CHALMERS | GÖTEBORGS UNIVERSITET

Hur mycket släkt är släktingar?

En studie i den genetiska likhetens variation

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet
Kandidatarbete inom civilingenjörsutbildningen vid Chalmers tekniska högskola

Vanessa Emanuelsson
Ida Petersson
Oskar Svensson

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2012

Hur mycket släkt är släktingar?

En studie i den genetiska likhetens variation

Examensarbete för kandidatexamen i matematik inom matematikprogrammet vid Göteborgs universitet

Vanessa Emanuelsson

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers tekniska högskola

Ida Petersson Oskar Svensson

Handledare: Staffan Nilsson
Examinator: Carl-Henrik Fant

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2012

Sammanfattning

När en recessiv sjukdom studeras i en släkt används jämförelser av familjemedlemmarnas arvsmassa. Med hjälp av datorsimuleringar som utgår från modellering av arvsförloppet kan information erhållas om hur mycket arvsmassa individerna har gemensamt. Denna information kan vara till nytta vid en fysisk kartläggning av individernas genom.

I detta projekt har ett Java-program konstruerats som på ett verklighetsnära sätt modellerar arvsförloppet. Tillsammans med Java-programmet har ett mer teoretiskt resonemang genomförts och implementerats i MATLAB, i syfte att få referensdata.

Java-programmet har använts för att undersöka den genetiska likheten mellan besläktade individer. Informationen som erhållits har använts för att approximera fördelningar för individernas genetiska likhet. Utifrån dessa uppskattningar fastslås att fördelningarna har relativt låg varians på grund av genomets extensiva totala genetiska längd. Det konstateras att individernas könskromosomer bidrar med skillnader i medelvärde. Dessutom fastställs att mäns och kvinnors olika genetiska längder bidrar med skillnader i varians.

Abstract

When a recessive disease is studied within a family, comparison of the family members' genome is used. With computer simulations based on modeling of the process of inheritance, information about the amount of shared DNA can be obtained. This information can be useful when mapping the individuals' genome.

In this project a Java program has been designed which in a realistic manner models the process of inheritance. To obtain reference data to the Java program, a more theoretical approach of modeling the inheritance process has been implemented in MATLAB.

The Java program has been used to investigate the genetic similarity of related individuals. The information obtained has been used to approximate the distributions of the individuals' genetic similarities. Based on these estimates, it can be stated that the distributions have relatively low variance due to the genome's extensive total genetic length. It is concluded that the individuals' sex chromosomes contribute to differences in mean values. In addition, the difference in men's and women's genetic lengths contributes to differences in variance.

Innehåll

Förord

Inledningsvis riktas ett stort Tack till projektets handledare Staffan Nilsson som varit till mycket hjälp under arbetsprocessen!

Under projektets utförande har gemensam dagbok och personliga loggböcker förts. I dessa framgår hur arbetsprocessen fortgått.

På grund av ett mycket väl fungerande samarbete har i princip hela projektet utförts gemensamt i helgrupp, det enda arbetet som utförts individuellt är skrivarbete. All framställd text har dock utvärderats och korrigerats av varje gruppmedlem. Därmed är det mycket svårt att ange huvudförfattare för olika avsnitt. Den person som skrivit de första versionerna av vissa avsnitt anges nedan.

Ida: Avsnitt 1.2, 1.3, 3.2, 3.3.4
Oskar: Avsnitt 2.2, 3.3.1, 3.3.2
Vanessa: Avsnitt 1.1, 1.4, 3.1, 3.3.3, 3.4

1 Inledning

Idag finns det möjlighet att kartlägga den mänskliga arvsmassan och jämföra två individers genetiska information. På så vis kan den genetiska likheten mellan släktingar beräknas. Om det finns en ärftlig sjukdom inom en familj kan en jämförelse av familjemedlemmarnas arvsmassa vara till nytta för att undvika ytterligare spridning av sjukdomen. Det är dock tidskrävande att fysiskt jämföra alla individers arvsmassa i en familj. För att underlätta processen går det, med hjälp av datorkraft, att modellera arvsförloppet och på så vis praktiskt beräkna sannolikheten för huruvida ett framtida barn kommer att drabbas av sjukdomen. I det här projektet används datorsimulering tillsammans med teori för att mellan besläktade individer undersöka fördelningar för genetisk likhet på grund av arv.

Den genetiska likheten mellan släktingar beror av hur arvsmassa förs vidare från förälder till barn. En individ delar alltid hälften av sin arvsmassa med sin far respektive mor [1]. Hur mycket individen delar med exempelvis sin morfar är dock inte lika självklart eftersom det beror av vilka delar av moderns arvsmassa personen ärver. I *genomsnitt* delar dock en person $(\frac{1}{2})^2 = \frac{1}{4}$ med varje mor- och farförälder, eftersom individen delar hälften av sin arvsmassa med vardera förälder som i sin tur delar hälften av sin arvsmassa med sina föräldrar. Genomsnittlig gemensam andel arvsmassa mellan två individer i ett rakt nedstigande led går därför att utöka till det generella uttrycket $(\frac{1}{2})^n$, där n är antalet generationer mellan de två individerna. *Fördelningen* för den genetiska likheten mellan två individer är dock mycket svårare att beräkna analytiskt och därför är datorsimulering ett lämpligt verktyg.

Projektet behandlar tillämpad statistik inom genetik och kommer därför att innehålla en hel del genetiska begrepp. En kortare sammanfattning av den nödvändiga genetiken presenteras därför i de kommande avsnitten. I de fyra första avsnitten, 1.1 - 1.4 har information hämtats från följande källor: Starr C, et al; 2010, [2], Martin D, et al; 2007, [3], Alberts B, et al; 2009, [4], Nilsson S; 2001, [5], Bennett R; 2011, [6], Hartl D, et al; 2006, [7], Frankham R, et al; 2002, [8].

1.1 Genetiska begrepp

Det kemiska ämnet deoxiribonukleinsyra (*DNA*) bygger upp *genomet* (den genetiska informationen) hos en individ [2]. Ämnets viktigaste funktion är att förvara instruktionerna som används för att konstruera bland annat proteiner. De delar av DNA-molekylen som ansvarar för tillverkningen av dessa kallas *gener*. Mellan generna finns ofta långa sekvenser som inte kodar för något protein. Dessa kallas ibland för *skräp-DNA* eftersom de inte har någon relevans för den genetiska koden.

Människor lagrar sitt DNA i *kromosomer* [3]. Genomet är för människor fördelat på 46 kromosomer som är uppdelade i 23 kromosompar. De första 22 kromosomparen kallas *autosomer* och det sista paret består av *könskromosomerna*; kvinnor har två X-kromosomer medan män har en X- och en Y-kromosom. Varje kromosompar består av en kromosom från individens mor samt en kromosom från individens far. Dessa kromosomer betecknas som *maternell* respektive *paternell* kromosom.

I kromosomerna finns de så kallade *nukleotiderna* som bland annat består av de fyra kvä-

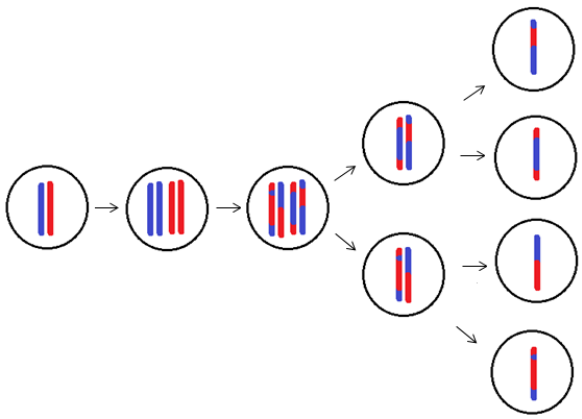
vebaserna tymin (T), cytosin (C), guanin (G) och adenin (A). Nukleotidsekvensen i genen översätts i steg till en aminosyrasekvens som bildar ett protein. På så sätt är ordningsföljden av dessa nukleotider i DNA-molekylen viktig eftersom den bestämmer vilken form proteinet får och därmed vilken funktion det har. DNA-molekylen är uppbyggd av två kedjor av nukleotider, vilka kopplas ihop på ett speciellt sätt. Tymin i den ena kedjan är alltid bunden till adenin i den andra, medan guanin alltid är bunden till cytosin. Två nukleotider ihopkopplade på detta sätt kallas ett *baspar*.

För att genomet ska föras vidare från cellgeneration till cellgeneration dubbleras DNA-molekylen vid celldelningen så att en kopia hamnar i varje dottercell. Denna process kallas för *replikation* och sker oftast utan problem. I vissa fall sker dock fel i kopieringen som leder till att genomet förändras, en *mutation* har då uppstått. Denna förändring i nukleotidsekvensen kan få till följd att den motsvarande sekvensen av aminosyror blir annorlunda. Mutationen har ofta ingen betydelse, men kan ibland förändra eller förhindra proteinets funktion. Detta leder till en *genetisk variation*.

Olika versioner av samma gen kallas för *alleler*. Individer som har två olika alleler av en gen sägs vara *heterozygota*, medan de som har två lika sägs vara *homozygota*. Fysiska egenskaper hos en individ, *fenotyp*, som till exempel utseende kan ärvas *dominant* eller *recessivt*. Egenskaper som ärvs dominant behöver endast ett anlag medan det vid recessiv ärvning krävs att båda allelerna bär samma anlag. *Genotyp* är den totala uppsättningen alleler hos en individ.

1.2 Meios - bildandet av könsceller

Våra könsceller är så kallade *haploider* [4], vilket innebär att de bara innehåller en uppsättning av kromosomer till skillnad från vanliga celler som är *diploider* och innehåller par av kromosomer. Celldelningen som framställer könsceller, *meios*, ser därför annorlunda ut jämfört med den delning övriga celler genomgår. Meiosen är snarlik för kvinnor och män och består av fyra stadier; en kopieringsfas, en överkorsningsfas samt två delningsfaser, se figur 1.



Figur 1: Meiosens stadier; kopieringsfasen, överkorsningsfasen och de två uppdelningsfaserna

Det hela börjar med en specialiserad cell som innehåller en kromosom från individens far och en kromosom från individens mor kallade paternell respektive maternell *homolog*. Dessa homologer dubbleras för att sedan genomgå en överkorsningsprocess. Under överkorsningen lägger sig tvillingparen längs med varandra där de sedan delar sig och parar ihop sig med varandra så att det bildas fyra nya homologer med information från både de tidigare maternella och paternella homologerna. Kopieringsfasen av meiosen avslutas genom att de fyra nya homologerna delar upp sig i två nya par.

Under nästkommande fas av meiosen delas först cellen i två med ett kromosompar i varje dottercell, likt en normal celldelning. Den sistkommande delningen skiljer sig från den normala celldelningen (*mitosen*) på så sätt att kromosomerna separeras och vardera homolog förs vidare till varsin könscell.

Sammanfattningsvis börjar meiosen med en ensam cell innehållande information från individens mor och far och avslutas med fyra haploidceller innehållande skilda kombinationer av information från individens föräldrar. Eftersom de fyra haploiderna är unika kan det i slutändan leda till att syskon är mycket olika varandra. Nästa omgång könsceller kan å andra sidan generera kromosomer som är väldigt lika de i förra omgången och syskon kan således också vara mycket lika varandra. Överkorsningsprocessen under meiosen ger alltså upphov till stor genetisk variation.

När nya homologer bildas under överkorsningsfasen kan de få mycket olika utseende. Vid startänden på en homolog är det lika stor sannolikhet att den börjar med *maternell* respektive *paternell* information. Överkorsningar som sker på homologen är sedan relativt slumpmässiga och modelleras i allmänhet som en Poisson-process eftersom nästkommande överkorsning inte beror av tidigare överkorsningar [5].

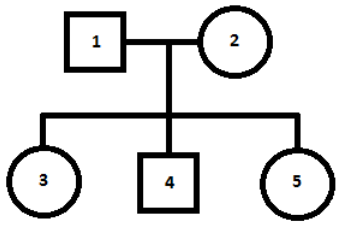
Hur många överkorsningar som sker på en homolog beror på om det är en kvinna eller en man som producerar könscellen, men även vilken kromosom det är. De 23 kromosomparen har nämligen olika *genetisk längd*, vilken beskriver hur många överkorsningar kromosomen i genomsnitt har, se appendix A. Kromosomer har alltså ett ytterligare längdbegrepp utöver den fysiska längden. Normalt sett sker en till tre överkorsningar på varje homolog beroende på den genetiska längden på ursprungskromosomerna. Kvinnor har större genetiska längder och därmed en intensivare meios än män. Det sker därför fler överkorsningar när en kvinna producerar sina könsceller. Den genetiska längden mäts i Morgan (M). På en kromosom som har den genetiska längden 1 M förväntas en överkorsning ske.

1.3 Pedigree - representation av ett släktträd

Ordet *pedigree* betyder stamtavla och används för att beskriva ett släktträd med hjälp av en bild samt eventuellt en matris [6].

Bilden för ett pedigree följer en standardstruktur där män definieras med en kvadrat och kvinnor definieras med en cirkel. Horisontella linjer mellan två personer indikerar att de är ett par och vertikala linjer indikerar relationen mellan föräldrar och deras barn, se figur 2. I fallet då en sjukdom studeras i ett pedigree markeras sjuka individer med en fylld cirkel

respektive kvadrat. Personer som är avlidna markeras med överstruken symbol.



Figur 2: Bilden för ett pedigree beskrivs med cirklar för kvinnor, kvadrater för män, horisontella linjer för föräldrapar och vertikala linjer för relation mellan föräldrar och barn

Pedigreematrisen består av fem standardkolumner enligt följande; familj, person, far, mor samt kön. Ett enkelt exempel på ett pedigree är den svenska kungafamiljen, det vill säga Kung Carl XVI Gustaf, Drottning Silvia och deras barn Kronprinsessan Victoria, Prins Carl Philip och Prinsessan Madeleine.

