CHALMERS

GÖTEBORGS UNIVERSITET

Prediktion av hårfärg och ögonfärg från genetiska markörer inom forensisk verksamhet

Prediction of hair color and eye color from genetic markers in forensic work.

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

Filippa Johansson Sporre
Gustav Järlstam
Ludwig Lewis
Robin Nilselid

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2025

Prediktion av hårfärg och ögonfärg från genetiska markörer inom forensisk verksamhet

Examensarbete för kandidatexamen i matematik, med inriktning matematisk statistik, inom Matematikprogrammet vid Göteborgs universitet

Robin Nilselid

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers

Filippa Johansson Sporre Ludwig Lewis

Kandidatarbete i matematik inom civilingenjörsprogrammet Globala System vid Chalmers

Gustav Järlstam

Handledare: Petter Mostad

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2025

Förord

Den här rapporten utforskar möjligheter och utmaningar med att predicera egenskaper kopplade till utseendet hos människor utifrån deras DNA med hjälp av logistisk regression, bayesiansk statistik och Monte Carlo metoder.

Vi vill tacka vår handledare Petter Mostad för sin vägledning under arbetets gång och för förslag av ämne för rapporten. Vi vill även tacka Andreas Tillmar och Rättsmedicinalverket som skapade det här projektet och tillät oss att använda deras data. Till sist vill tacka Maria Cervin-Ellqvist och övrig personal på fackspråk som har gett oss bra vägledning och många användbara tips. Även en stor eloge ska lyftas till Café $\langle \nu | \varphi \rangle$ med personal som försett oss med koffein, socker och glatt humör under hela processen.

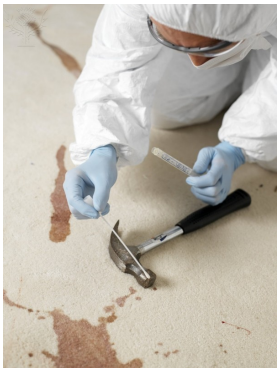
I tabellen nedan visas den huvudsakliga uppdelningen över hur rapporten har skrivits mellan gruppmedlemmarna. Varje gruppmedlem har även bidragit med återkoppling till resten av gruppen och mindre justeringar av texten. Utöver skrivandet av rapporten så har flertal modeller programmerats. Robin har haft huvudansvaret för modellerna som bara behandlar blå ögon, Ludwig har ansvarat för modellen som behandlar samtliga ögonfärger och Filippa har ansvarat för modellerna som predicerar hårfärg.

Avsnitt	**Rubrik**	**Huvudförfattare**	**Medförfattare**
	Förord	Ludwig	
	Populärvetenskaplig text	Gustav	Filippa
	Sammandrag / abstract	Gustav, Filippa	
1	Inledning	Gustav	
1.1	Bakgrund	Gustav	
1.2	Syfte	Gustav	Ludwig
1.3	Avgränsningar	Robin	Resterade
2.1	Logistisk regression	Robin	Ludwig
2.2	Bayensiansk inferens	Ludwig	
2.2.1	Likelihood-funktionen	Gustav	Ludwig
2.3	Stokastiska processer	Ludwig	Gustav
2.3.1	Markovkedjor	Ludwig, Robin	Gustav
2.4	Monte Carlo-metoden	Gustav	
2.4.1	Monte Carlo Integration	Gustav	
2.5	Markov Chain Monte Carlo (McMC)	Ludwig	Robin, Gustav
2.6	Metropolis-Hastings algoritm	Robin	Ludwig
2.7	Laplace approximation	Ludwig, Gustav	
2.8	Korsvalidering	Gustav	Robin
2.9	Förväxlingsmatriser	Ludwig	Gustav, Robin
3.1	Presentation och transformation av data	Gustav, Robin	
3.2.1	Härledning av modellgrund	Ludwig, Robin	
3.2.2	Grundläggande modell	Gustav, Ludwig	Robin
3.2.3	Numerisk modell	Ludwig	
3.2.4	Modell för blåa ögon	Gustav, Robin	Ludwig
3.2.5	Modell för samtliga ögonfärger	Ludwig	
3.3	Prediktionsmodeller för hårfärg	Filippa, Gustav	
4.1	Grundläggande modell och numerisk beräkning	Ludwig, Gustav	
4.2	Prediktionsmodell för blåa ögon	Gustav	Ludwig
4.3	Prediktionsmodell för samtliga ögonfärger	Ludwig	Gustav
4.4	Prediktionsmodell för brunt hår	Filippa	Ludwig
4.5	Prediktionsmodell för samtliga hårfärger	Filippa	Ludwig
4.6	Prediktionsmodell för blont, rött och svart hår	Filippa	Ludwig

5.1	Grundläggande modell och numerisk beräkning	Gustav	
5.2	Modeller för ögonfärg	Gustav, Ludwig	Filippa
5.3	Modeller för hårfärg	Filippa	Gustav, Ludwig
5.4	Osäkerheter och utmaningar	Gustav, Robin	Ludwig, Filippa
5.5	Potentiella användningsområden	Gustav, Filippa	Robin
5.6	Samhälleliga och etiska aspekter	Gustav	Filippa
5.6.1	Risk för diskriminering	Gustav	Filippa
5.6.2	Anpassad lagstiftning	Gustav	
5.6.3	Hantering av biometriska uppgifter	Gustav	
6	Slutsats	Filippa, Gustav, Ludwig	

Populärvetenskaplig presentation

Efter att ett brott har begåtts startas en brottsutredning. Under en brottsutredning samlas det in prover från brottsplatsen för att undersöka om det finns DNA-spår. Idag används nästan alltid en metod där DNA:t som hittats på brottsplatsen jämförs direkt med DNA:t från en misstänkt gärningsperson. Detta sker antingen genom att gärningspersonens DNA finns i en databas eller genom direkt jämförelse vid så kallas topsning.

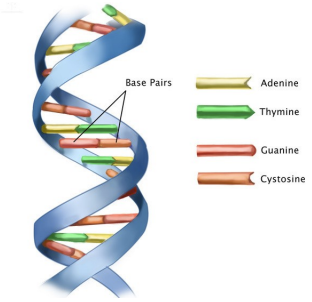


Figur 1: Insamling av forensiska bevis [1], Används med tillstånd från Britannica ImageQuest.

Men vad gör man när det inte finns något DNA att jämföra med? Genom att kombinera biologi och matematik skulle det vara möjligt att, utifrån DNA-spår som hittats på en brottsplats, få fram information om hur en person ser ut. Detta kan i sin tur hjälpa brottsutredningen genom att smalna av sökfältet till vissa utseendedrag, såsom ögonfärg eller hårfärg, hos den potentiellt misstänkta personen.

DNA finns i alla våra celler och fungerar som en instruktion för hur kroppen ska utvecklas och fungera. Alla människors DNA är nästintill identiska med varandra, vilket innebär att man behöver studera DNA:t noggrant för att hitta skillnader. Om man vill hitta en specifik genetisk markör som exempelvis avgör ögonfärg, behöver man jämföra specifika delar av DNA:t.

En genetisk markör är en sådan plats i DNA:t där det finns variation. DNA är uppbyggt av fyra olika kvävebaser: adenin (A), tymin (T), guanin (G) och cytosin (C). A parar sig alltid med T, och G med C. Man skulle kunna se på människans DNA som en bok som har skrivits av väldigt många gånger. När man snabbt bläddrar igenom böckerna ser de identiska ut. Men om man tittar närmare går det att upptäcka att det kan finnas små stavfel och i DNA:t skulle det vara så att det står A-T där det brukar stå G-C, det är en variation. Det är variationerna i dessa baspar som gör att vi får olika egenskaper.



Figur 2: DNA sträng med kvävebaser. [2], Används med tillstånd från Britannica ImageQuest.

Genom att samla in information från många människor, till exempel om deras ögonfärg eller hårfärg, och analysera deras DNA, är det möjligt att identifiera mönster. Detta innebär att det är fullt möjligt, utifrån variationerna i DNA:t från en okänd person, att göra en kvalificerad gissning om vissa utseendedrag. Det är lite som att kunna skissa fram ett suddigt porträtt av en person, endast baserat på vilka genvarianter som finns i dennes DNA.

För att kunna göra den här typen av förutsägelser använder man sig av statistiska modeller. Det är ungefär som att skapa en matematisk översättning mellan DNA och olika egenskaper. Modellen tittar på tidigare data, alltså DNA från personer där vi *redan vet* exempelvis deras ögonfärg, och utifrån detta lär sig modellen hur vissa variationer i DNA ofta hänger ihop med vissa utseendedrag.

När modellen sedan får in DNA från en person vi inte vet något om, kan den använda denna ”matematiska översättning” för att räkna ut vilka utseendedrag som är mest sannolika. Eftersom metoden bygger på sannolikhetsberäkningar innebär det att resultaten aldrig är helt säkra. Exempelvis kan en person ha störst sannolikhet för bruna ögon men ändå ha blå ögon. Därför är det viktigt att använda dessa beräkningar som ett hjälpmedel och inte ett absolut bevis.

Genom att använda statistiska modeller är det möjligt att, med bara några DNA-molekyler som utgångspunkt, börja ana konturerna av en människa. Detta är i sig kanske inte tillräckligt för att klara upp en utredning, men det kan vara det som krävs för att leda en brottsutredare åt rätt håll, och få utredningen att ta ett avgörande steg framåt.

Sammandrag

Just nu pågår studier om nya tekniker inom forensisk verksamhet som ska möjliggöra användning av DNA för att förutsäga fenotypiska egenskaper, såsom ögon- och hårfärg, från biologiskt material som hittats på brottsplatser. Dessa prediktioner kan vara särskilt värdefulla i utredningar där traditionell DNA-profilering inte ger tillräcklig information. I denna rapport har data från Rättsmedicinalverket använts, bestående av sex single-nucleotide polymorphisms (SNPs) associerade med ögonfärg och 22 SNPs associerade med hårfärg, insamlade från 85 individer. Syftet med rapporten är att utveckla en statistisk prediktionsmodell som pålitligt kan klassificera ögon- och hårfärg baserat på genetisk information med hjälp av Markov chain Monte Carlo (McMC)-metoder. Det utvecklades flera modeller under projektets gång men i huvudsak användes två modeller för ögonfärger och tre modeller för hårfärger. Prediktionsmodellerna för ögonfärg visade mycket god förmåga att särskilja personer med blå och bruna ögon. Däremot uppstod svårigheter vid identifiering av individer med gröna ögon eller ögonfärger som låg mellan blått och brunt. För hårfärg visade modellen en styrka i att identifiera personer med brunt hår, men hade begränsad förmåga att korrekt klassificera övriga hårfärger, exempelvis tenderade individer med blont, rött eller svart hår att felaktigt klassificeras som brunhåriga. Dessa resultat understryker behovet av vidare forskning med större och mer varierade datamängder för att förbättra modellens inlärningsförmåga och precision. Tekniken har stor potential att bidra till effektivare brottsutredningar genom att avgränsa antalet möjliga misstänkta, men det är också viktigt att beakta de osäkerheter som är förknippade med fenotypisk prediktion. I rapporten diskuteras faktorer som kan påverka prediktionernas tillförlitlighet, såsom tekniska begränsningar, kosmetiska förändringar, miljöfaktorer och trauman. Sammantaget indikerar resultaten att området är lovande, men att fortsatt forskning är nödvändig för att stärka metodens praktiska användbarhet.

Abstract

There are ongoing studies in forensic fields with focus on using DNA to predict phenotypical characteristics, such as eye- and hair colour, from biological materials found at crime scenes. Phenotypical prediction has the potential to guide police investigations when conventional DNA profiling is not providing enough information for criminal investigators to continue the investigation. This study utilises data from the National Board of Forensic Medicine (Rättsmedicinalverket), comprising the six most eye colour-informative and 22 hair colour-informative single nucleotide polymorphisms (SNPs), collected from 85 individuals. The aim is to develop a statistical prediction model capable of accurately classifying an individual's eye- and hair colour using Markov Chain Monte Carlo (McMC) probability estimation based solely on DNA data. Several models were developed during the course of the project, but primarily two models were used for eye color and three models for hair color. The prediction models for eye color demonstrated a great ability to distinguish between individuals with blue and brown eyes. However, the model had difficulties in identifying individuals with green eyes or eye colors that fell between blue and brown. For hair color, the model showed considerable strength in identifying individuals with brown hair but had limited ability to correctly classify other hair colors, for example were individuals with blonde, red, or black hair often falsly classified as having brown hair. These results show the need for further research with larger and more diverse datasets to improve the model's learning capability and precision. The technology holds great potential to contribute to more effective criminal investigations by narrowing down the number of potential suspects, but it is also important to consider the uncertainties associated with phenotypic prediction. The report discusses factors that can affect the reliability of the predictions, such as technical limitations, cosmetic alterations, environmental factors, and trauma. Overall, the results indicate that the field is promising, but continued research is necessary to strengthen the practical usefulness of the method.

