# Familiegenetikk og mutasjoner: betydningen av modellvalg

## Family genetics and mutations: the significance of model choice

Maria Berggreen

UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Institutt for kjemi, bioteknologi og matvitenskap

Masteroppgave 30 stp. 2013

# Abstract

Family genetics is used in forensics to find family relations between individuals, and especially in paternity cases. In a standard paternity case the child and alleged father will share at least one allele in every marker. This is not the case for non-standard paternity cases where there are small genetic differences that make it more complicated to tell if this man is the father of the child. The small genetic differences could arise from mutations, but in some circumstances it could be the case that a close relative of the man is the actual father. This makes it important to have a precise estimate for the probability of the fatherhood.

In this thesis it has been examined how the results in paternity tests are affected by different factors. The program Familias has been used to calculate the probabilities of the different propositions for the father. Four different mutation models have been used in Familias, and it has been examined how the number of alleles, excessive persons and the nomenclature of alleles have impact on these mutation models. This was done to determine which of the four models that are the most appropriate to use in calculations of paternity cases. In this thesis the number of alleles seemed to have the highest impact on the results, and a stable model with mutation probability decreasing with range was considered to be the most appropriate.

# Sammendrag

Familiegenetikk brukes i slektskapssaker i juridisk sammenheng, og først og fremst innen farskapssaker. I en standard farskapssak vil far og barn dele minst ett allel i hver markør. Ved en ikke-standard farskapssak er det små genetiske forskjeller mellom barn og den angivelige far som gjør det vanskelig å avgjøre om mannen er den biologiske faren til barnet. Disse genetiske forskjellene kan oppstå ved mutasjon, men dataene kan også i noen tilfeller tyde på at en nær slektning av mannen er barnets far, ettersom nære slektninger gjerne har relativt like DNA-profiler. Det er derfor viktig å ha mest mulig nøyaktige beregninger av sannsynligheten for farskapet.

I denne oppgaven er det blitt undersøkt hvordan forskjellige faktorer påvirker resultatene ved farskapstesting. Programmet Familias ble brukt til å beregne sannsynligheter for forskjellige hypoteser om hvem som er den biologiske faren til et barn i et reelt eksempel. Det er blitt brukt fire forskjellige mutasjonsmodeller i Familias, og det ble studert hvordan antall alleler, overflødige personer og nomenklatur av allelene påvirker resultatene i disse fire modellene. Dette ble gjort for å finne ut hvilken av de fire modellene som er mest hensiktsmessig å bruke i beregninger av farskapssaker. I denne oppgaven virket antall alleler som det mest utslagsgivende på resultatene, og en stasjonær mutasjonsmodell som tar hensyn til mutasjonsrekkevidde ble antatt å være den beste modellen.

# Forord

Denne masteroppgaven er utført ved Institutt for Kjemi, Bioteknologi og Matvitenskap ved Universitetet for Miljø- og Biovitenskap (UMB) i perioden februar til mai 2013. Det har vært en spennende og lærerik prosess.

Jeg vil rette en stor takk til:

- Veiledere Thore Egeland og Guro Dørum for svært god oppfølging. Dere har begge vært med på å øke min interesse for faget.
- Geir Bartz-Johannessen og Helge Skivenes for innspill og korrekturlesing.

Til slutt vil jeg takke alle Villabarna i Villa Løkkeberg for husly og støtte i masterperioden. Denne oppgaven hadde ikke blitt fullført uten dere.

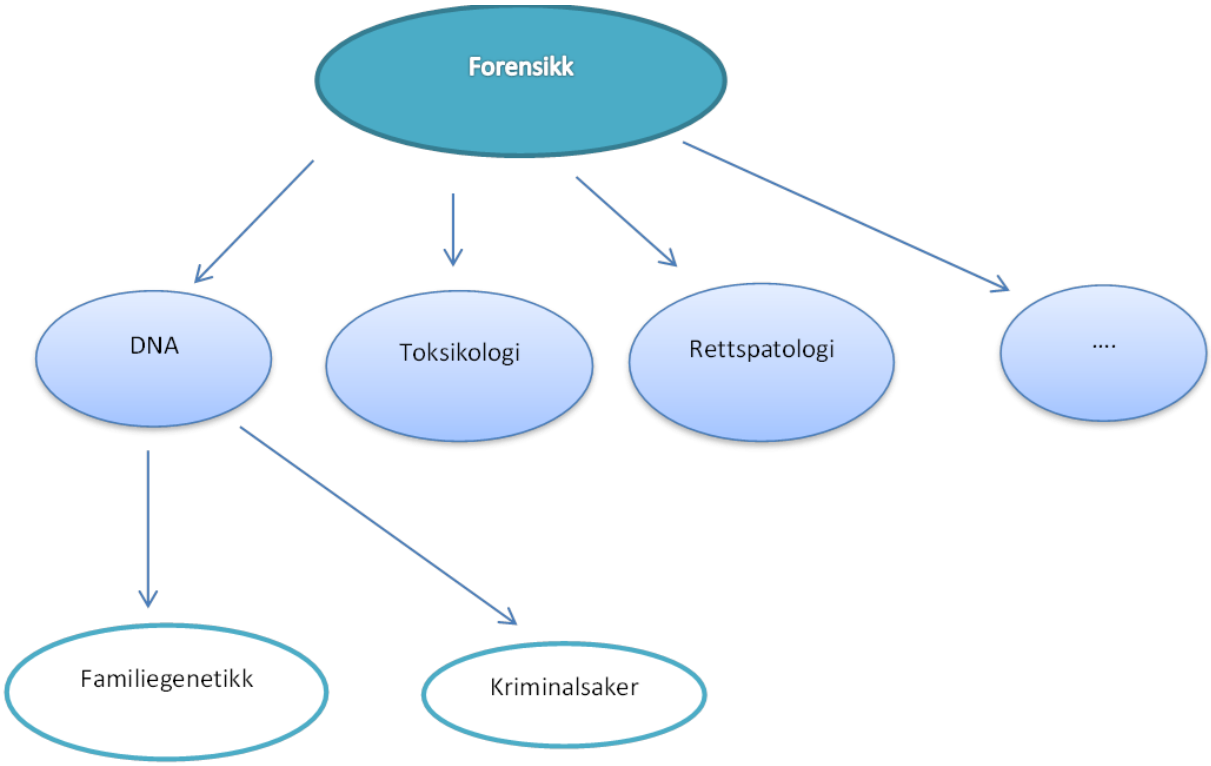
Ås, mai 2013.

Maria Berggreen

# Innhold

# 1 Innledning

Forensikk kommer fra det engelske ordet «forensics» og brukes om vitenskap i etterforskning og juridisk anvendelse. Ordet har ikke blitt godkjent i den norske ordboken enda, men blir brukt innenfor fagmiljøet. På norsk brukes adjektivet forensisk om rettslige anvendelser (lenke [3]). Tidligere har en gjerne brukt ordet rettsmedisin, men dette ordet blir ikke dekkende nok for alle de forskjellige områdene. Rettsmedisin vil si medisinsk kunnskap i juridisk anvendelse, men statistikk for eksempel er ikke en medisinsk gren. Forensikk blir dermed et mer beskrivende ord for de forskjellige fagfeltene, og brukes derfor i denne oppgaven.




**Figur 1.1:** Oversikt over noen av de forskjellige områdene innen forensikk. DNA, toksikologi og rettspatologi er undergrupper av forensikk, og familiegenetikk og kriminalsaker er undergrupper av DNA.

Forensikk kan deles inn i mange forskjellige områder, noen av undergruppene, som vises i Figur 1.1, er DNA, toksikologi og rettspatologi. Rettspatologi brukes når det er behov for å undersøke mistenkelige dødsfall, og for å finne ut hva som forårsaket dødsfallet. Toksikologi brukes ved kjemisk analyse av stoffer, og kan blant annet brukes for å undersøke en eventuell forgiftning eller ved mistanke om doping.

Det området som har blitt mer og mer brukt de siste årene innenfor rettslige forhold er DNA-bevis og DNA-profiler. Det brukes som bevis ved voldtektssaker, slektskapssaker og kan i tillegg brukes til å identifisere lik. DNA i rettsgenetiske anvendelser kan deles inn i undergruppene kriminalsaker (eller sporsaker) og familiegenetikk. Temaet for denne oppgaven er familiegenetikk, og identifikasjon er et nøkkelord i den sammenheng. Grunnen til at DNA kan brukes til å identifisere individer er at en DNA-profil viser stor variasjon fra person til person. For å lage en DNA-profil undersøker en spesifikke steder på DNA-molekyler, såkalte markører. Boken Butler (2012) er en standard referanse når det gjelder markører i rettsgenetiske anvendelser. Dersom et tilstrekkelig antall markører undersøkes, utgjør disse markørene til sammen en tilnærmet individuell profil for den enkelte personen. Ved for eksempel en voldtekt kan det samles DNA-bevis som deretter sammenliknes med en DNA-prøve fra den mistenkte, og dersom DNA-prøvene stemmer overens med hverandre sies det å være en match. Når DNA-profilene matcher er dette et svært sterkt bevis mot den mistenkte, men det trenger allikevel ikke være ensbetydende med at mannen er skyldig. Statistisk sett er det svært lite sannsynlig at en tilfeldig person vil ha den nøyaktig samme profilen, men hva om mannen har en enegget tvilling? Dette har vært tilfellet i en voldtektssak i Frankrike for ikke så lenge siden, se lenke [1] og [2].

I boken Buckleton, Triggs og Walsh (2005) beskrives de statistiske metodene som brukes i forensikk. For denne oppgaven er spesielt kapitlene 2 og 10 relevante. Det er ofte ønskelig å fremlegge statistiske bevis som baseres på DNA-beviset. Det kan beregnes sannsynligheter som taler for eller mot en hypotese, for eksempel om en mann er far til et barn eller ikke. Ved farskapssaker beregnes det gjerne en likelihood-ratio (LR) som taler for eller mot farskapet, og for å beregne denne LR-verdien kreves det *a posteriori*- og *a priori*-sannsynligheter.

*A priori*-sannsynligheten er sannsynligheten for en gitt hypotese før det er fremlagt et bevis, gjerne i form av DNA. Dette er altså hva en vet før DNA analysen foreligger. I en farskapssak kan en sjeldent gi sannsynligheter som veier for eller i mot et farskap før eventuelle genotyper er

undersøkt. Det blir derfor ikke lagt stor vekt på *a priori*-sannsynligheter i farskapssaker. *A posteriori*-sannsynligheten for farskap er sannsynligheten for at en person er far til et barn etter at det er gjort DNA-tester av barn og foreldre. En likelihood er en betinget sannsynlighet for dataene. Det kan for eksempel være sannsynligheten for et barns genotype når en vet at en mann er faren. En likelihood-ratio eller en sannsynlighetskvote sier noe om hvor sannsynlig en hypotese er i forhold til en annen. Dette blir beregnet ved å lage en brøk av to likelihood-verdier. For eksempel hvor sannsynlig dataene er gitt at en mann er far til et barn i forhold til sannsynligheten gitt at han ikke er far. Disse sannsynlighetene blir beskrevet mer detaljert i materialer og metoder.

Denne oppgaven handler farskapssaker, og beregningen av sannsynligheten for et eventuelt farskap. Området innenfor slektskapssaker i juridisk sammenheng kalles familiegenetikk.

## 1.1 Familiegenetikk

Familiegenetikk blir først og fremst brukt til DNA-testing ved avklaring av farskap, men det kan også brukes til å finne ut om personer er søsken (eller fjernere slektskap), eller om en kvinne er mor til et barn. Det blir først og fremst foretatt farskapstester i sivile saker der det er tvil om hvem som er den biologiske faren til barnet. Det kan for eksempel være usikkerhet om hvem som er forelderen til et barn hvis en kvinne har hatt flere partnere. Ved tvil om hvem som er far til barnet i en rettssak, kan det bli brukt familiegenetiske metoder for å bistå ved en eventuell sak for domstolen. DNA-testing og DNA-profiler kan også bli brukt som bevis i saker angående familiegjenforening av innvandrere og asylsøkere.

Ved en standard farskapssak vil alle markørene som blir testet hos far og barn stemme med hverandre. Dette betyr at et barn har fått et allel fra mor og et allel fra far og vil da dele minst ett allel med far i alle markørene som blir testet. Dette gjør sannsynligheten relativ enkel å beregne når dette er tilfellet. I andre slektskapssaker kan beregningene være litt mer kompliserte da søsken kan dele 0, 1 eller 2 alleler dersom de har samme mor og far. Ved fjernere slektskap blir det en litt større utfordring å beregne sannsynligheter for hvor mange alleler individene deler.

I denne oppgaven fokuseres det på ikke-standard farskapssaker. Mer spesifikt skal saker der mutasjoner kan ha forekommet studeres. Slike saker kan bli vanskelige, DNA-beviset kan bli svakt, fordi alternative forklaringsmodeller slik som at en nær slektning av utpekt far er sann far, kan fremstå som sannsynlige.

## 1.2 3 prinsipper for hypoteser

Formulering av hypoteser er et sentralt tema innen forensikk. Det dannes gjerne hypoteser ut ifra teori eller mistanker. Når hypotesen er formulert kan det beregnes hvor sannsynlig data er gitt hypotesen, ut ifra statistiske beregninger basert på observerte data. Formulering av en hypotese er i seg selv svært viktig, hvis hypotesen for eksempel skal presenteres i en rettssak må det legges vekt på hvordan en ordlegger seg, slik at hypotesen ikke blir feiltolket.