I detta exempel tillhör alla individer samma familj, därav samma siffra i första kolumnen i tabell 1. Numreringen av personer behöver inte följa någon struktur men vanligast är att personer från en äldre generation i familjen har lägre siffror än personer från en yngre generation.

Tabell 1: Matrisen för det pedigree som beskriver kungafamiljen, där kolumnerna representerar familj, person, föräldrar samt kön

	Familj	Person	Far	Mor	Kön
Carl XVI Gustaf	1	1	0	0	1
Silvia	1	2	0	0	2
Victoria	1	3	1	2	2
Carl Philip	1	4	1	2	1
Madeleine	1	5	1	2	2

Tredje och fjärde kolumnen bestäms av en persons föräldrar och är därför lika för Victoria, Carl Philip och Madeleine. Eftersom Carl XVI Gustaf är deras far markeras i detta fall kolumn tre med 1 och på samma sätt markeras kolumn fyra med 2 eftersom Silvia är deras mor. I det här exemplet utgör Carl XVI Gustaf och Silvia den äldsta generationen i trädet, därför markeras deras föräldrar som ”okända” med siffran 0. Slutligen används siffran 1 i kolumn fem för att visa att Carl XVI Gustaf och Carl Philip är män och 2 för att indikera att Silvia, Victoria och Madeleine är kvinnor.

Ett pedigree är ett bra verktyg för att följa genetiska sjukdomar inom familjer. Då utökas standardmatrisen med en extra kolumn där det noteras om individen är frisk med siffran 1 respektive har den aktuella sjukdomen med siffran 2.

Meioser är ett praktiskt sätt att mäta avståndet mellan två individer i ett pedigree. Mellan en förälder och dess barn skiljer det enbart en meios. Dessutom är denna meios deterministisk eftersom barn alltid ärver precis sina maternella respektive paternella kromosomer från respektive förälder. På motsvarande sätt skiljer det tre meioser mellan en individ och dess farfars far. Meioser kan utnyttjas för att modellera hur mycket DNA en individ delar med en förfader. Begreppet går med mindre förändringar att utnyttja även mellan exempelvis syskon eller kusiner.

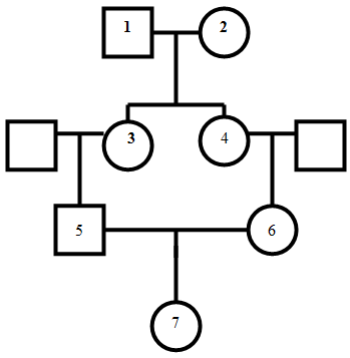
1.4 Genetisk inverkan vid inavel

Hos alla individer finns det skadliga alleler som är recessiva. Om dessa förekommer i heterozygot form tillsammans med en frisk allel så kommer de inte till uttryck. Sannolikheten för att de skadliga allelerna ska förekomma i homozygot form är liten. Vid inavel, då närbesläktade individer får barn tillsammans förekommer fler alleler i homozygot form eftersom samma anlag kan ärvas från fadern och modern. Detta leder till att den genetiska variationen minskar. I samband med till exempel sjukdomsutbrott ger två versioner av en gen bättre överlevnadsmöjligheter. Därav kan inavel orsaka exempelvis nedsatt immunförsvar, missbildningar och mentala handikapp.

Inavel mäts med hjälp av en inavelskoefficient som betecknas med F. Den kan variera från 0 till 100% och mäter sannolikheten att de två allelerna för varje gen är lika på grund av att de kommer från samma källa [7]. Även om den primära konsekvensen av inavel är ökad homozygositet, är F inte ett direkt mått på denna. De två allelerna kan vara lika av andra skäl, eftersom det finns en viss nivå av homozygositet i en vanlig befolkning.

1.4.1 Exempel på beräkning av inavelskoefficient

Hur inavelskoefficienten beräknas illustreras nedan med ett exempel där två individer är kusiner via två systrar, se figur 3. Kusinernas barn, person 7, kommer att ha en ökad homozygositet och därför beräknas inavelskoefficienten för henne.



Figur 3: Pedigree där en manlig och kvinnlig kusin har barn tillsammans

Den enklaste metoden för att beräkna inavelskoefficienten är *vägmetoden* [8]. Den består i att fastställa var och en av de möjliga vägarna från fadern till modern genom en gemensam förfader. Detta kommer att ge alla möjliga vägar för att ett barn ska få samma allel av båda sina föräldrar. Om det inte finns någon gemensam förfader är koefficienten 0. Om det finns mer än en gemensam förfader så bestämmer man vägarna för var och en av dessa och adderar dem.

I exemplet har kusinerna gemensamma morföräldrar. Vägen från fadern, person 5, till modern, person 6, genom morfar, person 1, respektive mormor, person 2, blir $5-3-1-4-6$ och $5-3-2-4-6$. Sannolikheten är alltid $\frac{1}{2}$ att en viss allel ska skickas vidare till nästa generation. Koefficienten blir då $(\frac{1}{2})^5+(\frac{1}{2})^5=\frac{1}{16}$, vilken alltså representerar sannolikheten att person 7 kommer att vara homozygot för en viss allel på grund av föräldrarnas gemensamma morföräldrar.

1.5 Syfte

Det huvudsakliga syftet med arbetet är att undersöka ”hur mycket släkt” det kan påstås att släktingar är. Detta alldagliga uttryck har i projektet kvantifierats med hjälp av ett Java-program som modellerar arvsförloppet. Fokus i arbetet ligger helt på genetisk nivå, yttre påverkan behandlas alltså inte. Resultatet består således av siffror för hur stor andel arvsmassa släktingar sannolikt delar med varandra.

Projektet är av intresse eftersom resultatet skulle kunna användas för att beräkna sannolikheter för att recessiva sjukdomar ska spridas till nästa generation. Om det går att jämföra gemensamma fragment mellan familjemedlemmar kan risken för sjukdomsspridning minimeras.

1.6 Avgränsningar

Vid varje meios antas antalet överkorsningar ske slumpmässigt enligt Poisson-processen. Det är enligt Gupta PK; 2007, [9] vedertaget att överkorsningsprocessen kan modelleras utifrån en Poisson-fördelning, vilken har egenskapen att den saknar minne. För varje punkt på kromosomen är det alltså lika stor sannolikhet för överkorsning oberoende av vad som hänt tidigare. Denna egenskap kan enkelt bevisas, se appendix B.

Under simuleringsfasen slumpas positioner för var överkorsningar kommer att ske med hjälp av en likformig fördelning. Här har dubbelöverkorsningar, det vill säga att två överkorsningar sker på samma position, exkluderats. Sannolikheten för att detta ska inträffa är minimal eftersom genomet är mycket långt.

Meios för kvinnans alla kromosomer simuleras, emellertid har simulering av mannens könskromosomer uteslutits. Av sin fader ärver alltså en individ en könskromosom som inte genomgått meios. Detta eftersom det sker så få överkorsningar mellan mannens X- och Y-kromosom att de är försumbara [10].

I projektet har det antagits att antalet mutationer som sker är noll. Om mutationer skulle tillgodoses skulle projektets storlek öka kraftigt och det skulle bli problematiskt att särskilja olika typer av mutationer. Dessutom skulle en punktmutation knappt påverka två individers genetiska likhet, även om den mot förmodan skulle leda till stor förändring av fenotypen hos en av individerna.

1.7 Metodöversikt

Grunden till projektet ligger i att kvantifiera hur lika släktingar kan vara.

- Först och främst har datorsimulering använts. Ett Java-program som läser in ett fördefinierat pedigree och som innehåller en rutin som simulerar meiosen vid bildande av könsceller har skrivits. Programmet syftar till att på ett realistiskt vis efterlikna arvsförloppet. Dessutom har jämförelser av två personers arvsmassa möjliggjorts för att kunna undersöka hur stor del av genomet som är gemensamt. En mer ingående beskrivning av den praktiska delen går att läsa i avsnitt 2.1.

- För att få referensdata till de simulerade resultaten har ett mer teoretiskt resonemang använts. Detta utförs med hjälp av statistiska beräkningar och logiska antaganden. Precis som Java-programmet syftar teorin till att så verklighetsnära som möjligt avbilda arvsförloppet. Teorin förtydligas även genom ett tillämpat exempel. Dessa finns att läsa i avsnitt 2.2.

Sammanfattningsvis har simuleringar och teori använts för att beräkna genetisk likhet mellan släktingar. Resultaten har nyttjats för att undersöka huruvida kusiner kan vara mer genetiskt lika än syskon. Hur inavel påverkar den genetiska likheten har också undersökts under frågeställningen om hur mycket mer genetiskt likt till exempel ett barn blir sina föräldrar om föräldrarna är kusiner. Dessutom har det undersökts på hur långt avstånd två individer kan vara släkt och fortfarande dela någon arvsmassa. Resultaten presenteras i avsnitt 3.

2 Genomförande

I detta avsnitt presenteras de verktyg som har använts för att undersöka hur mycket arvsmassa som individer har gemensamt givet en viss relation.

Först redogörs för det program som skrivits i Java, i avsnitt 2.1. Detta program har använts för att producera större delen av de resultat som presenteras i avsnitt 3. Programmet består av en ansenlig mängd kod, närmare 600 rader. Koden bifogas därför inte eftersom den skulle vara svår att överblicka.

För varje relation som undersökts har programmet simulerat arv genom ett fördefinierat pedigree. Relationerna har simulerats 100 000 gånger vardera. För varje simulering erhölls den andel gemensam arvsmassa som två individer delar. Datamängderna hanterades i MATLAB där de användes för att approximera sannolikhetsdistributionen för längden på individernas gemensamma DNA.

Det program som skrivits i Java implementerar ett rättframt och verklighetsnära sätt att simulera arvsprocessen. Programmet strävar efter att efterlikna det verkliga arvsförloppet i så stor utsträckning som möjligt. Därför bör programmet ge korrekta resultat. Tester har utförts för att försöka bekräfta detta, men de kan endast påvisa att programmet ger rimliga resultat. För att kunna motivera programmets korrekthet ytterliggare används ett mer teoretiskt sätt att representera arvsförloppet. Detta tillvägagångssätt presenteras i avsnitt 2.2 och resonemangets implementation i MATLAB beskrivs i avsnitt 2.2.1. Tillsammans presenteras och jämförs resultaten från Java-programmet och MATLAB-rutinen i avsnitt 3.1.

De mer teoretiska beräkningarna utförs endast för ett kromosompar. Resonemangets grundidé vilar på information som projektets handledare, Docent Staffan Nilsson, bidragit med. Resonemanget syftar till att hitta ett uttryck för fördelningen av den gemensamma längden arvsmassa som två individer har.

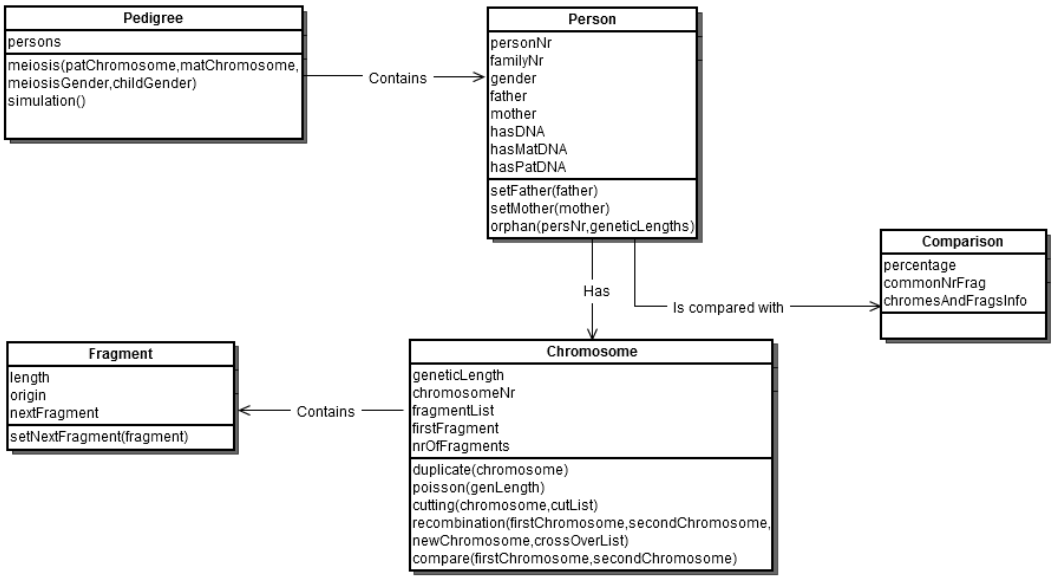
2.1 Beskrivning av Java-programmet

Det program som används för simulering av arvsprocessen enligt ett visst pedigree är skrivet i Java. Programmet är beskrivet nedan och för att visualisera de klasser som ingår används ett diagram, se figur 4.

Inledningsvis läser programmet in en text-fil innehållande en matris som beskriver ett pedigree. Matrisen kopieras sedan över till ett objekt i programmet där kolumnerna beskriver familj, person, far, mor samt kön på individerna. Matrisen används för att skapa ett objekt av klassen Pedigree. Denna klass innehåller huvudsakligen en lista i vilken personerna som ingår i släktträdet sparas.

För att kunna spara varje individs genetiska information på lämpligt sätt finns i Person-klassen pekare till tre heltal som indikerar vilken familj personen tillhör, vilket nummer personen har i strukturen samt vilket kön personen har. Ett objekt av Person-klassen pekar också till dess kromosomer och individens föräldrar. I Person-klassen finns även booleaner som anger huruvida personen redan har fått sitt DNA genererat från sina föräldrar. Till en

början saknas kromosominnehåll eftersom meiosen ännu inte simulerats, därav skapas tomma kromosomer och pekarna som ska peka till individens föräldrar tilldelas null. När ett objekt av klassen Pedigree initieras skapas lika många Person-objekt som det finns personer i släktträdet och dessa sparas i Pedigree-objektet. De tillskrivs även de värden som beskriver individens familj, personnummer, föräldrar och kön.

