UNIVERSITA' DEGLI STUDI DI FIRENZE

DIPARTIMENTO DI STATISTICA
"G. PARENTI"

CORSO DI DOTTORATO
Dottorato in Statistica Applicata – XVI

Tesi di Dottorato

DNA mitocondriale nella scienza forense: modelli e algoritmi

Coordinatore del corso di dottorato
Prof. Fabrizia Mealli

Candidato
Paola Berchialla

Relatore
Prof. F. M. Stefanini

Aprile 2004

Indice

Introduzione.....1

Capitolo 1

1.1 Il DNA nucleare e il DNA mitocondriale.....8

1.2 Il DNA nucleare e il DNA mitocondriale nell'indentificazione forense.....10

1.3 Esempio di analisi statistica per l'identificazione forense.....14

1.4 Principi dell'analisi statistica forense.....16

1.5 Analisi statistica per l'identificazione forense con il DNA mitocondriale.....17

Capitolo 2

2.1 Modello mutazionale.....30

2.2 Modello demografico: coalescenza standard.....31

2.3 Modello per Single-Nucleotide Polymorphism data.....33

2.4 Metodi di simulazione.....37

2.5 Modello demografico: modello di coalescenza con dimensione della popolazione variabile. 42

2.6 Modello per Single-Nucleotide Polymorphism data II.....45

2.4 Metodi di simulazione II.....47

Capitolo 3

3.1 Descrizione dei dati.....50

3.2 Dimensione della popolazione e parametri di crescita.....50

3.3 Tasso di mutazione e tempi di coalescenza.....52

3.4 Risultati.....53

3.5 Validità del modello....63
3.6 Esempio di calcolo della conditional match probability....64

Conclusioni....67

Appendice A: Modelli della genetica di popolazione.

A.1 Modello di Wright-Fisher....69
A.2 Modello della coalescenza standard....70
A.3 Infinitely Many Sites Model, K-allele model e Finite Sites Model....73

Appendice B: Approssimazione alla diffusione nella genetica di popolazione.

B.1 Approssimazione alla diffusione....75
B.2 Approssimazione alla diffusione nel modello neutrale di Wright-Fisher....76
B.3 Approssimazione alla diffusione nel modello di Wright-Fisher con mutazioni e selezione....78

Appendice C: Modello gerarchico per SNPs loci.

C.1 Introduzione al modello gerarchico per SNPs loci....80
C.2 Modello gerarchico per SNPs loci....81
C.3 Una motivazione al modello nell'ambito della genetica di popolazione....82

Appendice D: Glossario....86

Bibliografia....88

Introduzione

Il termine DNA (*acido desossiribonucleico*) è un vocabolo che ormai è entrato nel linguaggio comune. Le applicazioni pratiche nelle quali questa molecola è chiamata in causa sono infatti molte, e molti sono anche i problemi etici con le relative discussioni che spesso ne accompagnano l'impiego.

In questo lavoro, ci siamo occupati dell'utilizzo del DNA per l'identificazione statistica dei soggetti che è tipica dell'ambito forense ma trova applicazione anche nell'antropologia per esempio per la datazione di reperti archeologici. In particolare abbiamo concentrato la nostra attenzione sulle problematiche e le metodologie che hanno per oggetto il DNA mitocondriale (mtDNA).

Ad eccezione dei batteri, delle alghe azzurre e di alcune cellule altamente specializzate, come per esempio i globuli rossi, negli esseri viventi sono presenti due tipi di DNA: il DNA nucleare e il DNA mitocondriale. Variazioni di sequenza utili a distinguere il patrimonio genetico di un individuo da quello di un altro sono disponibili su entrambi i tipi di genoma.

Le tecniche per l'identificazione personale utilizzano in genere il DNA nucleare, nel quale, per la ricombinazione genetica del DNA dei genitori, le variazioni nelle sequenze sono più evidenti. Vi sono però alcuni casi in cui la lettura del DNA nucleare non è possibile, per esempio quando il tessuto disponibile è molto degradato. In situazioni di questo tipo si deve allora ricorrere ad analisi basate sul DNA mitocondriale.

Il polimorfismo del DNA mitocondriale riscuote interesse in ambito forense proprio per la caratteristica, che tale molecola possiede, di una bassa propensione alla degradazione, ma le problematiche di metodo e valutazione quantitativa dell'evidenza associata agli aplotipi micondriali sono ancora oggetto di studio (Butler e Levin, 1996).

L'oggetto della nostra tesi ruota attorno ad un problema pratico ben definito. Disponendo di due profili allelici di mtDNA completamente sovrapponibili, in che misura possiamo affermare che si tratta di individui appartenenti alla medesima linea mitocondriale, ovvero che si tratta di fratelli—ipotesi generalmente suggerita da evidenze che non sono di carattere biologico—e non invece di soggetti non correlati e casualmente uguali?

In termini diversi, è necessario dare una valutazione alla probabilità che due individui, scelti a caso in una popolazione comune di appartenenza—che può per esempio essere un gruppo etnicamente omogeneo—e pertanto non geneticamente correlati, condividano la medesima sequenza, o aplotipo mitocondriale, ipervariabile.

Allo stato dell'arte, l'approccio, denominato *match probability*, può essere riassunto nella seguente domanda: quante volte si è osservata, fra tutti i dati a disposizione, la sequenza che è l'oggetto dell'analisi? Tale approccio, che si avvale di database di sequenze di mtDNA di numerose popolazioni e sottopopolazioni mondiali moderne, non può fornire una stima statisticamente rigorosa. Infatti, mancando di un database che sia sufficientemente esteso della popolazione di riferimento a cui appartengono i soggetti, non è possibile disporre della distribuzione degli aplotipi mitocondriali caratterizzanti tale popolazione. A maggior ragione una simile considerazione vale se si lavora su reperti archeologici, avendo a che fare con popolazioni antiche e dunque non chiaramente definibili.

Nella nostra tesi cercheremo di dare una risposta a questi problemi presentando un metodo basato sulla verosimiglianza per il calcolo delle *match probabilities* relative a sequenze di DNA mitocondriale.

Abbiamo iniziato il capitolo 1 con una breve descrizione delle differenze principali tra il DNA nucleare ed il DNA mitocondriale senza avere la pretesa di fornire una esaustiva sintesi di carattere biologico ma con lo scopo principale di introdurre i termini tecnici che poi abbiamo utlizzato.

Abbiamo proseguito presentando i problemi che riguardano l'interpretazione dell'informazione derivante da un profilo di DNA nell'identificazione forense. In particolare abbiamo trattato dei problemi che riguardano l'inferenza statistica dell'evidenza relativa ad un profilo allelico.

La maggior parte di questi problemi ruotano attorno alle seguenti domande:

(*i*) quale rilevanza possiede il fatto che la frequenza del profilo varia tra la popolazione;

(*ii*) come incorporare l'evidenza del DNA con le altre evidenze;

(*iii*) in che maniera l'inferenza dipende da altre informazioni come per esempio quelle relative ad un possibile colpevole il cui profilo sia gi stato analizzato.

D. Balding e P. Donnelly in un loro articolo, (Balding e Donnelly, 1995), hanno presentato un contesto teorico coerente, e largamente accettato anche nelle corti durante la discussione di alcuni casi, nell'ambito del quale è possibile dare una risposta, almeno in via di principio, a tali domande.

Nel capitolo 1 abbiamo presentato l'approccio di Balding e Donnelly così come formulato nel contesto dell'analisi statistica forense che ha per oggetto sequenze di DNA nucleare e lo abbiamo poi riproposto con le dovute modifiche per il caso analogo del mtDNA nell'analisi delle ipotesi statistiche alternative relative a due soggetti che possono appartenere alla medesima linea mitocondriale.

Questo approccio si avvale non delle frequenze degli alleli ma di *match probabilities*, che sono probabilità condizionate, ovvero probabilità di osservare un profilo condizionatamente al fatto che questo sia già stato osservato. Tali probabilità dipendono dalle relazioni genetiche tra gli individui. Per esempio in (Cockerham, 1971) si trovano delle misure di relazioni tra insiemi di alleli che vengono utilizzate nel calcolo delle *match probabilities*.

Nel contesto tradizionale dell'analisi statistica forense, per calcolare la *match pro-*

bability quando un profilo X è osservato, si considera il caso in cui il colpevole sia una persona differente dal sospettato e si considerano le relazioni genetiche a seconda delle popolazioni a cui gli individui appartengono—se per esempio i due soggetti appartengono al medesimo gruppo etnico o sono imparentati tra loro.