For å kunne si noe om de observerte data må det stilles påstander eller hypoteser slik at det deretter kan uttrykkes hvilken usikkerhet eller sannsynlighet data har gitt de forskjellige hypotesene. I artikkelen «Statistisk vekting av DNA-funn i straffesaker» (Egeland, 2009, 193-194) blir det fremlagt følgende siterte prinsipper:

> «1. For å kunne bedømme usikkerheten i en påstand må den stilles opp mot minst én alternativ påstand
>
> 2. Vitenskapelig fortolkning baseres på spørsmål av typen «hvor sannsynlig er funnet, gitt en påstand?»
>
> 3. Vitenskapelig fortolkning må ta eksterne forhold i betraktning, i tillegg til konkurrerende forklaringer.»

Prinsipp 1 sier at det må stilles opp flere hypoteser mot hverandre. For eksempel hvis en vil finne ut om en DNA-profil tilhører en mann som er mistenkt i en rettsak må det først stilles en påstand om at DNA-profilen tilhører denne mannen, deretter må det også stilles en påstand om at DNA-profilen ikke tilhører mannen, men en tilfeldig person, eller en annen mistenkt hvis det finnes flere mistenkte i saken. Poenget er at det ikke er tilstrekkelig med kun én hypotese da det kan finnes flere mulige forklaringer til saken eller eventuelle DNA-bevis, og selv om en hypotese er

usannsynlig kan det hende at alle andre hypoteser eller forklaringsmodeller er enda mer usannsynlige. Dersom dette er tilfellet vil hypotesen som er minst usannsynlig være den mest sannsynlige hypotesen. Det er derfor viktig å ha flere hypoteser å sammenlikne med.

Prinsipp 2 vil i en farskapssak bety at en ser på sannsynligheten for genotypene som er funnet gitt at en mann er far eller ikke. Det angis ikke en direkte sannsynlighet for om en mann er far til barnet, men en sannsynlighet for barnets genotype gitt far og mors genotyper dersom mannen er far eller ikke.

I prinsipp 3 sies det at resultatet en får ut ifra data må sees i sammenheng. Dette prinsippet er nok mer gjeldene for rettsaker der en mistenkt kan ha et alibi. Ved en farskapssak kan det tenkes at forklaringen til en høy sannsynlighet for farskapet, dersom han ikke er barnets biologiske far, er at han har en svært lik DNA-profil til en bror.

## 1.3 Den frekventistiske metoden

Frekventistisk statistikk blir også gjerne kalt klassisk statistikk. Parameterne som brukes i frekventistisk statistikk er gjerne fikserte, men ukjente. I denne metoden blir det brukt objektive sannsynligheter, altså beregnes sannsynlighetene bare ut ifra dataene som observeres, og en finner da sannsynligheten for de observerte data gitt en hypotese (Buckleton, Triggs, Walsh, 2005).

Den frekventistiske metoden brukt i forensikk er lik generell sannsynlighetsteori der en går ut ifra en nullhypotese og en alternativ hypotese. Ved generell sannsynlighetsteori kan en ved et gitt signifikansnivå forkaste nullhypotesen, det er i midlertidig ikke vanlig å bruke denne formen for hypotesetesting innen familiegenetikk. Den frekventistiske metoden har forskjellige fremgangsmåter. En kan enten motbevise en sak ved å påstå at beviset er høyst usannsynlig dersom nullhypotesen er sann. Dette kan for eksempel være en DNA-match ved en ren tilfeldighet, der det er blitt tatt en DNA-prøve fra en tilfeldig person som ikke er skyldig. Hvis sannsynligheten for en tilfeldig match er svært lav, vil dette gi støtte til den alternative hypotesen som i dette tilfellet sier at det ikke er en tilfeldig match og at personen er skyldig. Jo lavere sannsynlighet for den gitte DNA-profilen, jo mer støtte til den alternative hypotesen.

Den andre fremgangsmåten går ut på at tilfeldige mennesker vil ha en svært stor sannsynlighet for å bli ekskludert som mistenkt, og hvis en person ikke er blitt ekskludert er sannsynligheten stor for at det er den skyldige. Imidlertid blir eksklusjon prinsipielt umulig i en farskapssak der mutasjoner kan forekomme.

## 1.4 Den Bayesianske metoden

Med den Bayesianske metoden beregnes sannsynligheten for en hypotese gitt data, altså det motsatte av den frekventistiske metoden. For å beregne sannsynligheter med den Bayesianske metoden må det tillegges *a priori*-sannsynligheter. Dette er sannsynligheter som blir satt før de observerte dataene er tatt i betraktning. *A priori*-sannsynligheten blir formet av en subjektiv oppfatning, altså hva en person anser som sannsynlig om en gitt situasjon før bevis eller beregninger er fremlagt. Ettersom *a priori* ikke baserer seg på det en observerer, for eksempel et DNA-bevis, kalles det for subjektive sannsynligheter.

I en farskapssak uttrykker *a priori*-fordelingen sannsynligheten for farskapet før genotypene til foreldre og barn er undersøkt. *A priori*-oddsen er ratioen mellom sannsynlighetene for hver av de to hypotesene. Dersom *a priori*-oddsen er lik 1 tilsvarer dette at sannsynligheten for den første hypotesen er like stor som sannsynligheten for den alternative hypotesen. I en farskapssak vil dette som oftest være tilfelle når en ikke har noen forutsetninger for å si at en mann er mer sannsynlig som far fremfor en annen.

*A posteriori*-sannsynligheten er sannsynligheten for hypotesen etter at data er observert, for eksempel etter DNA-testing ved farskapssaker. Den uttrykker da en sannsynlighet for farskapet gitt markørene som blir undersøkt (Buckleton, Triggs, Walsh, 2005).

Hvis en hypotese er at en mann er far til et barn, kan en ved denne metoden uttrykke en sannsynlighet for om han er den biologiske faren ut ifra genotypene som blir observert hos mannen, barnet og mor. Sannsynligheten kan skrives som P(far|data). I dette tilfellet vil *a priori*-sannsynligheten uttrykkes som P(far), og en kan da beregne *a posteriori*-sannsynligheten P(far|data) fra P(data|far) og P(far) ved å bruke Bayes' teorem. Bayes' teorem blir forklart mer detaljert i Avsnitt 2.2.3.

## 1.5 Motiverende eksempel. Hvorfor ta hensyn til mutasjoner

Det er ønskelig å finne ut om en mann er far til et gitt barn. Mannen testes for 21 forskjellige markører som hver inneholder mellom 8-21 alleler. Det foretas en farskapstest på bakgrunn av følgende hypoteser:

$$H_1: \text{Kandidat er far til barnet}$$
$$H_2: \text{Kandidat er ikke far til barnet}$$

Det kan bare påstås at den første hypotesen er sann dersom LR-verdien er høy nok.

**Tabell 1.1:** LR-verdier for hypotese 1 og 2 ved mutasjonsrate R=0 og R=0,001, mutasjonsrekkevidde r=0,5.

| | R=0 | R=0,001 |
|---|---|---|
| H2 | 1 | 1 |
| H1 | 0 | 4515433 |

Videre kan det tenkes at mannen passer i alle alleler i de 21 markørene bortsett fra ett eneste allel, altså deler mannen og barnet minst ett allel for 20 markører. Hvis det ikke blir tatt hensyn til eventuelle mutasjoner vil dette tilsvare at det antas en mutasjonsrate på R = 0. Ved å sette mutasjonsraten lik null utelukker en muligheten for at et allel kan ha mutert. Dette betyr at det ikke har noen innvirkningen på resultatet hvor mange markører som er like hos kandidaten og barnet så lenge det er ett enkelt allel som er ulikt.

Det kan sees i Tabell 1.1 at LR-verdien for $H_1 = 0$ når mutasjonsraten R = 0. PENTA_E er eneste markør som viser inkompatibilitet mellom kandidaten og barnet, men LR-verdien vil i dette tilfellet bli null uavhengig hvor mange markører som er kompatible mellom den angivelige far og barn når modellen ikke tar hensyn til mutasjoner.

Hvis det deretter brukes en modell i Familias for PENTA_E som tar hensyn til mutasjonsrater og avstand mellom allelene, vil dette endre resultatene. Mutasjonsraten kan for eksempel settes lik R = 0,001 og mutasjonsrekkevidde r = 0,5. Modellene og parameterne vil bli forklart senere i Avsnitt 2.3.

Ved disse parameterne gitt for PENTA_E blir resultatet som i Tabell 1.1, og det kan sees i tabellen at LR-verdien for den første hypotesen blir 4515433. Dette betyr i følge de nye beregningene at genotypedataene er 4515433 ganger mer sannsynlig dersom vi legger til grunn at mannen er far til barnet mot at han ikke er det. Resultatene avhenger altså sterkt av om det blir tatt hensyn til eventuelle mutasjonsrater eller ikke.

## 1.6 Hensikten

Hensikten med denne oppgaven er å se på farskapssaker der det kan ha skjedd en mutasjon. Dersom en nær slektning av kandidaten kan være far, kan man forvente at DNA-beviset ikke blir sterkt (LR-verdien blir relativt lav). I slike tilfeller blir valg av mutasjonsmodell spesielt viktig. Et sentralt tema for denne oppgaven blir dermed valg av mutasjonsmodeller og å undersøke konsekvensen av hvordan mutasjoner modelleres. Det brukes fire forskjellige mutasjonsmodeller for å beregne LR-verdier og *a posteriori*-sannsynligheter og det er altså ønskelig å finne ut hvordan de forskjellige modellene påvirker resultatene, og om en av modellene er mer hensiktsmessig å bruke fremfor de andre.

I oppgaven fokuseres det på eksempler på farskapssaker, de statistiske metodene som brukes i beregningene og markørene som brukes i farskapstester. For å beregne sannsynligheter for et farskap brukes det i denne oppgaven Familias, og en introduksjon av programmet kan sees i Avsnitt 2.3 i Materialer og metoder. I resultatdelen blir det sett på både eksempler og en reell farskapssak, hvordan en skal tolke verdiene og hvordan dette eventuelt skal legges frem i en rettssak. Spørsmålet «Hva er jobben til de sakkyndige, hvordan skal resultatene videreformidles?» diskuteres også.

Sannsynlighetene beregnes ut ifra alleler og deres frekvenser, og det er ønskelig å finne ut om det er nødvendig å ha med alle alleler i ett datasett når det bare er noen av allelene som finnes i

genotypene som er oppgitt. De resterende allelene kan dermed samles til ett allel og dette kalles da for et restallel.

To av modellene forutsetter at allelene er riktig sortert fordi mutasjoner forekommer oftere til naboalleler. Vi har grunn til å tro at feil bruk av programvare forekommer fordi brukere ikke har forstått betydningen av at alleler er riktig sortert. Dermed blir det interessant å studere feilen som introduseres hvis allelene ikke er riktig sortert når beregningene gjøres i Familias. Dette omtales senere i oppgaven som et sorteringsproblem (Avsnitt 2.3.3).

# 2 Materialer og metoder. Teori

## 2.1 Rettsgenetiske markører og mutasjoner

Sekvensering av hele genomer er i dag en tidkrevende og dyr prosess, og det er ikke allmenn praksis å få kartlagt et helt DNA ved for eksempel sykdom. Når det er behov for å kartlegge gener innen familiegenetikk, brukes det genetiske markører, og på grunn av stor variasjon hos disse markørene er dette tilstrekkelig for å identifisere individer og deres familierelasjoner. Genetiske markører er spesifikke posisjoner i DNA, og for å kunne brukes som referanse eller sammenlikning må markøren ha variasjon som kan påvises hos et individ. De mest brukte genetiske markørene er i dag er SNP (single nucleotide polymorphism) og STR (short tandem repeat).

Det menneskelige genomet varierer i svært liten grad fra person til person. Så mye som over 99 % av DNA-et kan sies å være identisk. Det er derfor utfordrende å finne genetiske markører som kan knytte slektninger sammen ut ifra deres genotyper. SNPs er en mye brukt markør, der det er to alleler som varierer i en gitt posisjon på DNA-molekylet. Når det er ønskelig å finne ut om personer er i slekt, er det lite gunstig å bruke en markør som består av bare to alleler. For å skille mellom forskjellige individer er det gunstig med flere alleler, det er altså ønskelig at markørene skal være mest mulig polymorfe. Dette fører til størst mulig variasjon mellom DNA-profiler, slik at sannsynligheten for å ha like DNA-profiler blir minst mulig. Innenfor rettsgenetikk blir det derfor først og fremst brukt STR. På grunn av at dette er genetiske markører som er polymorfe blir det en større sannsynlighet for mutasjoner enn ved eventuelle andre typer markører. Det er derfor svært viktig å ta høyde for mutasjonsraten innen familiegenetikk (Vigeland, Selmer, Egeland, 2012).

Det har opp gjennom tiden blitt brukt flere metoder for DNA-profilering. Den første var multilocus-testing som ble basert på tandemrepeterte sekvenser i DNA, også kalt minisatellitter. Minisatellittene har et varierende mønster fra individ til individ, såkalt DNA fingerprints. En annen metode som har blitt brukt i senere tid er Single Locus Probes. Denne metoden baserer seg

på varierende antall tandemrepeterte sekvenser, og i motsetning til multilocus-testing er det bare ett locus som blir undersøkt.