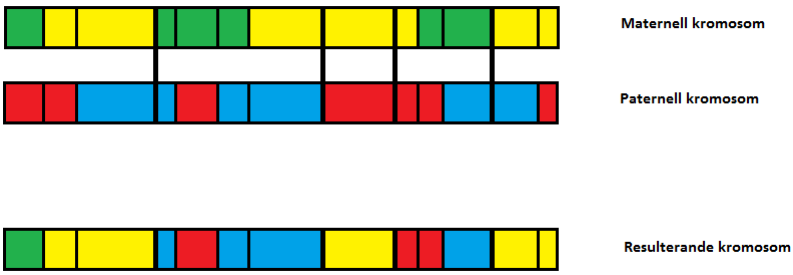


Figur 4: Ett diagram som beskriver Java-programmets struktur

Under simuleringsfasen då alla personer skall erhålla sin arvsmassa loopas personlistan till dess att alla individer har fått sina kromosomer. I loopen får personer som inte har några kända föräldrar *bas-kromosomer* som bara utgörs av ett enda fragment medan personer med kända föräldrar får sina kromosomer genererade genom en meios-funktion för vardera förälder. Individer som har föräldrar vars kromosomer ännu inte är kända kommer att lämnas till nästkommande varv i loopen. De kromosomer som tillhör ”föräldralösa” personer ges ett ursprung representerat av personens siffra, tagen med positivt eller negativt tecken för att särskilja paternellt respektive maternellt DNA.

För att kunna beskriva meiosen används de två klasserna Chromosome och Fragment. Ett objekt av klassen Chromosome innehåller en lista i vilken kromosomens alla fragment sparas. Varje kromosom innehåller även det antal fragment den består av, genetisk längd, kromosomnummer och en pekare till dess första fragment. Överkorsningar vid meiosen gör att fragmenten delas och blir fler och det är därför viktigt att spara längd samt ursprung i varje Fragment-objekt. Dessa behövs för att senare kunna jämföra släktingar och utröna hur många samt hur långa gemensamma fragment de delar. Varje fragment pekar även till kromosomens nästkommande fragment.

Pedigree-klassen innehåller en meios-metod som används för att generera en individs kromosomer utifrån dess föräldrars DNA. Denna metod tar in två kromosomer, ett kromosompar från en förälder, och bildar en ny kromosom. Denna nya kromosom blir en del av barnets paternella eller maternella DNA beroende på vilken förälder kromosomparet kom ifrån. Inledningsvis genereras ett slumptal från Poisson-fördelningen utifrån det förväntade antalet överkorsningar (den genetiska längden) som kromosomparet innehåller. Det heltal som erhålls är antalet överkorsningar som kommmer att ske vid meiosen. Utifrån detta tal genereras slumptal från en likformig fördelning som beskriver var på kromosomerna överkorsningarna ska ske. Sedan skapas ett nytt kromosompar som är identiskt med kromosomparet som givits som indata så att inga förändringar blir gjorda i förälderns kromosomer. Dessa kopior traverseras och positionerna vid vilka fragmenten slutar sparas. Detta för att sedan skära kopiorna så att fragmenten på båda kopior slutar på samma positioner. Båda kopiorna skärs också på de positionerna där överkorsningarna kommer ske. Avslutningsvis traverseras kopiorna parallellt och en ny kromosom byggs upp genom att fragment tas från kopiorna och läggs in i denna, se figur 5. I varje steg under traverseringen används objekt för att peka på ett fragment i varje kopia. Den nya kromosomen får fragment som är identiskt med ett av dessa beroende på hur många överkorsningar som passerats.



Figur 5: Överkorsningsfasen, visualiserad sådan som den behandlas i Java-programmet. Den nedre kromosomen är resultatet från överkorsningarna mellan de två övre. De tjockare linjerna beskriver var överkorsningar sker vid meiosen och de vanliga linjerna representerar vart fragment på kromosomerna slutar. Notera att efter skärningen kommer fragmenten på kromosomparet sluta på samma positioner

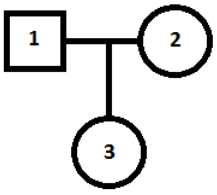
När simuleringsfasen är färdig finns all information som behövs för att jämföra två personers genetiska likhet. Detta utförs av en Comparison-klass. När ett objekt av denna klass skapas ges två personer som indata. Efter initieringen kommer objektet att innehålla information om hur mycket gemensamt DNA de två personerna har. Alltså sker själva jämförelsen i konstruktorn till Comparison-klassen. Detta görs med hjälp av en metod som jämför två kromosomer.

Det Comparison-objekt som initierats innehåller alltså till slut en andel som beskriver hur mycket gemensam arvsmassa de två individerna som ska jämföras har. Denna andel är beräknad utifrån den totala fysiska längden som de två individerna maximalt kan dela. En mans arvsmassa har kortare total fysisk längd än en kvinnas eftersom Y-kromosomen är kortare

än X-kromosomen. Detta kommer alltså att medföra att när en man och en kvinna jämförs beräknas andelen genom att dividera den totala gemensamma längden med totala längden på mannens arvsmassa.

2.1.1 Exempel på Java-programmets arbetsgång

För att beskriva programmet ännu lite tydligare bifogas nedan ett räkneexempel på hur ett pedigree behandlas. Exemplet är baserat på kronprinsessan Victoria, prins Daniel och den nyfödda tronarvingen prinsessan Estelle. Pedigree med tillhörande matris som representerar denna familj ses i figur 6. Prinsessan Estelle kommer självklart att ärva sitt paternella DNA från prins Daniel och sin maternella arvsmassa från kronprinsessan Victoria. Resultaten av jämförelser mellan individernas DNA är alltså i detta fall lätta att förutspå. Detta mycket enkla exempel bifogas alltså endast för att beskriva Java-programmets arbetsgång.



	Familj	Person	Far	Mor	Kön
Daniel	1	1	0	0	1
Victoria	1	2	0	0	2
Estelle	1	3	1	2	2

Figur 6: Pedigree med tillhörande matris som representerar prins Daniel, kronprinsessan Victoria och prinsessan Estelle

Inledningsvis läses matrisen in av programmet. Ett objekt av klassen Pedigree skapas och tre Person-objekt initieras och tillskrivs de värden som beskriver familjens nummer, personnummer, föräldrar och kön. Objektet som representerar prinsessan Estelle har alltså föräldrarna prins Daniel och kronprinsessan Victoria. Dessa har i sin tur föräldrar som inte ingår i trädet. Nu finns alltså de Person-objekt som kommer att behövas när arvsprocessen ska simuleras.

I detta skede genomlöps personlistan som finns i Pedigree-objektet. Prins Daniel behandlas först eftersom han i detta fall representeras med den lägsta siffran. Detta objekt har varken fått maternellt eller paternellt DNA och dessutom saknar det föräldrar. Därav genereras 46 bas-kromosomer med ursprung 1 eller -1 för paternellt respektive maternellt DNA och Person-objektet som representerar prins Daniel tillskrivs dessa kromosomer. Samma procedur följer för objektet som representerar kronprinsessan Victoria. Den enda skillnaden är att detta objekt får paternellt och maternellt DNA med ursprung 2 respektive -2. Sedan behandlas objektet som representerar prinsessan Estelle. Hon har föräldrar som ingår i trädet och alltså används meios-metoden för att generera hennes kromosomer. Först tillskrivs objektet sitt paternella DNA genom att kromosomerna som tillhör objektet som representerar prins Daniel genomgår meios. Sedan följer samma process för det maternella DNA:t som alltså genereras utifrån kronprinsessan Victorias kromosomer. Objektet som representerar

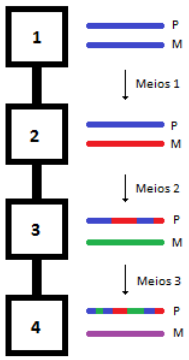
prinsessan Estelle har alltså paternella kromosomer uppbyggda av fragment med ursprung 1 eller -1 och maternella kromosomer uppbyggda av fragment med ursprung 2 eller -2.

Efter denna process har alla objekt fått sina kromosomer och utifrån denna arvsmassa kan jämförelser mellan personerna genomföras. Kronprinsessan Victoria och prins Daniel kommer ej att ha några gemensamma fragment medan prinsessan Estelle kommer att dela hälften av DNA:t med sin mor och hälften med sin far.

2.2 Teoretiskt resonemang kring arvsförloppet

Den grundläggande idén för de mer teoretiska beräkningarna är att översätta arvsförloppet till en process med ett antal tillstånd. Processen går från ett tillstånd till ett annat med ett antal övergångssannolikheter. Observera att i de teroretiska beräkningarna inkluderas inte könskromosomerna.

Först och främst behandlas arv i ett rakt nedstigande led, till exempel från en morförälder till ett barnbarn. Resonemanget som presenteras håller dock även när individer med andra relationer ska jämföras, till exempel syskon eller kusiner. Anledningen till att resonemanget presenteras för rakt nedstigande led är att detta fallet är det mest intuitiva. Endast ett kromosompar betraktas. Tankegången illustreras med ett exempel, fallet då en man ska ärva en viss längd från ett kromosompar i sin farfars fars DNA, se figur 7.



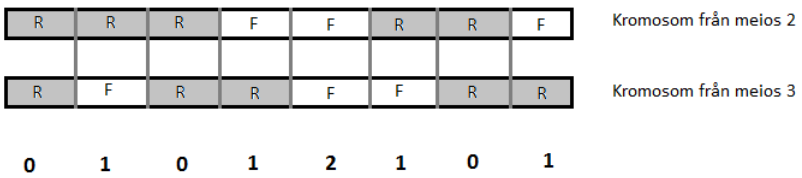
Figur 7: Arv i ett rakt nedstigande paternellt led. I detta förenklade pedigree som utelämnar mödrar har de blå fragmenten ärvts av person 1

Vid varje mans meios, som betecknas med pilar i figur 7, sker som bekant överkorsningar som gör att den nedärvda paternella kromosomen byggs upp av blå fragment som ärvts av person 1 och fragment från ingifta kvinnors arvsmassa. Då ett fragment förs vidare i en viss meios i detta rakt nedstigande paternella led betecknas detta som ”rätt” och då fragment från en ingift ärvs betecknas detta alltså som ”fel”. Att händelserna betecknas som rätt och fel är i avseendet att om arvet går rätt till på en position finns möjligheten att ett fragment som

ligger på samma position i den slutgiltiga kromosomen (hos person 4) kan komma från person 1. Observera alltså att ett fragment som betecknas med rätt inte nödvändigtvis behöver vara blått utan endast måste komma från faderns paternella kromosom för att möjligheten ska finnas att det är blått och kommer från person 1. Eftersom DNA från person 1 endast kommer kunna finnas i den paternella arvsmassan hos person 4 betraktas endast rätt och fel på de kromosomer som producerats i fädernas meioser.

Person 2 kommer självklart att ärva sin faders DNA. Detta första steg, meios 1, är alltså deterministiskt och alltid rätt. För att ett blått fragment i den paternella arvsmassan sedan skall ärvas ända ned till person 4 krävs två saker; i meios 2 måste överkorsningarna lägga sig så att det blå fragmentet förs vidare och samma sak måste gälla i meios 3. Med de införda beteckningarna krävs det alltså att det går rätt till på platsen där det blå fragmentet ligger i båda de ovan nämnda leden för att fragmentet skall ärvas.

Alltså måste det på ett visst intervall vara rätt i båda meioserna samtidigt för att ett fragment ska kunna ärvas hela vägen ned till person 4. Därför betraktas de båda producerade kromosomerna från meios 2 och 3 simultant, se figur 8. Intervall där paternellt DNA ärvs betecknas med R för rätt och resterande med F för fel. Både de överkorsningar som skett i meios 2 och de som skett i meios 3 ritas ut på de båda kromosomerna.

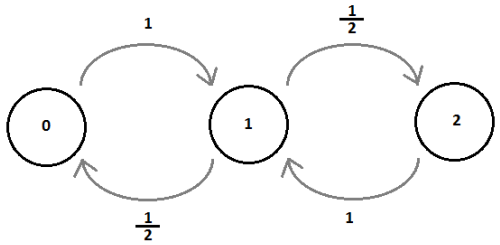


Figur 8: Här betraktas de producerade kromosomerna från meios 2 respektive 3 simultant. På intervallen mellan överkorsningarna markeras huruvida detta fragment ärvs från det paternella ledet eller inte med R respektive F. Nedanför kromosomerna markeras antal fel i vardera intervall

Genomlöpandet av den slutgiltiga kromosomen från vänster till höger definieras nu som en process. Avsnitten mellan överkorsningarna definierar intervall I. Totala antalet fel under ett intervall I är tillstånd i processen. För exemplet blir tillstånden alltså 0, 1 och 2. Sannolikheten att processen börjar i ett visst tillstånd är binomialfördelat med sannolikheten $\frac{1}{2}$. Detta eftersom sannolikheten att en kromosom börjar med ett fragment ärvt från rätt respektive fel person är $\frac{1}{2}$.

Efter varje överkorsning hamnar processen i ett annat tillstånd eftersom en av kromosomerna då byter vilken person den ärver från. På grund av antagandet att två överkorsningar ej kan ske på samma position kan bara en av kromosomerna skifta från att vara rätt till att vara fel eller tvärtom vid en överkorsning. Det betyder att tillstånden endast kan ändras med ett steg i taget, vilket framgår av figur 8. Övergångssannolikheterna mellan de olika

tillstånden beror av antalet inblandade meioser och vilket tillstånd processen befinner sig i. Processen för exemplet kan då illustreras som en Markov-kedja, se figur 9.