Innehåll

1 Inledning

År 1988 uppklarades ett brott för första gången med hjälp av DNA-bevis när Colin Pitchfork dömdes till livstids fängelse för två mord i Leicestershire, Storbritannien [3]. Detta fall markerade en viktig punkt i kriminalteknikens historia. Pitchfork kunde avslöjas efter det allra första mass-DNA-testet, där över 5 000 män testades. Metoden byggde på att jämföra DNA-profiler för att säkerställa en exakt matchning mellan biologiska spår från brottsplatsen och misstänkta individer. Introduktionen av DNA-bevis i rättssalen har haft en stor effekt på både brottsutredningar och rättsliga bedömningar [4]. Det möjliggjorde mer träffsäkra identifieringar av gärningspersoner samtidigt som det också kunde bidra till att fria oskyldiga från brottsmisstankar.

Den 24 april 2018 arresterades en misstänkt gärningsperson i det omtalade fallet ”Golden State Killer” i Kalifornien [5]. Den misstänkte Joseph James DeAngelo, en tidigare polis, var misstänkt för 12 mord och åtminstone 45 våldtäkter som skett mellan 1976 och 1986. Genom användningen av en ny teknik, nämligen genetisk släktforskning, lyckades polisen med hjälp av DNA-bevis från brottsplatsutredningar identifiera familjemedlemmar till den misstänkta genom en gratis DNA-databas. Efter denna rapportering uppstod en debatt i Sverige om ett motsvarande tillvägagångssätt kunde användas i svenska brottsutredningar [6].

En del av den här diskussionen har fokuserat på möjligheterna att utifrån DNA förutse utseendemässiga drag, vilket är något som idag undersöks av Rättsmedicinalverket i Linköping [7]. I en pilotstudie där ungefär 15 personer deltog gjordes just detta. Bilder skapades utifrån de medverkandes DNA och sedan fick en separat utvärderingsgrupp i uppgift att försöka matcha ihop de genererade bilderna med personerna som var med i studien. I ett mejl från Andreas Tillmar, som jobbar på Rättsmedicinalverket, menar han att det finns ett stort intresse inom rättsgenetiken att kunna göra prediktioner av fenotyper utifrån DNA-data [8]. Användningen för dessa prediktioner skulle främst vara för att hitta okända gärningspersoner eller för att kunna identifiera kvarlevor när den avlidnes identitet är helt okänd. I dagsläget är det vanligt att främst använda logistiska regressionsanalyser som tillvägagångssätt vid denna typ av prediktioner. Det finns dock ett växande intresse för att undersöka om andra angreppssätt kan ge förbättrade resultat, vilket ligger till grund för detta arbete.

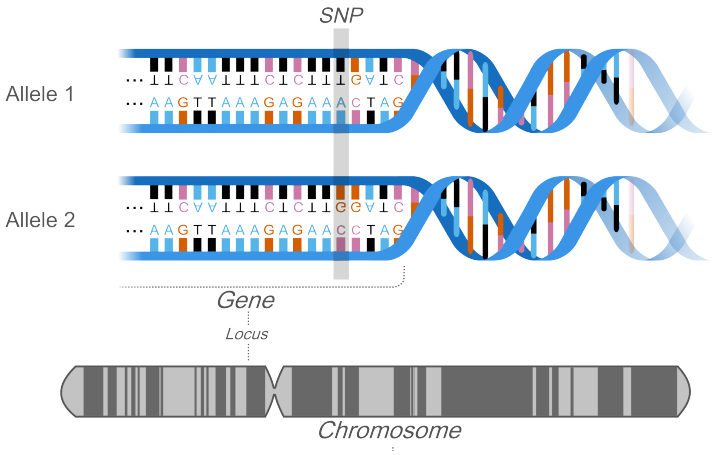
1.1 Bakgrund

Vid utredning av brott där DNA-spår har en avgörande roll används idag mikrosatelliter, också kallat Short Tandem Repeats (STR) [6]. Vanligtvis är STR-analys en standardmetod inom kriminaltekniken. STR är DNA-fragment i sekvenser av 1-6 baspar som upprepas i lång följd. STR-analys gör det möjligt att identifiera en individ genom att jämföra DNA-profiler från brottsplatsen i form av biologiska spår. Metoden är effektiv när det redan finns en misstänkt person som det går att jämföra mot eller när DNA-profilen matchas mot en befintlig profil i en kriminalteknisk databas.

Det finns dock fall där DNA-profilen som genererats från bevismaterialet inte matchar någon känd individ. När detta sker kan utredningen stanna upp, ett så kallat kallt fall. På grund av dessa begränsningar har det på senare år utvecklats metoder för att utvinna mer information ur DNA, det är möjligt att statistiskt kunna förutsäga vissa fysiska egenskaper baserat på en individs genetiska kod [9]. En sådan metod kallas för Forensisk DNA fenotypning (FDF) och innebär att man analyserar vissa specifika markörer som är kopplade till nedärvda synliga drag såsom ögonfärg eller hårfärg. Till skillnad från STR-analys, som endast kan skapa identifikationer genom direkt jämförelse med befintliga DNA-profiler, gör FDF det möjligt att få en fysisk beskrivning av en okänd person endast genom att kolla på deras genetiska kod. Detta kan ge utredare värdefull information när de saknar en matchande DNA-profil och därmed saknar en direkt misstänkt att arbeta vidare mot [10]. FDF ger dock inte en exakt bild av en persons utseende, utan den ger endast sannolikheter för olika drag [9].

En locus (plural, loci) är en specifik, fysisk plats på en kromosom där en viss gen eller genetisk markör är lokaliserad [11]. En single-nucleotide polymorphism (SNP) är en variation vid en enda position i DNA-sekvensen, det vill säga en basparsvariation i genomet [12]. En SNP kan förekomma vid en specifik locus, vilket innebär att variationen befinner sig på en exakt plats i DNA-sekvensen.

En SNP är alltså en variation som kan hittas vid en specifik locus på en kromosom och detta visualiseras i figur 3.



Figur 3: Illustration över baspars variationer i DNA:t. [13], CC-BY-SA 4.0.

Fenotyp är det observerbara uttrycket av en individs genotyp, exempelvis genom utseende eller symtom vid sjukdomar [14]. Genotypen påverkar produktionen av proteiner, cellfunktioner och biologiska processer. Genotypen bidrar på så sätt direkt till observerbara egenskaper och formar därmed fenotypen, tillsammans med utomstående miljöfaktorer [15].

FDF bygger på att man identifierar variationer i DNA-sekvensen, genom att analysera vissa utvalda SNP:er som ofta återfinns på specifika loci i genomet [16]. Genom att kombinera information från dessa markörer kan man statistiskt beräkna sannolikheten för att en individ uppvisar vissa observerbara egenskaper, vilket ger en indikation på den okända personens fenotyp.

Att kunna förutsäga en persons ögon- och hårfärg utifrån DNA handlar inte bara om att hitta rätt genetiska kopplingar, utan också om att förstå hur säkra dessa förutsägelser faktiskt är. Genom att använda statistiska metoder går det att uppskatta hur troligt det är att en viss DNA-profil stämmer överens med en viss egenskap, till exempel att en person har blå ögon [17]. Dessa sannolikheter hjälper inte bara till att skapa en profil av en okänd person, utan de kan också kombineras med annan information för att påverka bevisens styrka i en brottsutredning.

1.2 Syfte

Syftet med det här arbetet är att undersöka möjligheten att förutsäga vissa fenotypiska egenskaper, specifikt ögon- och hårfärg, enbart utifrån biologiska DNA-spår. Detta kommer att göras genom att bygga sannolikhetsmodeller som förutspår personers ögon- och hårfärg utifrån olika delar av deras DNA. I det här arbetet kommer det även att ske en analys av osäkerheter kopplade till sannolikhetsmodellerna och dess prediktioner. Rapporten hoppas därmed kunna bidra till hur denna sorts prediktioner kan användas i forensiska sammanhang men även andra områden. Därifrån följer även en analys över samhällsetiska frågor kring huruvida dessa modeller kan och kommer användas i praktiken.

1.3 Avgränsningar

Den här rapporten avgränsas av den tillgängliga data som tillhandahålls av Rättsgenetik i Linköping. Materialet består av två tabeller med genetisk information från 85 individer. Den första tabellen består av personernas observerade ögonfärg och genetiska variationer (SNP) vid sex specifika loci. Den andra tabellen är liknande, där finns individernas observerade hårfärg och SNP vid 22 olika loci. Därmed begränsas modellerna för prediktioner i detta arbete till ett visst antal individer och specifika delar av deras DNA. Ingen ytterligare data kommer att samlas in under

arbetets gång. Arbetet kommer enbart fokusera på att predicera hår- och ögonfärg, det kommer inte att inkludera andra fenotypiska eller genetiska egenskaper.

De matematiska modeller som används i detta arbete är främst logistiska regressionsmodeller för att predicera fenotypen samt Bayesiansk statistik för att ta fram parametrarna för den logistiska modellen. Andra prediktionsmetoder, såsom maskininlärning med artificiella neurala nätverk, kommer inte att användas.

Själva prediktionerna i det här arbetet begränsas till diskreta kategoriseringar av ögon- och hårfärg. För ögonfärg används kategorierna[1] blå, brun och intermed, medan hårfärg klassificeras som blond, brun, röd eller svart. Detaljerade variationer inom dessa färger såsom nyanser kommer inte att beaktas.

2 Teori

Ett sätt att förutsäga ögon- och hårfärg från DNA-spår är genom sannolikhetsmodeller. I detta avsnitt presenteras den matematiska teorin som utgör grunden för dessa statistiska modeller, samt de metoder som används för att bestämma deras parametrar.

2.1 Logistisk regression

Logistisk regression uppskattar sannolikheten att en händelse sker genom att omvandla ett reellt tal till en sannolikhet. En händelse kan, i det här arbetet, tolkas som att en person exempelvis har blå ögon eller inte. Eftersom det vi får ut av en logistisk regression är en sannolikhet kommer svaret alltid att ligga mellan 0 och 1. Om denna typ av modell utvecklas för fler än två utfall kommer resultatet från regressionen istället vara en vektor av sannolikheter. Varje position i vektorn motsvarar då sannolikheten för varje utfall, och summan av elementen i vektorn är därför 1.

En logistisk regressionsmodel för datan D är kopplad till en utfalls-slumpvariabel X. Kopplingen är att med all data D kan man konstruera statistiska modeller som beräknar sannolikheten för att observation $D_i = (D_{i,1}, ..., D_{i,p})$ har utfallet X. Detta genom att ta fram parametrarna θ_i till polynomet $\theta_0 + \theta_1 D_{i,1} + ... + \theta_p D_{i,p} = U$ som sedan används i den logistiska ekvationen [18, s. 449]. Den logistiska ekvationen för binära utfall är

$$\phi(U) = \frac{\exp(U)}{\exp(U) + 1} \in (0, 1) \tag{1}$$

och från denna fås sannolikheten för X som, om X är binärt, ges enligt

$$P(X = \text{Sant}|\theta) = \phi(\theta \cdot D_i),$$
$$P(X = \text{Falskt}|\theta) = 1 - \phi(\theta \cdot D_i).$$

Vidare kan ekvation (1) utvecklas till en multinomial logistisk ekvation när det finns tre eller flera utfall. Det går nu att ta fram sannolikheten för $s + 1$ olika utfall med

$$\phi(U) = \frac{1}{1 + \sum_{j=1}^{s} \exp(U_j)} \left(\exp(U_1), ..., \exp(U_s)\right) \text{ [19, s. 721].} \tag{2}$$

Sannolikheterna för $s + 1$ olika utfall för X fås nu enligt

$$P(X = j|\theta) = \begin{cases} \phi(D_i \cdot \theta)_j, & 1 \leq j \leq s \\ 1 - \sum_{k=1}^{s} \phi(D_i \cdot \theta)_k, & j = s + 1. \end{cases}$$

[1] Förklaring av kategorierna finns i kapitel 3.1

2.2 Bayesiansk Inferens

Inom Bayesiansk Statistik behandlas både parametrar och data som slumpvariabler [18, s. 3]. I praktiken innebär detta att om vi har observerat data $D = \{D_1, D_2, \ldots, D_n\}$ från någon fördelning med parameter θ så ser vi både D och θ som utfall av slumpvariabler. Den gemensamma fördelningen för D och θ är

$$\pi(D, \theta) = \pi(D|\theta)\pi(\theta).$$

Fördelningen $\pi(\theta)$ kallas för en a priori fördelning för θ och kan tolkas som ett grundantagande om θ. Den kan alltså ses som hur man tror parametrarna θ är fördelade och ska inte vara beroende eller baserad på D. Fördelningen $\pi(D|\theta)$ är den sannolikhetsmodell som beskriver hur data genereras givet parametern θ. Med hjälp av Bayes sats erhåller vi att

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}. \tag{3}$$

Den här fördelningen kallas a posteriori fördelningen för θ givet den observerade datan D och innehåller vår uppdaterade tro om θ efter att D har observerats. Oftast räcker det med att beräkna täljaren i ekvation (3) eftersom nämnaren inte beror på θ. Detta ger en funktion $f(\theta)$ som är proportionell mot a posteriori fördelningen, alltså har vi att

$$f(\theta) = \pi(D|\theta)\pi(\theta) \propto_\theta \pi(\theta|D). \tag{4}$$

2.2.1 Likelihood-funktionen

Likelihood-funktionen används för att hitta det mest sannolika värdet på en parameter, baserat på observerad data [18, kap 1.2.1]. Likelihood-funktionen definieras som sannolikheten för att observera data D, givet ett värde på parametern θ. Om varje observation D_i har ett binärt utfall $z_i \in \{0, 1\}$, där sannolikheten för varje utfall ges av logistisk regression, som definierades i kapitel 2.1, så kan Likelihood-funktionen skrivas som

$$\pi(D|\theta) = \prod_{i=1}^{n} \phi(\theta \cdot D_i)^{z_i}(1 - \phi(\theta \cdot D_i))^{(1-z_i)}. \tag{5}$$

För att förenkla beräkningar och undvika numeriska problem kan Likelihood-funktionen skrivas om till en log Likelihood

$$\log(\pi(D|\theta)) = \sum_{i=1}^{n} z_i \log(\phi(\theta \cdot D_i)) + (1 - z_i)\log(1 - \phi(\theta \cdot D_i)). \tag{6}$$

Genom att beräkna det θ som maximerar likelihood-funktionen för given data D fås θ_{ML} som är parametrarna som gör den observerade datan så sannolik som möjligt. θ_{ML} kallas *maximum-likelihood* värden och används i rapporten.