Den vanligste metoden for DNA-profilering som blir brukt i dag er STR-analyse. STR-analysen er mer nøyaktig enn de tidligere metodene da den kan finne differanser mellom alleler med én repetisjon. I tillegg trengs det mye mindre DNA-materiale for å gjennomføre analysen da den tar i bruk polymerase chain reaction (PCR) som i teorien kan amplifisere nok DNA fra én enkelt celle (Buckleton, Triggs, Walsh, 2005).

### 2.1.1 STR

Single tandem repeats (STR) er korte tandemrepeterte sekvenser som består av 2-7 baser. Disse repeterte enhetene har færre baser enn minisatellitter og kalles mikrosatellitter. De forskjellige STR-sekvensene opptrer først og fremst i ikke-kodende områder av genomet, såkalte introner. Ettersom de befinner seg mellom kodende gener kan de variere i størrelse og antall repetisjoner fra person til person uten å påvirke genuttrykket. Mikrosatellitter utgjør hele 3 % av det menneskelige DNA, og på grunn av de individuelle variasjonene kan det være hensiktsmessig å bruke STR som markører. En annen grunn til at STR er mye brukt som DNA-markører, er at de er relativt enkle å amplifisere med PCR på grunn av størrelsen som gjerne er mellom 100-400 basepar. Større markører som variable number tandem repeats (VNTR) er gjerne på 400-1000 basepar og blir lettere degradert under en PCR-amplifikasjon. Når markøren er mindre av størrelse, reduseres altså sjansen for at DNAet ødelegges under amplifiseringen.

Navnet til en STR-sekvens tilsvarer lengden på den repeterte enheten. For eksempel har trinukleotid-repetisjonen tre nukleotider som repeteres i sekvensen. Den mest brukte varianten er tetranukleotid-repetisjonen, da penta- og heksanukleotid-repetisjoner ikke forekommer så ofte hos mennesker (Butler, 2012).

### 2.1.2 Mutasjoner

En mutasjon er en endring av DNA-sekvensen, den kan enten oppstå spontant eller som følge av eksponering for mutagener. Mutasjoner kommer i forskjellige former, det kan være en punktmutasjon, en insersjon eller en delesjon. En punktmutasjon vil si at en base substitueres av en annen, det skjer altså en endring bare i den ene posisjonen til basen i sekvensen. En delesjon betyr at en del av sekvensen blir ødelagt eller fjernet, mens en insersjon er når det blir satt inn én eller flere baser. Når deler av en sekvens blir fjernet eller innsatt, vil dette endre hele leserammen til sekvensen. Hvis insersjoner eller delesjoner skjer i kodende gener, kan det dermed endre hele genuttrykket. De aller fleste mutasjoner er allikevel av mindre betydning. Somatiske mutasjoner kan oppstå når som helst i løpet av livet. Somatiske mutasjoner vil ikke bli videreført til neste generasjon da de kan forekomme i alle celler utenom kjønnsceller (Butler, 2012).

Mutasjoner som oppstår under meiosen vil derimot bli videreført til neste generasjon. Meiose er betegnelsen på dannelse av kjønnsceller, også kalt gameter. Disse gametene blir dannet når en diploid celle deler seg til haploide celler, slik at cellen inneholder halvparten av det genetiske arvematerialet, forenklet sagt. Barnet vil få et komplett sett med DNA ved å få en haploid celle fra mor og en haploid celle fra far, som danner en diploid celle. Det er disse kjønnscellene som kan ha mutert under meiosen, og ved en eventuell sammenlikning av barn og foreldres gener vil det kunne være forskjeller når mutasjoner har oppstått.

Det er viktig å ta høyde for mutasjoner innenfor genetiske markører, og spesielt i farskapssaker, da det oftere forekommer mutasjoner under meiosen. I tillegg er mutasjonsraten ved dannelse av gameter noe høyere hos menn enn hos kvinner. Ved dannelse av kjønnsceller ligger mutasjonsraten på rundt 0,5 % hos menn, men en mutasjonsrate vil ikke være konstant for alle loci, den er i høy grad varierende. Mutasjonsraten er altså relativt høy for STR-markørene som blir brukt, disse repeterte sekvensene har oppstått nettopp på grunn av mutasjoner der x antall baser har blitt repetert flere ganger (Buckleton, Triggs, Walsh, 2005; Vigeland, Selmer, Egeland, 2012).

Mutasjonsrater for Y-STR og autosomal STR ligger begge i området 1-4 mutasjoner pr 1000 meiose (Butler, 2012).

## 2.2 Statistiske beregninger i rettsgenetikk

Det finnes flere forskjellige statistiske metoder for å tolke DNA-bevis, og det er en viss uenighet om hvilken metode som er best egnet. I tillegg oppstår det et problem når det skal fremlegges DNA-bevis i rettssaker ettersom de færreste jurister har utdannelse innen statistikk. Enkelte metoder blir dermed vanskelige å anvende da det er problematisk å framlegge resultatene i en rapport eller rettssak på en forståelig måte. Konsekvensene av misforstått bevisfremlegging eller statistiske beregninger kan i verste fall føre til at uskyldige personer blir feilaktig dømt. Dette var tilfelle i London i 1999 (lenke [4]). De to mest brukte metodene innenfor rettsmedisin er den frekventistiske metoden og den Bayesianske metoden.

### 2.2.1 Hardy-Weinberg

Godfrey H. Hardy og Wilhelm Weinberg utarbeidet uavhengig av hverandre en matematisk modell på begynnelsen av 1900-tallet som beskriver forholdet mellom alleler og genotyper (Buckleton, Triggs, Walsh, 2005). Denne modellen kalles i dag Hardy-Weinberg (HW). Modellen forklarer hvordan frekvensen av alleler og genotyper opptrer i en «ideell populasjon». En ideell populasjon i denne sammenheng vil si en populasjon med mange individer (et stort nok utvalg) og tilfeldig reproduksjon uten mutasjoner. Videre sies populasjonen å være i Hardy-Weinberg-likevekt (HWE) når allelfrekvensene holdes konstant fra en generasjon til neste. Det er en forutsetning at genotypene til ubeslektede individer er uavhengige av hverandre i HW-modellen, og det er derfor viktig at populasjonen er stor nok når en bruker HW.

HW-modellen gjør det mulig å beregne sannsynligheter for genotyper i en populasjon ut ifra allelfrekvensene. Allelfrekvensene blir ofte betegnet som p og q=1-p for en diallelesik markør. En homozygot genotype uttrykkes som $p^2$ og en heterozygot genotype uttrykkes som $2pq$. Tilsvarende uttrykk gjelder for markører med mer enn to alleler. Dette gjør det mulig å beregne sannsynligheter for alle genotyper med de gitte allelfrekvensene i populasjonen. En fordel med HW-modellen er altså at det er tilstrekkelig å ha et datasett med frekvensene til allelene i populasjonen, der genotypene kan estimeres ut ifra dette.

### 2.2.2 Likelihood ratio (LR)

Likelihood-ratioen (LR) angir en verdi som tilsier hvor sannsynlig dataene er gitt en fastsatt hypotese mot en annen hypotese. For eksempel:

$H_1$: Mannen er far til barnet

$H_2$: En ukjent, ubeslektet mann er far til barnet

Verdien kalles en ratio ettersom det er to forskjellige likelihood-verdier som divideres på hverandre. Ut ifra brøken får en verdien som tilsvarer hvor mye mer sannsynlig den ene hypotesen er i forhold til den alternative hypotesen. For eksempel hvor sannsynlig data er gitt at en kandidat er den ekte faren til et barn, mot at han ikke er faren, altså når de er ubeslektet. Beregningen av LR-verdien baseres på regler for betingete sannsynligheter (Buckleton, Triggs, Walsh, 2005). Når genotypene til mor, far og barn er som i Figur 2.1 kan LR beregnes på følgende måte:

$$\mathrm{LR} = \frac{\mathrm{P(data|far)}}{\mathrm{P(data|ikke\ far)}} = \frac{1}{2p_B} \qquad \textbf{2.1}$$

der $p_B$ er allelfrekvensen til markøren. Denne verdien blir også kalt en farskapsindeks – PI (engelsk: Paternity Index). Farskapsindeksen beregnes altså ved å dividere sannsynligheten for data gitt at mannen er far på sannsynligheten for data gitt at mannen ikke er far, som i Formel 2.1. Jo større verdi farskapsindeksen har, jo mer støtte gir dataene til at hypotese 1 er riktig.

$\mathrm{P(data|far)}$ er sannsynligheten for data, det vil si genotypen til barnet, gitt at mor og far er foreldre til barnet, altså når det antas at hypotese 1 stemmer. Når det er gitt at mannen er far til barnet blir sannsynligheten for genotypen til barnet lik 1. Data som blir observert i dette tilfellet er genotypene til mor, far og barn. Sannsynligheten for dataene gitt den første hypotesen kan derfor uttrykkes på følgende måte:

$$\mathrm{P(data|H_1)} = \mathrm{P(G_F, G_M, G_B|H_1)} = \mathrm{P(G_F)P(G_M)P(G_B|G_F, G_M, H_1)} \qquad \textbf{2.2}$$

der $G_F$ er fars genotype, $G_M$ er mors genotype og $G_B$ er barnets genotype. $P(G_F, G_M, G_B|H_1)$ angir sannsynligheten for genotypene til kandidaten, mor og barn gitt at kandidaten faktisk er far til barnet. $P(G_B|G_F, G_M)$ angir sannsynligheten for barnets genotype gitt mor og fars genotype.

$P(\text{data}|\text{ikke far})$ er sannsynligheten for genotypen til barnet gitt at kandidaten ikke er far til barnet. Når det er gitt at mannen ikke er far til barnet, vil sannsynligheten for barnets genotype tilsvare frekvensen til allelet i barnets genotype. Sannsynligheten for data gitt at mannen ikke er far kan utledes som:

$$P(\text{data}|H_2) = P(G_F, G_M, G_B|H_2) = P(G_F|H_2)P(G_M|H_2)P(G_B|G_F, G_M, H_2)$$

$$= P(G_F)P(G_M)P(G_B|G_M, H_2) \qquad \textbf{2.3}$$

Den generelle formelen i 2.3 blir som i 2.2 ved at sannsynligheten for genotypene til mor og den angivelige far multipliseres i tillegg til sannsynligheten for barnets genotype. Forskjellen fra den første hypotesen er at det i hypotese 2 antas at mannen ikke er far, og dataene er dermed ikke betinget på mannens genotype.

Likelihood-ratioen kan også uttrykkes sammen med *a priori*- og *a posteriori*-sannsynligheter på følgende måte:

*a posteriori*-odds = likelihood-ratio * *a priori*-odds **2.4**

Likelihood-ratioen sier altså noe om forholdet mellom *a priori*- og *a posteriori*-oddsene. Jo større likelihood-ratio, jo større *a posteriori*-odds, og dette gir en sterkere støtte til den første hypotesen som påstår at kandidaten er far til barnet.

### 2.2.3 *A posteriori*-sannsynlighet

Den bayesianske fremgangsmåten har vært brukt i farskapstester siden 1930-årene, og har i senere tid blitt mer og mer brukt innen rettsmedisin. Som nevnt tidligere krever denne metoden *a priori*-sannsynligheter for å kunne beregne *a posteriori*-sannsynligheter, altså må det legges til

sannsynligheter før genotypene til de gjeldende personene er undersøkt. *A priori* kan kalles for en subjektiv sannsynlighet hvis det tillegges sannsynligheter som er i favør av en hypotese fremfor en annen før data er fremlagt (Hampel, 1998; Løvås, 2008).

Den Bayesianske metoden gjør det mulig å finne sannsynligheten for en hypotese gitt data. Det kan dermed uttrykkes en sannsynlighet for at en mann er far til et barn ut ifra genotypene til mannen, barnet og barnets mor. Denne sannsynligheten kalles *a posteriori* og kan skrives som P(far|data).