Nel nostro caso di analisi di sequenze mitocondriali, per prima cosa abbiamo definito i modelli teorici della genetica di popolazione la cui precisazione risulta importante per stabilire il contesto di lavoro.

Più precisamente abbiamo lavorato nell'ambito del *finite-sites model* che secondo molti studi, tra cui per esempio (Schneider e Excoffier, 1999), (Yang, 1996), (Drummond *et al.*, 2002), è il modello che si adatta meglio a descrivere il processo mutazionale che avviene nella molecola di DNA mitocondriale.

Nell'ambito del *finite-sites model* poi abbiamo distinto tra *one locus sites model* con due alleli e *one locus sites model* con 4 alleli. Il primo prevede che in ogni locus o sito della molecola di mtDNA ci possa essere uno di due possibili alleli. Il secondo invece prevede quattro alleli, e permette inoltre di distinguere tra due differenti tipi di mutazioni—transversioni e transizioni. Formalmente quest'ultimo modello appare più realistico rispetto al primo perché quattro sono le basi costitutive della molecola di DNA, sia nucleare che mitocondriale. Abbiamo però osservato che nei dati che abbiamo a nostra disposizione, in ogni sito polimorfico sono presenti solamente due varianti e questo fatto ci permette di ricondurci al caso del *one locus sites model* con due soli alleli. Per ciascuno dei due casi, comunque, abbiamo calcolato le relative *match probabilities*, nelle proposizioni 1.5.6 e 1.5.7.

Anche nel caso del DNA mitocondriale, abbiamo osservato che le *conditional match probabilities* coinvolgono dei parametri relativi al processo mutazionale ed al processo demografico. Entrambi questi processi riflettono l'insieme delle dipendenze e delle relazioni genetiche che esistono tra gli individui, e che sono racchiuse in un campione di sequenze

di mtDNA.

Nel capitolo 2 abbiamo definito un modello statistico con lo scopo di fare inferenza sia sui parametri mutazionali che sui parametri demografici che caratterizzano le popolazioni, come per esempio la dimensione effettiva della popolazione stessa o il tempo che separa un campione di individui da un loro antenato comune.

Lo spunto per la definizione del modello statistico che abbiamo presentato è derivato dal modello per dati di tipo SNPs (*single-nucleotide polymorphisms*) presentato da (Nicholson *et al.*, 2002).

Un locus SNP è una singola posizione nel DNA mitocondriale che presenta variazione tra differenti individui appartenenti ad una determinata popolazione. Mentre un nucleotide è identificato con una delle lettere A, C, G, T—associate rispettivamente alle basi contenenti Adenina, Citosina, Guanina e Timina—un SNP presenta due sole varianti che abbiamo indicato con 0 ed 1, rispettivamente la variante più comune e quella più rara.

Come abbiamo già avuto modo di affermare, considerare una sequenza di mtDNA come una successione di loci di tipo SNPs non è una semplificazione riduttiva, in quanto, almeno nei dati a nostra disposizione, i siti polimorfici presentano effettivamente due sole varianti.

Il modello gerarchico per SNPs introdotto da (Nicholson *et al.*, 2002) modellizza i dati in maniera binomiale introducendo come parametri non noti le frequenze degli alleli nella popolazione moderna e le relative frequenze in una popolazione ancestrale. Le frequenze della popolazione moderna sono legate a quelle della popolazione ancestrale attraverso una distribuzione normale la cui varianza è funzione di un coefficiente che può essere interpretato come il tempo che separa le due popolazioni. Minore è il tempo di separazione, maggiormente simili saranno le distribuzioni delle frequenze degli alleli nelle due popolazioni, quella ancestrale e quella moderna.

Rispetto al modello per SNPs così definito, al fine di poter fare inferenza sui parametri di nostro interesse che intervengono nel calcolo delle *match probabilities*, abbiamo implementato un processo mutazionale ed un processo demografico.

Poiché i dati a nostra disposizione presentano per ogni locus due sole varianti, abbiamo assunto come ipotesi di lavoro il *one finite-sites locus* con due soli alleli. All'interno di tale contesto, nella definizione del processo mutazionale, abbiamo introdotto un solo tasso di mutazione ed abbiamo modellato il relativo processo sulla falsariga del modello di Jukes-Cantor.

Per quanto riguarda il processo demografico, invece, esso attiene alla dimensione effettiva della popolazione ed al suo comportamento. Nel modello base ipotizzato che si mantenesse costante nel tempo, rispettando le assunzioni del modello della coalescenza standard, mentre negli altri modelli abbiamo assunto in un primo caso una crescita esponenziale per tutto il tempo di separazione tra la popolazione ancestrale e quella moderna, mentre in un secondo caso una crescita a due parametri. Quest'ultimo scenario demografico corrisponde ad una popolazione la cui numerosità o dimensione effettiva si mantiene costante fino a un tempo t_g a partire dal quale inizia a crescere in maniera esponenziale.

Per ciascuno dei modelli così definiti, abbiamo infine proposto i relativi algoritmi di campionamento basati sul metodo *Acceptance-Rejection* (Ripley, 1987) e che possono essere utilizzati al posto dei metodi Markov chain Monte Carlo.

Nel capitolo 3 abbiamo infine presentato i risultati ottenuti applicando il modello statistico definito nel capitolo 2 ad un database composto da 49 individui, appartenenti alla sottopopolazione toscana della popolazione italiana, che presentano 28 siti polimorfici. Avendo condotto un'analisi statistica bayesiana, abbiamo riportato i dati relativi alle posterior distributions.

Abbiamo poi concluso il capitolo riportando la valutazione numerica della *condi-*

tional match probability del caso pratico che abbiamo descritto, relativo a due aplotipi mitocondriali perfettamente sovrapponibili, per i quattro modelli.

Nello scrivere questa tesi, abbiamo adottato uno stile sintetico basato su definizioni operative dei modelli di genetica di popolazione che non presuppone una familiarità o una conoscenza degli argomenti correlati di biologia e genetica di popolazione. A questo proposito abbiamo rimandato per descrizioni più dettagliate nell'ambito della modellistica della genetica di popolazione alle appendici. Nelle appendici abbiamo anche aggiunto un breve glossario dei termini di biologia molecolare che abbiamo utilizzato, fornendo delle definizioni più rigorose dal punto di vista della biologia.

Capitolo 1

1.1 Il DNA nucleare e il DNA mitocondriale.

Ad eccezione dei batteri, delle alghe azzurre e di alcune cellule altamente specializzate, come per esempio i globuli rossi, negli esseri viventi sono presenti due tipi di DNA: il DNA mitocondriale (mtDNA) che si trova all'interno di organelli detti *mitocondri* ed ha una struttura molto semplice, ed il DNA nucleare, contenuto nel nucleo della cellula, più evoluto e complesso. Nella seguente figura sono illustrati questi due tipi di DNA insieme alle loro principali caratteristiche distintive.

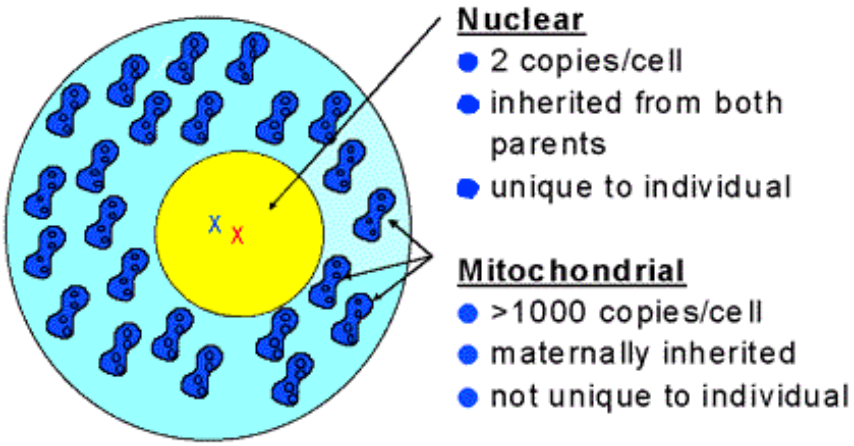

FIGURA 1.1.1 DNA nucleare e DNA mitocondriale.

Il DNA mitocondriale si trova nei mitocondri, organelli specializzati della cellula presenti in alcune migliaia di copie (800-2,500) nel citoplasma. La sua specifica funzione consiste proprio nella regolazione di questi organelli che presiedono all'attività energetica della cellula.