Figur 9: Genomlöpandet av den producerade kromosomen från meios 3 illustreras som en Markov-kedja

I detta exempel är övergångarna från tillstånd 0 och 2 triviala eftersom processen endast kan gå till tillstånd 1. I tillstånd 1 ärver den ena kromosomen fel och den andra rätt. Sannolikheten att gå från tillstånd 1 till 0 är alltså densamma som sannolikheten att nästa överkorsning tillhör den kromosomen som ärver fel. Kromosomen byter då till att ärva rätt i nästa intervall. Eftersom överkorsningarna antas vara likformigt utlagda längs kromosomerna är denna sannolikhet $\frac{1}{2}$. Samma resonemang gäller för övergångssannolikheten mellan tillstånd 1 och 2.

Det krävs som tidigare nämnt att de båda producerade kromosomerna från meios 2 och meios 3 ärver rätt samtidigt för att person 4 skall få arvsmassa från person 1. Alltså ärvs DNA av person 1 då processen befinner sig i tillståndet 0. Därav introduceras en stokastisk variabel B_0 vilken beskriver antalet besök i tillstånd 0 som processen gör. I exemplet är $B_0 = 3$ som framgår av figur 8.

Den producerade kromosomen från meios 3 betraktas nu. På fragmenten noteras endast vilket tillstånd processen är i under detta intervall, se figur 10.

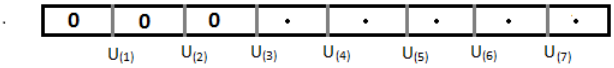
Figur 10: Den producerade kromosomen från meios 3. Mellan överkorsningarna betecknas hur många fel som skett på intervallet

Eftersom överkorsningarna antas vara likformigt fördelade över kromosomen är även fragmentlängderna likformigt fördelade. Därför har den totala längden av gemensamma fragment

med person 1, L, exakt samma *fördelning* även om tillstånden ligger i en annan ordning än tidigare. Alltså kan besöken i tillstånd 0 placeras först på kromosomen utan att detta påverkar fördelningen, se figur 11.

Figur 11: Den producerade kromosomen från meios 3, med processens besök i tillstånd 0 placerade först

Positionerna där överkorsningarna sker betraktas nu som värden $U_1, ..., U_n$ av en stokastisk variabel U, se figur 12. När dessa ordnas efter ökande storlek erhålls den så kallade *order-statistikan* för värdena [11]. Order-statistikan betecknas $U_{(1)}, ..., U_{(n)}$.



Figur 12: Den producerade kromosomen från meios 3, med order-statistika för överkorsningarna

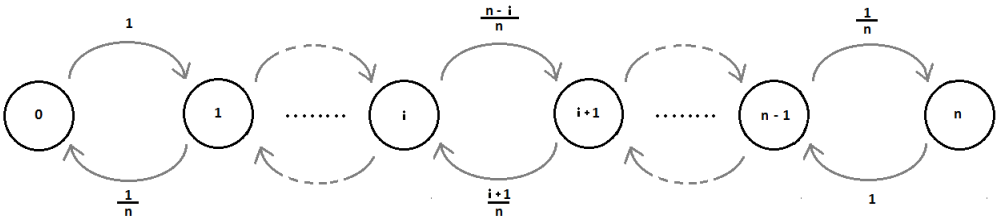
I och med detta fastslås att den totala längden av gemensamma fragment är exakt lika fördelad som order-statistikan med index B_0. Order-statistikan har en distribution som även beror på antalet överkorsningar, därför används notationen $U^N_{(B_0)}$ där N är totala antalet överkorsningar. Fördelningen för denna är känd och presenteras senare.

I exemplet är längden fördelad som $U^7_{(3)}$, eftersom $B_0 = 3$ och $N = 7$. Då har vi alltså för detta fall funnit fördelningen för den totala gemensamma längden. Notera att fördelningen hittades givet att totala antalet överkorsningar och antal besök i tillståndet noll var kända.

Hittills har resonemanget utgått ifrån ett exempel. För att kunna komma till ett allmänt fall måste resonemanget generaliseras.

I det allmänna fallet eftersöks fördelningen av den totala gemensamma längden DNA som en individ delar med en anfader. Antalet meioser som inte är deterministiska betecknas med n. Som för exemplet tidigare betraktas alla n meioserna simultant. Detta görs alltså analogt med figur 8, men för n meioser. För att individen ska dela ett avsnitt arvsmassa med anfadern krävs det att alla meioserna ärver från rätt person under ett intervall, alltså inte från en ingift person. En analog Markov-kedja definieras med samma tillstånd som ovan, alltså antalet fel

under ett intervall. Precis som ovan kan processen endast ta ett steg i taget, eftersom det inte kan ske två överkorsningar på exakt samma position. Övergångssannolikheterna beror på antalet meioser och vilket tillstånd som processen befinner sig i. Den allmänna Markov-kedjan med övergångssannolikheter visas i figur 13.



Figur 13: Markov-kedja som beskriver den allmänna processen vid n meioser

Det totala antalet överkorsningar, N, är i det allmänna fallet förstås okänt. Detta gäller även för antalet besök som processen gör i tillstånd 0, B_0. Eftersom B_0:s fördelning beror på N används betingade sannolikheter av formen $P(B_0 = j|N = i)$. Alla möjliga fall summeras och viktas med hur sannolika de är. Maximala antalet besök i tillstånd 0 blir $\frac{N+1}{2}$ avrundat uppåt till närmaste heltal, alltså $\lceil \frac{N+1}{2} \rceil$, eftersom processen alltid tar precis ett steg för varje överkorsning.

I exemplet tidigare visades att för ett bestämt totalt antal överkorsningar och ett bestämt antal besök i tillstånd 0 är den totala gemensamma längden fördelad som $U^N_{(B_0)}$. Fördelningen för den totala gemensamma längden, L, när värdena på B_0 och N är okända är då:

$$P(L \leq x) = \sum_{i=0}^{\infty} P(N = i) \sum_{j=0}^{\lceil \frac{i+1}{2} \rceil} P(B_0 = j|N = i) U^i_{(j)}(x) \tag{1}$$

Detta erhålls genom att summera alla möjliga fall och vikta dem med hur sannolika de är.

Det finns alltså ett slutet uttryck för fördelningen av den totala gemensamma längden arvsmassa som två individer har, L. Totala antalet överkorsningar är Poisson-fördelat med summan av intensiteterna hos alla meioser och $P(N = i)$ är således känd. Detta gäller inte $P(B_0 = j|N = i)$, ty denna måste approximeras med en empirisk fördelning. Orderstatistikan $U^i_{(j)}(x)$ är Beta-fördelad [11] med parametrarna j och $n - j + 1$ och har alltså täthetsfunktionen:

$$f_{U^i_{(j)}}(x) = \frac{i!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j} \tag{2}$$

Ur (1) kan då sannolikheten att ärva längden x eller mindre på en kromosom beräknas för alla $x \in (0, 1)$, där x representerar en andel av kromosomens längd.

Sannolikheten att ärva ingenting, $P(L = 0)$, på en kromosom kan i vissa fall beräknas exakt, till exempel då en individ och dess barnbarn ska jämföras. Mellan de två släktingarna finns

två meioser, en meios när första individen får barn och en meios när detta barn i sin tur får barn. Den första meiosen innehåller ingen slumpmässighet eftersom första individens barn kommer ärva precis hälften av förälderns arvsmassa. Längden som barnbarnet ärver från den första individen beror alltså endast på den andra meiosen. För att inget fragment ska ärvas ned till barnbarnet krävs att den producerade kromosomen börjar med att ärva från "fel" person och att inga överkorsningar sker. I detta fall är sannolikheten att ärva längden 0 alltså $\frac{1}{2}f_P(0)$, där f_P är täthetsfunktionen för Poisson-fördelningen med intensiteten lika med den genetiska längden hos kromosomen.

Även i fallet då en individ och dess barnbarns-barn ska jämföras kan sannolikheten att de inte har några gemensamma DNA-fragment beräknas exakt. I detta fall är två meioser av intresse. För att de två individerna inte ska ha någon gemensam arvsmassa krävs att den tidigare nämnda Markov-kedjan som beskriver genomlöpandet av den resulterande kromosomen hos barnbarns-barnet inte får några besök i tillstånd 0. Utseendet på den process som kedjan beskriver kommer då att bero på huruvida totala antalet överkorsningar N är jämnt eller udda. När N är jämn kommer processen kunna ha utseendet 1 2 1 ... 1 eller 2 1 2 ... 2. När N är udda kommer processen kunna ha utseendet 1 2 1 ... 2 eller 2 1 2 ... 1. Som nämnt tidigare är sannolikheten för processens första tillstånd binomialfördelad med faktorn $\frac{1}{2}$. Sannolikheten att börja i tillstånd 1 är alltså $\frac{1}{2}$ och sannolikheten att börja i tillstånd 2 är $\frac{1}{4}$. Processen går från tillstånd 1 till tillstånd 2 med sannolikheten $\frac{1}{2}$ och från 2 till 1 med sannolikheten 1.

När N är jämn kommer alltså sannolikheten för att processen ska anta formen 1 2 1 ... 1 att vara:

$$\frac{1}{2} \cdot \underbrace{\frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot ... \cdot 1}_{N \text{ stycken}} = \left(\frac{1}{2}\right)^{\frac{N+2}{2}}$$

Sannolikheten att processen antar formen 2 1 2 ... 2 kommer att vara:

$$\frac{1}{4} \cdot \underbrace{1 \cdot \frac{1}{2} \cdot 1 \cdot ... \cdot \frac{1}{2}}_{N \text{ stycken}} = \left(\frac{1}{2}\right)^{\frac{N+4}{2}}$$

När N är udda kommer sannolikheten för 1 2 1 ... 2 att vara:

$$\frac{1}{2} \cdot \underbrace{\frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot ... \cdot \frac{1}{2}}_{N \text{ stycken}} = \left(\frac{1}{2}\right)^{\frac{N+3}{2}}$$

Sannolikheten att processen får formen 2 1 2 ... 1 kommer att vara:

$$\frac{1}{4} \cdot \underbrace{1 \cdot \frac{1}{2} \cdot 1 \cdot ... \cdot 1}_{N \text{ stycken}} = \left(\frac{1}{2}\right)^{\frac{N+3}{2}}$$

Om alla dessa sannolikheter summeras för alla möjliga totala antal överkorsningar, N, erhålls sannolikheten att barnbarns-barnet inte har något gemensamt fragment med den första personen.

$$P(L=0) = \sum_{i \text{ jämn}} f_P(i) \cdot \left(\left(\frac{1}{2}\right)^{\frac{i+2}{2}} + \left(\frac{1}{2}\right)^{\frac{i+4}{2}}\right) + \sum_{i \text{ udda}} f_P(i) \cdot \left(2 \cdot \left(\frac{1}{2}\right)^{\frac{i+3}{2}}\right) \tag{3}$$

där $f_P(i)$ är sannolikheten att få i överkorsningar från Poisson-fördelningen då intensiteten är lika med den totala genetiska längden hos de två kromosomerna i de två meioserna. Med hjälp av Poisson-fördelningens täthetsfunktion och serieutveckling av exponentialfunktionen kan (3) förenklas till följande uttryck

$$P(L=0) = \frac{3}{4}e^{-\lambda} \cdot \frac{e^{\lambda\sqrt{\frac{1}{2}}} + e^{-\lambda\sqrt{\frac{1}{2}}}}{2} + \sqrt{\frac{1}{2}}e^{-\lambda} \cdot \frac{e^{\lambda\sqrt{\frac{1}{2}}} - e^{-\lambda\sqrt{\frac{1}{2}}}}{2} \tag{4}$$

där λ är Poisson-fördelningens intensitet.

2.2.1 Tillämpning av det teoretiska resonemanget i MATLAB

Det härledda uttrycket (1) för fördelningen av den totala gemensamma längden, L, har nyttjats i MATLAB. Målet var att för vissa relationer beräkna $P(L \leq x)$ för olika värden på x. Poisson- och Beta-fördelningen finns att tillgå i MATLAB, dock krävdes att den empiriska distributionen av B_0 givet N uppskattades. Detta gjordes med hjälp av den allmänna Markov-kedjan som presenterades i figur 13. Ett "spel" som efterliknar Markov-kedjan konstruerades. Spelet slumpar först fram ett initialtillstånd med hjälp av binomialfördelningen. Sedan slumpas likformiga tal som används för att låta processen vandra mellan tillstånden.

Eftersom övergångssannolikheterna beror av antalet inblandade meioser krävs att den empiriska fördelningen uppskattas en gång för varje antal meioser. Upp till fyra meioser behandlades. Det antogs vidare att det i varje meios kunde ske maximalt åtta överkorsningar, eftersom sannolikheten för att det ska bli fler är försvinnande liten för den genetiska längden som användes. Alltså uppskattades $P(B_0 = j|N = i)$ för upp till åtta överkorsningar när en meios behandlades, upp till 16 överkorsningar när två meioser behandlades och så vidare.

För varje antal meioser söktes alltså sannolikheten för att få j besök i tillstånd 0 givet ett totalt antal överkorsningar i. Det gjordes genom att köra spelet upprepade gånger för det antalet överkorsningar och notera hur stor del av gångerna som processen gjorde j besök i tillstånd 0. Denna procedur upprepades för alla i. Algoritmen för uppskattningen av den empiriska distributionen sammanfattas nedan.

- Ett antal meioser, n, bestäms.
- För varje n kan upp till $i = 8n$ överkorsningar ske. För varje i körs spelet upprepade gånger och för varje iteration noteras hur många besök processen gör i tillstånd 0.
- Antalet gånger som processen gjort ett visst antal besök i tillstånd 0 divideras med antalet gånger som spelet kördes för att få en uppskattning på $P(B_0 = j|N = i)$.