Anta følgende hypoteser:

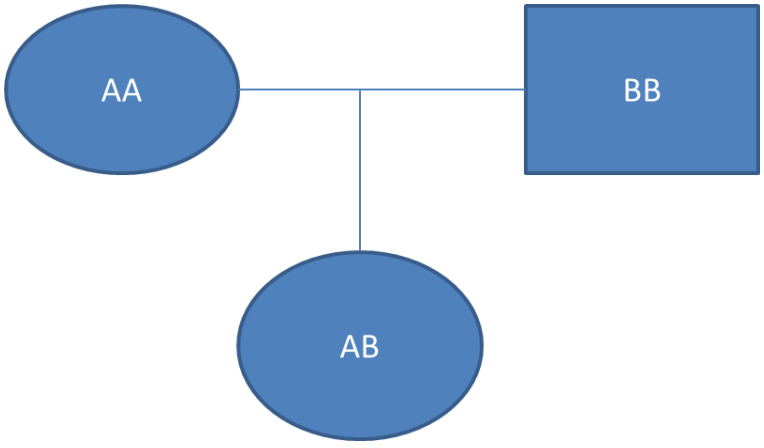
$H_1$: Mannen er far

$H_2$: Mannen er ikke far

Før genotypene til de gjeldende personene blir undersøkt må det oppgis *a priori*-sannsynligheter. I dette tilfellet antas det like stor sannsynlighet for de to hypotesene, da vi ikke har noe grunnlag for å si at det er mer sannsynlig at mannen er far til barnet enn at han ikke er det. Det er da 50 % sjanse for at mannen er far, og 50 % prosent sjanse for at barnets far er en ukjent mann, og ikke denne mannen. *A priori*-sannsynlighetene skrives som:

$$P(H_1) = 0{,}5, \qquad P(H_2) = 0{,}5$$

Dataene blir så betraktet og de oppgitte genotypene kan sees i Figur 2.1:




**Figur 2.1:** Figuren viser genotypene til mor, barn og den angivelige far, der mor har AA, far har BB og barn har AB.

Etter å ha undersøkt hvilke alleler mannen, barnet og mor har i en markør, kan en beregne sannsynligheten for genotypene under de to hypotesene. Dette blir gjort ved å bruke HW, og det forutsettes at mor og den angivelige faren er ubeslektet slik at deres genotyper er uavhengige av hverandre. Sannsynligheten for data når det er gitt at hypotese 1 stemmer, kan ut i fra Formel 2.2 skrives som:

$$P(\text{data}|H_1) = p_B^2 * p_A^2 * 1 \qquad \textbf{2.5}$$

der $p_A$ og $p_B$ er allelfrekvensene til allele A og B, respektivt. Når $H_1$ er gitt blir sannsynligheten for barnets genotype lik 1 ettersom vi ser bort fra mutasjoner i dette eksemplet, og under denne hypotesen vet at barnets alleler AB kommer fra far og mor. Det kan sees i Figur 2.1 at barnets mor er homozygot for AA og den angivelige faren er homozygot for BB, sannsynlighetene for deres genotyper blir dermed henholdsvis $p_A^2$ og $p_B^2$ i følge HW. Sannsynligheten for barnets genotype under hypotese 1 blir da lik sannsynligheten for foreldrenes genotyper.

Hypotese 2 sier at mannen ikke er far. Ut i fra Formel 2.3 kan sannsynligheten for data gitt hypotesen skrives som:

$$P(\text{data}|H_2) = p_B^2 * p_A^2 * p_B \qquad \textbf{2.6}$$

Forskjellen fra den første hypotesen er at det i hypotese 2 antas at mannen ikke er far og dermed blir sannsynligheten for barnets genotype lik allelfrekvensen til B, da det antas at dette allelet ikke kommer fra mannen men fra en ukjent person i populasjonen.

Det er nå uttrykt sannsynligheter for data gitt de to hypotesene, og vi kan bruke Bayes' teorem for å finne sannsynligheten for hypotesen gitt data. Bayes' teorem baseres på betinget sannsynlighet og multiplikasjonsregler ved flere hendelser, og det brukes følgende formel (Løvås, 2008):

$$P(B_i|A) = \frac{P(B_i) * P(A|B_i)}{P(A)} \qquad \textbf{2.7}$$

Bayes' teorem forutsetter at bare én av hendelsene $B_1, \ldots, B_n$ vil inntreffe. Formel 2.7 utrykker altså sannsynligheten for at hendelsen $B_i$ vil inntreffe gitt at hendelse A allerede har inntruffet. Sannsynligheten for at A inntreffer regnes på følgende måte:

$$P(A) = P(B_1) * P(A|B_1) + P(B_2) * P(A|B_2) + \cdots + P(B_n) * P(A|B_n) \qquad \textbf{2.8}$$

Ved hjelp av Formel 2.4 og 2.5 kan det nå uttrykkes en *a posteriori*-sannsynlighet for hypotese 1 gitt data:

$$P(H_1|data) = \frac{P(data|H_1) * P(H_1)}{P(data|H_1) * P(H_1) + P(data|H_2) * P(H_2)}$$

$$= \frac{p_B^2 * p_A^2 * 1 * 0{,}5}{p_B^2 * p_A^2 * 1 * 0{,}5 + p_B^2 * p_A^2 * p_B * 0{,}5} = \frac{1}{1 + p_B} \qquad \textbf{2.9}$$

I følge Formel 2.9 vil sannsynligheten for at mannen er den biologiske faren være størst når allelfrekvensen til B er lav. Når allelfrekvensen, $p_B$, går mot null vil $P(H_1|data)$ gå mot 1, altså mot en sannsynlighet på 100 %. En lav allelfrekvens for B, det *paternelle* allelet, vil altså gi en sterk indikasjon på at mannen er barnets far. Når allelfrekvensen går mot 1 vil sannsynligheten, $P(H_1|data)$, gå mot 0,5, og det er i denne sammenheng en ikke-informativ verdi. Mer generelt, hvis $P(H_1) = q$, så blir

$$P(H_1|data) = \frac{P(data|H_1) * P(H_1)}{P(data|H_1) * P(H_1) + P(data|H_2) * P(H_2)}$$

$$= \frac{pB^2 * pA^2 * 1 * q}{pB^2 * pA^2 * 1 * q + pB^2 * pA^2 * pB * (1 - q)} = \frac{q}{q + pB * (1 - q)} \qquad \textbf{2.10}$$

*A posteriori*-sannsynligheter kan også beregnes ut ifra likelihood-verdier. For å finne *a posteriori*-sannsynligheten, $A_i$, til en likelihood-verdi, $L_i$, brukes følgende formel:

$$A_i = \frac{L_i}{\sum_{i=1}^{n} L_i} \qquad \textbf{2.11}$$

der $L_i$ er en likelihood for en gitt markør eller et helt system. Likelihood-verdiene for hele systemet, altså totale likelihood-verdier, $L_n$, tilsvarer n antall hypoteser. Den totale likelihood-verdien for en hypotese beregnes ved å multiplisere likelihood-verdiene til hver markør sammen. Dette forutsetter uavhengige markører og at *a priori*-sannsynlighetene for hver hypotese er like. *A posteriori*-sannsynligheten for en hypotese, $H_i$, beregnes ved å dividere den totale likelihood-verdien på summen av alle likelihood-verdiene. Dette eksemplifiseres i Avsnitt 3.1.3.

## 2.3 Familias

Familias er utviklet av Petter Mostad og Thore Egeland i samarbeid med Bjørnar Olaisen og Margurethe Stenersen ved Norsk regnesentral og Bente Mevåg ved Rettsmedisinsk institutt. Programmet videreutvikles og vedlikeholdes nå av Daniel Kling (Folkehelseinstituttet og PhD-student ved UMB). Den første versjonen av Familias ble utgitt i 1995. Programmet kan lastes ned gratis på websiden http://familias.name/. Familias er et program som brukes til å beregne sannsynligheter ved farskapstesting eller DNA-identifisering, altså når genotypene til individene er kjent mens slektskapet mellom individene er ukjent (Egeland et. al., 2000). Familias er validert av Drábek (2009).

Programmet kan brukes til å kalkulere sannsynligheter og LR-verdier som beskrevet i Avsnitt 1.4, 2.2.2 og 2.2.3. For å utføre disse beregningene må det defineres alleler i de forskjellige markørene som blir brukt, og hvert allel må ha en gitt allelfrekvens. Videre må det defineres personer, hvilke genotyper de har, og hvilket kjønn de har. Det kan også legges inn hypotetiske personer i programmet som ikke er DNA-testet og dermed ikke har en gitt genotype, disse personene blir lagt inn i systemet for å definere familietreet. For eksempel må begge foreldrene til barn defineres dersom barna er helsøsken. Når personene og deres genotyper er lagt inn i programmet, kan det defineres forskjellige hypoteser, for eksempel om en mann er far til et barn.

Programmet vil angi sannsynligheter for de forskjellige hypotesene, hvilken hypotese som er mest sannsynlig og hvor mye mer sannsynlig den er i forhold til de andre hypotesene. Det kan vises sannsynligheter for et sammenlagt markørsystem, eller ulike markører kan evalueres hver for seg. I Familias er det i tillegg til standard farskapssaker mulig å se på mer kompliserte tilfeller der det kan ha forekommet mutasjon(er).

### 2.3.1 Mutasjonsmodeller

Det finnes fire forskjellige mutasjonsmodeller i Familias, disse fire modellene vektlegger mutasjonsrate og mutasjonsrekkevidde (dvs størrelsen på mutasjonen) på ulike måter. Størrelsen på mutasjonen og hvor ofte mutasjoner forekommer blir altså behandlet forskjellig i mutasjonsmodellene. To av modellene tar ikke hensyn til mutasjonsrekkevidden. I tillegg er det to modeller som er stasjonære, dette vil si at den tilhørende markovkjeden er stasjonær. Stasjonære modeller forutsetter altså at allelfrekvensen holdes konstant fra generasjon til generasjon, og resultatene i disse modellene vil derfor ikke påvirkes av introduksjon av flere individer i et familietre. Dette blir som en stasjonær markovkjede der det antas at det ikke oppstår flere mulige alleler, og allelet (tilstanden) i en generasjon er bare avhengig av den forrige generasjonen.

**Modell 1: Equal probability**

Modell 1 vektlegger alle mutasjoner likt med mutasjonsrate R. Dette betyr for eksempel at hvis mutasjonsraten er 0,5, så er det 50 % sannsynlighet for mutasjon (urealistisk høyt, men brukt for eksemplets skyld). Det er lik sannsynlighet for mutasjon til alle alleler. Det blir med andre ord ikke tatt hensyn til avstanden mellom allelene (hvor avstand betyr antall repetisjoner som skiller allelene), og modellen kan derfor sies å være relativt forenklet. Sannsynligheten for at det ikke forekommer en mutasjon blir da $1 - R$, da det enten ikke skjer en mutasjon og allelet forblir det samme, eller det kan mutere til hvilket som helst av allelene, inkludert det samme allelet.

Sannsynligheten for en mutasjon fra et gitt allel til et annet allel blir $\frac{R}{N-1}$, der N-1 er antall alleler det gitte allelet kan mutere til.

**Modell 2: Probability proportional to frequency**

I denne modellen er sannsynligheten for mutasjon til et allel proporsjonal med frekvensen til allelet. Sannsynligheten for å mutere til det gitte allelet blir da større dersom allelet forekommer ofte, og sannsynligheten synker dersom allelfrekvensen er lav. Hvis en for eksempel har to alleler; 1 og 2, der allel 1 har en frekvens på 0,01 og frekvensen til allel 2 er 0,05, vil sannsynligheten for en mutasjon fra et tilfeldig allel til allel 2 være 5 ganger større enn en mutasjon til allel 1. Som i modell 1 tar denne mutasjonsmodellen ikke hensyn til avstanden mellom allelene. Denne modellen er stasjonær, det vil si at den ikke vil påvirkes av inkorporering av flere personer som legges til systemet uten genotype.

**Modell 3: Probability decreasing with range (equal)**

I denne modellen blir det tatt hensyn til mutasjonsrekkevidden, altså hvor stor avstand det er mellom allelene før og etter mutasjon. Det er kjent at et allel vil ha en større tendens til å mutere til alleler som ligger nær. For eksempel vil et allel med fire repetisjoner ha større sannsynlighet for å mutere til alleler med tre eller fem repetisjoner fremfor alleler som har færre eller flere repetisjoner. Sannsynligheten for mutasjonen reduseres altså når avstanden mellom allelene øker. Modell 3 og 4 (under) krever en parameter som angir hvor mye sannsynligheten synker for hver repetisjon. Parameteren heter mutation range (mutasjonsrekkevidde) i Familias, og den kan settes mellom 0,1-0,5. Det er anbefalt å bruke mutasjonsrekkevidde r = 0,1. Dette tilsier at sannsynligheten for mutasjonen synker med en tiendedel for hver repetisjon.

**Modell 4: Probability decreasing with range (stable)**

I modell 4 reduseres sannsynligheten for mutasjon mellom to alleler når avstanden mellom dem øker, som i modell 3. Modell 3 og 4 krever altså at det blir lagt inn en parameter for mutasjonsrekkevidde, denne samsvarer med hvor mye sannsynligheten synker jo større avstand det er mellom allelene. Verdien til parameteren må ligge mellom 0 og 1. Modell 4 er en stasjonær versjon av modell 3, dette betyr at sannsynlighetene ikke vil endres selv om det legges til flere personer i familietreet, for eksempel om det legges inn besteforeldre til barnet familietreet i en situasjon der man ikke har genotypene til besteforeldrene.

### 2.3.2 Mutasjonsmatriser

Hver modell beskriver en mutasjonsmatrise. En mutasjonsmatrise angir sannsynligheten $m_{ij}$ for en mutasjon fra allel *i* til allel *j*. Den enkleste mutasjonsmatrisen beskrives av modell 1, der alle mutasjoner er like sannsynlige:

$$M = \begin{bmatrix} 1-R & \frac{R}{N-1} & \frac{R}{N-1} \\ \frac{R}{N-1} & 1-R & \frac{R}{N-1} \\ \frac{R}{N-1} & \frac{R}{N-1} & 1-R \end{bmatrix}$$

Denne mutasjonsmatrisen kan for eksempel beskrive tre alleler, allel 1, 2 og 3. I matrisen over er N=2. Sannsynligheten for at allel 1 ikke muterer og forblir allel 1 er tilsvarende $m_{11}$ altså 1-R som kan sees i første rad i første kolonne i mutasjonsmatrisen. Om allelet skulle mutere fra 1 til 2 er sannsynligheten for dette $m_{12}$ og $m_{21}$ som kan sees i første rad i andre kolonne og andre rad i første kolonne; $\frac{R}{N-1}$.