È organizzato in una molecola circolare (anello) a doppio filamento delle dimensioni di 16,569 coppie di basi (Anderson et al., 1981). I due filamenti sono distinguibili

l'uno dall'altro per il differente peso molecolare e sono indicati convenzionalmente con H (*Heavy*) quello più ricco di Adenina e Guanina (basi o nucleotidi A e G) e con L (*Light*) quello complementare, più ricco di Timina e Citosina (basi T e C).

La regione dell'anello nota come *D-loop* contiene due sottoregioni ipervariabili, denotate come HV1 e HV2, rispettivamente formate dalle coppie di basi comprese tra le posizioni 16,000–16,430 e 40–440. Il *D-loop* è illustrato nella seguente figura.

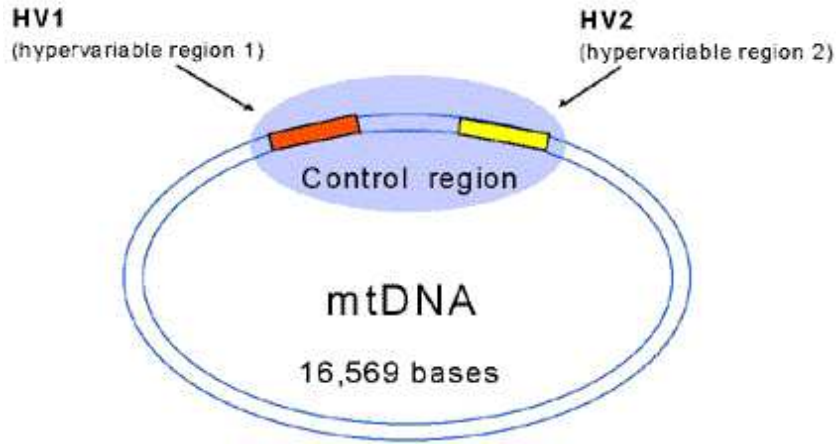

FIGURA 1.1.2 D-loop della molecola mitocondriale.

Il DNA nucleare è una molecola di grosse dimensioni più evoluta e complessa del DNA mitocondriale. Secondo il modello di Crick e Watson, la molecola del DNA nucleare è formata da due filamenti ciascuno dei quali costituito da una catena di basi o nucleotidi—come per il mtDNA, indicate con le lettere A (adenina), G (guanina), T (timina) e C (citosina)—appaiati e avvolti a spirale uno attorno all'altro a formare la cosiddetta doppia elica. La successione di nucleotidi costituisce il patrimonio genetico dell'individuo e varia da soggetto a soggetto ma è identica per tutte le cellule della stessa persona.

Nonostante le dimensioni enormemente ridotte rispetto a quelle del DNA nucleare (1/200,000 rispetto al corredo aploide costituito da 3 miliardi di coppie di basi), il DNA mitocondriale costituisce circa lo 0.5% del genoma grazie all'elevato numero di copie di

mitocondri presenti nel citoplasma.

Differente è anche il modo di trasmissione tra genitori e figli dei due tipi di DNA. Infatti, mentre il DNA nucleare viene trasmesso come combinazione del DNA nucleare di entrambi i genitori, il DNA mitocondriale viene trasmesso in blocco da madre a figlio ed è caratterizzato dall'assenza di eventi di ricombinazione genetica, cioè scambi di materiale genetico tra cromosomi. Vedremo in seguito come questo fatto, a causa di una probabilità non trascurabile di mutazioni, non significhi comunque che tutti gli individui che appartengono ad una medesima discendenza per linea materna abbiano un identico DNA mitocondriale. La molecola di DNA mitocondriale, infatti, evolve secondo un complesso processo di sostituzione caratterizzato principalmente dai seguenti fenomeni:

- la composizione delle basi non è uniforme;
- le transizioni (le sostituzione di tipo $T \leftrightarrow C,\ A \leftrightarrow G$) si verificano con una frequenza maggiore delle transversioni ($T,\ C \leftrightarrow C,\ A,\ G$);
- il numero di transizioni di pirimidine nella catena L (le sostituzioni di tipo $T \leftrightarrow C$) eccede il numero di transizioni di purine ($A \leftrightarrow G$).

1.2 Il DNA nucleare e il DNA mitocondriale nell'indentificazione forense.

Abbiamo visto che il genoma di un individuo è composto dal DNA nucleare e dal DNA mitocondriale. Inoltre per questi due tipi di DNA vengono trasmessi dai genitori ai figli in maniera differente. Infatti il DNA nucleare viene trasmesso come combinazione del DNA nucleare di entrambi i genitori e varia da soggetto a soggetto essendo identico solo in gemelli omozigoti. Il DNA mitocondriale, invece, viene trasmesso inalterato da madre a figlio, senza alcuna ricombinazione genetica. A causa di una probabilità non

trascurabile di mutazioni, però, non tutti gli individui che appartengono ad una medesima discendenza per linea materna hanno un identico DNA mitocondriale.

Le tecniche per l'identificazione personale tramite DNA utilizzano in genere il DNA nucleare nel quale, come conseguenza della ricombinazione genetica del DNA dei genitori, le variazioni nelle sequenze sono più evidenti.

Per la lettura del DNA nucleare si utilizza l'analisi dei loci STR (*Short Tandem Repeat*). I loci STR sono particolari posizioni occupate da un gene su un cromosona per i quali la tecnologia disponibile consente di trascurare la possibilità di errori di misura. Su di essi si va a leggere il numero di volte per cui una ben individuata sequenza di nucleotidi si ripete (Foreman et al., 1997).

Talvolta la lettura del DNA nucleare non è possibile. Infatti può accadere di dover analizzare reperti biologici in cui il DNA nucleare è scarso o degradato, come per esempio un capello, un dente o un frammento d'osso. In situazioni di questo tipo, si deve ricorrere ad analisi basate sul DNA mitocondriale. Essendo infatti presente in copie multiple all'interno di ogni cellula, a parità di numero di cellule del campione da esaminare il numero di copie di un gene localizzato sul DNA mitocondriale è di molto superiore al numero di copie di un gene localizzato sul DNA nucleare—per quest'ultimo nella maggior parte dei geni si dispone di due copie per cellula—ed è più resistente nel tempo, consentendo anche di effettuare analisi su resti ossei risalenti a migliaia di anni fa (Ward et al., 1991).

L'identificazione tramite DNA nucleare si basa sul fatto che esso, con l'eccezione dei gemelli omozigoti, è pressoché unico. La struttura che lo compone risulta essere irripetibile. Se per un individuo potessimo analizzare l'intero filamento di DNA, questo lo farebbe distinguere da chiunque altro, risultando trascurabile la probabilità di osservare due persone con DNA nucleare identico. Inoltre tutte le cellule di un individuo hanno lo stesso DNA. Ciò significa che da una goccia di sangue o un frammento di pelle si può

identificare con certezza l'individuo a cui tale campione biologico appartiene.

L'analisi del DNA mitocondriale inoltre presenta degli specifici vantaggi quando l'unico DNA di riferimento è quello parentale. Un'importante caratteristica del genoma mitocondriale è infatti la sua trasmissione per via materna. Nel caso in cui si abbia un riferimento parentale, l'identificazione è allora supportata dall'identità del DNA mitocondriale, a meno di mutazioni, nei parenti per linea materna.

Supponiamo che sia stato commesso un crimine e sulla scena del delitto si sia rivelata la presenza di una goccia di sangue. Supponiamo anche che il sangue rinvenuto sia differente da quello della vittima e le circostanze facciano ritenere che esso appartenga a qualcuno che era presente al momento in cui il reato è stato commesso e che ci siano ragioni per credere che questi sia il colpevole. Infine supponiamo che il profilo genetico estratto da un campione biologico, che può essere per esempio una goccia di sangue o di saliva, di un sospetto coincida con quello rinvenuto sulla scena del delitto appartenente al colpevole, situazione che prende il nome di *matching*—che è il termine che utilizzeremo anche in seguito—di genotipo per il DNA nucleare o di aplotipo per il DNA mitocondriale.

Il problema naturale che segue ad un tale approccio consiste nel valutare l'evidenza contro il sospettato contenuta nel *matching* dei due profili. In termini probabilistici, una risposta a questo problema è la seguente: la probabilità che il sospettato, non essendo il colpevole, abbia il medesimo profilo rinvenuto sul luogo del delitto è data dalla frequenza nella popolazione del profilo stesso. Alti valori di questa frequenza sono a favore del sospettato e dunque della sua innocenza, mentre bassi valori sono a suo sfavore.