När den empiriska distributionen $P(B_0 = j|N = i)$ uppskattats finns alla verktyg som behövs för att beräkna $P(L \leq x)$ för olika relationer och för olika värden på x. Beräkningarna gjordes utifrån (1) med vissa specialfall, då N respektive B_0 var noll. Genom att beräkna $P(L \leq x)$ för olika värden på x erhölls resultat som senare kommer att presenteras och jämföras med de resultat som det producerade programmet i Java gav. De resultat som jämförs är gjorda för endast ett kromosompar, med genetisk längd 1 Morgan.

När två individer skiljs åt av en eller två icke-deterministiska meioser kan sannolikheten att de inte delar någon arvsmassa beräknas förhållandevis enkelt, som nämnt i slutet på avsnitt 2.2. Även detta gjordes med MATLAB. De resulterande värdena för $P(L = 0)$ användes som referensvärden till de värden på samma sannolikhet som beräknades med Java-programmet och MATLAB-rutinen.

3 Resultat

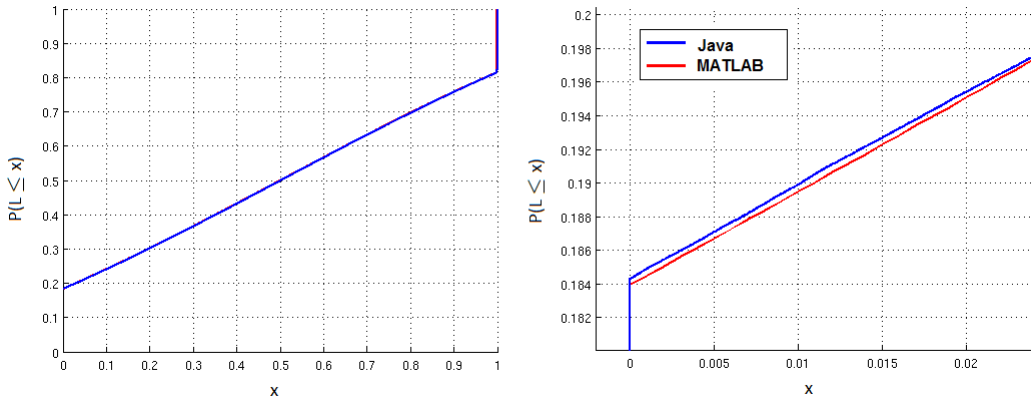
Resultatavsnittet avser att belysa hur olika släktskap påverkar genetisk likhet mellan individer. Här presenteras även resultat över hur inavel påverkar den genetiska likheten för olika relationer. Inledningsvis behandlas arvets inverkan på en kromosom samt den avtagande genomsnittliga genetiska likheten i ett rakt nedstigande släktled.

3.1 Arvsprocessens inverkan på en kromosom

Hittills har två olika sätt att jämföra två personers DNA beskrivits. Det första består av Java-programmet som givet ett pedigree simulerar arvsprocessen för att därefter jämföra två personers arvsmassa. För att kontrollera programmet utfördes ett mer teoretisk resonemang som tillämpades i MATLAB.

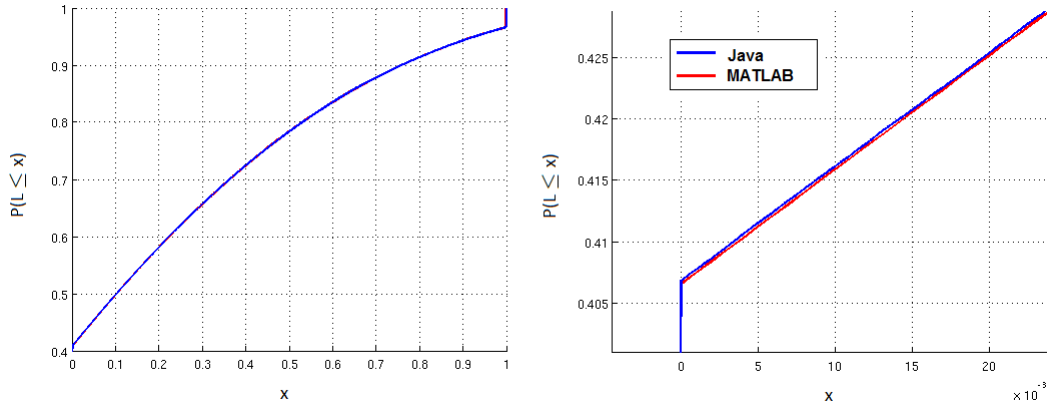
Resultaten från de två metoderna jämfördes genom att sannolikheten att ärva en viss längd från ett kromosompar givet ett antal meioser ritades ut i en gemensam bild. I detta fall var kromosomens genetiska längd 1 Morgan.

Sannolikheten att ett barnbarn ärver något från en far- eller morförälder undersöktes först. Observera att i detta fall är antalet intressanta meioser inte två utan bara ett. Med intressanta avses här icke-deterministiska meioser. När denna sannolikhet ritades ut hamnade kurvorna ovanpå varandra vilket indikerar att Java- och MATLAB-programmen ger konsekventa resultat, se figur 14.

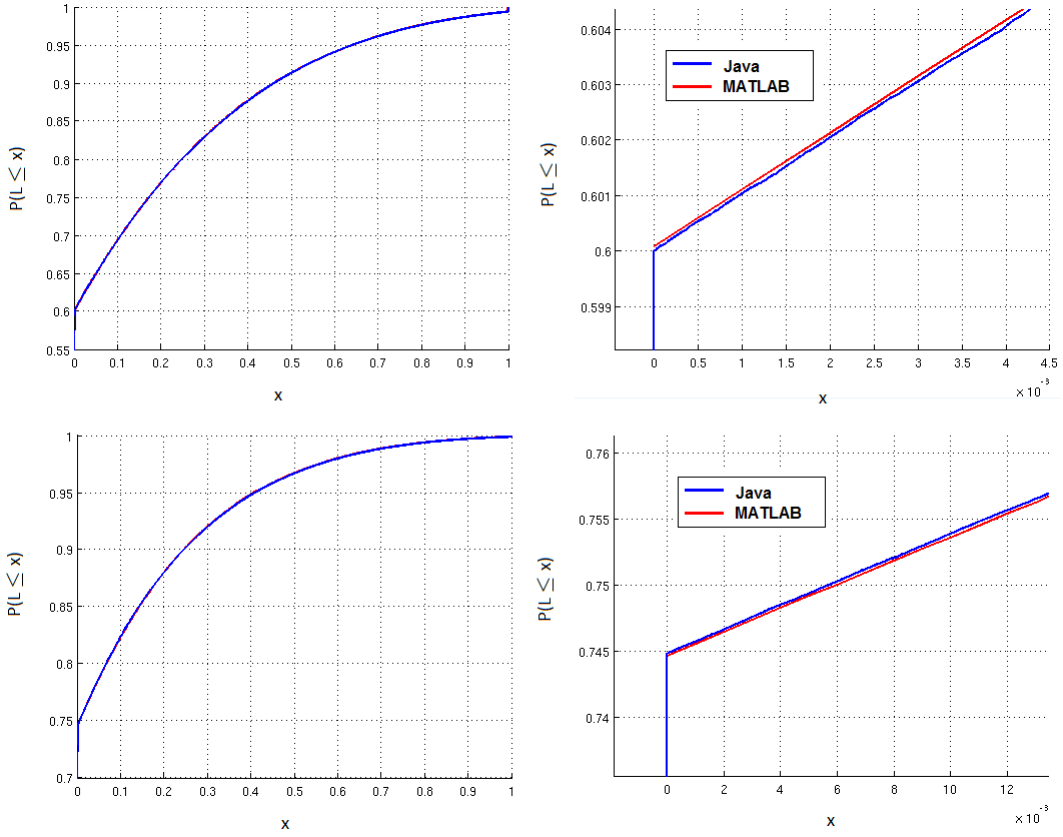


Figur 14: Sannolikheten att ärva en viss kromosomlängd givet en icke-deterministisk meios. Av höger bild framgår att sannolikheten att inte ärva något är ungefär 0,184

Då antalet intressanta meioser ökas från ett till två studeras fallet då ett barnbarns-barn ska ärva något. Även för detta fall hamnar kurvorna ovanpå varandra, se figur 15. Detta gäller även då antalet meioser ökas, se figur 16.



Figur 15: Sannolikheten att ärva en viss kromosomlängd givet två meioser. Av höger bild framgår att sannolikheten att inte ärva något är ungefär 0,406

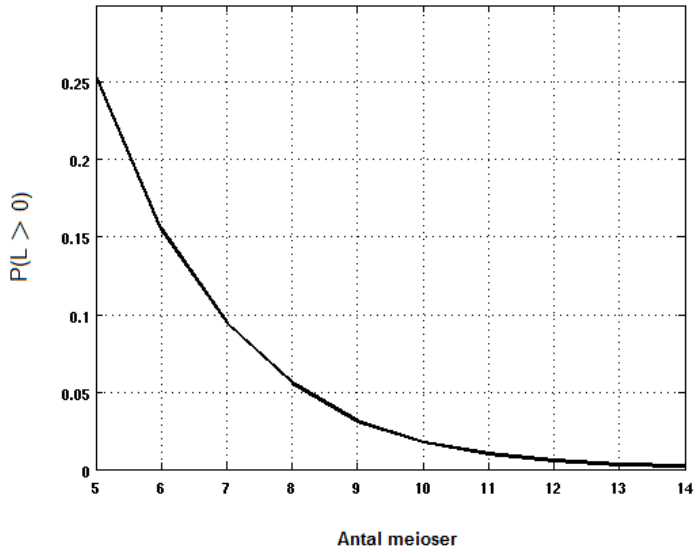


Figur 16: Resultaten av Java- och MATLAB-programmen då tre och fyra meioser studerades. De högra bilderna visar en förstoring vid sannolikheten att ärva längden 0

Från tidigare beräkningar (avsnitt 2.2) är sannolikheten att inte ärva något känd. Detta gäller dock endast vid en eller två meioser. Sannolikheterna är då 0,1839 respektive 0,4063. Detta är konsekvent med de resultat som visas i figur 14 och 15.

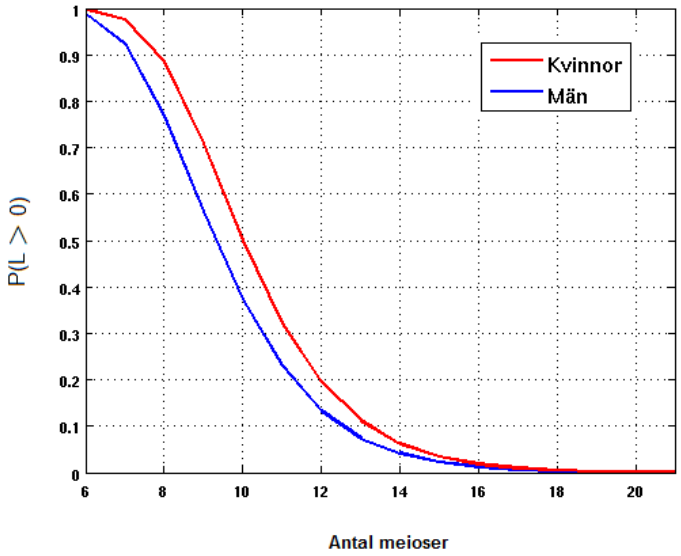
3.2 Möjligt genetiskt arv efter ett stort antal meioser

Sannolikheten att två besläktade individer har någon gemensam arvsmassa då de skiljs åt av ett stort antal meioser undersöktes med hjälp av Java-programmet. Först behandlades endast en kromosom, alltså beräknades sannolikheten att två individer delar någonting på en viss kromosom. Denna kromosom gavs en genetisk längd på 1 Morgan, resultatet visas i figur 17. Sannolikheten att de delar något på kromosomen avtar relativt snabbt mot noll.



Figur 17: Sannolikheten att två individer som skiljs åt av ett antal meioser delar någon arvsmassa på en kromosom med den genetiska längden 1 Morgan

Samma sannolikhet som ovan beräknades också på två individers 22 första kromosomer. Alltså behandlades alla kromosomer utom könskromosomerna. Dock inverkar personernas kön i form av olika genetiska längder på kromosomerna. Fallen med endast män i ett rakt nedstigande led respektive endast kvinnor i ett rakt nedstigande led studerades i figur 18. I dessa fall avtar sannolikheten långsammare, vilket är väntat eftersom det finns fler möjligheter att dela någon arvsmassa när fler kromosomer betraktas. Kvinnors kromosomer har större genetisk längd vilket bidrar till att de delar genetisk information på ett längre avstånd än vad män gör.



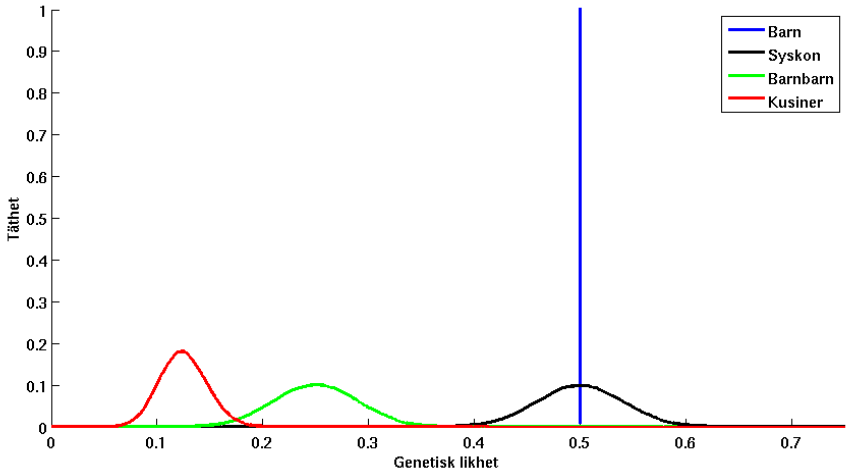
Figur 18: Sannolikheten att två män som skiljs åt av ett antal meioser i ett helt manligt släktled respektive två kvinnor i ett helt kvinnligt släktled delar någon arvsmassa på de 22 autosomerna

3.3 Genetisk likhet i vanliga släktskap

I detta avsnitt studeras den genetiska likheten mellan olika vanliga relationer. Relationerna som undersöks är mellan föräldrar och deras barn, syskon, barnbarn och far- eller morförälder samt kusiner. Fördelningarna för dessa relationers genetiska likhet har approximerats utifrån datasimuleringar med Java-programmet som beskrevs i avsnitt 2.1.