### 2.3.3 Sorteringsproblemet

Allelene som anvendes i Familias kan navngis med tall som tilsier hvor mange repetisjoner allelet består av. Ved bruk av modell 3 og 4 er det viktig at allelene er sortert i riktig rekkefølge slik at

modellen tar hensyn til den faktiske avstanden mellom dem. Et problem som kan oppstå er når allelene navngis fra for eksempel 7-11. Programmet vil da sortere allelene på følgende måte: 10, 11, 7, 8, 9. Grunnen til at dette problemet oppstår er at programmet bruker en standard metode for alfabetisk sortering. Dette fører til at programmet ikke tar hensyn til hvor mange sifre det er i tallet, men ser på det første sifferet i tallet for å alfabetisere det, i dette tilfellet; 10, 11, 7, 8, 9. Når allelene er navngitt på denne måten vil modell 3 og 4 tilsi at avstanden mellom 11 og 7 er like stor som avstanden mellom 7 og 8. Modellen sier dermed at en mutasjon fra 11 til 7 er like sannsynlig som en mutasjon fra 8 til 7. Problemet kan unngås ved å sette 0 foran tall som inneholder ett siffer. I dette tilfellet bør altså allelene navngis som: 07, 08, 09, 10 og 11. Programmet vil nå sortere dem i riktig rekkefølge.

### 2.3.4 Fullstendig og minimalt allelsystem

I resultatdelen blir det anvendt fullstendig og minimalt allelsystem i noen av Familias-eksemplene. Et fullstendig system vil si alle mulige alleler i en markør som er lagt inn i systemet med sine frekvenser, det vil si alle allelene i databasen. Allelfrekvensene skal alltid summeres til 1. Når en mor, far og et barn er genotypet for allelelene i en markør kan genotypene til sammen bestå av minimum ett enkelt allel (hvis alle tre er homozygote) og maksimum seks forskjellige alleler hvis alle tre er heterozygote med forskjellige alleler. I et minimalt system vil de andre allelene i systemet som ikke inngår i personenes genotyper bli slått sammen til ett enkelt restallel, og allelfrekvensene vil fortsatt summeres til 1. Det blir brukt fullstendig og minimalt system i resultatdelen for å se hvordan dette påvirker resultatene.

# 3 Resultater

## 3.1 Familias-eksempler

Familias-eksemplene i denne oppgaven er basert på konstruerte data, med unntak av det reelle eksempelet i Avsnitt 3.1.3 som er basert på data hentet fra Rettsmedisinsk institutt i Oslo (upublisert). Datasettene med genotyper og deres tilhørende allelfrekvenser framgår av Familias-filene, disse filene er tilgjengelig fra Maria Berggreen. I disse Familias-eksemplene ser vi bort i fra eventuelle genotypingsfeil, og ved inkompatibilitet mellom foreldre og barn blir det antatt at dette skyldes mutasjon.

### 3.1.1 Eksempel 1

Dette er et eksempel som omfatter to markører med henholdsvis tre og fire alleler. Eksempel 1 inneholder for enkelhetens skyld ikke mutasjoner, og bakgrunnen for eksempelet er å vise hvordan LR-verdien til to hypoteser beregnes. Genotypene til mor, far og barn i markør S1 er som i Avsnitt 2.2.3, se Figur 2.1.

I dette eksempelet er det tre personer; mor, barn og en kandidat for far. Det er ønskelig å beregne en LR-verdi som kan brukes til å avgjøre om kandidaten er far til barnet. Markøren som brukes i dette eksempelet blir kalt S1 og inneholder allelene A, B og C. Mor er homozygot for A, kandidaten for far er homozygot for B og barnet har en heterozygot genotype AB, som vist i Figur 2.1. Allelfrekvensene for A og B er henholdsvis $p_A = 0{,}05$ og $p_B = 0{,}05$.. Det defineres følgende hypoteser:

$H_1$: Kandidat er far til barnet

$H_2$: Kandidat er ikke far til barnet

Likelihood-ratio beregnes ved Formel 2.1 :

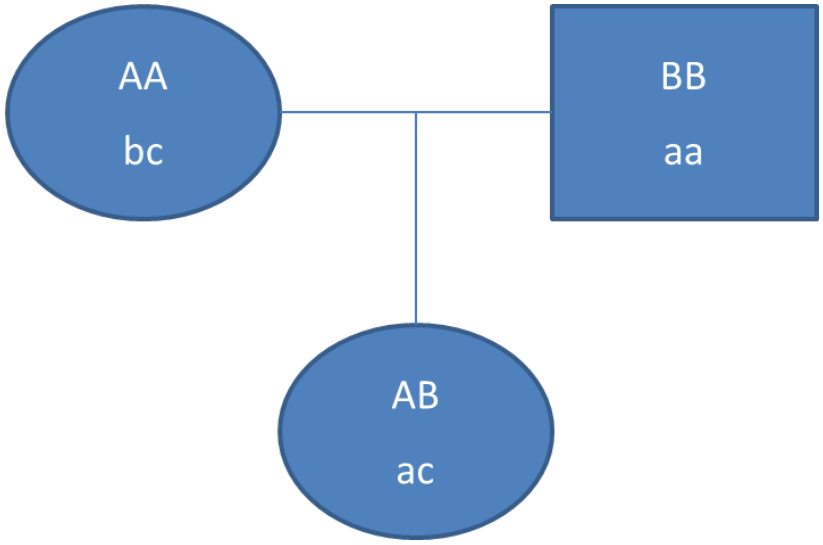
$$\mathrm{LR} = \frac{1}{0{,}05} = 20$$

LR-verdien angir dataene som 20 ganger mer sannsynlig under hypotesen om at kandidat er far i forhold til den alternative hypotesen om at han ikke er far. Deretter blir det undersøkt for en ny markør som kalles S2. S2 inneholder allelene a, b, c og d, med allelfrekvenser som vist i Tabell 3.1.

**Tabell 3.1:** viser allelene i markøren S2 og de tilhørende allelfrekvensene

| allel | a | b | c | d |
|---|---|---|---|---|
| allelfrekvens | 0,1 | 0,1 | 0,1 | 0,7 |

Kandidat for far er homozygot for a, mor er heterozygot for bc og barnet er heterozygot for ac, som i Figur 3.1.




**Figur 3.1:** Figuren viser genotypene til mor, far og barn for både markørsystem 1 og markørsystem 2.

Ettersom markørene antas å være uavhengige, kan LR-verdien for hver av de to markørene multipliseres for å finne den totale LR-verdien. Likelihood-ratio for S1 og S2:

$$\left(\frac{1}{p_A}\right) * \left(\frac{1}{p_a}\right) = \left(\frac{1}{0{,}05}\right) * \left(\frac{1}{0{,}1}\right) = 200$$

Når LR-verdien regnes ut ifra begge markørene S1 og S2 blir dataene 200 ganger mer sannsynlig gitt at den første hypotesen stemmer, altså at mannen er far, mot at han ikke er far. Både kandidaten og barnet har allelene B og a henholdsvis i de to markørene som blir brukt. Det er dermed 200 ganger mer sannsynlig at barnet har fått allelene fra mannen enn fra en ukjent mann.

### 3.1.2 Eksempel 2

I Eksempel 2 illustreres effekten av de ulike mutasjonsmodellene, og hvordan resultatene påvirkes av å introdusere restallel og utypede overflødige personer. I dette eksemplet er det kun to personer med kjente genotyper: en kandidat til far og et barn. Det defineres to hypoteser som i Eksempel 1.

$H_1$: Kandidat er far til barnet

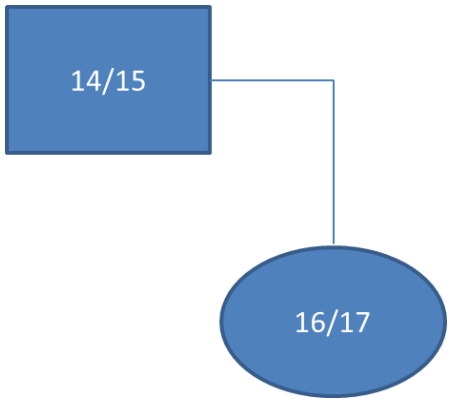
$H_2$: Kandidat er ikke far til barnet

Det finnes i dette tilfellet 8 alleler, og allelfrekvensene er gitt i Tabell 3.2.

**Tabell 3.2:** Viser allelfrekvensene til de åtte allelene 14-21.

| allel | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|
| frekvens | 0,072 | 0,082 | 0,212 | 0,292 | 0,222 | 0,097 | 0,02 | 0,003 |

Kandidaten for barnets far har allelene 14 og 15, og barnet har allelene 16 og 17, og dette vises i Figur 3.2. Det må ha skjedd en mutasjon dersom kandidaten faktisk er barnets far.




**Figur 3.2:** Genotypene til far og barn, der den angivelige far har allelene 14/15 og barnet har allelene 16/17.

Sannsynligheten for data gitt hypotese 1, altså at mannen er faren til barnet, blir så beregnet med de fire forskjellige modellene for å se om de gir forskjellige LR-verdier. Det antas en mutasjonsrate på R = 0,005.

**Tabell 3.3:** LR-verdier for de fire forskjellige modellene i Familias, mutasjonsrate R = 0,005 og mutasjonsrekkevidde r = 0,5.

| Modell | LR-verdi |
|---|---|
| 1 | 0,0029 |
| 2 | 0,0063 |
| 3 | 0,0047 |
| 4 | 0,0064 |

Alle de fire modellene gir ganske lave LR-verdier for at mannen kan være barnets far. Det kan sees i Tabell 3.3 at modell 1 gir den laveste LR-verdien, denne LR-verdien tilsvarer at data er mer sannsynlig dersom faren til barnet ikke er denne mannen. Modell 4 gir den høyeste av de fire modellene. Alle de fire modellene støtter oppunder den alternative hypotesen som sier at mannen og barnet er ubeslektet, og at barnets far antakeligvis er en ukjent ubeslektet mann.

Modell 2 vekter ikke alle mutasjonene likt, det vil i dette tilfellet tilsi at den mest sannsynlige mutasjonen er fra et hvilket som helst allel til allel 17 da det har den høyeste allelfrekvensen. Den minst sannsynlige mutasjonen er til allel 20 som har den laveste allelfrekvensen. Barnet har allelene 16 og 17, og dermed allelet som det er mest sannsynlig å mutere til i denne modellen. Dette samsvarer med at modell 2 gir den høyeste sannsynligheten for at mannen kan være barnets far, selv om det ikke er en spesielt høy sannsynlighet.

Den tredje modellen i Familias tar både mutasjonsrate og mutasjonsrekkevidde med i beregningen. Som navnet antyder (Probability decreasing with range) gir den lavere sannsynlighet for mutasjon jo større avstand det er mellom allelene. Med mutasjonsrekkevidde r = 0,5 halveres sannsynligheten for hvert trinn det muteres fra det opprinnelige allelet. For eksempel vil sannsynligheten halveres hvis allelet muterer fra allel 14 til 15, og deretter halveres enda en gang om allelet muterer videre til allel 16. Modellen gir lik reduksjon for hvert trinn, og differensierer derfor ikke mellom allelene, bare avstanden mellom dem.

Den fjerde modellen, Probability decreasing with range (stable), tar hensyn til de samme parameterne som i modell 3. I motsetning til modell 1 og 3 er modell 2 og 4 som tidligere nevnt stasjonære. Hvis vi for eksempel legger til en mor og en bestemor uten genotype, vil LR-verdiene i modell 3 altså endre seg, mens tallene ikke vil endres i modell 4.

Det er bare fire av totalt åtte mulige alleler som er gitt i personenes genotyper. De fire siste allelene, 18-21, kan dermed slås sammen til ett restallel for å undersøke om dette påvirker LR-verdien. En får da et fullstendig system som inneholder alle åtte alleler med deres tilhørende allelfrekvenser, og et minimalt system som inneholder fem alleler totalt. Det minimale systemet består da av allelene 14-17 og et restallel med de resterende allelfrekvensene. Hvordan LR-verdien varierer ut ifra hvor mange alleler som er tatt med i beregningen ved forskjellige mutasjonsrater kan sees i Tabell 3.4. Det beregnes også en ratio for minimalt og fullstendig system for å enklere kunne undersøke forskjellen mellom dem, ratioverdien gjelder for alle tre mutasjonsratene i systemet.

**Tabell 3.4:** LR-verdier for de fire modellene, ved fullstendig og minimalt allelsystem, med henholdsvis mutasjonsrate R=0,001, R=0,0025, R=0,004 og mutasjonsrekkevidde r=0,5, og ratio for minimalt/fullstendig system for de fire modellene.

| | Fullstendig allel-system | | | Minimalt allel-system | | | Ratio (Min/Full) |
|---|---|---|---|---|---|---|---|
| Modell | R = 0,001 | R = 0,0025 | R = 0,004 | R = 0,001 | R = 0,0025 | R = 0,004 | |
| 1 | 0,000582 | 0,001454 | 0,002326 | 0,001018 | 0,002544 | 0,004071 | 1,75 |
| 2 | 0,001252 | 0,003130 | 0,005008 | 0,001350 | 0,003374 | 0,005399 | 1,08 |
| 3 | 0,000946 | 0,002366 | 0,003786 | 0,001013 | 0,002533 | 0,004052 | 1,07 |
| 4 | 0,001280 | 0,003200 | 0,005121 | 0,002511 | 0,006278 | 0,010045 | 1,96 |

Det kan sees i Tabell 3.4 at den høyeste LR-verdien i det fullstendige systemet er ved den høyeste mutasjonsraten R=0,004 i modell 4. Dette betyr at sannsynligheten for dataene gitt hypotese 1 øker når en gir en høyere mutasjonsrate. Dette er logisk da en høyere mutasjonsrate vil si at det antas høyere forekomst av mutasjoner. Når kandidaten har et allel 15 og barnet et allel 16, vil dette være den mest sannsynlige mutasjonen ifølge modell 4, da avstanden mellom dem bare er en repetisjon, altså den minste mulige avstanden.