2.3 Stokastiska Processer

En stokastisk process är en samling av indexerade slumpvariabler $X_t, t \in I$ som delar samma utfallsrum S [20, s. 6]. Indexeringsmängden I representerar oftast tiden, här används $I = 1, 2, 3, ...$ för en diskret tidsprocess. I denna rapport används stokastiska processer för att lära modellen om parametervärden. Slumpvariablerna i processen kan vara oberoende eller så kan deras fördelningar vara betingade på föregående värden.

2.3.1 Markovkedjor

Markovkedjor är en särskild typ av stokastiska processer med egenskapen att deras värde i varje steg endast beror på värdet i det tidigare steget. Processens värden är helt oberoende av andra tidigare värden [20, s. 65]. En följd av slumpvariabler $(X_n)_{n=0}^{\infty}$ är en Markovkedja i diskret tid, om det för alla $x_1, x_2, ..., x_n, x_{n+1} \in \mathbb{R}$ och $n \in \mathbb{N}$ gäller att

$$P(X_{n+1} = x_{n+1}|X_n = x_n, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1}|X_n = x_n).$$

Det betyder att sannolikheten för vad som händer härnäst ($P(X_{n+1} = x_{n+1})$) endast beror på det nuvarande tillståndet (X_n), och inte på något tidigare steg.

2.4 Monte Carlo-metoden

Monte Carlo-metoden är en numerisk metod som används för att lösa problem genom att använda slumpmässiga tal [20, s. 10]. Den grundläggande idén är att upprepa ett slumpmässigt experiment många gånger och använda resultatens frekvens för att uppskatta sannolikheten för en viss händelse.

Säg att vi vill uppskatta sannolikheten för en händelse B genom att upprepa ett experiment flera gånger. Då kan sekvensen $X_1, X_2, \ldots, X_k$, definieras som

$$X_k = \begin{cases} 1, & \text{om B sker på det } k\text{:te försöket,} \\ 0, & \text{om B inte sker på det } k\text{:te försöket,} \end{cases}$$

för $k \geq 1$. Då vi har gjort n försök är $(X_1 + X_2 + ... + X_n)/n$ andelen försök då B inträffade. Varje X_k är en identiskt fördelad slumpvariabel med väntevärde $E(X_K) = P(B)$.

Från den starka versionen av de stora talens lag får vi,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = P(B), \text{ med sannolikhet 1.}$$

Detta innebär att om vi gör tillräckligt många försök, kommer andelen gånger som B inträffar att närma sig den verkliga sannolikheten $P(B)$. För stora n, uppskattar Monte Carlo metoden $P(B)$ som

$$P(B) \approx \frac{X_1 + ... + X_n}{n}.$$

2.4.1 Monte Carlo-Integration

Monte Carlo-integration är en metod som används för att approximera väntevärdet av en funktion när direkta beräkningar är svåra [18, s. 4]. Väntevärdet av en funktion $f(X)$ med avseende på en sannolikhetsfördelning $\pi(X)$ ges av integralen

$$E[f(X)] = \int f(X)\pi(X)dX.$$

Eftersom denna integral ofta är svår att beräkna exakt kan vi istället approximera den med hjälp av n oberoende utfall $\{X_t\}_{t=1}^n$ av X. Genom att använda approximationen

$$E[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t), \tag{7}$$

som kallas Monte Carlo integration kan vi få ett approximativt väntevärde. Populationsmedelvärdet av $f(X)$ uppskattas med stickprovsmedelvärdet. Eftersom utfallen X_t är oberoende, ser stora talens lag till att uppskattningen blir mer exakt ju större urvalet n är.

2.5 Markov Chain Monte Carlo (McMC)

Markov chain Monte Carlo (McMC) är ett samlingsbegrepp för metoder som tillåter simulering av komplexa och högdimensionella sannolikhetsfördelningar [20, s. 181]. Målet med McMC är att, givet en sannolikhetsfördelning π, simulera en slumpvariabel X vars gränsfördelning är π. Gränsfördelningen kan antingen vara diskret eller kontinuerlig.

2.6 Metropolis-Hastings algoritm

Den vanligaste McMC metoden är Metropolis-Hastings algoritmen, i den simulerar man en slumpvariabel $\theta_\cdot = (\theta_{\cdot,0}, \theta_{\cdot,1}, ..., \theta_{\cdot,p})$ genom att ta fram N olika iterationer av parametern θ_i där $i = 2, ..., N$ [18, kap. 1]. Varje steg θ_i i kedjan väljs genom att dra en kandidatpunkt θ^* från

en förslagsfördelning $q(\theta^*|\theta_{i-1})$. Den första iterationen θ_1 är en initial gissning men väljs ofta med Maximum-Likelihood metoden från ekvation (6).

Om kandidatpunkten blir nästa iteration i kedjan eller inte bestäms av ekvationen

$$\theta_i = \begin{cases} \theta^* & \text{om } u \leq \rho(\theta^*, \theta_{i-1}), \\ \theta_{i-1} & \text{om } u > \rho(\theta^*, \theta_{i-1}), \end{cases}$$

där $i = 2, ..., N$, u dras från Unif$(0, 1)$ och accepteringssannolikheten $\rho(\theta^*, \theta_{i-1})$ ges av

$$\rho(\theta^*, \theta_{i-1}) = \min\left(1, \frac{f(\theta^*)q(\theta_{i-1}|\theta^*)}{f(\theta_{i-1})q(\theta^*|\theta_{i-1})}\right) \in [0, 1]$$

där ρ är acceptanssannolikheten om man accepterar kandidatpunkten eller inte. För att beräkna ρ så behöver man funktionen f som väljs som samma f i ekvation (4), där $\pi(\theta|D)$ är densiteten för θ givet vår observerade data D och $\pi(\theta)$ är densiteten för θ utan någon betingning. Kedjan blir alltså en Markovkedja eftersom varje iteration endast beror på den tidigare iterationen. Med denna kedja är det möjligt att använda Monte Carlo integrering från ekvation (7) för att approximera θ's fördelning. Värt att notera är fallet då fördelningen för θ^* är symmetrisk. Exempelvis om $\theta^* = \theta_{i-1} + \epsilon$, där $\epsilon \sim N(0, \sigma^2)$. Då är $q(\theta_{i-1}|\theta^*) = q(\theta^*|\theta_{i-1})$ vilket innebär att ρ reduceras till

$$\rho = \min\left(1, \frac{f(\theta^*)}{f(\theta_{i-1})}\right) \in [0, 1].$$

2.7 Laplace approximation

Laplace-approximationen är en approximationsmetod som kan användas till att approximera en fördelning $\pi(x)$ givet en funktion $f(x) \propto \pi(x)$ [21, kap 4.7]. Metoden bygger på en andra ordningens Taylorutveckling av $g(x) = \log f(x)$ kring en punkt $x = x_0$.

Vi utvecklar $g(x)$ kring $x = x_0$

$$g(x) \approx g(x_0) + (x - x_0)\left.\frac{\partial g(x)}{\partial x}\right|_{x=x_0} + \frac{(x - x_0)^2}{2}\left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=x_0}.$$

Om vi väljer $x_0 = x^*$ där $x^* = \arg\max g(x)$, så försvinner den första derivatan eftersom den är lika med noll i x^*. Då blir approximationen

$$g(x) \approx g(x^*) + \frac{(x - x^*)^2}{2}\left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=x^*}.$$

Vi kan därmed uppskatta $f(x)$ som

$$f(x) \approx \exp\left(g(x^*) + \frac{(x - x^*)^2}{2}\left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=x^*}\right).$$

Detta påminner om tätheten i en normalfördelning. Genom att definiera variansen som

$$\sigma^2 = -\left(\left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=x^*}\right)^{-1},$$

får vi slutligen att $\pi(x)$ approximeras av en normalfördelning centrerad i x^* med varians σ^2, alltså

$$\pi(x) \approx N(x^*, \sigma^2).$$

Laplace-approximationen innebär alltså att vi approximerar $\pi(x)$ med en normalfördelning centrerad vid dess maximum, och där spridningen beror på "kurvigheten" i logaritmen av tätheten vid detta maximum. I rapporten används Laplace approximationen för att approximera posteriori fördelningen $\pi(\theta|D)$, approximationen används sedan som förslagsfunktion i Metropolis-Hastings algoritmen. Detta är nödvändigt eftersom om olika värden i θ spelar olika mycket roll så fungerar en förslagsfunktion som tar lika stora steg åt alla håll dåligt och därmed kommer kedjan att konvergera långsamt till den önskade fördelningen.

2.8 Korsvalidering

Korsvalidering är en av de enklaste och mest använda metoderna för att undersöka prediktionsförmågan hos statistiska modeller. Korsvalidering går ut på att man delar upp datan i k lika stora bitar, där man sedan tränar k olika modeller [22, kap. 7.5.2]. Vid varje iteration tränas modellen på $k-1$ delar av datan och valideras på den återstående delen. Detta upprepas k gånger, så att varje del används exakt en gång som valideringsmängd. Resultatet blir medelvärdet av prestandan efter att de k modellerna har validerats. Det finns olika korsvalideringsmetoder men, i de fall där man har få datapunkter brukar Leave-One-Out korsvalidering (LOOCV) vara en bra metod. LOOCV går ut på att man sätter $k = n$, alltså lika med antalet observationer. Varje modell tränas på $n-1$ datapunkter och valideras på den datapunkt som utelämnats. LOOCV används i rapporten för att validera modellerna.

2.9 Förväxlingsmatriser

Förväxlingsmatriser är en sorts matris som används för att presentera prediktionsresultat från klassificeringsmodeller. En förväxlingsmatris för en klassificeringsmodell för k olika klasser är en $k \times k$ matris C där element $C_{i,j}$ anger antalet datapunkter där den korrekta klassen är i och klassen modellen predicerar är j [22, kap 7.4.3]. I rapporten används förväxlingsmatriser för att presentera och utvärdera prestandan hos modellerna för hår och ögonfärg.

Tabell 1: Exempel på en 2x2 förväxlingsmatris

Pred → Sann ↓	**Falsk**	**Sann**
Falsk	C_{11}	C_{12}
Sann	C_{21}	C_{22}

Resultaten från förväxlingsmatriser går att använda för att bedöma en modells prestanda, motivering för detta presenteras i appendix A.

3 Metod

I följande avsnitt presenteras metoder som använts vid konstruktionen av de statistiska modeller som predicerar ögon- och hårfärg hos personer utifrån deras DNA. Samtliga modeller bygger på materialet som presenterats i teoriavsnittet.

3.1 Presentation och transformation av data

Datamängderna som det här arbetet har utgått från består av ögon- och hårfärger samt basparsvariationer på specifika fysiska platser (loci) i DNA-sekvensen hos 85 personer. För att matematiskt kunna arbeta med datan behövde den transformeras från baspar till numeriska värden. De angivna ögonfärgerna i datan är blå, brun och "intermed", där intermed syftar på intermediära ögonfärger. Det är alltså ögonfärger som inte kan klassas som varken blå eller bruna. De angivna hårfärgerna i datan är brun, blond, svart och röd. För att underlätta vid programmering och samtal runt detta döptes samtliga loci om till en siffra som representerade dess placering i datamängden. De nya

namnen blev då exempelvis ”locus 1” istället för locus rs12203592. De nya namnen visas i tabellerna nedan.