Varje relation simulerades som tidigare nämt 100 000 gånger. Simuleringsresultatet bestod således av andelarna gemensamt DNA vid varje iteration och datamängderna behandlades med MATLAB. Intervallet $[0, 1]$ delades upp i delintervall med längden 0,01. Utifrån varje simuleringsresultat beräknades antalet gånger som den gentiska likheten låg i ett visst delintervall. Detta blir alltså en approximation av täthetsfunktionen. Efter att ha jämnats ut med hjälp av en MATLAB-funktion ritades värdena ut.

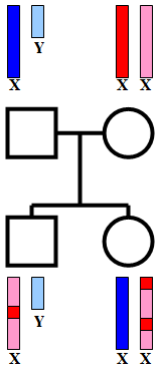
För att ge en första överblick visas fördelningarna för alla relationernas genetiska likheter i samma bild, se figur 19. I denna figur inkluderas inte könskromosomerna. I figuren framgår att graferna är mycket centrerade kring sina respektive medelvärden.



Figur 19: I figuren visas graferna som representerar relationerna mellan en person och sitt barn, syskon, barnbarn och kusin (utan könskromosomer)

3.3.1 Föräldrar och barn

Relationen mellan en förälder och ett barn belyser framförallt könskromosomernas inverkan på hur mycket gemensamt DNA två individer har. I övrigt bidrar inte denna relation med särskilt mycket information eftersom andelen gemensamt DNA är helt deterministisk. Den arvsmassa som en individ delar med sin mor består självklart alltid av de maternella kromosomerna och analogt delas de paternella kromosomerna med fadern. Om endast de autosomala kromosomerna behandlas, alltså om könskromosomerna utesluts, kommer varje individ att dela exakt hälften av vardera förälders arvsmassan. Det beror på att genomet då består av två stycken likvärdiga uppsättningar av 22 kromosomer.



Figur 20: Exempel på överkorsningsfördelning på könskromosomerna mellan föräldrar och deras barn. Observera att inga överkorsningar sker på faderns könskromosomer

Detta gäller inte då könskromosomerna tas med i beräkningen eftersom X- och Y-kromosomen har olika fysisk längd. Ett exempel på överkorsningsfördelning mellan könskromosomerna visas i figur 20. Y-kromosomen är betydligt kortare än X-kromosomen och en man kommer således att dela en mindre andel DNA med sin far (49,18%) än vad han delar med sin mor (50,82%). Detta eftersom uppsättningen paternella kromosomer är kortare än den maternella arvsmassan. En kvinna kommer att dela 50,82% med sin far och 50,00% med sin mor. Notera att en kvinna ärver exakt lika lång fysisk längd av vardera förälder och att procentsatserna skiljer sig endast på grund av hur andelarna beräknas, se avsnitt 2.1. Värdena för de genetiska likheterna samt standardavvikelse visas i tabell 2.

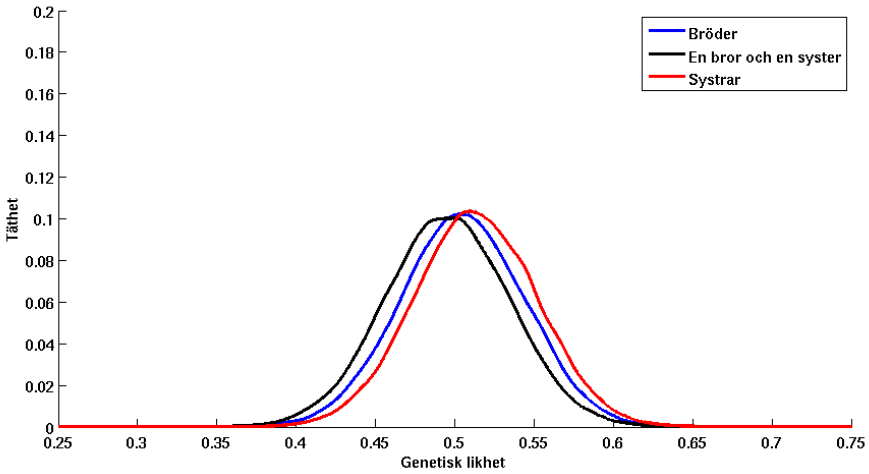
Tabell 2: Den genetiska likheten mellan föräldrar och barn

Väntevärden samt standardavvikelse för relationen mellan föräldrar och deras barn		
	Väntevärde	Standardavvikelse
Far och son	0,4918	$0{,}2816*10^{-12}$
Far och dotter	0,5082	$0{,}3010*10^{-12}$
Mor och son	0,5082	$0{,}3010*10^{-12}$
Mor och dotter	0,5000	$0{,}0001*10^{-12}$

3.3.2 Syskon

Till skillnad från relationen mellan föräldrar och barn är det inte självklart hur mycket DNA två syskon kommer att ha gemensamt. Detta beror istället på var överkorsningarna skedde i föräldrarnas meioser vid bildandet av de könsceller som kom att bli vardera syskon. Behandlas endast autosomerna kommer syskon att dela i genomsnitt 50% oavsett kön, med en standardavvikelse på ungefär 4%.

När könskromosomerna inkluderas kommer systrar att i genomsnitt ha mest gemensamt DNA, 51,30%. Bröder kommer att dela i genomsnitt 50,48% av arvsmassan och ett syskonpar av olika kön kommer att dela 49,52%. Systrar delar mest eftersom de har ärvt en identisk X-kromosom från sin far. Bröder har exakt samma Y-kromosom men eftersom den är kortare än X-kromosomen har den mindre inverkan på andelen gemensamt DNA. Eftersom inga överkorsningar sker mellan X- och Y-kromosomerna kan dessa inte innehålla någon gemensam arvsmassa i den använda modellen. Därför kommer en bror och en syster att ha i genomsnitt mindre än hälften av DNA:t gemensamt. De approximerade fördelningarna visas i figur 21. För standardavvikelsen av alla tre relationerna se tabell 3.



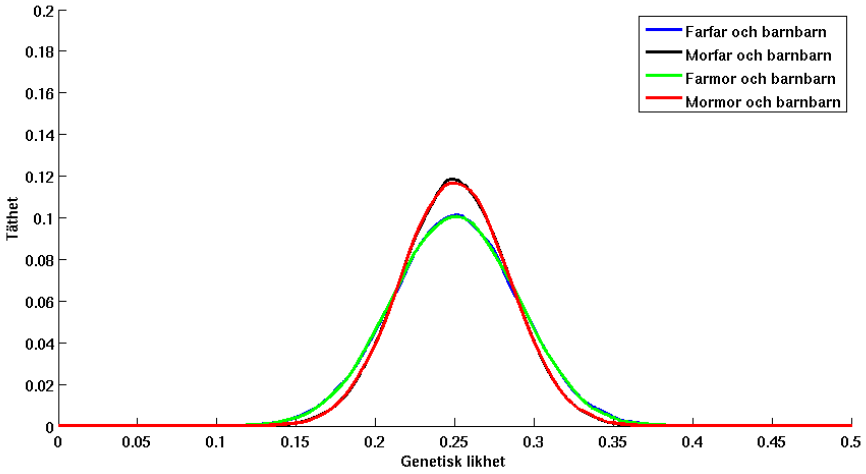
Figur 21: Fördelningen för den genetiska likheten mellan olika syskonpar, med könskromosomer

Tabell 3: Tabell över den genetiska likheten mellan syskon

Väntevärden samt standardavvikelse för relationen mellan syskon		
	Väntevärde	Standardavvikelse
Bröder	0,5048	0,0391
Bror och syster	0,4952	0,0392
Systrar	0,5130	0,0385

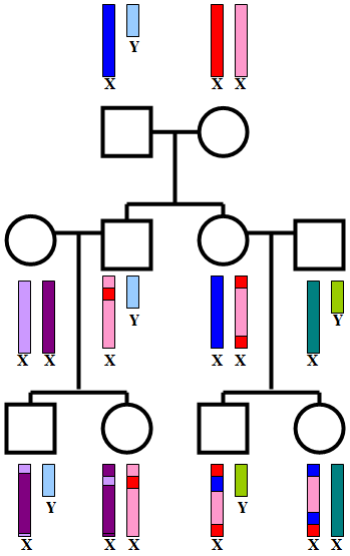
3.3.3 Far- respektive morföräldrar och barnbarn

Inledningsvis studeras de fyra relationerna då far- och morföräldrar jämförs med ett barnbarn utan könskromosomer. I dessa fall ligger den genomsnittliga genetiska likheten på 25%. Standardavvikelsen för morföräldrarna är 3,36% medan farföräldrarnas är 3,92%. Skillnaden på 0,56 procentenheter gör att morföräldrarnas kurvor blir mer centrerade kring medelvärdet i jämförelse med farföräldrarnas, se figur 22.

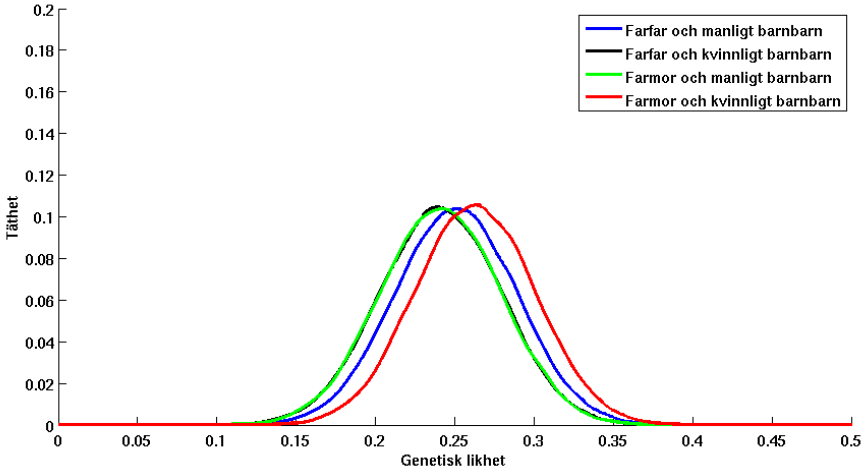


Figur 22: Fördelningen för genetisk likhet mellan far- och morföräldrar och deras barnbarn utan könskromosomer

Då könskromosomerna inkluderas studeras åtta olika relationer, vilket ger mer variation i den genetiska likheten. Hur könskromosomerna ärvs illustreras i figur 23. En farmor delar i genomsnitt lite mer än de tidigare 25% av sitt DNA med sina kvinnliga barnbarn. Genomsnittet är 26,27%, vilket är en följd av könskromosomernas inverkan. Ett kvinnligt barnbarn får en X-kromosom som är en blandning av farmors båda X-kromosomer och detta gör att hon delar lite extra DNA med sin farmor. Ett manligt barnbarn får däremot ingen könskromosom från sin farmor och delar därmed i genomsnitt 24,10% med henne. Det kvinnliga barnbarnet får inte heller någon könskromosom från sin farfar och delar därför i genomsnitt 24,10% med honom. Ett manligt barnbarn får sin farfars exakta Y-kromosom men eftersom denna är betydligt kortare än X-kromosomen kommer detta endast att bidra med att genomsnittet ökar från 25,00% till 25,06%. För farföräldrarnas standardavvikelser se tabell 4. Fördelningarna visas i figur 24.



Figur 23: Illustration av hur könskromosomerna ärvs från far- och morföräldrar till barnbarn

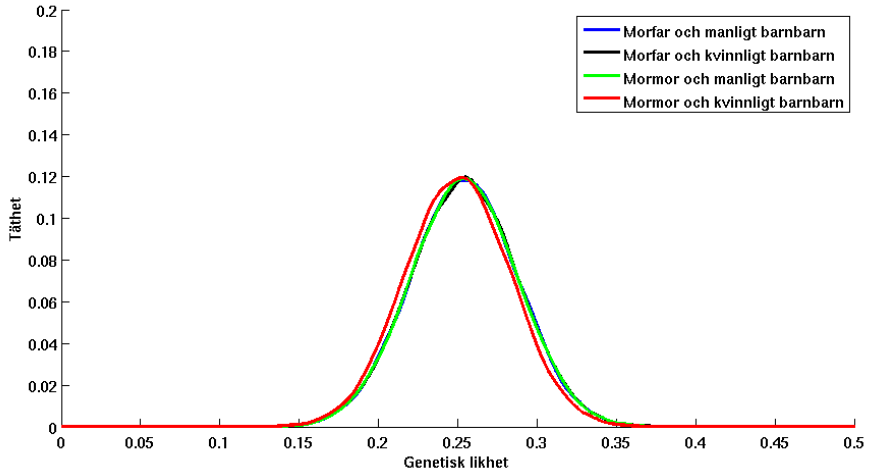


Figur 24: Fördelning för den genetiska likheten för farföräldrar och barnbarn när könskromosomer inkluderas

Tabell 4: Genetisk likhet för farförälder och barnbarn

Väntevärden samt standardavvikelse för relationen mellan farföräldrar och deras barnbarn		
	Väntevärde	Standardavvikelse
Farmor och manligt barnbarn	0,2410	0,0379
Farmor och kvinnligt barnbarn	0,2627	0,0373
Farfar och manligt barnbarn	0,2506	0,0380
Farfar och kvinnligt barnbarn	0,2410	0,0380

Till skillnad från farföräldrarna för morföräldrarna alltid vidare en liten del av sina könskromosomer. Mormor för vidare en blandning av sina två X-kromosomer till sin dotter medan morfar för vidare sin exakta X-kromosom. På så sätt kommer dottern alltid att skicka vidare en blandning av mormors och morfars X-kromosomer till sina barn. En mormor och hennes kvinnliga barnbarn delar i genomsnitt 24,99%. I de övriga fallen ligger den genetiska likheten mellan 25,40%-25,41% och standardavvikelsen för de fyra relationerna visas i tabell 5. De approximerade fördelningarna visas i figur 25.