I det minimale systemet finner en den samme tendensen der den høyeste LR-verdien også finnes i modell 4 med den høyeste mutasjonsraten. I dette systemet kan det sees i tabellen at LR-verdien for modell 1 og 4 nesten er doblet i forhold til det fullstendige systemet med alle åtte alleler. Resultatene blir dermed i stor grad påvirket ved å lage et restallel, og en får relativt sett større sannsynligheter. I begge systemene finnes den laveste LR-verdien i modell 1 ved den laveste mutasjonsraten. En lav mutasjonsrate tilsier en mye lavere sannsynlighet for at det kan ha oppstått en mutasjon, og modell 1 vektlegger alle mutasjoner likt. Det blir dermed ikke sett på avstanden mellom allelene, og det blir ikke tatt hensyn til at allel 15 og 16 ligger nær hverandre. Det kan i tillegg sees i tabellen at den største forskjellen mellom det fullstendige og det minimale systemet finnes i modell 4. Modell 4 blir altså mest påvirket ved å legge inn et restallel, i dette tilfellet, selv om det er en stasjonær modell. Den er da stasjonær i forhold til eventuelle ekstra personer som blir lagt inn i systemet, men den er ikke stasjonær i forhold til hvor mange alleler som blir lagt inn.

Videre kan det defineres en tredje hypotese; vil det påvirke LR-verdiene om det introduseres andre slektninger av barn eller kandidaten i systemet? Det introduseres derfor en mor og far til kandidaten uten genotyper som kalles BM og BF.

$H_3$: Kandidat er far til barnet gitt at BM og BF er foreldre til kandidaten.

Det antas at modell 2 og modell 4 ikke vil påvirkes av de nye personene da de er stasjonære modeller. De nye LR-verdiene kan sees i Tabell 3.5. Fullstendig familietre inneholder barn, far og fars foreldre, og minimalt familietre inneholder bare far og barn.

**Tabell 3.5:** LR-verdier for fullstendig familietre med besteforeldre og minimalt familietre uten besteforeldre. Mutasjonsratene er R=0,001 og R=0,004 og mutasjonsrekkevidde r = 0,5

| | Fullstendig familietre | | Minimalt familietre | |
|---|---|---|---|---|
| Modell | R = 0,001 | R = 0,004 | R = 0,001 | R = 0,004 |
| 1 | 0,0005824 | 0,0023396 | 0,0005815 | 0,0023262 |
| 2 | 0,0012519 | 0,0050076 | 0,0012519 | 0,0050076 |
| 3 | 0,0009475 | 0,0038030 | 0,0009464 | 0,0037857 |
| 4 | 0,0012802 | 0,0051206 | 0,0012802 | 0,0051206 |

Det kan sees i Tabell 3.5 som antatt at modell 2 og 4 er stasjonære og LR-verdiene endres dermed ikke når en introduserer besteforeldrene i familietreet. Modell 1 og 3 får en liten endring av LR-verdiene, men det har ikke noe vesentlig utslag for resultatet totalt sett i dette eksempelet. I dette tilfellet er resultatene i modell 3 de som blir mest påvirket av introduksjon av overflødige personer. Alle LR-verdiene forblir lave både i fullstendig og minimalt system, og fullstendig og minimalt familietre.

### 1.1.3 Reelt eksempel

Det tredje Familias-eksempelet er et reelt eksempel med ti markører. I dette eksempelet illustreres sorteringsproblemet, og i tillegg undersøkes muligheten for at en nær slektning av kandidaten kan være barnets biologiske far.

**Tabell 3.6:** Oversikt over allelene til far, mor og barn i de ti markørene som de er blitt genotypet for.

| Markør | Far | Mor | Barn |
|---|---|---|---|
| D3S1358 | 15,16 | 13,15 | 15,15 |
| VWA | 18,18 | 18,18 | 18,18 |
| D16S539 | 12,12 | 12,13 | 12,13 |
| D2S1338 | 21,21 | 20,25 | 21,25 |
| D8S1179 | 11,13 | 11,15 | 11,13 |
| D21S11 | 21,32.2 | 31,31.2 | 31.2,32.2 |
| D18S51 | 12,12 | 16,16 | 13,16 |
| D19S433 | 12,15.2 | 13,16 | 13,15.2 |
| FGA | 23,24 | 21,24 | 21,23 |
| TH01 | 9.3,9 | 8,9.3 | 9.3,9.3 |

Dette er en farskapssak der foreldre og tre barn er testet for sine genotyper. I tillegg til foreldre og barn er det lagt inn besteforeldre, en bror og en halvbror av den angivelige faren i Familias, som kan sees i Figur 3.3. Disse personene er lagt inn i systemet uten gitte genotyper. Det er blitt testet for ti forskjellige markører, og to av barna stemte overens med begge foreldrene i alle ti markørene. Ett barn stemte bare i ni av ti markører, genotypen til dette barnet og foreldrene blir derfor undersøkt nærmere, og det er bare disse tre DNA-profilene som blir brukt. I markør D18S51 har den angivelige faren genotype 12,12 og moren til barnet har genotype 16,16, mens barnet har genotypen 16,13, og deler dermed ikke et allel med mannen i denne markøren. Dette kan sees i Tabell 3.6. Det kan ut ifra dette defineres forskjellige hypoteser; det kan ha oppstått en mutasjon fra allel 12 til 13 (dette er den mest sannsynlige mutasjonen) og mannen er den riktige

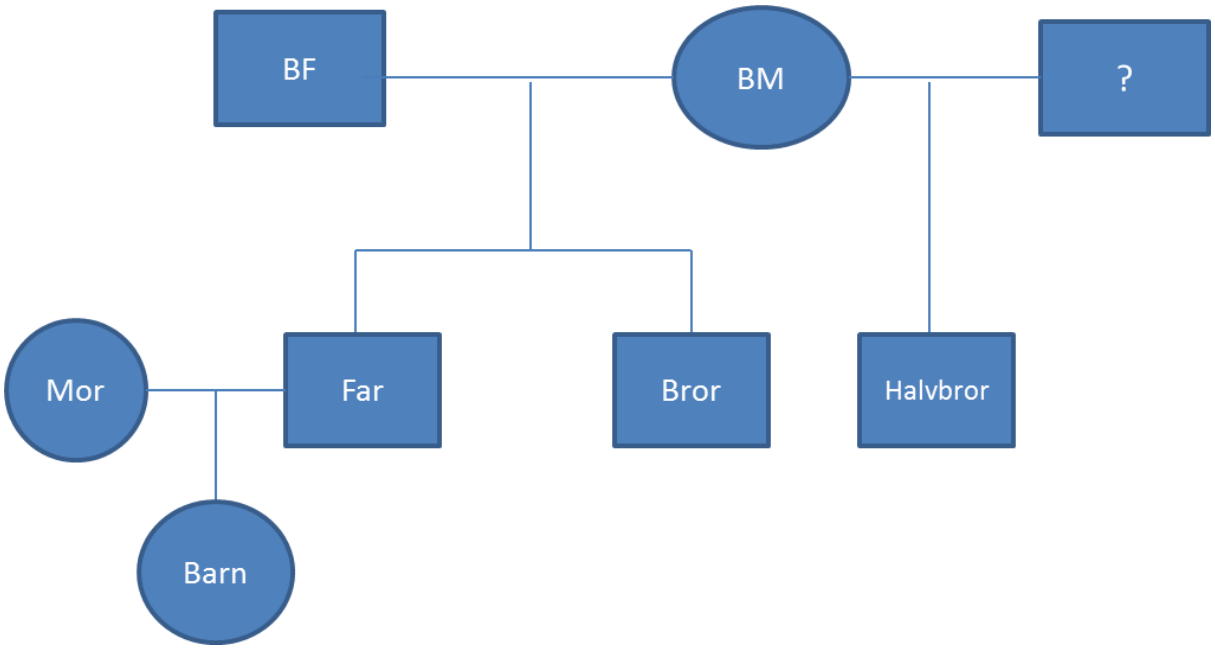
faren, på en annen side kan det være en annen tilfeldig mann som er far til barnet, eller kan det være en nær slektning, for eksempel en bror, av den angivelige faren. Hypotesene er som følger:

$H_1$: Mannen er barnets far

$H_2$: En ukjent, ubeslektet mann er far

$H_3$: Mannens bror er far

$H_4$: Mannens halvbror er far




**Figur 3.3:** Det fullstendige familietreet som svarer til den første hypotesen, der bestefar (BF) og bestemor (BM) er foreldre til far og bror, bestemor er mor til halvbror, og Mor og Far er barnets foreldre.

Det beregnes så *a posteriori*-sannsynligheter for de fire hypotesene i Familias ut i fra de gitte DNA-profilene i Tabell 3.6, og sannsynlighetene vises i Tabell 3.7.

**Tabell 3.7:** Sannsynligheter for de fire forskjellige hypotesene, med mutasjonsrate R = 0,002530 for markøren D18S51 og mutasjonsrekkevidde r = 0,1. Det fullstendige systemet inneholder alle allelene i de ti markørene, og det minimale systemet inneholder et restallel i markøren D18S51.

| | Fullstendig | | | | Minimal | | | |
|---|---|---|---|---|---|---|---|---|
| Modell | kandidat | bror | halvbror | ubeslektet | kandidat | bror | halvbror | ubeslektet |
| 1 | 0,0765 | 0,8602 | 0,0632 | 0,0001 | 0,3365 | 0,6186 | 0,0448 | 0,0001 |
| 2 | 0,1846 | 0,7598 | 0,0556 | 0,0001 | 0,2601 | 0,6895 | 0,0503 | 0,0001 |
| 3 | 0,4094 | 0,5508 | 0,0398 | 0,0001 | 0,5749 | 0,3968 | 0,0283 | 0,0000 |
| 4 | 0,2188 | 0,7279 | 0,0532 | 0,0001 | 0,6338 | 0,3421 | 0,0241 | 0,0000 |

I det fullstendige systemet med alle allelene til stede er det størst sannsynlighet for at broren til mannen er den riktige faren i alle fire modellene. Det er lavest sannsynlighet for at faren til barnet er en tilfeldig mann, og kandidaten har en høyere sannsynlighet for å være far enn en eventuell halvbror.

I det minimale systemet er det tatt med allelene som opptrer i de gitte genotypene, de resterende allelene i systemet er blitt satt sammen til et restallel, og det er altså totalt fire alleler for markøren D18S51 i denne beregningen. De to første modellene angir broren til å være den mest sannsynlige faren, mens de to siste modellene gir høyest sannsynlighet for at kandidaten er far. Det er i tillegg angitt en enda mindre sannsynlighet for at faren er en ukjent, ubeslektet mann i de to siste modellene i dette minimale systemet.

Den høyeste sannsynligheten for hypotese 3, at kandidatens bror er far, er 0,8602 med modell 1 i det fullstendige systemet. Den laveste sannsynligheten for denne hypotesen er 0,3421 med modell 4 i det minimale systemet. Kandidaten har lavere sannsynlighet enn broren for å være far i alle modeller i det fullstendige systemet. I det minimale systemet varierer sannsynlighetene en del, og kandidaten har høyest sannsynlighet for å være far i modell 3 og 4. En ubeslektet mann og en eventuell halvbror har generelt lav sannsynlighet for å være far i dette tilfellet. Det er altså ikke noe som støtter hypotesene om ubeslektet og halvbror i de to systemene.

Resultatene er relativt sprikende i Tabell 3.7, og det blir videre undersøkt mer detaljert for de ti forskjellige markørene som er blitt brukt i beregningene, med fullstendig og minimalt system. Heretter ser vi kun på resultatene fra modell 4.

**Tabell 3.8:** viser likelihood-verdiene for de fire forskjellige hypotesene ved hver av de ti forskjellige markørene ved fullstendig og minimalt system i modell 4.

| | | Likelihood | | | |
|---|---|---|---|---|---|
| System | Markør | Kandidat | Bror | Halvbror | Ubeslektet |
| | D3S1358 | 5,60E-05 | 4,32E-05 | 3,68E-05 | 3,04E-05 |
| | TH01 | 1,40E-03 | 1,18E-03 | 1,07E-03 | 9,62E-04 |
| | D21S11 | 3,64E-05 | 2,09E-05 | 1,31E-05 | 5,34E-06 |
| Fullstendig | D18S51 | 1,07E-07 | 1,52E-05 | 2,26E-05 | 3,00E-05 |
| Minimal | D18S51 | 6,97E-07 | 1,60E-05 | 2,30E-05 | 2,99E-05 |
| | D16S539 | 3,71E-03 | 2,69E-03 | 2,19E-03 | 1,68E-03 |
| | VWA | 1,74E-03 | 1,05E-03 | 7,02E-04 | 3,56E-04 |
| | D8S1179 | 1,98E-04 | 1,66E-04 | 1,49E-04 | 1,33E-04 |
| | FGA | 4,18E-04 | 2,67E-04 | 1,91E-04 | 1,16E-04 |
| | D2S1338 | 1,56E-05 | 8,03E-06 | 4,23E-06 | 4,38E-07 |
| | D19S433 | 2,58E-05 | 1,35E-05 | 7,35E-06 | 1,20E-06 |
| Fullstendig | Total | 6,5661E-41 | 2,1977E-40 | 1,5858E-41 | 2,2721E-44 |
| Minimal | Total | 4,28E-40 | 2,31E-40 | 1,61E-41 | 2,26E-44 |

Likelihood-verdiene i Tabell 3.8 er tatt med i resultatene da de inneholder informasjonen som er nødvendig for å estimere sannsynlighetene som senere blir presentert i Tabell 3.9. Det er kun i markør D18S51 det er forskjell i likelihood-verdiene mellom fullstendig og minimalt system. Det er derfor tatt med en egen linje med verdien til D18S51 i det minimale systemet, de andre markørene vil ha den samme verdien som i det fullstendige systemet.