Tabell 2: Ögon loci

Ögon loci	
rs12203592	locus 1
rs1393350	locus 2
rs12896399	locus 3
rs1800407	locus 4
rs16891982	locus 5
rs12913832	**locus 6**

Tabell 3: Hår loci

Hår loci			
rs28777	locus h1	rs12203592	locus h2
rs4959270	locus h3	rs683	locus h4
rs1042602	locus h5	rs12821256	locus h6
rs2402130	locus h7	rs1800407	locus h8
rs312262906	locus h9	rs1805005	locus h10
rs1805006	locus h11	rs2228479	locus h12
rs11547464	locus h13	rs1805007	locus h14
rs201326893	locus h15	rs1110400	locus h16
rs1805008	locus h17	rs885479	locus h18
rs1805009	locus h19	rs2378249	locus h20
rs16891982	locus h21	rs12913832	locus h22

Omskrivningen av ögon- och hårfärgerna gjordes genom att beteckna varje enskilt utfall med ett tal. Vilket tal som väljs för vilken färg spelar i sig ingen roll men för enkelhetens skull valdes de naturliga talen i stigande storleksordning, relationen eller i detta fallet storleksordningen på talen har i sig heller ingen påverkan av analysen. Exempelvis i en modell som ger sannolikheter för hårfärgerna brunt, blont, rött och svart anges hårfärgerna med respektive värden: 0, 1, 2 och 3.

För att transformera basparsvariationerna från DNA-datan till numeriska värden beräknades frekvensen av vissa specifika baspar vid varje locus. För att förtydliga, varje locus, exempelvis locus 2, innehåller en viss uppsättning baspar från två alleler. För just locus 2 kan den innehålla någon kombination av basparen (G,C) eller (A,T). Den numeriska omvandlingen görs nu genom att räkna hur ofta basparet (i det här fallet (G,C)) förekommer hos varje individ. Detta ger tre möjliga resultat beroende på antalet gånger ett visst baspar förekommer på denna plats:

1. Värde 0: Om en individ inte har några av det specifika basparet (G,C) på någon av sina alleler.
2. Värde 1: Om en individ har basparet (G,C) på en av sina alleler, men inte på den andra.
3. Värde 2: Om en individ har basparet (G,C) på båda sina alleler.

Alltså, för locus 2 kan följande kombinationer vara möjliga:

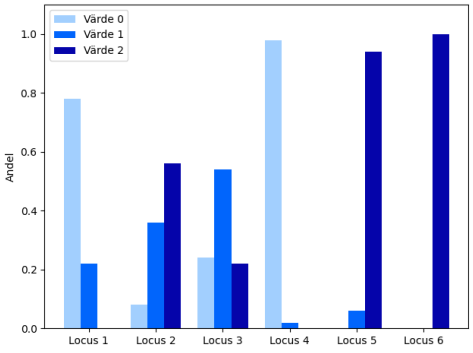
- **(A,T)** och **(A,T)** → Värde 0 (Inget G,C)
- **(G,C)** och **(A,T)** → Värde 1 (En G,C och en A,T)
- **(G,C)** och **(G,C)** → Värde 2 (Två G,C-par)

Detta ger en numerisk representation av varje locus, där värdena 0, 1 eller 2 återspeglar hur ofta ett specifikt baspar förekommer på den platsen.

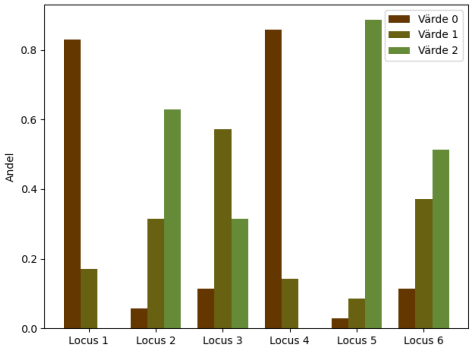
All data normaliserades med hjälp av z-normalisering, detta görs för att se till att variablerna är jämförbara med varandra och att all data har samma skala [22, kap 11.5.4]. Varje datapunkt x omvandlades enligt ekvationen

$$z = \frac{x - \mu}{\sigma} \tag{8}$$

där μ är medelvärdet och σ^2 är stickprovsvariansen för samtliga värden i samma kolumn. Resultatet är att varje variabel får ett medelvärde på 0 och en standardavvikelse på 1, vilket bidrar till en stabilitet i modellen och ser till att variabler med en större numerisk skala inte dominerar beräkningar.



Figur 4: Visualisering över datavärden för individer med blå ögon.



Figur 5: Visualisering över datavärden för individer med icke-blå ögon.

Figur 4 och 5 visar hur ofta ett specifikt baspar förekommer för varje locus för personer med blå respektive icke-blå ögon. Visualiseringen i figur 4 visar att det finns en tydlig koppling mellan att ha värde noll på locus 4, värde 2 på locus 5 och värde två på locus 6 och att ha blå ögon. Men om detta jämförs med figur 5 syns det tydligt att det är samma kriterier för locus 4 och 5 som är viktiga för att en person ska ha icke-blå ögon. Det locus som skiljer sig mest mellan blå och icke-blå ögon och alltså locus 6.

3.2 Prediktionsmodeller för ögonfärg

Arbetet började med att konstruera en prediktionsmodell för ögonfärg eftersom den skulle bli mindre komplex än hårfärgsmodellen. Detta eftersom datamängden för ögonfärg innehöll färre datapunkter, endast sex stycken loci.

3.2.1 Härledning av modellgrund

Modeller byggs genom att bestämma fördelningen $\pi(D_{ny}|D)$ där D är angiven data och D_{ny} är DNA-data för en ny person vars ögonfärg ska förutsägas. Genom att introducera parametern $\theta \in \mathbb{R}^n$, som i det här arbetet symboliserar vikten av den numeriska representationen för varje loci, ges ekvationen

$$\pi(D_{\text{ny}}|D) = \int_{-\infty}^{\infty} \pi(D_{\text{ny}}, \theta|D)d\theta = \int_{-\infty}^{\infty} \pi(D_{\text{ny}}|\theta)\pi(\theta|D)d\theta,$$

som med ekvation (3) ger oss

$$\pi(D_{\text{ny}}|D) = \int_{-\infty}^{\infty} \pi(D_{\text{ny}}|\theta)\frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}d\theta.$$

Denna integral går att beräkna genom att applicera logistisk regression för att bestämma $\pi(D_{\text{ny}}|\theta)$ och $\pi(D|\theta)$.

$$\pi(D_{ny}|\theta) = \phi(D_{\text{ny}} \cdot \theta).$$

Likelihoodfunktionen $\pi(D|\theta)$ är samma som i ekvation (5) och $\pi(D)$ ges av

$$\pi(D) = \int_{-\infty}^{\infty} \pi(D|\theta)d\theta.$$

Det slutgiltiga uttrycket blir därmed

$$\pi(D_{\text{ny}}|D) = \int_{-\infty}^{\infty} \phi(D_{\text{ny}} \cdot \theta)\frac{\pi(D|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} \pi(D|\theta^*)d\theta^*}d\theta. \tag{9}$$

Integralen i ekvation (9) kan approximeras med hjälp av Monte Carlo-integration från avsnitt 2.4.1. Uttrycket blir då

$$\pi(D_{\text{ny}}|D) \approx \frac{1}{n}\sum_{i=1}^{n}\pi(D_{\text{ny}}|\theta_i), \tag{10}$$

där θ_i är stickprov ur posteriorifördelningen $\pi(\theta|D)$.

3.2.2 Grundläggande modell

Som en utgångspunkt för arbetet, och för att få en grundläggande förståelse för datans struktur, utvecklades en enkel prediktionsmodell. Ett av delmålen med den här modellen var att identifiera vilka av de sex givna loci som har störst inverkan på fenotypen blåa ögon.

Tidiga observationer i vårt arbete, som illustreras i figur 4 och 5, visade att det var locus 6 som hade en tydlig påverkan över modellens förmåga att förutsäga ifall en person borde ha blå ögon eller inte. Detta fynd stöttades även av tidigare forskning som också har visat att locus 6 har en stark koppling till ögonfärg [17], [23], [24]. Därför fattades beslutet att börja med en modell som endast använder locus 6.

Modellen konstruerades genom att transformera datan enligt avsnitt 3.1 och färgerna klassificerades endast i två kategorier: blåa ögon och icke-blåa ögon. För att få fram startpunkter användes metoden som presenteras i avsnitt 2.2.1 genom att maximera log likelihood-funktionen från ekvation (6), maximum-likelihood värdena kallas $\theta_{1,:} = (\theta_{1,0}, \theta_{1,1})$, där $\theta_{1,0}$ representerar skärningspunkten och $\theta_{1,1}$ representerar koefficienten för locus 6. Därefter tillämpades Metropolis-Hastings algoritmen, vilken förklaras i avsnitt 2.6, med $\theta_{1,:}$ som startpunkt för att få nya iterationer $\theta_{i,:}$ och därmed en Markov kedja. Nya kandidatpunkter genererades genom att sätta $\theta^* = \theta_{j-1} + \epsilon$ där ϵ är en vektor med längd två och vars element är oberoende och normalfördelade runt 0. Resultatet för modellen togs fram genom att beräkna sannolikheten för blå ögon givet de möjliga värdena för locus 6, alltså 0, 1 eller 2.

3.2.3 Numerisk modell

För att verifiera att det inte fanns några beräkningsfel i den grundläggande modellen skapades även en modell som kunde beräkna integralen i ekvation (9) numeriskt. Men på grund av den inre integralen i nämnaren, är det endast möjligt att genomföra denna beräkningen numeriskt i lägre dimensioner då den blir alltför beräkningstung i höga dimensioner. Beräkningen utgick därför från det locus som har störst påverkan på modellen, locus 6, och använde således två dimensioner för θ, en för skärningspunkten och en som koefficient för locuset.

3.2.4 Modell för blåa ögon

Efter att den grundläggande modellen fungerade relativt bra utvecklades den för att kunna hantera alla sex loci. Startpunkterna togs fram på samma sätt som i den grundläggande modellen fast denna gång i en sjudimensionell vektor $\theta_{1,:} = (\theta_{1,0}, \ldots, \theta_{1,6})$. Från Metropolis-Hastings algoritm togs kandidatpunkter fram genom att sätta $\theta^* = \theta_{j-1} + \epsilon$ där ϵ har samma funktion som innan men nu är den istället en sjudimensionell vektor. För att begränsa storleken på modellparametrarna och därmed minska risken för överanpassning användes en normalfördelad prior i modellen. Komponenterna i priorfördelningen för parametern θ är fördelade enligt

$$\theta_i \sim N(0, \sigma^2),$$

där σ är standardavvikelsen. Väntevärdet valdes till 0 eftersom parametrarna uppfattades som små. Standardavvikelsen σ valdes till 5, vilket tillåter variation i parametrarna utan att tillåta alltför stora värden. Denna avvägning ger modellen flexibilitet att anpassa sig till data, men samtidigt en viss begränsning för att undvika att den överanpassar sig till brus och blir instabil. För den slutgiltiga valideringen av modellen användes LOOCV, vilket förklaras i avsnitt 2.8.

3.2.5 Modell för samtliga ögonfärger

För att utöka modellen till att omfatta alla ögonfärger användes den generaliserade logistiska ekvationen i ekvation (2) med $s = 2$. Parametern θ är nu en 2×7 matris där varje rad motsvarar en ögonfärg och varje kolumn ett locus. Dessa två rader av θ-värden används för att ge sannolikheten för att en viss person har ögonfärgen blå respektive intermed. För att få sannolikheten för att en person ska ha bruna ögon användes det faktum att summan av sannolikheterna för de olika utfallen är 1, vilket gör att sannolikheten för bruna ögon kan beräknas med

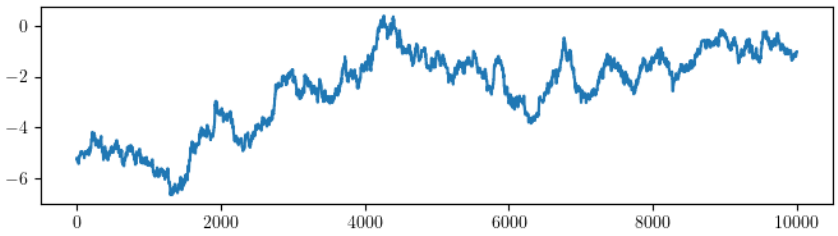
$$\begin{aligned} P(\text{Blå}|\theta, D_i) &= \phi(D_i \cdot \theta)_1 =: P_1 \\ P(\text{Intermed}|\theta, D_i) &= \phi(D_i \cdot \theta)_2 =: P_2 \\ P(\text{Brun}|\theta, D_i) &= 1 - \phi(D_i \cdot \theta) - \phi(D_i \cdot \theta)_2 =: P_3. \end{aligned}$$

Detta kan skrivas med en log likelihood från avsnitt 2.2.1

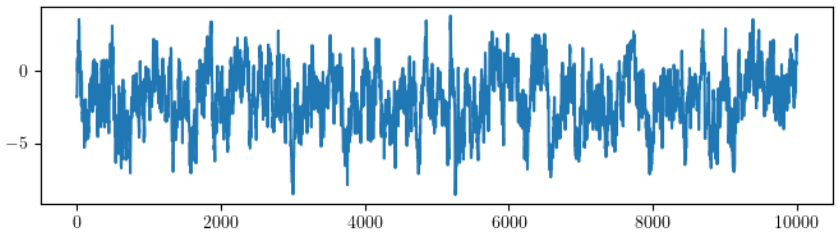
$$\log(\pi(D|\theta)) = \sum_{i=1}^{85} \log(P_{z_i}),$$

där z_i är 1, 2 eller 3 om person i har blå, intermed eller bruna ögon. Från denna bestämdes θ_1 genom maximering med avseende på θ. Som tidigare valdes nya kandidatpunkter genom att sätta $\theta^* = \theta_{j-i} + \epsilon$, där ϵ är en vektor med 14 punkter vars element är normalfördelade enligt $N(0, \sigma)$. För att hjälpa kedjan att konvergera snabbare valdes priorfördelningen $N(0, 3)$ för θ.Modellen testades med LOOCV prediktioner.