Figur 25: Fördelning för den genetiska likheten för morföräldrar och barnbarn när könskromosomer inkluderas

Sammanfattningsvis då alla 23 kromosomerna betraktas är den genetiska likheten störst mellan en farmor och hennes kvinnliga barnbarn. Den genetiska likheten är minst då farmor och farfar inte för vidare sina könskromosomer till sina manliga respektive kvinnliga barnbarn.

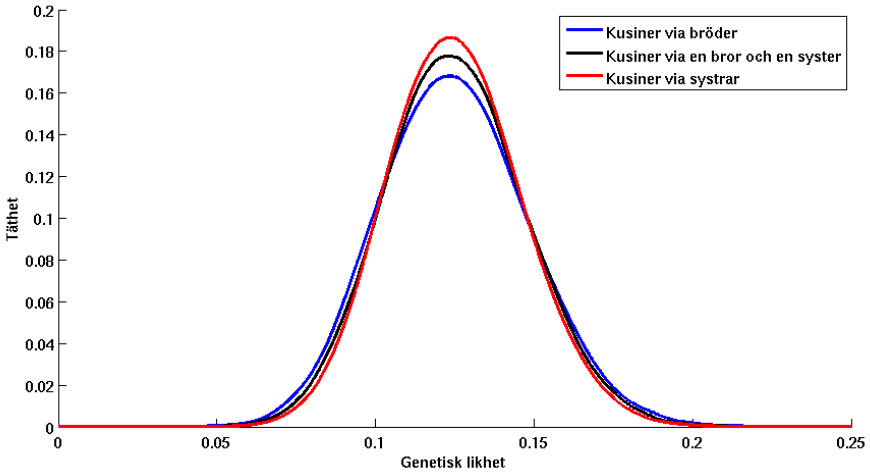
Tabell 5: Genetisk likhet för morförälder och barnbarn

Väntevärden samt standardavvikelse för relationen mellan morföräldrar och deras barnbarn		
	Väntevärde	Standardavvikelse
Mormor och manligt barnbarn	0,2540	0,0335
Mormor och kvinnligt barnbarn	0,2499	0,0330
Morfar och manligt barnbarn	0,2540	0,0334
Morfar och kvinnligt barnbarn	0,2541	0,0336

3.3.4 Kusiner

Kusiner kan vara släkt på hela tio olika sätt beroende på vilka kön kusinparet har samt hur deras föräldrar är syskon. De olika kusin-relationerna innebär olika stora förutsättningar för genetisk likhet.

Inledningsvis studeras kusinrelationer där alla inblandade individer antas sakna könskromosomer. Den genomsnittliga genetiska likheten mellan kusiner är under dessa förutsättningar 12,50% vilket kan ses i figur 26.

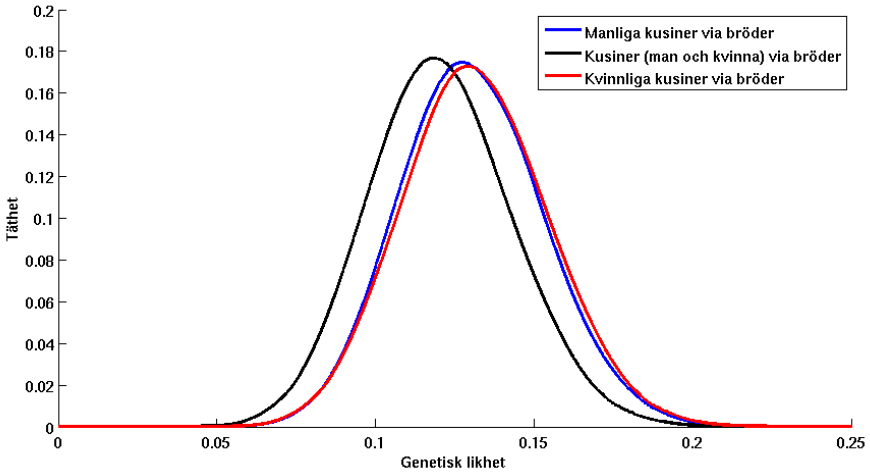


Figur 26: Fördelningen för den genetiska likheten mellan olika kusinpar utan könskromosomer

Om könskromosomernas inverkan tillgodoses i modellen blir de genomsnittliga likheterna olika och mer distinkta. Den största sannolikheten för genetisk likhet återfinns exempelvis i relationen mellan två kvinnliga kusiner vars fäder är bröder. Kvinnorna delar i genomsnitt 13,15% av sitt DNA. På motsvarande sätt är det minst genomsnittlig likhet mellan en man och kvinna där deras fäder är bröder eller där mannens far och kvinnans mor är syskon. De

delar i genomsnitt enbart 12,04% av sitt DNA. Skillnaden mellan kusinrelationernas genomsnittliga genetiska likhet är alltså hela 1,11 procentenheter.

Då kusinerna är släkt via två bröder kan det i figur 27 samt tabell 6 utrönas att utfallet mellan kvinnor och män är liten ty bröderna delar samma Y-kromosom, vilken då även de manliga kusinerna delar. I det kvinnliga fallet är brödernas X-kromosomer inte identiska men i genomsnitt 50% varför det genomsnittet också gäller för de kvinnliga kusinerna. Detta förklarar således den lägre genetiska likheten mellan en manlig och en kvinnlig kusin eftersom de fått brödernas Y- respektive X-kromosom.

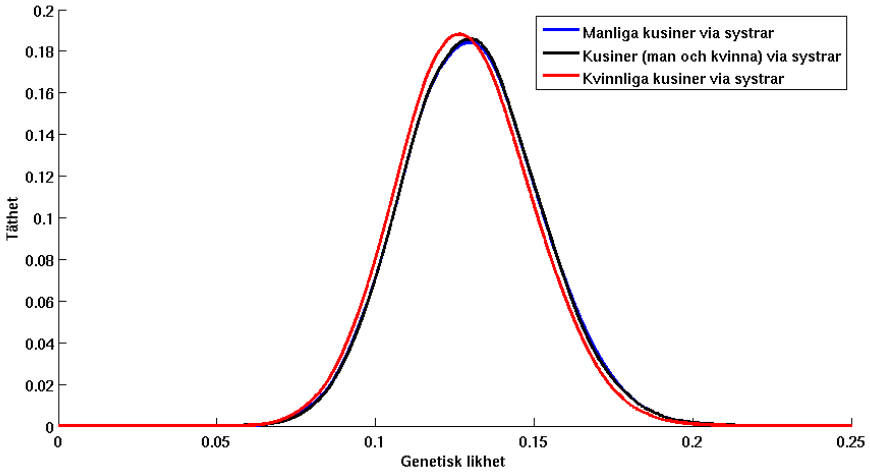


Figur 27: Fördelningen för den genetiska likheten för kusinpar via två bröder

Tabell 6: Genetisk likhet för kusinpar via två bröder

Väntevärden samt standardavvikelse för kusiner via bröder		
	Väntevärde	Standardavvikelse
Manliga kusiner	0,1303	0,0225
Kusiner (man-kvinna)	0,1204	0,0226
Kvinnliga kusiner	0,1315	0,0230

I fallet då kusiner är släkt via systrar är skillnaden mellan de olika kusinrelationerna diffusare, se figur 28 samt tabell 7. Män delar störst andel genetisk information eftersom X-kromosomen ger större bidrag i förhållande till deras olika Y-kromosomer. Detsamma gäller för kusiner där den ena är man och den andra kvinna. Detta eftersom de maximalt kan dela de 22 första kromosomerna samt en X-kromosom. För kvinnliga kusiner ger den X-kromosom som kommer från någon av systrarna endast halva bidraget och därav lägre genetisk likhet jämfört med de andra kusinparen.

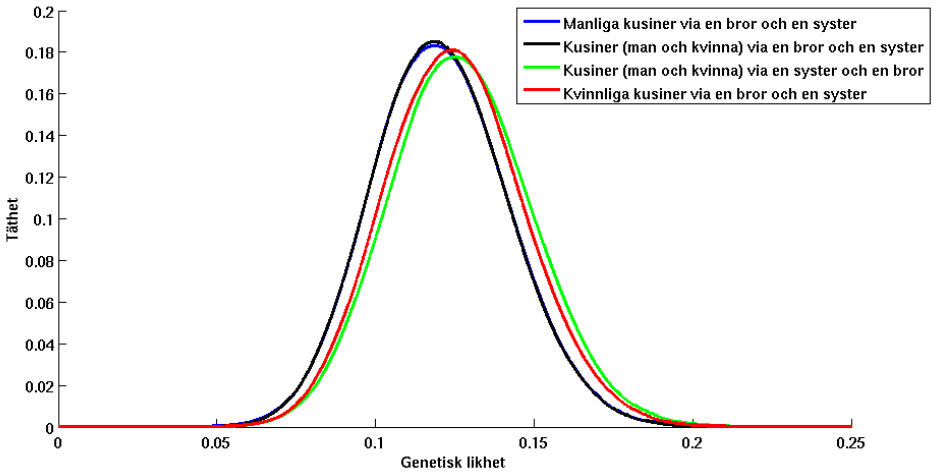


Figur 28: Fördelningen för den genetiska likheten för kusinpar via två systrar

Tabell 7: Genetisk likhet för kusinpar via två systrar

Väntevärden samt standardavvikelse för kusiner via systrar		
	Väntevärde	Standardavvikelse
Manliga kusiner	0,1303	0,0213
Kusiner (man-kvinna)	0,1303	0,0211
Kvinnliga kusiner	0,1282	0,0209

Då kusinrelationen uppstår via ett syskonpar bestående av en man och kvinna blir den genetiska likheten något lägre jämfört med övriga fall, se figur 29 samt tabell 8. Största likheten återfinns mellan en man och en kvinna där mannens mor är syster till kvinnans far.



Figur 29: Fördelningen för den genetiska likheten för kusinpar via en bror och en syster

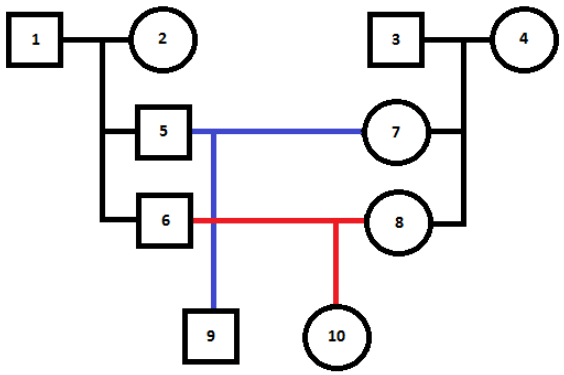
Tabell 8: Genetisk likhet för kusinpar via en bror och en syster

Väntevärden samt standardavvikelse för kusiner via en bror och en syster		
	Väntevärde	Standardavvikelse
Manliga kusiner	0,1205	0,0213
Kusiner (man-kvinna)	0,1204	0,0212
Kusiner (kvinna-man)	0,1270	0,0221
Kvinnliga kusiner	0,1251	0,0217

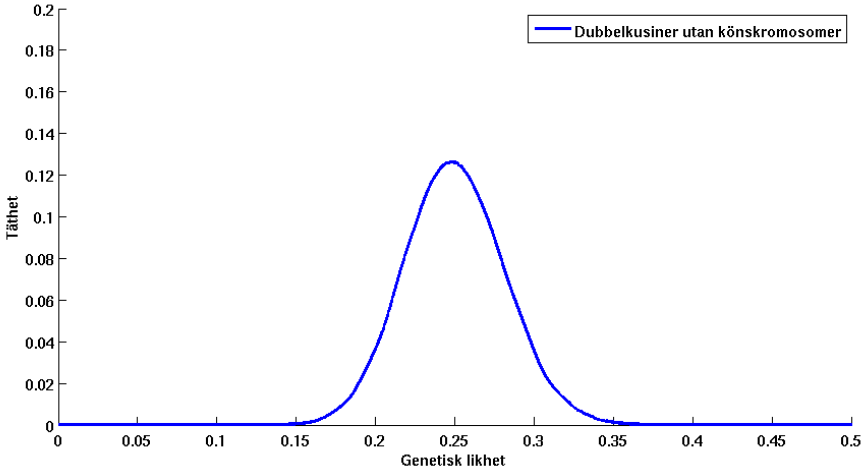
Under arbetets gång undersöktes huruvida kusiner kan vara mer lika än syskon. Detta visade sig dock vara ytterst osannolikt, vilket konstaterades genom att med hjälp av Java-programmet beräkna hur ofta kusiner blev mer genetiskt lika än syskon. På 10 miljoner simuleringar blev inte kusiner mer lika än syskon en enda gång, något som tydligast ses i figur 19 där det inte sker någon överlappning mellan syskon- och kusin-kurvorna.

Även dubbelkusiners genetiska likhet jämfördes med syskon. Dubbelkusiner är kusiner vars föräldrar är syskon på båda sidor, se figur 30. Om könskromosomer exkluderas för alla i familjen blir den genetiska likheten mellan kusinerna lika oavsett vilka kön både syskonen samt kusinerna har, se figur 31. Precis som för vanliga kusiner kördes Java-programmet 10

miljoner gånger utan att dubbelkusinerna blev mer lika än syskon en enda gång. Den genetiska likheten för dubbelkusinerna då könskromosomer uteslutits ligger mellan 15% och 35% med ett genomsnitt på 25%. Dubbelkusiner kan alltså inte heller bli mer lika än syskon ty syskons genetiska likhet sträcker sig inte lägre än 40% och dubbelkusiners genetiska likhet inte högre än 35%.



