skal man igen regne den marginale sandsynlighed ud for at finde $P(D) = P(S) * P(D|S) + P(\neg S) * P(D|\neg S)$. Det sidste led på højresiden beskriver det tilfælde, hvor den anklagede er så uheldig at være uskyldig, men at hans DNA-materiale alligevel findes på gerningsstedet.

**Et praktisk eksempel**
I retssagen mod Adams i 1996, hvor han var anklaget for voldtægt (se links) var det eneste bevis mod ham, at hans DNA fandtes på gerningsstedet. Offeret kunne ikke genkende ham, og Adams havde også et alibi, som dog afhang af hans kæreste. Anklagernes argument var, at chancen for at en tilfældig mands DNA-profil matchede DNA-materialet fra gerningsstedet var $P(D)=1/200.000.000$. Dette virker ved første indtryk som et overvældende godt bevismateriale. Adams' forsvarere valgte på den anden side at inddrage Bayes formel i et håb om at vise, at det kan være problematisk at dømme en person udelukkende ud fra DNA-materiale.
Adams' forsvarere så på det bevismateriale i retssagen, som ikke var knyttet til DNA-undersøgelser og påviste, at sandsynligheden for at Adams var skyldig ud fra dette materiale blot var $P(S)=1/3.600.000$. Dette tal fandt de frem til ved at stille nogle spørgsmål/hændelser, hvorpå jurymedlemmerne kunne hæfte deres egen sandsynlighedsvurdering. Forsvarerens bud på disse sandsynligheder var:

- 75 % sandsynlighed for at gerningsmanden var fra lokalområdet og i 18-60.
- 90 % sandsynlighed for at den anklagede ikke ville blive genkendt af offeret, hvis den anklagede var uskyldig.
- 25 % sandsynlighed for at den anklagedes alibi holdt vand, hvis han var skyldig. 50 % i tilfældet, hvis den anklagede var skyldig.

Disse oplysninger kan vha. Bayes formel bruges til til at give en vurdering af sandsynligheden for, hvorvidt Adams var skyldig:
$P(S|D) = P(D|S) * P(S)/P(D)$
$= 1 * (1/3.600.000) / ( (1/3.600.000) * 1 + (1-1/3.600000) * (1/200.000.000) ) = 0,9823.$

Altså er sandsynligheden for at Adams var gerningsmanden reduceret til at være omkring 54/55. Stadig rimelig sandsynligt, men alligevel en væsentlig forbedring, idet der her dog må siges at findes en realistisk sandsynlighed for, at Adams er uskyldig (ca. 1,8 %).
Jurymedlemmerne kunne selvfølgelig komme med deres egen vurdering over sandsynlighederne i de tre punkter her ovenfor. På denne måde bliver Bayes formel et redskab hos juryen til at regne på sandsynligheden for, hvorvidt den anklagede var skyldig. Det viste sig at være en del bekymring og forvirring over, hvorvidt denne anvendelse af Bayes formel lod sig gøre på en rimelig måde. Det endte dog med, at juryen fik et spørgeskema, hvor de skulle angive forskellige vurderinger i %. I spørgeskemaet var der så inkluderet en udregningsformel. Sagen endte med, at Adams blev dømt skyldig.

**Biologidelen**
Bayesiansk matematik er en type matematik, der blandt andet finder anvendelse i retssager. En vigtig del af den biologiske del af opgaven er, at forklare og diskutere baggrunden for dette.

Relevante aspekter i den forbindelse er:

- DNAs opbygning og funktion.
- Gennemgang af forskellige metoder til bestemmelse af DNA-sekvenser, fx RFLP (som faktisk ikke bestemmer DNA-sekvenser, men alligevel er meget brugt) og egentlig sekvensering. Herunder diskussion af disse metoders sikkerhed, fx med fokus på muligheder for fejl introduceret under PCR.
- I forbindelse med RFLP kunne eleverne udregne sandsynligheder for hvor ofte specifikke DNA-sekvenser for restriktionsenzymer kan forventes at optræde tilfældigt.
- Matematisk set kan DNA anskues som en kode med fire mulige tegn. På hvilke måder er denne anskuelse rigtig og forkert biologisk set?
- Gennemgang af mekanismer, der skaber forskelle og ligheder i arvematerialet, herunder gennemgang og evt. udregning af ligheder indenfor familier, i etnisk homogene/heterogene samfund (fx kunne Hardy/Weinberg-ligevægt inddrages i den forbindelse).
- Diskussion af hvorfor nogle DNA-sekvenser er meget konservative og hvorfor nogle udviser stor variation.
- Diskussion af fordele og ulemper ved identifikation vha. fænotypiske egenskaber (fx køn, udseende og blodtype) i forhold til genotypiske egenskaber (altså DNA-sekvenser).
- Generelle overvejelser over præcision i laboratoriearbejde –er det rimeligt at antage laboratorieresultater som matematiske sandheder?

Vigtigt biologiske aspekt:
I den matematiske del bruges produktreglen (at sandsynligheden for to sandsynligheder kan findes som produktet af disse) flittigt – dette forudsætter at begivenhederne er uafhængige. Vi ved jo at DNA-sekvenser ikke er uafhængige fx indenfor en familie og mellem koblede gener hos et individ. Diskussion af, om og hvornår man med rimelighed kan antage uafhængighed mellem gener er derfor centralt (Devlin nævner netop denne problematik i sine artikler, se links).