Efter testning och undersökning av modellen upptäcktes dålig blandning inom Markovkedjorna vilket innebär att kedjorna inte utforskar fördelningen för parametern θ på ett bra sätt. För att åtgärda detta ändrades förslagsfunktionen för nya kandidatpunkter. Istället för att sätta $\theta^* = \theta_{j-1} + \epsilon$, $\epsilon \sim N(0, \sigma^2)$ så användes en multivariat normalfördelning centrerad i θ_{j-1} vars kovariansmatris kom från en Laplace approximation av posteriorifördelningen för θ vilket presenteras i avsnitt 2.7. En jämförelse av kedjornas första dimension med de olika förslagsfunktionerna visas i Figur 6. Det syns tydligt att den nya förslagsfunktionen rör sig snabbare genom fördelningen.



(a) Kedja utan Laplace-approximerad förslagsfunktion.



(b) Kedja med Laplace-approximerad förslagsfunktion

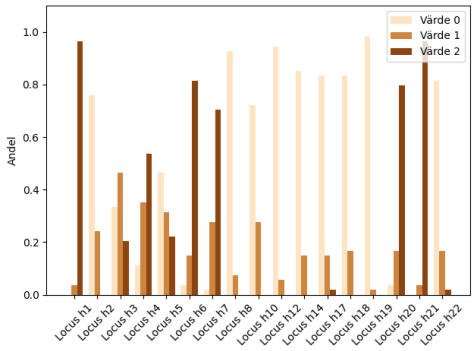
Figur 6: Jämförelse av kedjorna för dimension 1 i θ med och utan Laplace approximationen som förslagsfunktion.

3.3 Prediktionsmodeller för hårfärg

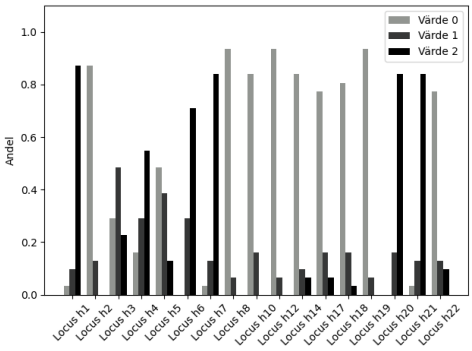
Efter att ha fått en förståelse för metodiken genom utvecklingen av modellerna för ögonfärger anpassades modellerna för att hantera hårfärger. Samma modellgrund används för att predicera

hårfärg. Den största skillnaden i att predicera hårfärg jämfört med ögonfärg är att det nu finns fyra olika utfall istället för tre och att den har fler parametrar då den använder sig av 22 istället för sex stycken loci.

Istället för att skapa en grundläggande modell för att hitta vilken eller vilka loci som är mest viktiga, användes istället en visualisering av datan. Genom visualiseringen, som visas i figur 10 i appendix B, upptäcktes det att vissa locus hade samma värde för alla personer i den givna datamängden oavsett hårfärg. Därför exkluderades samtliga av dessa loci från modellen då de inte hade någon påverkan på modellens prestation. Dessa loci var locus h9, locus h11, locus h13, locus h15 och locus h16. Dessutom upptäcktes det att alla personer utom en hade samma värde i locus h1 som i locus h21. Därför valdes även att locus h1 skulle exkluderas.



Figur 7: Visualisering över värden för brunt hår



Figur 8: Visualisering över värden för icke-brunt hår

Även modellen som predicerar samtliga hårfärger utgick från modellerna för ögonfärger. Båda modellerna använder sig av samma ovannämnda loci, den enda skillnaden är att antalet klasser som ska prediceras nu är tre, brunt, blont och rött hår där sannolikheten för svart hår beräknas på samma sätt som i avsnitt 3.2.5.

Till skillnad från modellerna för ögonfärger utvecklades även en modell som analyserar samtliga hårfärger exklusive brunt hår. Modellen fungerar likadant som modellen som predicerar samtliga hårfärger förutom att personer med brunt hår exkluderades från datan.

4 Resultat

I detta avsnitt presenteras modellernas prestanda genom tabeller och förväxlingsmatriser. Först presenteras den grundläggande modellen och resultatet från numeriska beräkningar. Efter det presenteras modeller som predicerar ögonfärg och slutligen presenteras modeller som predicerar hårfärg.

4.1 Grundläggande modell och numeriska beräkningar

Eftersom den grundläggande modellens resultat verifierades med numeriska beräkningar kommer resultatet från båda dessa modeller att presenteras samtidigt.

Tabell 4: Sannolikheter för blåa ögon givet värde i locus 6.

Värde	Sannolikhet
0	0.00063
1	0.0031
2	0.897

Tabell 5: 2x2 Förväxlingsmatris för den grundläggande modellen.

Pred → Sann ↓	**Icke-blåa ögon**	**Blåa ögon**
Icke-blå ögon	17	18
Blåa ögon	0	50

Resultatet från de numeriska beräkningarna visar att samtliga som har värde 2 har en hög sannolikhet att klassificeras som blåögda personer, medan de som har värde 1 eller 0 har en väldigt låg sannolikhet att klassificeras som blåögda. I praktiken innebär detta att alla individer som har värde 2 för locus 6 klassificeras som blåögda. Den grundläggande modellens resultat presenteras i en förväxlingsmatris som visas i tabell 5. Den visar att modellen är bra på att upptäcka blå ögon, men har problem med att särskilja blå och icke-blå ögon och överklassificerar icke-blåögda som blåögda i viss mån.

4.2 Prediktionsmodell för blåa ögon

Förväxlingsmatrisen i tabell 6 visar resultatet från tio upprepade analyser av LOOCV där, varje kedja har en längd på 10 000. Alla medelvärden (μ) och standardavvikelser (σ) från de upprepade analyserna presenteras i förväxlingsmatrisen på följande vis $\mu \pm \sigma$. Resultatet visar att modellen fortfarande är bra på att predicera om en person har blå ögon, dock med något sämre resultat än den grundläggande modellen. De låga standardavvikelserna visar att modellen är stabil i sina prediktioner. McMC-kedjor och modellens prestanda presenteras i appendix B.

Tabell 6: Förväxlingsmatris för modell för blåa ögon.

Pred → Sann ↓	**Icke-blå (0)**	**Blå (1)**
Icke-blå (0)	18.4 ± 0.49	16.6 ± 0.49
Blå (1)	2.0 ± 0.48	48 ± 0.48

4.3 Prediktionsmodell för samtliga ögonfärger

Förväxlingsmatrisen i tabell 7 visar hur ofta modellen förväxlar de olika ögonfärgerna.

Tabell 7: Förväxlingsmatris för modell med samtliga ögonfärger.

Pred → Sann ↓	**Blå (0)**	**Intermed (1)**	**Brun (2)**
Blå (0)	47.7 ± 0.64	2.3 ± 0.64	0.0 ± 0.0
Intermed (1)	16.0 ± 0.0	5.9 ± 1.22	3.1 ± 1.22
Brun (2)	0.0 ± 0.0	4.1 ± 0.7	5.9 ± 0.7

Tabell 7 visar resultatet från tio upprepade analyser av LOOCV, där varje kedja har en längd på 10 000. Modellen visar en hög precision för att korrekt klassificera blå ögon, med en låg förväxlingsgrad med intermed och ingen alls med brun. För intermed ögon kan vi se att majoriteten av individerna har klassificerats ha blå ögon av modellen och att det är samma 16 individer som förväxlas varje gång eftersom standardavvikelsen är 0. Modellen har även problem att skilja på bruna och intermed ögon vilket syns i matrisens högra nedre del. Standardavvikelserna visar att modellen är stabil i sina prediktioner och att det finns ganska små variationer i förväxlingsgraden. De simulerade kedjorna visas, upp till dimension tre, i figur 9 i appendix B.

4.4 Prediktionsmodell för brunt hår

Prediktionsmodellen för brunt hår resulterar i förväxlingsmatrisen, som återges i tabell 8. Matrisen visar att modellen har svårt att korrekt predicera icke-brunt hår, vilket leder till både många falska negativa och falska positiva prediktioner.

Tabell 8: Förväxlingsmatris för modell för brun hårfärg.

Pred → Sann ↓	**Icke-brunt hår**	**Brunt hår**
Icke-brunt hår	5.5 ± 1.28	25.5 ± 1.28
Brunt hår	15.5 ± 1.03	38.5 ± 1.03

Tabell 9: Prestandamått för modell för brunt hår.

Accuracy	52%
Precision	60%
Recall	71%
Specificity	18%

Tabell 15 i appendix B visar prestandamåtten för modellen som predicerar brunt hår. Här kan ses en tydlig skillnad mellan precision och specificitet, vilket tyder på att modellen ofta identifierar personer med brunt hår men sällan korrekt identifierar dem utan brunt hår.

4.5 Prediktionsmodell för samtliga hårfärger

Förväxlingsmatrisen för modellen som predicerar samtliga hårfärger visas i tabell 10. Även här framgår att modellen ofta predicerar brunt hår för personer som i själva verket har en annan hårfärg. Samtliga personer med blont hår prediceras ha brunt hår, medan modellen predicerar rätt vid fyra tillfällen och fel vid tio för svart hår. Personer med rött hår prediceras rätt ungefär hälften av gångerna och prediceras ha brunt hår resterande gånger. Modellen uppnår en accuracy på 67.06%.

Tabell 10: 4x4 Förväxlingsmatris för prediktionsmodell för samtliga hårfärger.

Pred → Sann ↓	**Brunt hår**	**Blont hår**	**Rött hår**	**Svart hår**
Brunt hår	43.4 ± 1.56	4.4 ± 0.66	1.8 ± 0.75	4.4 ± 0.92
Blont hår	8.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	2.0 ± 0.0
Rött hår	3.3 ± 0.9	0.0 ± 0.0	3.7 ± 0.9	0.0 ± 0.0
Svart hår	10.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	4.0 ± 0.0

Tabell 11 visar tydligt att modellen har betydligt bättre förmåga att predicera brunt hår jämfört med övriga hårfärger. För brunt hår uppnår modellen en precision på 67% och en recall på 80%, vilket innebär att majoriteten av de som prediceras ha brunt hår faktiskt har det, samt att nästan alla personer med brunt hår korrekt identifieras. För blont hår är dock både precisionen och recall 0%, vilket indikerar att modellen helt misslyckas med att identifiera blonda personer, varken några riktiga positiva träffar hittas eller några korrekta prediktioner görs. Modellen lyckas i viss mån identifiera röd- och svarthåriga men inte med samma säkerhet som brunhåriga. Resultatet visar alltså en tydlig snedfördelning där modellen är partisk mot att identifiera brunt hår.

4.6 Prediktionsmodell för blont, rött och svart hår

När brunt hår exkluderas från modellen förbättras resultaten markant. Förväxlingsmatrisen för modellen som predicerar endast icke-bruna hårfärger återges i tabell 12. En jämförelse mellan

Tabell 11: Prestandamått för samtliga hårfärger.

Hårfärg	Precision	Recall
Brunt hår	67%	80%
Blont hår	0%	0%
Rött hår	67%	53%
Svart hår	38%	29%

prestandamåtten för modellen med brunt hår i tabeller 11 och 13 visar att modellen som exkluderar brunt hår är betydligt bättre på att särskilja de icke-bruna hårfärgerna. Precisionen är högre för både blont och svart hår men något sämre för rött hår. Recallvärdena kunde dock vara bättre då modellen endast identifierar 31% av alla med blont hår, 77% av alla med rött hår och 59% av alla med svart hårt, men identifierar ändå fler jämfört med tidigare modeller. Sammanfattningsvis visar denna jämförelse att modellen utan brunt hår presterar betydligt bättre för de icke-bruna färgerna än vad den ursprungliga modellen gjorde.

Tabell 12: 3x3 Förväxlingsmatris för prediktionsmodell för blont, rött och svart hår.

Pred → Sann ↓	**Blont hår**	**Rött hår**	**Svart hår**
Blont hår	3.1 ± 0.3	2.7 ± 0.46	4.2 ± 0.4
Rött hår	1.6 ± 0.66	5.4 ± 0.66	0.0 ± 0.0
Svart hår	5.0 ± 0.0	0.8 ± 0.4	8.2 ± 0.4

Tabell 13: Prestandamått för samtliga hårfärger när brunt hår är exkluderat.