Figur 30: Pedigree över en familj med dubbelkusiner



Figur 31: Fördelningen för den genetiska likheten för dubbelkusiner utan könskromosomer

3.4 Genetisk likhet vid inavel

Det producerade Java-programmet användes på pedigrees med olika grader av inavel. Fallen då kusiner, en far och hans dotter, en mor och hennes son samt en bror och en syster får barn tillsammans behandlades. Sedan jämfördes hur mycket arvsmassa som detta barn delar med sin mor respektive far samt syskon. Utan inavelns inverkan ska den genomsnittliga genetiska likhet för dessa relationer ligga omkring 50%, men för dessa pedigrees blir den högre.

Inavelskoefficienten för fallet då kusiner får barn tillsammans beräknades tidigare till 6,25%, se avsnitt 1.4.1. Alltså bör detta barn dela kring 56,25% med sina närmsta familjemedlemmar. Då programmet simulerade detta testades fallen då kusiner via systrar, via en bror och en syster samt via bröder får barn. Resultatet för kusinrelationerna gav en genomsnittlig genetisk likhet i intervallet 55,21%-57,16%.

För barn till ett syskonpar erhölls genomsnittliga likheter mellan 73,93%-76,32% beroende på könet på föräldrarna och barnen. Ökningen på cirka 25% beror förstås på att barnets mor- och farföräldrar är desamma. Inavelskoefficienten i detta fall är 25% vilket stämmer bra överens med ökningen av genetisk likhet. Snarlik genetisk likhet erhölls för fallet då en far och en dotter eller en mor och en son får barn tillsammans och även då är inavelskoefficienten 25%.

För alla de simulerade relationerna stämmer inavelskoefficienten bra med den uppmätta ökningen av genomsnittlig genetisk likhet. Avvikelserna i resultatet beror på könskromosomernas inverkan.

4 Diskussion

Två skilda tillvägagångssätt har använts för att beräkna sannolikheter angående hur mycket gemensam arvsmassa två besläktade individer har. Först och främst efterliknades arvsprocessen genom ett Java-program där personer och kromosomer behandlades som objekt. Sedan implementerades ett mer teoretiskt resonemang i MATLAB, där endast ett kromosompar behandlades. Programmet som producerats i Java har använts för att simulera arvsprocessen både för ett ensamt kromosompar och för individers hela genom. De resultat som genererades när endast ett kromosompar betraktades jämfördes i avsnitt 3.1 med de resultat som erhölls ur det teoretiska resonemanget. Notera att både Java- och MATLAB-programmen utgår från antagandet att antalet överkorsningar sker slumpmässigt enligt Poisson-fördelningen vid överkorsningsfasen i meiosen. Detta är som tidigare nämnt vedertaget. Det är alltså inte antagandets riktighet som resultatet i avsnitt 3.1 önskar påvisa, utan snarare att Java-programmet på ett korrekt sätt efterliknar arvsprocessen.

Det framgår i avsnitt 3.1 att Java-programmet ger konsekventa resultat när det jämförs med de siffror som erhållits ur det mer teoretiska resonemanget med hjälp av MATLAB. Detta tyder på att programmet korrekt simulerar arvsprocessen för ett kromosompar. Dessutom ger Java-programmet korrekta medelvärden för den genetiska likheten hos de vanliga släktskapen som presenterats i avsnitt 3.3. Resultaten är alltså rimliga även när hela genomet simuleras.

Sannolikheterna att två individer delar någon arvsmassa avtar förhållandevis långsamt när hela genomet betraktas, som framgår av avsnitt 3.2. Detta beror på att hela genomet har en stor total genetisk längd. Mellan varje kromosompar sker oftast en till tre överkorsningar. På hela genomet sker alltså ett mycket stort antal överkorsningar och det finns således många möjligheter för ett fragment att föras vidare. Dock är sannolikheten ungefär 50% att två individer inte delar någon arvsmassa redan när de skiljs åt av 10 meioser, se figur 18. Alltså kan det i många fall ifrågasättas hur stor vikt det ligger i vissa påståenden om att två individer är släkt. Efter 20 meioser kan det med största sannolikhet hävdas att två individer inte delar någon arvsmassa, trots att de är släkt.

En annan effekt av genomets extensiva totala genetiska längd är det faktum att de approximerade fördelningarna som presenterats i avsnitt 3.3 är mycket centrerade kring sitt medelvärde. De har alltså låg varians och således låg standardavvikelse. Detta eftersom meiosens intensitet, som beror på den genetiska längden, gör att arvsmassan blir uppdelad i många små fragment. Om genomet skulle ha en mindre genetisk längd skulle färre överkorsningar ske i meioserna och DNA:t skulle bli mindre uppdelat. Det skulle leda till en högre sannolikhet att en person ärver oblandade kromosomer av sina föräldrar. Dessa kromosomer skulle alltså bestå helt av förälderns paternella eller maternella DNA. Således skulle variansen öka. En minskad genetisk längd skulle därför leda till att ett barn oftare skulle kunna bli till exempel mycket mer lik sin mormor än sin morfar.

En följd av att fördelningarna för genetiska likheter har låg varians är det faktum att kusiner i princip aldrig blir mer lika än syskon. Teoretiskt är det möjligt, eftersom kusiner kan dela upp till hälften av sitt genom. Detta skulle kunna ske i extrema fall om syskonparet som är kusinernas föräldrar är mycket genetiskt lika. Syskon kan rent fysiskt dela upp mot 100% av

sin arvsmassa men eftersom genomet har stor genetisk längd sker det oerhört sällan. Som framgår av figur 21 i avsnitt 3.3.2 varierar normalt den genetiska likheten hos syskon mellan 40% och 60%. Inte heller när dubbelkusiner behandlades hittades några fall då kusiner var likare än syskon. Dock är det mycket möjligt att kusiner skulle kunna vara mer utseendemässigt lika än syskon. Hur lång del av arvsmassan som kodar för utseende är oklart, men den längden är självklart kortare än hela genomet.

Som tidigare nämnt har kvinnors och mäns kromosomer olika genetiska längder, kvinnors kromosomer är längre. Det får exempelvis till följd att fördelningen för den genetiska likheten mellan kusiner via två systrar blir mycket centrerad kring sitt medelvärde. Detta följer av samma resonemang som ovan, att om meiosen är intensivare minskar variansen. Den minskade variansen för kusiner via två systrar kan ses i figur 26 i avsnitt 3.3.4. Samma fenomen återfinns när morföräldrar och barnbarn jämförs med farföräldrar och barnbarn. I detta fall har den genetiska likheten för morföräldrar och barnbarn lägre varians än likheten för farföräldrar och barnbarn. Detta kan observeras i figur 22 i avsnitt 3.3.3. Den första meiosen som sker, mellan till exempel mormodern och modern, innehåller ingen slumpmässighet och bidrar därför inte med någon varians. I den andra meiosen, mellan föräldern och barnet, avgör förälderns kön hur många överkorsningar som sker. I kvinnans fall sker fler överkorsningar och alltså blir variansen lägre.

Den andra skillnaden som uppstår på grund av kön kommer av könskromosomernas inverkan. Eftersom Y-kromosomen är betydligt kortare än X-kromosomen kommer en son att dela en kortare längd DNA med sin far än med sin mor. En dotter kommer att dela en lika lång fysisk längd arvsmassa med vardera förälder. Dock kommer den längden att utgöra en större andel av faderns DNA eftersom en man har kortare total fysisk längd på sina kromosomer, på grund av Y-kromosomens längd. Således kommer en kvinna dela större *andel* DNA med sin far än med sin mor eftersom andelarna beräknas utifrån vad de maximalt kan dela. Könskromosomerna spelar även in när barnbarn och far- eller morföräldrar jämförs. Till exempel så ärvs en farfars Y-kromosom intakt ned till ett manligt barnbarn och de kommer därför att dela lite mer än det generella genomsnittet. Samma sak gäller för manliga kusiner via bröder.

De resultat som erhållits för pedigrees med inavel stämmer bra överens med den ökade sannolikheten för homozygositet som beräknas med inavelskoefficienten. Detta tyder på att Java-programmet hanterar sådana pedigrees korrekt.

Om Java-programmets riktighet kan motiveras till den grad att det anses tillförlitligt kan det till exempel vara till nytta för att undersöka fall av recessiv sjukdomsspridning. Ponera att två avlägset besläktade individer delar en recessiv sjukdom. Deras genom kartläggs och ett långt fragment som delas av individerna hittas. Frågan är då om det kan finnas andra kortare fragment som också delas av de sjuka eller om det kan fastslås att sjukdomsgenen finns i detta fragment. Utifrån det pedigree som beskriver deras släktskap kan sannolikheten att de delar någon annan arvsmassa beräknas med hjälp av det producerade Java-programmet. Om sannolikheten att de två individerna har något annat gemensamt DNA är mycket liten kan misstanken stärkas att det funna DNA-avsnittet är upphov till sjukdomen.

Genom att undvika detta skadliga fragment vid provrörsbefruktning kan sjuka individers föräldrar få fler barn utan risk för att de ska bli sjuka. Dessutom kan friska människor testa om de är bärare av den skadliga genen. Det är då också möjligt att utföra fosterdiagnostik för att undersöka om ett ofött barn är bärare av sjukdomen.

5 Slutsats

Modelleringen av arvsförloppet har lett fram till ett bra verktyg att undersöka och jämföra den genetiska likheten för besläktade individer. Utifrån den information som erhålls från modellen kan fördelningen för den genetiska likheten för två individer approximeras. Dessa fördelningar har för de vanligaste relationerna låg varians och i de flesta fallen kommer alltså den genetiska likheten ligga mycket nära genomsnittet. Könskromosomerna bidrar med förändrat medelvärde på de olika fördelningarna. Mäns och kvinnors kromosomers olika genetiska längder bidrar med skillnader i varians.

Programmet kan även användas för att beräkna hur många meioser som krävs mellan två besläktade individer för att dessa inte skall ha någon gemensam arvsmassa. Sådan information kan användas vid beräkningar gällande recessiv sjukdomsspridning.

Det program som skrivits i Java uppvisar övertygande resultat och anses därför vara tillförlitligt.

Referenser

[1] Klein J, Takahata N. Where Do We Come From?: The Molecular Evidence for Human Descent. Illustrated edition. Springer-Verlag Berlin and Heidelberg GmbH Co.K; 2002.

[2] Starr C, Evers C, Starr L. Biology: Concepts and Applications. Eighth edition. Wadsworth Publishing Co Inc; 2010.

[3] Martin D, Solomon E, Berg L. Biology. Eighth edition. Brooks/Cole; 2007.

[4] Alberts B, Bray D, Hopkin K, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Essential Cell Biology. Third edition. New York: Garland Science; 2009.

[5] Nilsson S. Which Genes are Involved?. Göteborg: Chalmers University of Technology and Göteborg University; 2001.

[6] Bennett R. The practical guide to the genetic family history. Second edition. John Wiley & Sons; 2011.

[7] Hartl D, Clark A. Principles of Population Genetics. Fourth edition. Sinauer Associates; 2006.

[8] Frankham R, Ballou J, Briscoe D. Introduction to Conservation Genetics. Second edition. Cambridge University Press; 2002.

[9] Gupta P.K. Genetics Classical to Modern. First edition. Rakesh Kumar Rastogi for Rastogi Publications; 2007.

[10] Roth S. Genetics Primer for Exercise Science and Health. Illustrated edition. Human Kinetics Publishers; 2007.

[11] David H, Nagaraja H. Order Statistics. Third edition. John Wiley and Sons; 2003.

[12] Cassandras C, Lafortune S. Introduction to Discrete Event Systems. Second edition. Kluwer Academic Publishers; 1999.

[13] Ott J. Analysis of Human Genetic Linkage. Third edition. The Johns Hopkins University Press; 1999.

A Genetiska längder

Genetiska längder i cM

Kromosom	Man	Kvinna
1	221	376
2	193	297
3	186	289
4	157	274
5	149	267
6	142	222
7	144	244
8	135	226
9	130	176
10	144	192
11	125	189
12	136	232
13	107	157
14	106	151
15	84	149
16	110	152
17	108	152
18	111	149
19	113	121
20	104	120
21	66	77
22	78	89
X	0	193
Y	0	-

För tabellvärdena har Ott J; 1999, [13] använts.

B Den minneslösa Poisson-processen

Betrakta en stokastisk variabel Z. Denna är Poisson-fördelad om och endast om den tid som förflutit mellan två på varandra förekomster av en händelsen har en exponential-fördelning.

En viktig egenskap hos exponential-fördelningen är att den saknar minne [12]. Alltså om X är exponential-fördelad gäller att:

$$P(X \leq x + y | X > x) = P(X \leq y) \text{ för varje } x \geq 0$$

Bevis:

$$\begin{aligned} P(X \leq x + y | X > x) &= \frac{P(X \leq x + y \cap X > x)}{P(X > x)} \\ &= \frac{P(x < X \leq x + y)}{P(X > x)} \\ &= \frac{F(x + y) - F(x)}{1 - F(x)} \\ &= \frac{1 - e^{-\lambda(x+y)} - (1 - e^{-\lambda x})}{e^{-\lambda x}} \\ &= \frac{e^{-\lambda x} - e^{-\lambda(x+y)}}{e^{-\lambda x}} \\ &= 1 - e^{-\lambda y} \\ &= F(y) = P(X \leq y) \end{aligned}$$

X är den tid som passerar innan en viss händelse inträffar. Ovanstående egenskap säger att sannolikheten att händelsen inträffar under en tidsperiod av längd y är oberoende av hur mycket tid x som redan passerat utan att händelsen inträffar.