Il problema che immediatamente segue a questo primo, riguarda il metodo di stima della frequenza del profilo nella popolazione. Per profili genetici basati su singoli loci, si può stimare la frequenza di ogni genotipo sulla base di un campione di profili estratto dalla popolazione. La numerosità del campione dipenderà allora dalla variabilità dei loci, ossia dovrà essere maggiore se i loci sono variabili, ovvero se ammettono differenti

genotipi. Un altro problema consiste poi nella definizione della popolazione da cui estrarre il campione.

Naturalmente all'aumentare del numero dei loci utilizzati per l'identificazione personale, aumenta anche il numero di genotipi possibili. La pratica corrente nell'ambito del DNA nucleare, negli Stati Uniti, consiste nell'utilizzare un set di tredici *Short Tandem Repeat* (STR).

Nel caso del DNA mitocondriale, invece, la pratica corrente consiste nel confrontare il profilo sequenziato con i profili contenuti in database di riferimento. Infatti i termini del quesito sopra esposto possono essere così sintetizzati, per l'identificazione tramite mtDNA. Attesa l'evidenza, e dunque il *matching* di due aplotipi mitocondriali, in quale misura è possibile affermare che tale dato deriva dal fatto che i reperti biologici appartengono entrambi alla medesima persona, oppure verificano la circostanza che si tratti di soggetti appartenenti alla medesima linea materna? In termini diversi, è necessario comprendere quale sia la probabilità che due individui, presi a caso nella popolazione e pertanto non geneticamente correlati, condividano la medesima sequenza ipervariabile del mitocondrio.

La risposta a tale quesito non banale. Osserviamo anzitutto che, a priori, il dato di compatibilità basato su sequenze di DNA mitocondriale, allo stato dell'arte, non può avere una stima statisticamente rigorosa in quanto manca un database sufficientemente esteso per tutte le popolazioni di riferimento. Non è quindi disponibile la distribuzione degli aplotipi mitocondriali caratterizzanti tutte le popolazioni. A maggior ragione tale considerazione vale se si tratta di una popolazione antica e non chiaramente definibile, come nel caso di analisi di reperti archeologici.

L'approccio concettuale precedentemente esposto per il DNA nucleare, nel caso del DNA mitocondriale può quindi essere riassunto nella seguente domanda: quante volte si è osservata, fra tutti i dati a disposizione, la sequenza che è l'oggetto dell'analisi?

1.3 Esempio di analisi statistica per l'identificazione forense.

Supponiamo che sulla scena di un delitto sia stato rilevato un reperto biologico (una goccia di sangue, oppure un capello) e supponiamo che la sequenza mitocondriale ottenuta coincida con quella di un sospettato S. Per dare una valutazione probabilistica dell'ipotesi che S sia il colpevole è necessario valutare la probabilità che la traccia o il profilo rinvenuto appartenga ad un altro possibile colpevole.

Indichiamo con G_s e G_c i profili genetici rispettivamente del sospettato e della traccia rilevata sulla scena del crimine, e supponiamo che questi due profili siano identici. Indichiamo con la lettera E l'evidenza di questi due profili: $E = (G_s, G_c)$ e con la lettera I tutto il resto dell'informazione disponibile tramite l'indagine.

Formuliamo le due seguenti ipotesi mutuamente esclusive:

H_p: la traccia rinvenuta appartiene al sospettato;

H_d: la traccia rinvenuta proviene da un individuo diverso dal sospettato appartenente ad una specifica popolazione di possibili colpevoli;

Per valutare queste due ipotesi si utilizza un approccio di tipo bayesiano che permette di ottenere una probabilità a posteriori. Al contrario, l'inferenza classica fornisce solamente un rapporto di verosimiglianza che può dare luogo ad errate interpretazioni dei risultati.

Prima dell'analisi del DNA, la probabilità di H_p è condizionata solo all'informazione I relativa all'indagine, pertanto $P(H_p|I)$ è una probabilità a priori. Dopo l'analisi del mtDNA, calcoleremo invece la probabilità a posteriori $P(H_p|I, E)$, ed analogamente per l'ipotesi alternativa H_d.

Utilizzando quindi il teorema di Bayes abbiamo la seguente equazione:

$$\frac{P(H_p|I,E)}{P(H_d|I,E)} = \frac{P(E|H_p,I)}{P(E|H_d,I)} \times \frac{P(H_p|I)}{P(H_d|I)} \tag{1.3.1}$$

La (1.3.1) può essere formulata anche nei seguenti termini:

$$\text{Odds a posteriori su } \mathrm{H_p} = \mathrm{LR} \times \text{Odds a priori su } \mathrm{H_p} \tag{1.3.2}$$

L'interesse è rivolto a quel fattore che trasforma l'odds a priori in odds a posteriori e che viene spesso indicato come valore dell'evidenza:

$$\frac{P(E|H_p, I)}{P(E|H_d, I)} \tag{1.3.3}$$

e che nella (1.3.2) abbiamo indicato come LR (*Likelihood Ratio*) ma che nell'ambito dell'inferenza bayesiana è il *Bayes Factor*. Seguendo la notazione di (Weir, 2001) utilizzeremo il termine di LR.

Il *Likelihood Ratio* o rapporto di verosimiglianza assume la seguente forma:

$$\begin{aligned} \mathrm{LR} = \frac{P(G_s, G_c|H_p, I)}{P(G_s, G_c|H_d, I)} &= \frac{P(G_c|G_s, H_p, I)}{P(G_c|G_s, H_d, I)} \times \frac{P(G_s|H_p, I)}{P(G_s|H_d I)} \\ &= \frac{1}{P(G_c|G_s, H_d, I)} \end{aligned} \tag{1.3.4}$$

L'ultimo passaggio è giustificato dal fatto che nell'ipotesi H_p in cui la traccia rinvenuta sia quella del sospettato, segue $P(G_c|G_s, H_p, I) = 1$ sulla base dell'assunzione che i profili G_s e G_c coincidono quando provengono dalla medesima persona—vengono trascurati quindi gli errori strumentali. Inoltre la probabilità di osservare il profilo G_s non dipende dalla particolare ipotesi considerata.

Il denominatore del rapporto di verosimiglianza $P(G_c|G_s, H_d, I)$ prende il nome di *conditional match probability* e rappresenta la probabilità di osservare il profilo rinvenuto in una qualunque persona che non sia il sospettato.

A volte il *Likelihood Ratio* si trova scritto nella forma

$$\mathrm{LR} = \frac{P(G_s, G_c|H_p, I)}{P(G_s, G_c|H_d, I)} = \frac{P(G_s|G_c, H_p, I)}{P(G_s|G_c, H_d, I)} \times \frac{P(G_c|H_p, I)}{P(G_c|H_d I)} \tag{1.3.5}$$

in cui i ruoli di G_C e G_s sono ribaltati rispetto alla (1.3.4).

L'equazione (1.3.4) viene utilizzata nel caso denominato *suspect anchored perspective*, cioé quando viene ritrovata sulla scena del delitto una traccia non appartenente alla vittima e che potrebbe provenire dal sospettato. L'equazione (1.3.5), invece, fa riferimento al caso detto *scene anchored perspective*, in cui una traccia biologica che potrebbe appartenere alla vittima viene rinvenuta sul corpo del sospettato.

1.4 Principi dell'analisi statistica forense.

Alla base dell'interpretazione statistica dell'evidenza genetica vi sono tre principi (Evett e Weir, 1998):

- *primo principio:* per valutare l'incertezza di una qualsiasi proposizione (ipotesi) data, è necessario considerare almeno una proposizione (ipotesi) alternativa;
- *secondo principio:* l'interpretazione deve essere basata su domande del tipo "qual è la probabilità dell'evidenza data la proposizione (ipotesi)?";
- *terzo principio:* l'interpretazione è condizionata non solo dalle proposizioni (ipotesi) ma anche dall'insieme delle circostanze nell'ambito delle quali tali ipotesi devono essere valutate.

Il *primo principio* è legato alla computazione del *likelihood ratio*. Implicitamente formulando un tale principio si accetta che la frequenza del profilo genetico non sia direttamente rilevante ai fini dell'inferenza forense.