Hårfärg	Precision	Recall
Blont hår	32%	31%
Rött hår	61%	77%
Svart hår	66%	59%

5 Diskussion

5.1 Grundläggande modell och numeriska beräkningar

Den grundläggande modellen visade ett bra resultat, men den var för påverkad av den data som den tränades på. Modellen lärde sig i praktiken endast att om en individ har värde 2 på locus 6 så förutsägs den personen ha blå ögon. Detta stämmer dock inte i verkligheten, det stämmer inte heller om man endast kollar på den tillgängliga data som finns. I datamängden finns det 18 personer med värde 2 men som klassificeras som intermed. Det är dessa individer som kan ses i figur 5 som klassas som icke-blå och har värde 2. Trots att resultatet från modellen framstår som bra är det inte ett önskvärt beteende i en prediktiv modell. Från figurer 4 och 5 går det att se att alla som har värde 2 kommer klassificeras som blåögda.

5.2 Modeller för ögonfärg

Resultatet från modellen som predicerar blå ögon och inkluderar alla loci är något sämre jämfört med den grundläggande modellen. Detta beror förmodligen på att modellen blir påverkad av de övriga fem loci som lades till. Dessa fem loci kan eventuellt vara mer informativa för att förutsäga brun eller intermed men är inte så informativa för blåa ögon.

Modellen som predicerar samtliga ögonfärger har vissa svårigheter med att skilja blå och brun från intermed. Detta innebär att modellen ibland felaktigt klassificerar en brun- eller blåögd individ som intermed eller tvärtom, vilket syns i tabell 7. Däremot är modellen betydligt bättre på att

skilja mellan blåögda och brunögda personer. Att modellen har problem med att skilja på blå och intermed samt brun och intermed är inte förvånande då det även är väldigt svårt att skilja dessa åt genom att bara undersöka datan. Det finns exempelvis två individer med identisk DNA uppsättning men den ena klassas som blå och den andra som intermed. Det hade därför varit värdefullt att ha tillgång till visuella representationer av de fall som klassas som intermed. Detta för att lättare kunna validera ifall modellens klassificering faktiskt motsvarar verkligheten.

Vid närmare studier av de 16 personer som har intermed ögon men alltid klassificeras som blåögda av modellen syns att de alla har en sak gemensamt. Alla har värde 2 på locus 6. I resultatet för de enkla ögonfärgsmodellerna, kap 4.1, noterades att locus 6 i särklass är det viktigaste locuset för att avgöra om en person har blåa ögon eller inte och därför är det inte så konstigt att modellen gör detta felet. Med fler individer eller SNPs i datan skulle det kanske vara möjligt att klassificera dessa individer bättre, men det skulle även kunna finnas yttre anledningar till att dessa personer har intermed ögon.

5.3 Modeller för hårfärg

Modellen som predicerar brunt hår har svårt att särskilja icke-brunt hår från brunt hår. En möjlig förklaring till detta skulle kunna vara att den tillgängliga datan har en överrepresentation av personer med brunt hår, vilket innebär att modellen har tränats mer på denna hårfärg än på de övriga. Samma problem återkommer i modellen som predicerar samtliga hårfärger, även den har stora svårigheter att skilja på de övriga hårfärgerna med brunt hår vilket är tydligt i tabell 10. Modellen har mest problem med att hitta blont hår. Detta skulle kunna bero på att anlaget för blont hår är en recessiv gen och både mörkt och rött hår är en dominant genetisk egenskap [25].

En person som har svart hår kan bära på anlaget för blont men anlaget för svart hår dominerar. På grund av detta finns det en risk att modellen förknippar de blonda anlagen med andra färger. Tillsammans med faktumet att det endast är tio individer som har blont hår, leder detta till att modellen inte får träna tillräckligt mycket på att identifiera blonda (recessiva) anlag.

Modellen som bara hanterar blont, rött och svart hår får bättre prediktioner än de tidigare för de relevanta färgerna. Detta beror antagligen på att den överrepresenterade klassen brunt hår har tagits bort vilket tillåter modellen att lära sig mer om de specifika anlag som krävs för exempelvis blont hår. Dock presterar modellen fortfarande ganska dåligt. En möjlig förklaring till detta är att datamängden är ännu mindre, endast 31 datapunkter. Trots användning av LOOCV för utvärdering är detta väldigt lite data att utgå från. En annan möjlig förklaring är att det finns andra genotyper som påverkar hårfärgerna som inte är med i datamängden.

En annan, något mer spekulativ, anledning till att modellerna inte presterar tillfredsställande skulle kunna vara att hårfärger kan ändras med ålder. Vissa personer som i vuxen ålder uppfattas ha brunt hår hade som barn blont, eller ibland vitt hår. Detta är även något som Andreas Tillmar, från Rättsmedicinalverket i Linköping, nämner i en intervju [7]. Han menar att detta, tillsammans med andra miljöfaktorer som exempelvis solexponering, är en av svårigheterna med att predicera hud- och hårfärg. Om detta är fallet skulle det innebära att några datapunkter för individer med brunt och blont hår eventuellt har ett visst överlapp och att det är en av anledningarna för modellens svårigheter att skilja på just brunt och blont hår.

En ytterligare spekulation till varför modellen har svårt att skilja hårfärgerna åt kan vara att det även är svårt att skilja dessa åt i verkligheten. På en skala, är det exempelvis inte helt tydligt var blont hår övergår till brunt hår eller var brunt hår övergår till svart hår. Det är möjligt att detta reflekteras i genetiken och övergången mellan de olika färgerna inte är särskilt tydliga.

5.4 Osäkerheter och utmaningar

När det gäller statistik och prediktioner kommer det alltid att finnas en viss osäkerhet kring resultaten. Även om en modell uppnår hög precision är det viktigt att se prediktionerna som sannolikheter snarare än absolut sanning. Modeller ger en uppskattning av de mest troliga resultaten baserat på tillgänglig data, men de kan aldrig förutsäga ett resultat med fullständig säkerhet. Detta innebär att även om modellen indikerar en hög sannolikhet för ett visst resultat, kommer det alltid att

finnas en risk för felaktigheter eller avvikelser från de förväntade resultaten. Ett exempel på detta i arbetet skulle kunna vara att ögonfärgen hos en viss person kan påverkas av andra SNP:er som inte inkluderades i denna rapport. Ingen modell kan fånga alla variationer i den verkliga världen. Det finns alltid okända faktorer eller influenser som inte syns i DNA:t som kan påverka utseende såsom skador, åldrande, kosmetiska ingrepp eller andra åkommor, och därför kommer det att finnas en viss grad av osäkerhet i varje prediktion. Ett känt exempel på yttre påverkan av hur en persons ögonfärg uppfattas är artisten David Bowies vänstra öga. Bowies vänstra pupill var permanent utvidgad efter en skada i ett slagsmål [26]. Hans vänstra öga uppfattas därför ofta som helt svart vilket skapar en stark kontrast mot hans blåa högra öga. På grund av osäkerheterna med yttre påverkan så är det viktigt i ett rättsligt sammanhang att denna typ av modeller inte används som bevismaterial i sig, utan som ett verktyg som utredare kan använda sig av som stöd i den initiala utredningsprocessen.

En av de största osäkerheterna i det här arbetet är datan i sig som består av angivna ögon- och hårfärger samt loci och SNP. Ögon- och hårfärgerna är också klassificerade av personerna själva och det finns inte något sätt att verifiera hur de faktiskt ser ut. Det finns alltså en osäkerhet över hur varje person tolkar sin egna ögon- och hårfärg. En studie från år 2000 visade att människor generellt uppfattar färger på liknande sätt, men att det finns stor variation i hur nyanser tolkas [27]. Detta innebär att subjektiva bedömningar kan påverka både träningen av modellen och hur man förstår dess resultat beroende på vem det är som tolkar det. Detta kommer att påverka modellerna i det här arbetet, som i sin tur kommer att påverka prediktionerna. Det är alltså oklart för oss hur tydliga gränserna är mellan de olika färgklasserna.

I datan som anger ögonfärg finns det även en osäkerhet i klassificeringen, det finns en ögonfärg som klassas intermed. Det är alltså en färg som inte är brun eller blå, detta kan betyda att ögonfärgen är grön, blå-grön, brun-grön eller blå med bruna fläckar. Allt som inte är just blå eller brun. I det här arbetet blir en sådan färg problematisk då det inte är möjligt att verifiera exakt hur de specifika intermedfärgerna ser ut. Även modellen får svårt att predicera intermed som en färg då vissa eventuellt är närmare blå än brunt eller tvärtom. Hårfärgerna hos personerna är självskattade enligt Fischer-Saller skalan. Men det är, precis som med ögon-datan, osäkert var gränsen mellan de olika nyanserna går. Att hårfärgerna skattas enligt en skala är rimligt, men eftersom datan redan var klassificerad innan den överlämnades till det här arbetet är det svårt att veta var gränserna dragits.

En annan osäkerhet gäller DNA-spår där endast en liten mängd DNA finns tillgängligt. När DNA-mängden är låg kommer analyserna att bli mer osäkra vilket i sin tur ökar risken för feltolkningar. Det kan även uppstå problem vid insamlingen av DNA från brottsplatsen. Med detta menas att DNA vanligtvis är kontaminerat på en brottsplats, oftast är DNA-spår tagna från blod, hud, hår eller liknande, vilket kan ha smuts i sig. Det kanske till och med kan vara så att det är ihopblandat med andras DNA på en brottsplats. För att motverka detta brukar polisen även ta DNA från målsägande för att använda som ”elimineringsDNA” [6]. Ett annat problem är även att DNA kan bli kontaminerad vid hanteringen, antingen redan vid insamling eller vid analyser.

5.5 Potentiella användningsområden

Det här arbetet har från början utgått från att tekniken ska kunna användas inom forensisk verksamhet och det finns flera områden inom forensisk verksamhet som skulle kunna dra nytta av en sådan teknik. Ett av de viktigaste uppgifterna tekniken kan hjälpa till med är att ge kalla fall en chans till uppklarning. Genom att använda tekniken i fall där det saknas en känd gärningsperson, men där biologiskt material har säkrats från brottsplatsen, kan utredare få nya spår att följa och därmed en ny riktning i en annars ouppklarad utredning.

En annan potentiell tillämpning inom forensisk verksamhet är inom identifiering av offer i samband med brottsutredningar, naturkatastrofer eller terrorattacker. Detta kan gälla exempelvis fall där kroppen är svårt skadad till följd av dödsorsaken eller där det saknas anhöriga i landet som kan hjälpa till vid identifiering. Om ingen anhörig finns tillgänglig för att identifiera kroppen, kan en prediktion av personens fenotyp användas för att skapa en sannolik bild eller beskrivning av personen. Denna bild eller beskrivning kan sedan spridas via media eller polisens nätverk, vilket i

sin tur kan leda till tips från allmänheten. Det kan också vara så att brottsoffret saknar identifikationsbevis och inte är registrerad i svenska databaser, men ändå finns med i andra internationella register. Här kan en fenotypbaserad profil spridas till utländska myndigheter och därefter ge dem ett bättre underlag för att göra kopplingar till vad som efterlyses.

Ytterligare en potentiell användning av tekniken är att initialt smalna av utredningar som saknar en direkt misstänkt. Om brottsplatsen innehåller DNA från gärningspersonen kan fenotypiska prediktioner hjälpa polisen att prioritera vilka individer som bör utredas först, särskilt i väldigt breda eller resurskrävande utredningar. Även här måste man vara kritisk till modellen. Då den aldrig kan vara helt säker är det viktigt att inte blint lita på resultatet av analysen. Det finns en risk att viktiga spår missas, eller att utredningen leds bort från den faktiska gärningspersonen på grund av felaktiga antaganden. Tekniken bör därför ses som ett komplement till och inte en ersättning för vanliga utredningsmetoder. Vidare användningsområden är att skapa digitala fantombilder. Här krävs mer information än bara ögon- och hårfärg men det skulle vara möjligt att, genom att analysera DNA från okända individer, rekonstruera deras utseende och därigenom underlätta identifieringsprocessen. Dessa fantombilder skulle kunna användas i de fall där det inte finns något ögonvittne.

Ett användningsområde som inte är kopplat till forensisk verksamhet är att fenotypiska prediktioner i framtiden kan användas vid insemination eller IVF (in vitro-fertilisering). Vid assisterad befruktning kan föräldrar ibland välja mellan olika donatorer och vid IVF sker befruktningen i ett provrör. Med hjälp av fenotypisk prediktion kan man tänka sig att tekniken skulle kunna användas för att skapa en uppskattning av hur ett barn potentiellt kan se ut. Detta skulle exempelvis kunna ske genom att jämföra barnets DNA och potentiella utseende för respektive donator. Detta skulle kunna ge föräldrar en mer ”visuell” förståelse för genetiska utfall och kan vara av intresse för de som önskar att barnet liknar dem själva.

5.6 Samhälleliga och etiska aspekter

Användningen av DNA-baserad fenotypisk prediktion för att bestämma egenskaper såsom hår- och ögonfärg innebär både potentiella fördelar och nackdelar. Dessa aspekter är viktiga att noggrant utvärdera för att säkerställa en rättssäker och etiskt försvarbar implementering av tekniken.