Disse likelihood-verdiene for hver enkelt markør kan så regnes om til *a posteriori*-sannsynligheter. Det er fire hypoteser for hver markør, og likelihood-verdiene for hver av disse hypotesene kan benevnes med $L_1, L_2, L_3$ og $L_4$. For å finne *a posteriori*-sannsynlighetene som svarer til likelihood-verdiene i Tabell 3.8 brukes Formel 2.11.

**Tabell 3.9:** *a posteriori*-sannsynligheter for de fire forskjellige hypotesene ved hver av de ti forskjellige markørene ved fullstendig og minimalt system i modell 4.

| | | *A posteriori*-sannsynligheter | | | |
|---|---|---|---|---|---|
| System | Markør | Kandidat | Bror | Halvbror | Ubeslektet |
| | D3S1358 | 0,3365 | 0,2596 | 0,2212 | 0,1827 |
| | TH01 | 0,3036 | 0,2559 | 0,2320 | 0,2086 |
| | D21S11 | 0,4806 | 0,2759 | 0,1730 | 0,0705 |
| Fullstendig | D18S51 | 0,0016 | 0,2238 | 0,3328 | 0,4418 |
| Minimal | D18S51 | 0,0100 | 0,2299 | 0,3305 | 0,4296 |
| | D16S539 | 0,3612 | 0,2619 | 0,2132 | 0,1636 |
| | VWA | 0,4522 | 0,2729 | 0,1824 | 0,0925 |
| | D8S1179 | 0,3065 | 0,2570 | 0,2307 | 0,2059 |
| | FGA | 0,4214 | 0,2692 | 0,1925 | 0,1169 |
| | D2S1338 | 0,5513 | 0,2838 | 0,1495 | 0,0155 |
| | D19S433 | 0,5392 | 0,2821 | 0,1536 | 0,0251 |
| Fullstendig | Total | 0,2179 | 0,7294 | 0,0526 | 0,0001 |
| Minimal | Total | 0,6335 | 0,3426 | 0,0239 | 0,0000 |

Den totale *a posteriori*-sannsynligheten til en hypotese beregnes ved å multiplisere likelihood-verdiene til hver enkelt markør, og deretter blir de totale likelihood-verdiene omgjort til totale *a posteriori*-sannsynligheter ved bruk av Formel 2.11. For at sannsynlighetene kan multipliseres sammen forutsettes det at markørene er uavhengige av hverandre.

I Tabell 3.9 er det oppgitt en *a posteriori*-sannsynlighet for hypotese 3 (Bror) på 0,7294 i det fullstendige systemet. Dette beregnes ut ifra Tabell 3.8 på følgende måte:

$$\frac{2{,}1977\mathrm{E}-40}{(6{,}5661\mathrm{E}-41)+(2{,}1977\mathrm{E}-40)+(1{,}5858\mathrm{E}-41)+(2{,}2721\mathrm{E}-44)}=0{,}7924$$

I ni av ti markører er det størst sannsynlighet for at kandidaten er faren ifølge Tabell 3.9. I markøren D18S51 er det derimot minst sannsynlighet for at han er faren. I denne markøren er det størst sannsynlighet for at det er en ubeslektet mann som er den riktige faren. I de ni andre markørene er dette hypotesen som er minst sannsynlig.

Ved minimalt system er det bare sannsynlighetene for D18S51 som endres. For bror, halvbror og ubeslektet er det veldig lite endringer, det blir litt høyere sannsynlighet for at bror eller halvbror er far, mens sannsynligheten for ubeslektet synker så vidt. Sannsynligheten for at kandidaten er faren stiger derimot fra 0,0016 til 0,01.

I Avsnitt 2.3.3 står det litt om sorteringsproblemet i Familias. Markøren D18S51 inneholder 20 alleler i dette datasettet. Før disse allelene er sortert i riktig rekkefølge vil programmet angi følgende rekkefølge: 10, 11, 12, 13, 14, 15, 16, 17, 18, 18.3, 19, 20, 21, 22, 23, 24, 25, 27, 7, 9. Dette er problematisk da modell 3 og 4 i Familias vil anta en like stor sannsynlighet for mutasjonene 25 → 27 og 27 → 7. Når alle allelene navngis med to siffer blir rekkefølgen i Familias riktig: 07, 09, 10, 11, 12, 13, 14, 15, 16, 17, 18, 18.3, 19, 20, 21, 22, 23, 24, 25, 27.

Det undersøkes videre om en ny navngivning av allelene, slik at de sorteres i riktig rekkefølge, vil påvirke resultatene.

**Tabell 3.10:** *A posteriori*-sannsynlighetene for hypotesene som i Tabell 3.7 med sorterte alleler.

| | Fullstendig | | | | Minimal | | | |
|---|---|---|---|---|---|---|---|---|
| Modell | kandidat | bror | halvbror | ubeslektet | kandidat | bror | halvbror | ubeslektet |
| 1 | 0,0765 | 0,8602 | 0,0632 | 0,0001 | 0,3365 | 0,6186 | 0,0448 | 0,0001 |
| 2 | 0,1846 | 0,7598 | 0,0556 | 0,0001 | 0,2601 | 0,6895 | 0,0503 | 0,0001 |
| 3 | 0,4082 | 0,5519 | 0,0399 | 0,0001 | 0,5749 | 0,3968 | 0,0283 | 0,0000 |
| 4 | 0,2188 | 0,7279 | 0,0532 | 0,0001 | 0,6338 | 0,3421 | 0,0241 | 0,0000 |

Når allelene i de ti markørene er navngitt slik at de blir sortert i riktig rekkefølge blir resultatene i dette tilfellet veldig lite endret, som det kan sees i Tabell 3.10. I det fullstendige systemet er det litt sterkere støtte til hypotese 3 som foreslår broren til den opprinnelige kandidaten som far, i modell 3. Det er som forventet ingen endring i modell 1 og modell 2 da disse mutasjonsmodellene ikke tar hensyn til mutasjonsrekkevidde. Endringene i modell 4 i det fullstendige systemet er så små at det ikke er synlig med bare 4 desimaler.

# 4 Diskusjon

Målet med denne oppgaven var å se på den statistiske beregningen av farskapssaker der det ikke er klart hvem som er faren til barnet, om det kan ha oppstått en mutasjon eller om en nær slektning av kandidaten er den biologiske faren. Programmet Familias ble brukt til å beregne sannsynligheter for de forskjellige familietrærne, målet var å finne ut hvilken av de fire mutasjonsmodellene i Familias som er best egnet til å beregne sannsynlighetene. Det ble brukt minimale og fullstendige systemer for alleler og familietrær for å se hvordan dette påvirker resultatene i de fire forskjellige mutasjonsmodellene. I tillegg ble det undersøkt om navngiving av alleler slik at de sorteres i riktig rekkefølge hadde betydning for resultatene.

## 4.1 Reelt eksempel

I Tabell 3.7 fra det reelle eksempelet kan det sees at det er størst sannsynlighet for at broren til den opprinnelige kandidaten er faren ved bruk av det fullstendige systemet av alleler. I det minimale systemet, hvor de ubrukte allelene er slått sammen til et restallel, har derimot kandidaten høyest sannsynlighet i modell 3 og 4. Dataene er altså relativt varierende, og det kan være forskjellige grunner til dette. Barnet og den angivelige fars genotyper stemmer overens i ni av ti markører. Det er altså bare én markør som ikke stemmer, og det er da naturlig å tenke at det kan være en mutasjon. Men hvor sannsynlig er egentlig denne mutasjonen? Er den så usannsynlig at det istedenfor kan være en eventuell bror som har en DNA-profil svært lik mannen som også stemmer overens med barnets?

I Tabell 3.8 er det oppgitt likelihood-verdier for de fire forskjellige hypotesene for alle de ti markørene, disse verdiene blir så regnet om til *a posteriori*-verdier som vises i Tabell 3.9. Begge tabellene inneholder minimalt og fullstendig allelsystem for markør D18S51. Her kan det sees som tidligere nevnt at barnet og den angivelige faren er kompatible i ni av ti markører, altså er det størst sannsynlighet for at denne mannen er faren i alle markørene unntatt D18S51. I markør D18S51 er det derimot størst sannsynlighet for at faren til barnet er helt ubeslektet med mannen, og av de fire hypotesene er $H_1$ den minst sannsynlige. I denne markøren er det altså så lav

sannsynlighet for at mannen er far til barnet at det totalt sett blir for lite grunnlag for å påstå at denne hypotesen kan være sann. I tillegg gir data liten grunn til å tro at hypotesen om at faren er en ubeslektet mann er riktig. På grunnlag av dette vil det kanskje være mer trolig at det er en nær slektning av mannen som kan være faren, og dette gir støtte til hypotesen om at mannens bror kan være den riktige faren.

I det minimale systemet i Tabell 3.7 og 3.10 gir modell 3 og 4 støtte til hypotesen om at kandidaten kan være faren til barnet. Dette strider mot resultatene i det fullstendige systemet og det kan tenkes at det i dette tilfellet er restallelet som er problematisk. Det kan se ut til at resultatene ut ifra dette påvirkes i stor grad av hvor mange alleler som er til stede i systemet, og kanskje dette er noe som må holdes konstant i en farskapstest.

## 4.2 Restallel

I denne oppgaven virket restallelet til å ha størst påvirkning på resultatene, dette betyr altså at et varierende antall alleler i markørene kan utgjøre en forskjell på hvilke resultater en får. Dette kan sees blant annet i Tabell 3.9 og Tabell 3.10. Hvis dette er en tendens generelt ved farskapstester bør det kanskje være et konstant antall alleler (altså at man bruker allelene og deres frekvenser slik de er observert i tilstrekkelig store databaser) ved hver markør for å unngå sprikende resultater. I denne sammenheng er det dermed viktig at alle alleler for markøren er kartlagt, og det kan virke som det er hensiktsmessig å ha det fullstendige allelsystemet med i beregningen, selv om ikke alle allelene inngår i de gitte genotypene. I denne oppgaven gav et minimalt allelsystem helt andre sannsynligheter enn det fullstendige allelsystemet, og viste seg dermed å være en usikker metode. Resultatene ved bruk av den minimal strategien avhenger altså av hva som tilfeldigvis er observert i en sak og det kan virke urimelig.

## 4.3 Minimalt og maksimalt familietre

Mutasjonsmodell 2 og 4 er som tidligere nevnt stasjonære modeller, og vil dermed ikke påvirkes av at flere individer blir inkorporert i et hypotetisk familietre. Dette vises i Familias-eksempelet i Avsnitt 3.1.2. Det kan også sees i Tabell 3.5 at endringen i LR-verdien ikke er veldig stor når det legges inn besteforeldre til barnet i familietreet. På en annen side bør ikke resultatene bli påvirket av disse overflødige personene i familietreet da alle fullstendige familietrær nødvendigvis vil inneholde foreldre og besteforeldre. Et barn kan ikke eksistere uten foreldre. En hensiktsmessig modell bør derfor være stasjonær i forhold til minimalt og maksimalt familietre, da det som oftest vil bli brukt minimale familietrær ved farskapssaker. Dette kommer av at det vil være mer tungvint å lage et mer fullstendig familietre, og det ville gjort prosedyren ved farskapssaker mer tidkrevende og mindre effektiv.

## 4.4 Sortering

Sortering av alleler viste seg ikke å ha stor påvirkning på resultatene i denne oppgaven, det gav allikevel en liten endring i modell 3 i Tabell 3.10. En grunn til at det ikke gav utslag i dette tilfellet kan muligens være at allelene i markør D18S51 i de gitte genotypene var 12,13 og 16, og disse allelene vil dermed allerede være sortert i riktig rekkefølge. Det kan tenkes at sortering av alleler vil ha større effekt dersom allelene i de gitte genotypene, for en markør der det kan ha forekommet en mutasjon, er mellom 1-9, da det er disse som vil bli sortert i feil rekkefølge av programmet. Prinsipielt bør alleler navngis med minst to siffer slik at de vil sorteres riktig. En av grunnene til dette er at modellen for eksempel ikke bør vektlegge en mutasjon fra 1 til 10 likt som en mutasjon fra 11 til 10, da det i virkeligheten er svært stor forskjell på disse mutasjonene. Det reelle eksempelet i Avsnitt 3.1.3 tar bare utgangspunkt i ett enkelt datasett, og det kunne ha vært interessant å se på effekten av sortering ved flere datasett. Dette kunne for eksempel blitt gjort ved simulering, og det hadde trolig vært mulig å trekke enda sterkere konklusjoner ut ifra dette.