L'approccio legato alla valutazione delle frequenze dei profili genetici, che non richiede la computazione dei *likelihood ratios*, può essere così sintetizzato. Il colpevole—rappresentato da una traccia biologica rinvenuta sul luogo del crimine—possiede un certo profilo genetico il quale coincide con quello del sospettato. Inoltre il profilo genetico sequenziato è raro. Sulla base di queste evidenze, si tratta di decidere se la traccia

biologica rinvenuta ed associata al colpevole appartenga o meno al sospettato, ovvero se il sospettato è il colpevole. Intuitivamente quanto più il profilo analizzato è raro, tanto maggiore si è inclini a ritenere che il colpevole ed il sospettato siano la medesima persona.

Un tale approccio presenta delle limitazioni. Infatti risulta difficile, in un contesto del genere, incorporare l'evidenza relativa all'analisi del DNA al resto dell'evidenza legata all'investigazione. A questa necessità di valutare l'evidenza nel suo insieme fa riferimento il *terzo principio*.

(Balding e Donnelly, 1995) hanno presentato un'analisi dell'inferenza nell'identificazione forense in cui il ruolo centrale è demandato ad i *likelihood ratios* e conseguentemente alle *conditional match probabilities* piuttosto che alle frequenze dei profili. Nel loro metodo, che abbiamo descritto nella sezione precedente, la domanda finale a cui si risponde non è più se il sospettato è il colpevole, ma se il sospettato è la persona a cui è associato il campione ritrovato sul luogo del crimine. Ovvero l'espressione "è colpevole"diviene "è la sorgente della traccia rilevata". Questa impostazione risponde all'esigenza formulata dal *secondo principio*.

1.5 Analisi statistica per l'indentificazione forense il DNA mitocondriale.

In questa sezione introduciamo una notazione leggermente differente da quella finora utilizzata. Indichiamo con X_1 ed X_2 gli aplotipi mitocondriali di due individui—con un abuso di notazione, con X_1 ed X_2 indicheremo anche gli individui stessi. In particolare, con riferimento alle sezioni precedenti, possiamo pensare ad X_1 come al profilo G_c ritrovato sul luogo di un crimine e che si suppone appartenere al colpevole, ed a X_2 come al profilo G_s relativo al sospettato. Indichiamo inoltre con la lettera H l'ipotesi che X_1 ed X_2 siano lo stesso individuo, mentre con $\bar{H}$, invece, l'ipotesi complementare in cui X_1 ed X_2 non siano la medesima persona.

Introduciamo questa notazione per svincolarci dagli esempi precedentemente considerati, propri di un'analisi che ha per oggetto il sequenziamento del DNA nucleare di una traccia biologica ritrovata sulla scena di un crimine, ed esaminare invece il problema di nostro interesse che riguarda due sequenze mitocondriali identiche. Queste possono riferirsi non solo a tracce rivenute sulla scena di un crimine ma anche, per esempio, ad antichi reperti archeologici. Inoltre come conseguenza della modalità di trasmissione del DNA mitocondriale—l'mtDNA viene ereditato per via materna in assenza di fenomeni di ricombinazione genetica (vedi §1.1)—due sequenze identiche possono riferirsi anche all'ipotesi H che X_1 e X_2 siano fratelli.

Abbiamo visto che il denominatore del rapporto di verosimiglianza $P(X_1 = X|X_2 = X, \bar{H})$ noto anche come *conditional match probability* è la probabilità che un membro X_1 della popolazione condivida lo stesso aplotipo di un altro individuo X_2. Se i profili di X_1 e X_2 non coincidono, allora il corrispondente rapporto di verosimiglianza è zero. Per calcolare il rapporto di verosimiglianza, invece, quando si osserva un *match* tra i due aplotipi, bisogna considerare l'ipotesi alternativa $\bar{H}$ in cui le tracce sequenziate si riferiscono a due differenti individui.

Supponiamo che se i due profili provengono dallo stesso individuo, allora necessariamente coincidono. In questa ipotesi abbiamo visto con la (1.3.4) che il rapporto di verosimiglianza assume la seguente forma:

$$\mathrm{LR}(X) = \frac{1}{P(X_1 = X|X_2 = X, \bar{H})} \tag{1.5.1}$$

La validità di questa ipotesi risiede nel problema legato all'eteroplasmia. Questo fenomeno consiste nella presenza di diversi genotipi mitocondriali nel medesimo individuo. Vi sono molti dubbi sulla variabilità inter-generazionale dell'eteroplasmia. Secondo alcuni studi (Howell et al., 1992; Bendall e Sykes, 1995), può rimanere tale all'interno di un gruppo familiare per più generazioni ma anche facilmente e velocemente

modificarsi in una situazione omoplastica.

Una spiegazione al comportamente inter-generazionale di questo fenomeno è il modello a collo di bottiglia di (Hauswirth e Laipis, 1985). Questo modello suggerisce che a qualche stadio della formazione della cellula riproduttiva femminile, il numero di molecole di mtDNA, oppure dei mitocondri, si riduca drasticamente (*bottleneck*) per poi tornare ai livelli precedenti grazie ad uno stadio successivo caratterizzato da un forte aumento dei cicli replicativi (*over replication*). Sarebbe perciò quest'ultimo stadio il responsabile delle variazioni inter-generazionali del grado di eteroplasmia. Esso infatti terrebbe conto solo della distribuzione delle frequenze alleliche nella popolazione ridotta e non in quella più ampia di partenza. Sarebbe perciò possibile che una variante mutata, presente in bassissimi livelli prima del collo di bottiglia, si fissasse rapidamente all'interno della linea materna.

Relazioni genetiche tra gli individui di una popolazione.

Nel calcolo della *conditional match probability* di X nella (1.5.1), intervengono le relazioni genetiche tra i due individui X_1 e X_2.

Supponiamo che X_1 e X_2 appartengano ad una popolazione $\mathcal{P}$ e supponiamo di disporre di un database $\mathcal{D}$ relativo a tale popolazione. $\mathcal{D}$ sarà costituito da un campione di n sequenze relative ad n individui scelti a caso nella popolazione $\mathcal{P}$ e non imparentati tra loro. Nei casi-lavoro in campo forense, tipicamente si dispone di una tale informazione. Infatti sono generalmente disponibili delle sequenze di DNA mitocondriale relative ad individui non imparentati tra loro ed appartenenti ad un identico gruppo etnico—per esempio la popolazione del sospettato (Wilson et al., 2003).

I due individui X_1 e X_2, appartenendo alla medesima popolazione $\mathcal{P}$, condividono delle relazioni genealogiche. Quando si confrontano delle sequenze omologhe, infatti, il pattern di somiglianze contiene delle informazioni sull'evoluzione delle sequenze stesse.

Tenendo conto di una varietà di circostanze, queste sequenze contengono quindi elementi informativi sui rapporti di parentela che intercorrono tra di loro e su quanto è lontano nel tempo un loro progenitore comune.

In assenza di fenomeni di ricombinazione, ogni sequenza ha un solo genitore nella generazione precedente. Questo è il caso del DNA mitocondriale, a differenza del DNA nucleare in cui invece il numero di progenitori cresce risalendo indietro nel tempo poiché ogni sequenza è associata alla ricombinazione di due sequenze.

Le proprietà statistiche di una genealogia dipendono da fattori quali la numerosità della popolazione ed anche dalla presenza di forze selettive che possono favorire la trasmissione di particolari alleli. Oltre alle relazioni di tipo demografico descritte, nel determinare le proprietà di una genealogia interviene un processo mutazionale. Verosimilmente il processo genealogico dipende fortemente dal modello mutazionale. Infatti, per esempio, se il tasso di mutazione di un allele è molto basso, tutte le sequenze di un campione potrebbero essere identiche ed in tale circostanza non si disporrebbe di alcuna informazione sulla genealogia del campione stesso.

Consideriamo una popolazione diploide di numerosità costante N, e consideriamo un locus con due alleli A_1 e A_2. Supponiamo che il processo mutazionale sia neutrale, ovvero che nessuno dei due alleli sia favorito nella trasmissione da generazione a generazione.