5.6.1 Risk för diskriminering

Teknikens osäkerheter kan innebära risker när det kommer till rättssäkerheten, framförallt om den används som ett huvudsakligt bevis i en brottsutredning. Det finns en risk att en prediktion av en profil leder till att utredningar riktas felaktigt. Detta kan leda till att fel person utreds eller att andra alternativ förbises och att fallet förblir ouppklarat.

Något att beakta i utvecklingen av en teknik som kan predicera ögon- och hårfärg är risken för partiskhet, särskilt när det gäller olika etniska grupper. En liknande studie, vilken använde sig av samma loci som i det här arbetet, visade att prediktionen av ögonfärg var mer säker för européer och närliggande populationer [9]. En trolig orsak till detta är att den utvalda genotyp-datan huvudsakligen kommer ifrån personer med ett europeiskt ursprung.

Denna typ av partiskhet kan leda till diskriminering genom att individer från vissa populationer inte får användning av teknikens fulla nytta. Om en prediktionsmodell är mindre träffsäker för personer med ett visst ursprung kan det leda till att deras fysiska egenskaper inte identifieras med samma säkerhet. Detta innebär att tekniken kan vara mindre användbar för att exempelvis identifiera brottsoffer eller saknade personer i dessa grupper. En annan central aspekt är integriteten och den potentiella diskrimineringen som kan uppstå. Om den genetiska informationen inkluderar fler drag såsom hudfärg eller geografiskt ursprung, kan det leda till att vissa genetiska grupper utsätts för ökad övervakning och orättvist misstänkliggörande [16].

Ur ett juridiskt perspektiv skulle en sådan vinkling potentiellt strida mot dataskyddsförordningens (GDPR) princip om korrekthet, som föreskriver att behandling av personuppgifter ska vara rättvis, skälig och proportionerlig i förhållande till den registrerade individen [28]. Principen om rättvisa innebär att personuppgifter ska behandlas på ett sätt som den registrerade rimligen kan förvänta

sig och att hanteringen inte får vara diskriminerande, varken i metod eller i resultat. Om en DNA-baserad prediktionsmodell systematiskt ger sämre resultat för vissa etniska grupper, kan detta betraktas som en diskriminerande effekt av personuppgiftshanteringen, trots att syftet med tekniken inte är att diskriminera. För att motverka dessa problem är det viktigt att framtida studier försöker att ge en mer representativ träningsdata som täcker en bredare genetisk variation genom att samla in DNA från en större del av världens befolkning. Det bör även finnas en tydlig kommunikation i metodik och felmarginal. Detta är framför allt viktigt i rättsliga sammanhang där tekniken kan påverka beslut om misstankar och utredningar.

Men diskrimineringar kan även ske utanför den forensiska verksamheten, tidigare nämndes att en potentiell användning av tekniken skulle kunna vara att predicera hur sitt framtida barn kommer att se ut. Det är viktigt att poängtera att en sådan tillämpning väcker många etiska frågor. Det finns en risk att tekniken används för att göra val baserade på personliga ideal. Ett sådant beteende skulle kunna förstärka normer kring att det finns mer ”önskvärda” utseenden. Om dessa tankar blir mer utbredda skulle det kunna leda till en minskad mångfald i fenotyp och att vissa ”mindre önskvärda” utseenden diskrimineras. Gällande optimeringen av personers genitiska egenskaper finns det så kallade ”designer babies” där föräldrar genom genetiska ingrepp kan välja önskvärda egenskaper hos sina barn, något som även det ställer frågor kring etiken med dessa typer av metoder. En annan aspekt är frågan om det faktiskt är moraliskt försvarsbart att helt och hållet kontrollera skapandet av liv och huruvida detta utsätter foster för fara.

5.6.2 Anpassad lagstiftning

I takt med att tekniker för att predicera fenotypiska egenskaper utvecklas uppstår ett behov av att anpassa den befintliga lagstiftningen. Den nuvarande svenska lagstiftningen är anpassad efter användningen av DNA i identifieringssyfte, exempelvis genom STR-profiler i kriminaltekniska databaser. Att använda tekniker för att förutsäga fysiska egenskaper utan att identifiera en specifik individ regleras inte av dagens lagstiftning. Om en helt ny teknik implementeras kan det leda till en juridisk gråzon, där det saknas tydliga riktlinjer för hur och när dessa verktyg får användas. För att säkerställa en rättssäker och etiskt försvarbar användning av dessa tekniker behövs tydliga ramverk som reglerar hur information från fenotypiska prediktioner ska tolkas och integreras i brottsutredningar. Ett sådant ramverk vore exempelvis att tekniken regleras till att endast användas som ett utredningsstöd och inte som bevis i domstol.

Svensk lagstiftning är under ständig utveckling. Den 1 Juli 2025 träder en ny lag i kraft som möjliggör användning av DNA-baserade släktforskningsdatabaser i brottsutredningar för att identifiera potentiella gärningspersoner [29]. Denna lag kommer att regleras genom att den begränsas till utredning av särskilt grova brott, såsom mord och grov våldtäkt. Lagändringen innebär att biometriska uppgifter kan samlas in, registreras och användas i betydligt större utsträckning än tidigare. Det är Nationellt forensiskt centrum (NFC) som ansvarar för att ta fram en process för DNA-baserad släktforskning. Liknande juridiska ramar skulle kunna utvecklas och användas för fler typer av användningsområden som exmpelvis den typ av modell som tagits fram i det här arbetet.

5.6.3 Hantering av biometriska uppgifter

Eftersom teknikens syfte är att förutsäga fenotypiska egenskaper utifrån DNA i forensiskt syfte, blir hanteringen av dessa uppgifter en central fråga. DNA är en biometrisk uppgift vilka anses vara känsliga personuppgifter i GDPR och dessa kräver ett särskilt starkt skydd [30].

Artikel 9 i GDPR reglerar behandling av känsliga personuppgifter [31]. Det finns dock en del undantag som möjliggör hantering av dessa uppgifter, exempelvis när det är nödvändigt för syften som rör rättsväsendet, såsom brottsutredningar eller fastställande av rättsliga anspråk. För att dessa undantag ska vara tillämpliga krävs att behandlingen sker med tydliga rättsliga grunder, och under strikt reglerade former.

Behandling av biometriska uppgifter ställer höga krav på datasäkerhet. För att skydda individens integritet ska uppgifterna pseudonymiseras och krypteras. Organisationer måste även löpande

säkerställa att systemens konfidentialitet, integritet och tillgänglighet upprätthålls. Regelbundna säkerhetsgranskningar är nödvändiga för att minimera risken för obehörig åtkomst eller dataläckor.

En annan viktig aspekt är individens rätt till radering. Enligt artikel 17 i GDPR har en person rätt att få sina personuppgifter raderade utan dröjsmål [32]. Detta gäller bland annat när personuppgifterna inte längre behövs för de ändamål för vilka personuppgifterna samlats in.

Polisen har ett särskilt register som innehåller DNA från alla personer dömda för brott med påföljd som inte är böter [33]. Enligt rättegångsbalken får polisen samla in DNA från personer som är skäligen misstänkta för brott som kan ge fängelse som påföljd. Dessa uppgifter får sparas i DNA-registret så länge som personen finns kvar i belastningsregistret. Men var skulle Forensisk DNA fenotypings profiler (FDF-profiler) falla in? Det finns ett specifikt register som kallas för fingeravtrycks- och signalementsregistret, där behandlas uppgifter om misstänkta eller dömda personer [33]. Sådana uppgifter får dock behandlas i sjuttio år efter registreringen om de avser utredningar om grova brott.

6 Slutsats

Samtidigt som DNA-baserad fenotypisk prediktion har potential att förbättra brottsutredningar, medför den också betydande vetenskapliga, etiska och juridiska utmaningar. Tekniken måste tillämpas med försiktighet och i kombination med andra utredningsmetoder för att säkerställa att dess användning inte leder till felaktiga domar eller diskriminering. Fortsatt forskning och noggranna etiska överväganden är nödvändiga för att balansera teknikens möjligheter med dess risker.

Ett förslag för att kunna motverka tolkningsosäkerheten som uppstår av de självskattade färgerna är att låta flera människor annotera datan innan den skickas ut. Det hade även kunnat vara en möjlighet att ha fler klasser, speciellt på hår då färgerna jämförs med en skala och där gränsen för de olika färgerna inte är helt tydlig. Det hade exempelvis kunnat vara en kategori för mörkblont och rödbrunt. För ögon hade man kunnat skilja på gröna ögon och blå ögon med bruna fläckar och vice versa istället för samla ihop alla dessa till klassen intermed ögon. Ett förslag för att undvika problemen som uppkommer med recessiva anlag, som blont hår, är att utforska beslutsträdsmodeller. Dessa skulle eventuellt fånga upp de hierarkiska förhållanden som finns bland anlagen bättre än våra statistiska modeller.

En annan viktig aspekt för framtida studier är att använda en större datamängd och en jämnare balans mellan de olika utseendedragen. För att modellen ska kunna göra bättre prediktioner krävs en stor mängd data att träna på. Datan som använts i detta arbete har varit begränsad och framförallt för kategorier som inte är blå ögon eller brunt hår, vilket har lett till svårigheter när det kommer till att predicera dessa. För att kunna göra en ordentlig datainlärning hade det krävts betydligt större mängd data.

Det finns även ett behov av mer forskning om andra loci på DNA:t då vi upptäckte att för ögonfärgerna ser vi personer som har olika färger men samma kombination av baspar i datan. Bland hårfärgerna upptäckte vi att vissa loci inte hade någon variation alls mellan de olika individerna. I en större mängd data hade man eventuellt kunnat upptäcka variationer hos dessa.

Fenotypisk prediktion är baserad på statistiska sannolikheter snarare än absoluta deterministiska samband. Detta innebär att även om en viss genetisk profil indikerar en hög sannolikhet för en viss ögon- eller hårfärg, finns det alltid en grad av osäkerhet. Miljöfaktorer som trauman, åldrande, kosmetiska ingrepp och sjukdomar kan förändra en individs utseende på sätt som tekniken inte har en möjlighet att förutse. Forensisk prediktion kan ge en sannolikhet för vissa genetiska drag, men den är begränsad i det faktum att den inte kan förutse kalkylerade förändringar som exempelvis hårfärgning eller användning av kontaktlinser.

Tekniken har potential att nyttjas i forensiska sammanhang genom att smalna av antalet misstänkta, hjälpa till med identifiering av avlidna och leda brottsundersökningar i rätt riktning. För att så rättssäkert som möjligt kunna implementera tekniken är det däremot av stor vikt att ha förståelse för dess begränsningar och risker.

Referenser

[1] *Forensic Evidence*, [Fotografi], 2016. URL: quest-eb-com.eu1.proxy.openathens.net/images/132_1255305 (hämtad 2025-04-11).

[2] S. Sutton och R. Managed, *DNA*, [Fotografi], 2021. URL: quest-eb-com.eu1.proxy.openathens.net/images/139_3828894 (hämtad 2025-03-23).

[3] R. AARli, "The Status and Meaning of Criminal Procedure: An exploration of the reception of DNA evidence in the criminal process", *Bergen Journal of Criminal Law & Criminal Justice*, årg. 1, nr 1, s. 63–74, 2013. DOI: 10.15845/bjclcj. URL: http://dx.doi.org/10.15845/bjclcj.v1i1.524.

[4] N. R. Council, D. on Earth, L. Studies, C. on Life Sciences, C. on DNA Forensic Science och A. Update, "The evaluation of forensic DNA evidence", 1997.

[5] C. J. Guerrini, J. O. Robinson, D. Petersen och A. L. McGuire, "Should police have access to genetic genealogy databases? Capturing the Golden State Killer and other criminals using a controversial new forensic technique", *PLoS biology*, årg. 16, nr 10, e2006906, 2018. URL: https://doi.org/10.1371/journal.pbio.2006906.

[6] S. medicinsk-etiska råd, "Kort om DNA och brottsutredning", *Kort om*, 2021. URL: https://smer.se/wp-content/uploads/2021/04/smer_dna_brott_tga.pdf (hämtad 2025-01-29).

[7] Rättsmedicinalverket, "Forskningsprojekt om ansiktsprediktion med hjälp av DNA", 2023. URL: https://www.rmv.se/aktuellt/forskningsprojekt-om-ansiktsprediktion-med-hjalp-av-dna/ (hämtad 2025-03-01).

[8] A. Tillmar, Privat kommunikation, Maj 2025.

[9] S. Walsh, L. Chaitanya, L. Clarisse m. fl., "Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage", *Forensic Science International: Genetics*, årg. 9, s. 150–161, 2014, ISSN: 1872-4973. DOI: 10.1016/j.fsigen. URL: https://www.sciencedirect.com/science/article/pii/S1872497313002536 (hämtad 2025-01-29).