## Variationsmuligheder:

**Biologi:**
De skitserede muligheder er allerede langt mere end der kan inddrages i én opgave – variationsmuligheder kommer derfor i udvælgelsen.
Der ud over kan man fokusere på forskellige metoder til DNA-analyse: RFLP, sekvensering af enkeltgener, brug af DNA-prober, forskellige typer blotting eller andre metoder. Særligt interessant kunne være at inddrage de til enhver tid seneste teknikker.

**Matematik:**

- Fokus på cold hits (DNA fundet i forbryderkartoteker og altså ikke blandt mistænkte).
- Fokus på sammenkædning af mistænkte og DNA-spor fundet på et gerningssted.
- Diskussion af faktiske sager og matematikken brugt i de sager – der er i den forbindelse en del kontroversielle afgørelser fra de amerikanske domstole, som kan inddrage og danne godt diskussionsgrundlag.
- Fokus på de mere generelle aspekter af bayesianske netværk og opstilling af disse og altså mindre fokus på de specifikke juridiske aspekter.
- Et mere ambitiøst oplæg kunne inkludere beviser for nogle af de formler og sætninger, som bliver anvendt i den sandsynlighedsteoretiske del.
- Betingede sandsynligheder virker ofte kontraintuitive (et kendt eksempel er Monty Hall problemet) og en stor del af litteraturen om Bayesiansk matematik i retssager, handler om *fejlagtig* brug. En mulig opgave kan derfor være at give eleven et konkret oplæg (fx http://www.dcs.qmul.ac.uk/~norman/papers/jury_fallacy.pdf) og give eleven til opgave at forklare, hvad de forståelsesmæssige faldgruber er og hvorfor matematikken eventuelt er rigtig (og selvfølgelig tage kritisk stilling til om matematikken *er* rigtig).

# Litteratur

Finn V Jensen: ”Introduction to Bayesian Networks”. Institut for matematik og datalogi, Aalborg Universitet 1993.
Michael Sørensen: ”En introduktion til sandsynlighedsregning”, Københavns Universitet 2000 (Bayes formel på s. 25)

## Websider:

Hvis du har problemer med nogle af nedenstående links så copy/paste dem direkte ind adressefeltet på i din browser. Alle links er tjekket august 2016.

http://people.math.aau.dk/~svante/Cafe05.html (side med links til dansk litteratur om bayesiansk sandsynlighedsregning og bayesianske netværk)
http://www.hugin.com/Products_Services/Products/Demo/ (Demoversion af programmet ”Hugin”)
http://en.wikipedia.org/wiki/Bayesian_statistics (Et afsnit, der beskriver matematikken bad anvendelsen af Bayes formel i retssager)
http://www.eecs.qmul.ac.uk/~norman/BBNs/BBNs.htm (En fin og relativt lettilgængelig introduktion til Bayesisk sandsynlighedsregning)
http://www.inference.phy.cam.ac.uk/mackay/itila/book.html (link til bog om emnet)
http://www.bbc.co.uk/dna/h2g2/A801695 (en kort introduktion til bayesiansk matematik og problemstillinger)
http://www.dcs.qmw.ac.uk/%7Enorman/BBNs/BBNs.htm (opslagsværk om bayesiansk matematik)
http://en.wikipedia.org/wiki/Bayes%27_theorem (gennemgang af bayes formel)

**Konkrete sager og mere specifikke elementer:**

http://en.wikipedia.org/wiki/Prosecutor%27s_fallacy (gennemgang af matematik og sager med problematiske aspekter vdr. Bayesiansk matematik)
http://www.dcs.qmul.ac.uk/~norman/papers/jury_fallacy.pdf (En fin artikel, der beskæftiger sig med typiske eksempler på misforstået anvendelse af Bayesisk matematik i retten. Viser hvordan anvendelse af computerprogrammer kan klargøre anvendelsen af Bayesisk sandsynlighedsregning.)
https://www.maa.org/external_archive/devlin/devlin_09_06.html (Interessante overvejelser, hvor bl.a. ”cold hit” fremgangsmåden kritiseres. Problematikken omkring antagelsen om uafhængighed mellem to DNA-sekvenser nævnes også.)
http://homepages.law.asu.edu/~kayed/pubs/evid/dreyfus.pdf (I fodnoterne er der referencer til flere kontroversielle retssager, hvor DNA-materiale har haft betydning.)
http://www.time.com/time/magazine/article/0,9171,838296,00.html (Eksempel på, hvordan sandsynlighedsregning kan misbruges i retssager. Kræver abonnement.)
http://en.wikipedia.org/wiki/People_v._Collins#fn_4_back (Eksempel på, hvordan sandsynlighedsregning kan misbruges i retssager)
http://en.wikipedia.org/wiki/Howland_will_forgery_trial (et eksempel på brug af sandsynlighedsregning i en retssag)
http://en.wikipedia.org/wiki/Sally_Clark (et eksempel på problematisk brug af sandsynlighedsregning i en retssag)
http://en.wikipedia.org/wiki/Regina_versus_Denis_John_Adams (Retssagen mod Adams)
http://www.maa.org/devlin/devlin_10_06.html (Igen kritiseres brugen af DNA-materiale i retssalen bl.a. ud fra uafhængighedsantagelsen)