I due seguenti principi, che sono alla base di molti modelli della genetica di popolazione, sono fondamentali per una agevole trattazione matematica dei processi demografici e mutazionali

Principio 1.5.1. *In un modello neutrale, poiché le mutazioni non hanno effetto sul processo riproduttivo, è sempre possibile separare il processo mutazionale da quello genealogico.*

Per tradurre in termini pratici di questo principio, consideriamo una popolazione di N individui che si riproduce secondo il modello neutrale di Wright-Fisher (vedi appendice), ovvero una popolazione in cui le generazioni sono discrete, non si sovrappongono ed ogni nuova generazione si forma campionando N individui dalla generazione precedente secondo uno schema di estrazione con ripetizione. Il numero di discendenti con cui ogni individuo concorre nella formazione della nuova generazione ha una distribuzione binomiale con parametro N (numero delle prove) e con probabilità $1/N$ (probabilità di essere scelto tra gli N individui) mentre la distribuzione congiunta del numero di discendenti generato dagli N individui della generazione corrente è una simmetrica multinomiale. Un esempio di realizzazione di questo modello si ha nella seguente figura.

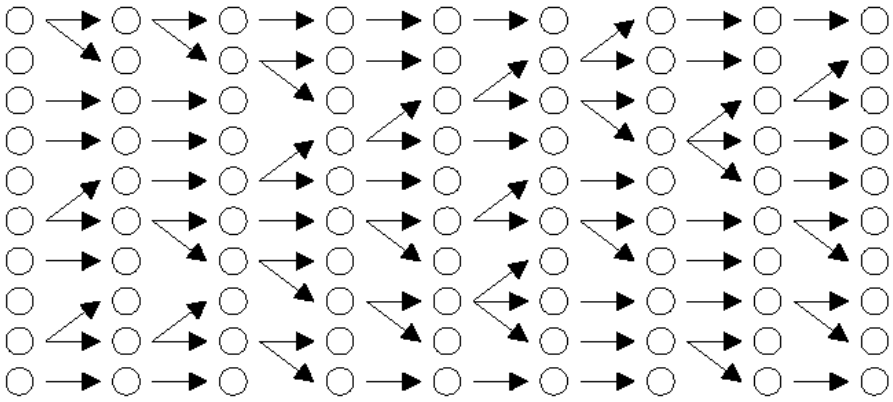

FIGURA 1.5.1 Processo genealogico secondo il modello Wright-Fisher.

Seguendo l'evoluzione del processo genealogico nel tempo dalla generazione più antica a quella corrente, si verificano delle biforcazioni quando un individuo ha due o più discendenti, mentre la sua linea evolutiva si interrompe quando non ne ha. Se però ripercorriamo la genealogia partendo dall'ultima generazione e risalendo indietro nel tempo, allora possiamo tracciare la linea evolutiva di ciascun individuo con una matita senza mai staccarla dal foglio. Avremo così un numero di linee che decresce fino a diventare eventualmente una sola quando viene raggiunto il progenitore comune più recente del

gruppo di N individui (*most recent common ancestor* o MRCA) come mostrato nella seguente figura da cui risulta anche che la maggior parte della storia genealogica della popolazione è irrilevante.

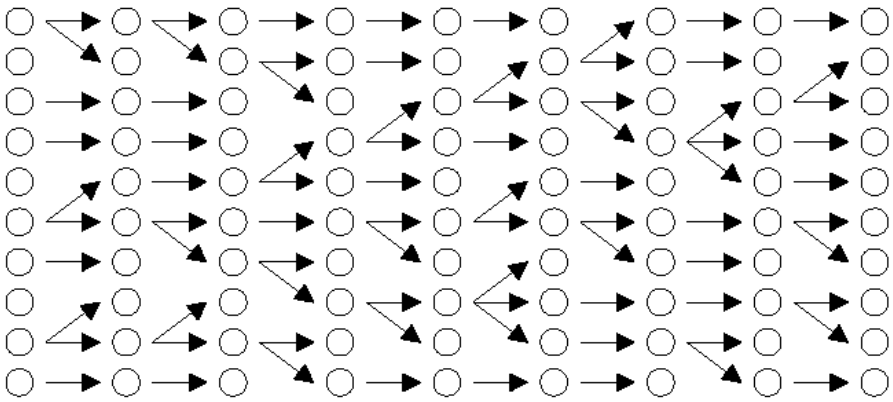

FIGURA 1.5.2 Processo genealogico secondo il modello Wright-Fisher percorso indietro nel tempo.

Data la realizzazione del progesso genealogico, si può quindi imporre un processo mutazionale. Per esempio si assegnano gli stati allelici alla generazione più antica e questi vengono trasmessi alla generazione successiva con la regola che i figli ereditano lo stato del genitore a meno di una mutazione che si verifica con una certa probabilità ad ogni generazione. In particolare per quanto abbiamo visto è sufficiente assegnare lo stato allelico all'MRCA di un gruppo di individui.

Principio 1.5.2. *In un modello neutrale è possibile modellare la genealogia di un gruppo di individui risalendo indietro nel tempo e trascurando il resto della popolazione.*

Il principio 1.5.2 è una conseguenza dell'assunzione di neutralità selettiva in base alla quale per ogni individuo di una generazione è come se si scegliesse a caso il genitore dalla generazione precedente. Da questa osservazione deriva il fatto che la genealogia di un gruppo di individui può essere generata semplicemente tracciando le linee da figlio a

padre (indietro nel tempo) di generazione in generazione.

In definitiva dai due principi enunciati consegue che l'effetto congiunto della riproduzione (casuale) e delle mutazioni (casuali e neutrali) nel determinare la composizione genetica di un gruppo di individui può essere modellizzata prima ricostruendo indietro nel tempo la loro genealogia e poi imponendo una regola mutazionale che ripercorre la genealogia avanti nel tempo di generazione in generazione.

Calcolo della conditional match probability.

Qui di seguito daremo delle definizioni di carattere operativo, rimandando alle appendici per una più esaustiva trattazione nell'ambito della genetica di popolazione e della biologia.

Definizione 1.5.3. *Data una sequenza di* DNA *mitocondriale di lunghezza L ed appartenente ad una popolazione $\mathcal{P}$, definiamo sito segregante della sequenza un locus che presenta variabilità all'interno della popolazione.*

Definizione 1.5.4. *Definiamo finite-sites model un modello per sequenze di DNA che non presentano eventi di ricombinazione ed in cui ogni sito segregante è un sito che ha subito almeno una mutazione. Un one locus finite-sites model è un finite-sites model con un solo locus.*

L'ipotesi di lavoro di un *finite-sites model* è giustificata da molti studi (Ward et al. 1991; Wakeley, 1993; Tamura e Nei, 1993; Yang, 1996) che dimostrano come i nucleotidi che si trovano nel D-loop della molecola di DNA mitocondriale violano le assunzioni dell'*infinite-sites model* (Kimura e Crow, 1964) in cui ogni mutazione interviene su di un nuovo locus. Per rendere i dati relativi a sequenze mitocondriali conformi

all'infinite-sites ipotesi, sono stati implementati degli algoritmi (GRIFFITHS e TAVARÉ, 1994) che rimuovono i siti nei quali si verifica la violazione.

Definizione 1.5.5. *Diciamo K-allele model un finite-sites model in cui ogni locus è caratterizzato da uno di K possibili alleli.*

Dimostriamo i seguenti risultati.

Proposizione 1.5.6. *Consideriamo un one-locus finite site model con due alleli A_1 ed A_2. Sia π la frequenza relativa dell'allele A_i, $i = 1$ o 2, e $1 - \pi$ la frequenza relativa dell'allele complementare. Sia poi $\theta = 2N\mu$, dove μ è la probabilità di mutazione per nucleotide per generazione ed N la numerosità effettiva di una popolazione aploide. La probabilità che due individui X_1 e X_2 non imparentati tra loro, scelti a caso nella popolazione e con un progenitore comune al tempo t espresso in unità di generazioni abbiano lo stesso allele è la seguente:*

$$P(X_1 = A_i | X_2 = A_i, \theta, t) = \left(1 - \frac{\theta t}{2}\right)\left(\pi + (1 - \pi)\frac{\theta t}{2}\right). \tag{1.5.2}$$

Dimostrazione. Nell'ipotesi del *finite-sites model* ogni sito è colpito da almeno una mutazione e dunque si possono verificare anche delle retromutazioni. In questo caso, due individui possono avere ereditato l'allele comune A_i dal loro progenitore a meno di un numero pari di mutazioni che si sono annullate (o eventualmente nessuna mutazione); oppure, nel caso in cui il progenitore avesse l'allele complementare, possono averlo ereditato per via di un numero dispari di mutazioni. Pertanto abbiamo la seguente probabilità:

$$P(X_1 = A_i | X_2 = A_i, \mu, N, t) = \pi \sum_{j=0}^{\lfloor tN/2 \rfloor} (1 - \mu)^{\lfloor tN \rfloor - 2j} \mu^{2j} +$$

$$(1-\pi)\sum_{j=1}^{\lfloor tN/2\rfloor}(1-\mu)^{\lfloor tN\rfloor-2j+1}\mu^{2j-1} \tag{1.5.3}$$

dove $\lfloor tN\rfloor$ è il più grande intero minore o uguale a tN, ed è il numero di generazioni che separano i due individui dal loro progenitore comune.