Studier har vist at mutasjoner hos STR stort sett er trinnvise mutasjoner der avstanden som oftest bare er én enkelt repetisjon (Dupuy et al., 2004). Videre kan det se ut til å være en korrelasjon mellom allelets størrelse og om mutasjonen vil føre til en økning eller tap av en repetisjonsenhet.

Store alleler viste en tendens til å miste en repetisjon, mens små alleler som oftest fikk en ekstra repetisjon. I studiet til Dupuy et. al. observerte de også mutasjoner der et allel fikk en økning på tre repetisjoner, men den vanligste mutasjonen i dette studiet var en økning på én repetisjon.

## 4.5 Hva er oppgaven til den sakkyndige?

I sivile farskapssaker sendes det DNA-tester inn til et laboratorium, resultatene blir så utredet av sakkyndige, og det skrives en sakkyndigrapport som sendes tilbake til rekvirenten. I sakkyndigrapporten formuleres det en konklusjon som sier at resultatene enten taler for eller mot farskapet. I noen tilfeller må en sakkyndig også bistå i rettssaker.

I farskapssaker der det brukes DNA-bevis beregnes det som oftest LR-verdier eller sannsynlighetsbrøker som det også kan kalles på norsk. Det beregnes først en likelihood, altså en sannsynlighet, ut ifra DNA-profilene. En likelihood for en gitt hypotese kan divideres på en likelihood for en alternativ hypotese, og til sammen utgjør dette en LR-verdi. Det kreves *a priori*-sannsynligheter for å beregne en *a posteriori*-sannsynlighet med den Bayesianske metoden som nevnt i Avsnitt 1.4 og 2.2.3. Noe som kan være problematisk med denne beregningen er *a priori*-sannsynligheten, altså sannsynligheten for de forskjellige hypotesene før en har fremlagt DNA-bevis eller foretatt beregninger.

En sakkyndig skal forholde seg til resultatene en finner ved beregninger basert på data, og dermed kan en sakkyndig ikke angi noe som teller for eller mot før beviset er fremlagt. Så langt det er mulig skal en sakkyndig ikke ha noen oppfatning om saken før det er foretatt beregninger, og *a priori*-sannsynlighetene for de forskjellige hypotesene blir dermed antatt å være like store. Videre er det ikke en sakkyndigs oppgave å etterforske utover det resultatet viser. Dette er kanskje noe lettere i en farskapssak enn for eksempel i en drapssak. Det som derimot kanskje kan være litt vanskeligere i en farskapssak er å ikke spekulere når det tilsynelatende er noe rart med resultatene en får. Som i det reelle eksempelet som er brukt i denne oppgaven, der det kan virke som det er en høy sannsynlighet for at broren til kandidaten er barnets far. Dersom farskapssaken som er sendt inn derimot bare etterspør sannsynligheten for om mannen er far, bør en sakkyndig da si ifra om at det er høyere sannsynlighet for at det er mannens bror? Dette er nok ikke den

sakkyndiges jobb da det vil falle inn under kategorien etterforskning. Videre er det fortsatt mulig å angi en sannsynlighet for at mannen er barnets far, men dette blir da bare en del av det endelige svaret på farskapssaken. Det er i dag altså ikke problematisk å beregne en LR-verdi som taler for eller mot et farskap, men problematikken ligger i hvordan resultatet skal presenteres i den sakkyndiges rapport.

## 4.6 Tall eller ord?

Finnes det en riktig måte å fremlegge resultatet på etter at sannsynligheten for farskapet er beregnet? I Norge finnes det ikke en standardisert måte å fremlegge bevis innenfor forensikk enda. Dette gjør fremlegging av resultater enda mer komplisert. Er det i det hele tatt mulig for en sakkyndig å presentere tallene uten å formulere en konklusjon i denne saken?

Det kan tenkes at det ikke er et tilfredsstillende svar på en farskapstest at LR-verdien taler mot farskapet, men heller ikke for en ubeslektet mann. En mulighet er å presentere LR-verdien og en konklusjon som sier at det er en svak indikasjon for at mannen er barnets far. Det bør i det minste formuleres om tallene taler for eller mot farskapet slik at det ikke er opp til personen som får svaret å tolke det alene, da det gjerne kan feiltolkes. I dette tilfellet ville det kanskje vært konkludert med at tallene ikke gir en sterk indikasjon på farskapet.

## 4.7 Forbehold om nære slektninger

Ved svar på en farskapstest er det en gitt standard at det blir tatt forbehold om at det er en viss sannsynlighet for at den ekte faren kan være en nær slektning av kandidaten. Grunnen til dette er at søskens genotyper stort sett er svært like, men sjeldent identiske med mindre de er eneggede tvillinger. Når det viser seg at genotypen til et barn og den angivelige faren ikke stemmer 100 % overens, kan det på ene siden ha oppstått en mutasjon, men det kan også være en mulighet for at det er mannens bror som har en genotype som stemmer bedre med barnets. Det største problemet med dette forbeholdet er at det ikke kan testes om det er mannens bror som har den «riktige»

genotypen med mindre det blir etterspurt. Genotypene kan dermed ikke sammenliknes og konklusjonen som gis er da ikke like tilfredsstillende som den kanskje kunne ha vært. På en annen side er det kanskje ikke den sakkyndiges oppgave å foreslå at en eventuell bror bør DNA-testes da det for eksempel kan være snakk om utroskap som den angivelige faren ikke er klar over. Dette blir i så tilfelle et etisk problem. På grunn av dette kan det bli tatt et forbehold om at det kan være en nær slektning, men det bør ikke bli oppgitt i en eventuell konklusjon med mindre det blir etterspurt.

## 4.8 Videre arbeid. Hva er riktig mutasjonsmodell?

For å kunne gi pålitelige svar i en farskapssak er det viktig at modellen som brukes til de statistiske beregningene er mest mulig egnet. Mutasjonsmodell 1 som vektlegger alle mutasjoner likt blir nok i mange tilfeller litt for enkel, da det i virkeligheten ikke er like stor sannsynlighet for alle mutasjoner i et system med mange alleler. To alleler med en avstand på ti repetisjoner mellom seg har ikke like stor sannsynlighet for å mutere til hverandre som to alleler som differensieres med bare én repetisjon.

Modell 2 angir en proporsjonalitet mellom allelfrekvens og eventuelle mutasjoner til det gitte allelet. Den tar ikke hensyn til hvilket allel det muteres fra, og dermed heller ikke avstanden mellom allelene, i likhet med modell 1. En fordel med de to første modellene i Familias er at det ikke er nødvendig å oppgi parameteren mutasjonsrekkevidde, da det ikke har betydning i disse modellene som ikke ser på avstanden mellom allelene.

Modell 3 og 4 krever at det blir lagt inn en parameter for mutasjonsrekkevidde. Dette gjør modellene litt mer kompliserte enn modell 1 og 2 da det må angis et tall som sier hvor mye mutasjonsraten reduseres for hver repetisjon. De to modellene er kanskje de mest realistiske i forhold til at de tar hensyn til hvilket allel mutasjonen går fra, i motsetning til modell 1 og 2. Forskjellen mellom modell 3 og 4 er som tidligere nevnt at modell 4 er stasjonær, mens modell 3 ikke er stasjonær. Dette betyr at resultatene i modell 3 ikke vil være konstant når det legges til flere personer i systemet, som ikke er genotypet. Det problematiske med disse ikke-stasjonære modellene er at sannsynligheten for om en mann er far til et barn vil endres dersom det legges inn

i systemet som ekstra informasjon at mannen selv har en far og en mor. Det sier seg selv at en person ikke kan eksistere uten en far og en mor, og dette burde dermed ikke påvirke sannsynligheten for at mannen er barnets far. Det kan derfor virke som modell 4 kan være et bedre alternativ da det er en stasjonær modell og ikke vil gi varierende resultater når det legges inn såkalte overflødige personer i systemet.

## 4.9 Parameterestimering

Mutasjonsraten må angis i alle de fire forskjellige mutasjonsmodellene i Familias. Mutasjonsraten angir hvor sannsynlig det er at mutasjoner forekommer, og det er tidligere nevnt at mutasjonsraten ved meiose hos menn ligger rundt 0,005. En mutasjonsrate vil allikevel ikke være konstant for alle markører, men for å beregne nøyaktige mutasjonsrater for hver enkelt markør er det behov for omfattende studier med store mengder data. Studiet til Dupuy et. al. støtter oppunder dette, da det ble observert signifikante forskjeller i mutasjonsrater både mellom loci og innad i locus. Dette studiet konkluderte med at mutasjonsraten bør være spesifikk for det gitte locuset, altså markøren, og i tillegg allelspesifikt da det ble påvist store forskjeller i mutasjoner hos alleler av ulik størrelse.

I Avsnitt 2.1.1 blir det nevnt at mutasjonsraten for menn ved dannelse av kjønnsceller er ca. 0,5 %, dette har vist seg å være noe høyere enn hos kvinner. Dette kan virke logisk da det dannes mange flere kjønnsceller hos menn enn hos kvinner. Denne forskjellen bør i så fall kartlegges nærmere da det kanskje burde vært gitt forskjellige parametere for menn og kvinner; Familias tillater ulike modeller og parameterverdier for menn og kvinner.

Mutasjonsrekkevidde er en parameter som bare må oppgis i modell 3 og 4 i Familias. Denne parameteren gjør beregningen mer komplisert da mutasjonsrekkevidden i seg selv må beregnes. På en annen side er det viktig å ta hensyn til avstanden mellom allelene i en eventuell mutasjon, som tidligere nevnt. Dette tallet, mutasjonsrekkevidden r, kan være vanskelig å estimere basert på tilgjengelige data. Selv med nærmere 2000 farskapssaker, slik som i artikkelen Dupuy (2004), blir det relativt få mutasjoner og lite grunnlag for å estimere r.

# 5 Konklusjon

Bruk av restallel viste seg å ha størst utslag for resultatene i denne oppgaven, og metoden viste seg å være usikker da det minimale systemet gav andre resultater enn det fullstendige. Antall alleler i en farskapssak bør være basert på det som er observert i databasen for å unngå denne usikkerheten. Overflødige personer i et familietre påvirker modell 1 og 3. Dette usikkerhetsmomentet kan unngås ved å bruke modell 2 og 4. Riktig sortering av alleler hadde ikke stor påvirkning på resultatene i denne oppgaven. Alleler bør allikevel navngis med minst to siffer da det bør tas hensyn til avstanden mellom alleler i en mutasjon. Studier har vist at det er størst sannsynlighet for at et allel muterer til de allelene som ligger nærmest. Modell 4 i Familias kan virke som den mest hensiktsmessige å bruke ved beregninger i farskapssaker med mutasjoner, da den ikke påvirkes av overflødige personer i familietreet samtidig som den tar hensyn til størrelsen på en eventuell mutasjon.

## Referanser


Buckleton JS, Triggs CM, Walsh SJ. *Forensic DNA evidence interpretation.* CRC Press, 2005.

Butler JM. *Advanced topics in forensic DNA typing: methodology.* Academic Press, 2012.

Drábek J. Validation of software for calculating the likelihood ratio for parentage and kinship. *Forensic Science International: Genetics*, 3:112-118, 2009.

Dupuy Myhre B, Stenersen M, Egeland T, Olaisen B. Y-Chromosomal Microsatellite Mutation Rates: Differences in Mutation Rate Between and Within Loci. *Human mutation*, 23:117-124, 2004.

Egeland T. Statistisk vekting av DNA-funn
i straffesaker. *Tidsskrift for Strafferett*, 2: 190-204, 2009.

Egeland T, Mostad PF, Mevag B, Stenersen M: Beyond traditional paternity and identification cases. Selecting the most probable pedigree. *Forensic Sci Int* , 110:47-59, 2000.

Hampel F. On the foundations of statistics: A frequentist approach. Seminar für Statistik, *Eidgenössische Technische Hochschule* (ETH). 1998

Løvås GG. *Statistikk for universiteter og høgskoler*.Universitetsforlaget, 2. utgave, 2008.

Vigeland MD, Selmer KK, Egeland T. Medical statistics; Statistical methods in genetics. Book chapter in Medical statistics in clinical and epidemiological research, ed Marit B. Veierød, Stian Lydersen, Petter Laake, Gyldendal Akademisk 2012.


## Lenker


[1] Ervik, Kristian (2013). «Eneggede tvillinger skaper hodebry for fransk politi», 12.februar. <http://www.tv2.no/nyheter/utenriks/eneggede-tvillinger-skaper-hodebry-for-fransk-politi-3985066.html> [Lesedato: 03.04.13]

[2] NTB (2013). «Forsøker å skille mellom tvillinger i voldtektssak», 10.februar. <http://www.aftenposten.no/nyheter/uriks/Forsoker-a-skille-mellom-tvillinger-i-voldtektssak-7117504.html#.UVryvZMqzy0> [Lesedato: 03.04.13]

[3] Store norske leksikon (2009). «forensisk», 14.februar.

<http://snl.no/forensisk> [Lesedato: 19.04.13]

[4] Thair Shaikh (2007). «Sally Clark, mother wrongly convicted of killing her sons, found dead at home», 17.mars. <http://www.guardian.co.uk/society/2007/mar/17/childrensservices.uknews>[Lesedato: 21.04.13]

[5] Bjørnar Olaisen, Store norske leksikon (2009). «Rettsgenetikk», 13. februar. <http://snl.no/.sml_artikkel/rettsgenetikk> [Lesedato: 1.1.2013]

Familias-filene kan fås fra Maria Berggreen på forespørsel.