[10] S. Matheson, "DNA Phenotyping: Snapshot of a Criminal", *Cell*, årg. 166, s. 1061–1064, 2016. DOI: 10.1016/j.cell. URL: http://dx.doi.org/10.1016/j.cell.2016.08.016.

[11] E. Sidransky, "Locus", 2025. URL: https://www.genome.gov/genetics-glossary/Locus (hämtad 2025-03-06).

[12] A. J. Brookes, "The essence of SNPs", *Gene*, årg. 234, nr 2, s. 177–186, 1999. URL: https://doi.org/10.1016/S0378-1119(99)00219-X (hämtad 2025-03-06).

[13] K. Mäkelä, *Simple illustration showcasing basic genetics terminology in a hierarchical manner.* [Illustration], 2022. URL: https://upload.wikimedia.org/wikipedia/commons/2/2d/AllelemodWlocus%26karyotype_eng.png (hämtad 2025-03-07).

[14] M. K. Wojczynski och H. K. Tiwari, "Definition of phenotype", *Advances in genetics*, årg. 60, s. 75–105, 2008. URL: https://doi.org/10.1016/S0065-2660(07)00404-X (hämtad 2025-03-06).

[15] D. Adams, "Genotype", 2025. URL: https://www.genome.gov/genetics-glossary/genotype (hämtad 2025-03-06).

[16] B.-J. Koops och M. Schellekens, *Forensic DNA phenotyping: regulatory issues*, 2008. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=975032 (hämtad 2025-03-03).

[17] W. Branicki, F. Liu, K. van Duijn m. fl., "Model-based prediction of human hair color using DNA variants", *Human genetics*, årg. 129, s. 443–454, 2011. URL: https://doi.org/10.1007/s00439-010-0939-8.

[18] W. Gilks, S. Richardson och D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Chapman Hall, 1996.

[19] W. H. Greene, *Econometric Analysis*. Pretince Hall, 2003.

[20] R. P. Dobrow, *Introduction to Stochastic Processes With R*. John Wiley Sons, Inc, 2016, ISBN: 9781118740712. DOI: 10.1002/9781118740712. URL: http://dx.doi.org/10.1002/9781118740712 (hämtad 2025-02-24).

[21] M. Blangiardo och M. Cameletti, *Spatial and Spatio-Temporal Bayesian Models with R - INLA*. John Wiley Sons, 2015.

[22] S. S. Skiena, *The Data Science Design Manual*. Springer International Publishing, 2017. URL: https://doi.org/10.1007/978-3-319-55444-0 (hämtad 2025-05-06).

[23] H. M. Wallace, A. Jackson, J. Gruber och A. Thibedeau, "Forensic DNA databases–Ethical and legal standards: A global review", *Egyptian Journal of Forensic Sciences*, årg. 4, nr 3, s. 57–63, 2014. URL: https://doi.org/10.1016/j.ejfs.2014.04.002.

[24] K. N. B. m. S. Walsh F. Liu, "IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information", *Forensic Science International: Genetics*, årg. 5, nr 3, s. 170–180, 2011.

[25] G. C. Davenport och C. B. Davenport, "Heredity of hair color in man", *The American Naturalist*, årg. 43, nr 508, s. 193–211, 1909. URL: https://www.journals.uchicago.edu/doi/pdf/10.1086/279048 (hämtad 2025-05-09).

[26] K. Hunt, *The remarkable story behind David Bowie's most iconic feature*, 2016. URL: https://www.dailymail.co.uk/health/article-3394180/The-remarkable-story-David-Bowie-s-iconic-feature.html (hämtad 2025-05-09).

[27] M. A. Webster, E. Miyahara, G. Malkoc och V. E. Raker, "Variations in normal color vision. II. Unique hues", *Journal of the Optical Society of America A*, årg. 17, nr 9, s. 1545–1555, 2000.

[28] Integritetsskyddsmyndigheten, "Principen om korrekthet och diskriminerande algoritmer", 2024. URL: https://www.imy.se/verksamhet/dataskydd/innovationsportalen/vagledning-om-gdpr-och-ai/gdpr-och-ai/principen-om-korrekthet-och-diskriminerande-algoritmer/ (hämtad 2025-03-09).

[29] V. Strömberg och T. Mossinge-Norheim, "Klart: Polisen får använda dna-baserad släktforskning vid vissa brott", *SVT*, 2025. URL: https://www.svt.se/nyheter/inrikes/klart-polisen-far-anvanda-dna-baserad-slaktforskning-vid-vissa-brott (hämtad 2025-04-18).

[30] Integritetsskyddsmyndigheten, "Känsliga personuppgifter", *Introduktion till dataskyddsförordningen*, 2021. URL: https://www.imy.se/privatperson/dataskydd/introduktion-till-gdpr/vad-ar-personuppgifter/kansliga-personuppgifter/ (hämtad 2025-04-18).

[31] Europaparlamentet och E. unionens råd, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Article 9", *OJ L 119*, 2016. URL: https://gdpr-info.eu/art-9-gdpr/ (hämtad 2025-04-16).

[32] Europaparlamentet och E. unionens råd, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Article 17", *OJ L 119*, 2016. URL: https://gdpr-info.eu/art-17-gdpr/ (hämtad 2025-04-16).

[33] Polisen, "Polisens register", 2024. URL: https://polisen.se/lagar-och-regler/behandling-av-personuppgifter/polisens-register/ (hämtad 2024-04-24).

Användning av AI

Rapporten har använt sig av AI på flera sätt. Det främsta användningsområdet har varit kodning. Vi har använt AI-verktyg som OpenAis ChatGPT och Mistrals LeChat för att felsöka kod, omstrukturera kod för att öka dess effektivitet och för att skapa visualiseringar snabbt. AI-verktyg har även använts för att dubbelkolla meningsuppbyggnader och formuleringar i flera delar av rapporten, främst för avsnitt 1 och 3. AI-verktyg har inte använts för att samla information eller för att skapa nytt material.

A Appendix – Teori

A.1 Prestandamått av förväxlingsmatriser

Med en förväxlingsmatris för en modell på en viss datamängd finns det olika prestandamått man kan beräkna för att tolka och döma modellen [22]. Prestandamåtten som vi använder i rapporten är följande:

- **Accuracy** definieras som andelen korrekta prediktioner. Alltså för en $n \times n$ förxälingsmatris blir det $\frac{C_{11}+...+C_{nn}}{\sum_i \sum_j C_{ij}}$.
- **Precision** för klass j mäter hur ofta modellen hade rätt när den predicerade klass j, definitionen ges av $\frac{C_{jj}}{C_{1j}+...+C_{nj}}$.
- **Recall** är andelen observationer av klass j som modellen gissade rätt på och definieras som $\frac{C_{jj}}{\sum_{i=1}^{n} C_{ji}}$.
- **Specificity** är ett mått på hur ofta man haft rätt när man gissat falsk och för en $n \times n$-matris definieras det för klass j som $\frac{TN_j}{TN_j+FP_j}$ där $TN_j = \sum_{i \neq j} \sum_{k \neq j} C_{ik}$ och $FP_j = \sum_{i \neq j} C_{ji}$.

B Appendix – Figurer och tabeller

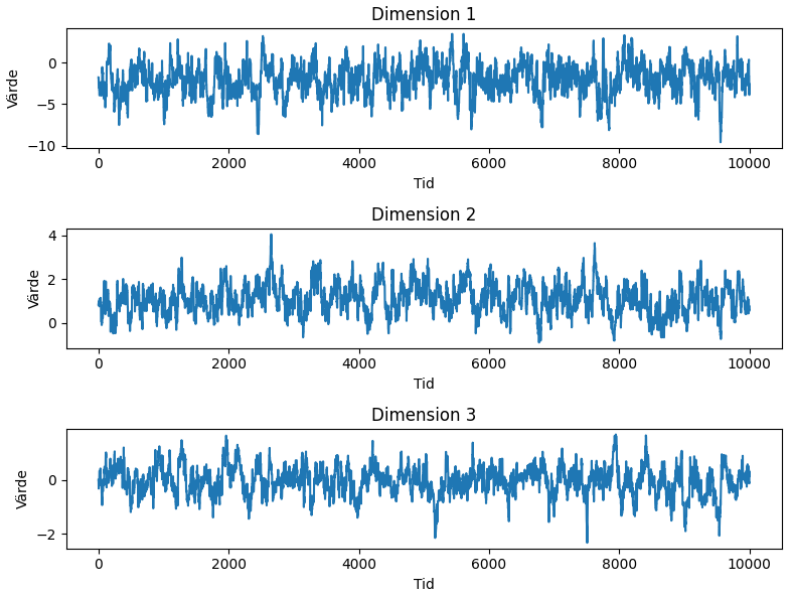
Kedjor och modellernas prestation

Prediktionsmodell blå ögon

Accuracy	78%
Precision	75%
Recall	94%
Specificity	54%

Tabell 14: Prestandamått för blå ögon

Prediktionsmodell samtliga ögonfärger

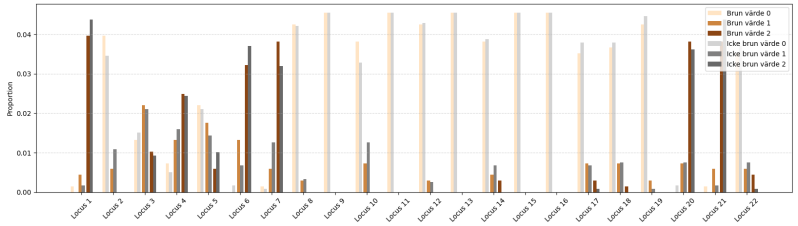


Figur 9: Markovkedjor för tre av 14 dimensioner

Prediktionsmodell brunt hår

Tabell 15: Prestandamått för enkel modell för brunt hår

Accuracy	52%
Precision	60%
Recall	71%
Specificity	18%



Figur 10: Visualisering av värden i alla loci för brunt hår och icke-brunt hår

C Appendix – Kod

Nedan visas kod för den grundläggande modellen. Först presenteras de matematiska funktionerna sedan visas hur kedjan simuleras.

```
import numpy as np
from scipy.optimize import minimize
from scipy.stats import norm


# Läs data
loci_data = np.read("loci_data.npy")
eye_data = np.read("eye_data.npy")


# Sannolikhetsfunktion
def phi(x):
    return np.exp(x) / (np.exp(x) + 1)


# Välj loci
selected_loci = [0, 1, 2, 3, 4, 5, 6]

# Log likelihood funktion

def log_likelihood(theta, loci_data=loci_data, eye_data=eye_data):
    z = np.dot(loci_data[:, selected_loci], theta)
    phi_vals = phi(z)
    return np.sum(
        eye_data * np.log(phi_vals)
        + (1 - eye_data) * np.log(1 - phi_vals))


# Normalfördelad prior
def log_prior(theta):
    return sum(norm.logpdf(theta, loc=np.zeros(len(theta)), scale=5))

# Beräkna maximum likelihood med minimize
n_locis = len(selected_loci)
theta_1 = minimize(lambda t: -log_likelihood(t), np.zeros(n_locis)).x

# Printa maximum likelihood
print(f"Maximum likelihood för intercept: {theta_1[0]}")
for i in range(1, len(theta_1)):
    print(f"Maximum likelihood för _ {i}: {theta_1[i]}")


# Accepteringssannolikhet
# Logaritmerad för att ge numerisk stabilitet
def rho(theta_new, theta_old, loci_data=loci_data, eye_data=eye_data):
    numerator = log_likelihood(theta_new, loci_data, eye_data)
        + log_prior(theta_new)
    denominator = log_likelihood(theta_old, loci_data, eye_data)
        + log_prior(theta_old)

    return np.exp(numerator - denominator)
```

```
# Antal steg i kedjan
N = 10000
# Varians
var = 0.2
# Matrix med nollor för att spara kedjan
theta_vals = np.zeros([N, n_locis])

# Beräkna slumpade värden som används i iterationerna
norm_vals = np.random.normal(0, var, [N, n_locis])
unif_vals = np.random.uniform(0, 1, N)

# Sätt första värdet till maximum likelihood värdena
theta_vals[0, :] = theta_1

n_accepted = 0
for i in range(1, N):
    # Föreslä nytt theta
    theta_new = theta_vals[i - 1, :] + norm_vals[i, :]

    # Beräkna acceptansnivä
    accpetance_prob = rho(theta_new, theta_vals[i - 1, :])

    if unif_vals[i] < accpetance_prob:
        # Acceptera nya theta
        theta_vals[i, :] = theta_new
        n_accepted += 1

    else:
        # Sätt nya theta till gamla theta
        theta_vals[i, :] = theta_vals[i - 1, :]
# Burn in för konvergens
burn_in = N // 2
burned_theta_vals = theta_vals[burn_in:, :]

# Ny persons DNA vars ögonfärg vi vill predicera
# Plats noll är alltid 1 för intercept
new_DNA = [1, 0, 2, 1, 0, 2, 2]

# Phi av theta gänger nya DNA värden
phi_vals = phi(np.dot(burned_theta_vals, new_DNA))
# Medelvärdet av alla phi-värden ger prediktionen
prediction = np.mean(phi_vals)

print(f"Prob of having blue eyes: {prediction}")
```