Nella prima sommatoria della (1.5.3) abbiamo

$$\sum_{j=0}^{\lfloor tN/2\rfloor}(1-\mu)^{\lfloor tN\rfloor-2j}\mu^{2j} = (1-\mu)^{\lfloor tN\rfloor}\sum_{j=0}^{\lfloor tN/2\rfloor}\left[\left(\frac{\mu}{1-\mu}\right)^2\right]^j =$$

$$\frac{1-\left(\frac{\mu}{1-\mu}\right)^{\lfloor tN\rfloor+2}}{1-\left(\frac{\mu}{1-\mu}\right)^2}\cdot(1-\mu)^{\lfloor tN\rfloor} = \frac{(1-\mu)^{\lfloor tN\rfloor+2}-\mu^{\lfloor tN\rfloor+2}}{1-2\mu} \tag{1.5.4}$$

essendo $\sum_{j=0}^{\lfloor tN/2\rfloor}(1-\mu)^{-2j}\mu^{2j}$ la ridotta $\lfloor tN/2\rfloor$-*esima* di una serie geometrica di ragione $\mu^2\cdot(1-\mu)^{-2}$ convergente se e solo se $\mu<1/2$, condizione soddisfatta per μ dell'ordine di 10^{-5} o 10^{-6}, come usualmente ipotizzato per il tasso di mutazione per nucleotide per generazione (Tavaré et al., 1996). Inoltre poiché μ è sufficientemente piccolo, vale la seguente approssimazione:

$$(1-\mu)^{\lfloor tN\rfloor+2}-\mu^{\lfloor tN\rfloor+2}\simeq 1-(\lfloor tN\rfloor+2)\mu. \tag{1.5.5}$$

Applicando l'approssimazione del processo di diffusione per $N\to\infty$ al modello neutrale di Wright-Fisher, ed assumendo quindi l'esistenza del $\lim_{N\to\infty}2N\mu$ che denotiamo con θ, otteniamo la seguente relazione:

$$\lim_{N\to\infty}\frac{1-(\lfloor tN\rfloor+2)\mu}{1-2\mu} = \lim_{N\to\infty}1-\frac{\lfloor tN\rfloor\mu}{1-2\mu} = 1-\frac{\theta t/2}{1-2\mu}\simeq 1-\frac{\theta t}{2}. \tag{1.5.6}$$

Anche la seconda sommatoria della (1.5.3) ha la forma di una ridotta di una serie geometrica convergente per $\mu<1/2$ e dunque per essa vale la seguente relazione:

$$\sum_{j=1}^{\lfloor tN/2\rfloor}(1-\mu)^{\lfloor tN\rfloor-2j+1}\mu^{2j-1} = (1-\mu)^{\lfloor tN\rfloor}\sum_{j=1}^{\lfloor tN/2\rfloor}\frac{\mu^{2j-1}}{(1-\mu)^{2j-1}} =$$

$$(1-\mu)^{\lfloor tN \rfloor}\left[\frac{1}{1-\frac{\mu}{1-\mu}}-\sum_{j=0}^{\lfloor tN/2 \rfloor}\frac{\mu^{2j}}{(1-\mu)^{2j}}\right]\simeq(1-\mu)^{\lfloor tN \rfloor}\left[1-\sum_{j=0}^{\lfloor tN/2 \rfloor}\frac{\mu^{2j}}{(1-\mu)^{2j}}\right] \quad (1.5.7)$$

Utilizzando ancora l'approssimazione del processo di diffusione quando la numerosità della popolazione tende ad infinito e tenendo conto del risultato precedente, si ottiene il seguente:

$$\lim_{N\to\infty}(1-\mu)^{\lfloor tN \rfloor}\left[1-\sum_{j=0}^{\lfloor tN/2 \rfloor}\frac{\mu^{2j}}{(1-\mu)^{2j}}\right]=(1-\frac{\theta t}{2})\cdot\left[1-\left(1-\frac{\theta t}{2}\right)\right]=$$

$$\left(1-\frac{\theta t}{2}\right)\frac{\theta t}{2} \quad (1.5.8)$$

Dunque dalle (1.5.6) e (1.5.8) otteniamo passando al limite la seguente relazione,

$$\lim_{N\to\infty}P(X_1=A_i|X_2=A_i,\mu,N,t)=P(X_1=A_i|X_2=A_i,\theta,t)=$$

$$\pi(1-\frac{\theta t}{2})+(1-\pi)\frac{\theta t}{2}(1-\frac{\theta t}{2}) \quad (1.5.9)$$

da cui la (1.5.2). □

Proposizione 1.5.7. *Consideriamo un one-locus K-allele finite site model con $K=4$ con un modello mutazionale che prevede due tipi di mutazione con tasso μ e ν e siano poi $\theta_\mu=2N\mu$ e $\theta_\nu=2N\nu$, dove N è la numerosità effettiva di una popolazione aploide. Sia π la frequenza relativa dell'allele A_i, $i=1,2,3,4$. La probabilità che due individui X_1 e X_2 non imparentati tra loro, scelti a caso nella popolazione e con un progenitore comune al tempo t espresso in unità di generazioni abbiano nel locus considerato lo stesso allele è la seguente:*

$$P(X_1=A_i|X_2=A_i,\theta,t)=\pi_i\left((1-\frac{\theta_\mu t}{2})-\frac{1}{2}\theta_\nu t\right)+(1-\pi_i)\left(\frac{\theta_\mu t}{2}+\frac{\theta_\nu t}{2}\right). \quad (1.5.10)$$

Dimostrazione. Il processo di mutazione può essere modellizzato come un processo di Markov con matrice di transizione

$$P = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 1-(\mu+\nu) & \nu/2 & \mu & \nu/2 \\ \nu/2 & 1-(\mu+\nu) & \nu/2 & \mu \\ \mu & \nu/2 & 1-(\mu+\nu) & \nu/2 \\ \nu/2 & \mu & \nu/2 & 1-(\mu+\nu) \end{pmatrix} \end{array} \tag{1.5.11}$$

dove abbiamo indicato con le lettere dell'alfabeto $\{A, C, G, T\}$ i quattro possibili alleli.

La potenza n-esima della matrice P è la matrice delle probabilità di transizione dopo n passi. Poiché P è diagonalizzabile esiste una matrice V composta da una base di autovettori degli autospazi associati alla matrice P tale che $P = VDV^{-1}$ dove D è la matrice diagonale con gli autovalori di P. Risolvendo l'equazione caratteristica della matrice P si ottiene $D = \mathrm{diag}\{1-2\mu-2\nu, 1-2\mu-2\nu, 1-2\nu, 1\}$, mentre

$$V = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix} \tag{1.5.12}$$

Pertanto

$$P^n = VD^nV^{-1} = \begin{pmatrix} a_n & b_n & c_n & b_n \\ b_n & a_n & b_n & c_n \\ c_n & b_n & a_n & b_n \\ b_n & c_n & b_n & a_n \end{pmatrix} \tag{1.5.13}$$

dove

$$a_n = \frac{1}{2}(1-2\mu-\nu)^n + \frac{1}{4}(1-2\nu)^n + \frac{1}{4} \tag{1.5.13a}$$

$$b_n = -\frac{1}{4}(1-2\nu)^n + \frac{1}{4} \tag{1.5.13b}$$

$$c_n = -\frac{1}{2}(1-2\mu-\nu)^n + \frac{1}{4}(1-2\nu)^n + \frac{1}{4} \tag{1.5.13c}$$

La probabilità di osservare, tra due individui separati da $\lfloor tN \rfloor$ generazioni dal loro progenitore comune, lo stesso allele è quindi la probabilità di non osservare differenze se

hanno ereditato l'allele comune del loro progenitore; oppure di osservare una differenza nel caso complementare:

$$P(A_i|A_i, \lfloor tN \rfloor \text{ transizioni}) = \pi_i \cdot a_{\lfloor tN \rfloor} + (1 - \pi_i) \cdot (1 - a_{\lfloor tN \rfloor}) \tag{1.5.14}$$

Applicando i risultati dell'approssimzione alla diffusione per grandi popolazioni, si ottiene, poiché μ e ν sono generalmente molto piccoli, il seguente risultato:

$$\begin{aligned} a_{\lfloor tN \rfloor} = \frac{1}{2}(1 - 2\mu - \nu)^{\lfloor tN \rfloor} + \frac{1}{4}(1 - 2\nu)^{\lfloor tN \rfloor} + \frac{1}{4} = \\ \frac{1}{2}(1 - (2\mu + \nu)\lfloor tN \rfloor) + \frac{1}{4}(1 - 2\nu\lfloor tN \rfloor) + \frac{1}{4} \end{aligned} \tag{1.5.15}$$

da cui per N tendente ad infinito

$$a_{\lfloor tN \rfloor} \to (1 - \frac{\theta_\mu t}{2}) - \frac{1}{2}\theta_\nu t \tag{1.5.16}$$

dalla quale segue la (1.5.10). □

Abbiamo visto che nel calcolo della *conditional match probability* intervengono le relazioni genetiche tra i due individui X_1 e X_2. Quando i dati $\mathcal{D}$ relativi alla popolazione $\mathcal{P}$ sono disponibili, possiamo stimare i parametri relativi al processo demografico ed al processo mutazionale definendo un appropriato modello statistico e quindi calcolare la *posterior match probability* di interesse. Infatti, indicando con X la sequenza mitocondriale relativa agli individui X_1 ed X_2, possiamo osservare che:

$$P(X_1 = X|X_2 = X, \bar{H}, \mathcal{D}) = \int_{\theta,t} P(X_1 = X|X_2 = X, \bar{H}, \theta, t)P(\theta, t|\mathcal{D})\ d\theta dt \tag{1.5.17}$$

dove $P(X_1 = X|X_2 = X, \bar{H}, \theta, t)$ è noto dalla proposizione 1.5.6 o dalla proposizione 1.5.7, mentre $P(\theta, t|\mathcal{D})$ è la distribuzione a posteriori dei parametri genealogici e richiede la specificazione di un modello statistico appropriato.

Si può calcolare un'approssimazione della *posterior match probability* (1.5.17) utilizzando il metodo Monte Carlo per il quale se una variabile aleatoria Y ha densità f_Y,

allora la media di ogni funzione dello spazio L^2, misurabile $g(Y)$ può essere approssimata simulando un campione $(Y^{(1)}, Y^{(2)}, \ldots, Y^{(M)})$ dalla distribuzione di densità f_Y e calcolandone la media:

$$E(g(Y)) = \int g(y) f_Y(y)\ dy \simeq \frac{1}{M} \sum_{k=1}^{M} g(Y^{(k)}) \tag{1.5.18}$$

La (1.5.18) fornisce una buona approssimazione per M sufficientemente grande—formalmente l'errore tende a zero per M tendente ad infinito.

Applicando la (1.5.18) alla *posterior match probability* espressa nella (1.5.17) ed associata al profilo X si avrà

$$P(X_1 = X | X_2 = X, \bar{H}, \mathcal{D}) \simeq \frac{1}{K} \sum_{k=1}^{K} P(X_1 = X | X_2 = X, \bar{H}, \theta^{(k)}, t^{(k)}) \tag{1.5.19}$$

dove $(\theta^{(k)}, t^{(k)}) \sim P(\theta, t|\mathcal{D})$.

Capitolo 2

2.1 Modello mutazionale.

Indichiamo con π_j la frequenza della j-esima base o j-esimo nucleotide, dove i valori $j = 1, 2, 3, 4$ corrispondono alle quattro basi A, C, G e T e consideriamo il più semplice modello di sostituzione che prevede un tasso di mutazione u identico per tutte le basi.

Quando si verifica una mutazione, la base mutante si trasforma in una nuova base di tipo j con probabilità costante π_j. Rientra in questo schema anche il caso particolare in cui la mutazione non è osservabile e che si verifica quando la base iniziale e quella mutata sono dello stesso tipo.

Sia u il tasso di mutazione per nucleotide per generazione. La probabilità che non si verifichino mutazioni dopo T generazioni è $(1-u)^T$ e pertanto la probabilità α che ci sia stata mutazione è la seguente:

$$\alpha = 1 - (1-u)^T \approx 1 - \exp(-uT) \tag{2.1}$$

La probabilità di una mutazione dalla base i alla base j dopo T generazioni possiamo allora scriverla (Felsenstein, 1981) come

$$P_{jj}(T) = (1-\alpha) + \alpha\pi_j \tag{2.2a}$$

$$Pij(T) = \alpha\pi_j, \quad j \neq i \tag{2.2b}$$

Quando $\pi_j = 1/4$ per $j = 1, 2, 3, 4$, le probabilità (2.2a) e (2.2b) definiscono il modello mutazionale di Jukes-Cantor (v. appendice) con una differente interpretazione del tasso di mutazione. Il tasso di mutazione u nelle (2.2), identico per tutte le basi, è infatti uguale a 4/3 volte il tasso di mutazione μ, anch'esso identico per tutte le basi, del modello di Jukes-Cantor.

2.2 Modello demografico: coalescenza standard.

La coalescenza è un modello stocastico che rappresenta le relazioni ancestrali tra un campione di n sequenze di DNA. Il modello ha due importanti caratteristiche: è matematicamente trattabile ed approssima la distribuzione delle genealogie delle sequenze nell'ambito di un'importante classe di modelli *neutrali* della genetica di popolazione, incluso il modello di Wright-Fisher per popolazioni ad accoppiamento casuale (*random mating*) e numerosità effettiva N costante nel tempo.

Al fine di ottenere le approssimazioni di questi modelli, l'unità del tempo di coalescenza deve essere interpretata come N/σ^2 generazioni, dove σ^2 denota la varianza del numero di discendenti di ciascun individuo nella generazione successiva (Kingman, 1982). Noi considereremo $\sigma^2 = 1$ (Donnelly, 1996) che permette di misurare il tempo di coalescenza unicamente in termini di numero di generazioni.

In generale N indica il numero di cromosomi in una popolazione diploide. Pertanto nel caso di dati relativi al DNA mitocondriale (popolazione apoloide), la numerosità N della popolazione corrisponde a $2N$ individui: N maschi ed N femmine.

Come abbiamo detto, la coalescenza è un modello stocastico che rappresenta le relazioni ancestrali tra un campione di n sequenze di DNA, corrispondenti ad n distinti individui. Graficamente possiamo visualizzarlo ricorrendo ad un albero genealogico. Un esempio di albero di coalescenza per un campione di 5 individui è illustrato nella seguente figura 2.2.1

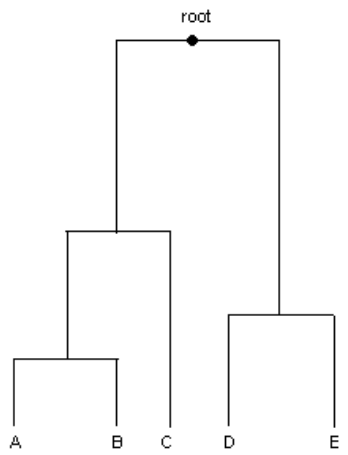

FIGURA 2.2.1 Esempio di albero di coalescenza per un campione di 5 individui

Il tempo di coalescenza percorre all'indietro il processo genealogico. Il tempo $t_0 = 0$ indica il presente che corrisponde alle foglie dell'albero, mentre t_k, $k = 1, 2, \ldots, n-1$, denota il tempo del k-esimo evento di coalescenza—un evento di coalescenza si verifica quando due rami si uniscono. In particolare t_{n-1} denota il tempo della radice dell'albero detta *most recent common ancestor*, brevemente indicata con MRCA.

Nel modello standard, gli intervalli di tempo $t_k - t_{k-1}$ tra due successivi eventi di coalescenza hanno distribuzioni esponenziali indipendenti:

$$P(t_k > t | t_{k-1} = t') = \exp\left(\gamma_k (t' - t)\right), \qquad \gamma_k = \binom{n+1-k}{2} \tag{2.2.2}$$

per $t > t'$. Ad ogni evento di coalescenza, tutte le coppie dei rimanenti rami hanno la medesima probabilità di coalescere, nell'ipotesi di assenza di fenomeni di selezione che privilegiano la trasmissione di determinati genotipi.

Al modello standard della coalescenza sono associate le seguenti proprietà caratteristiche:

(i) un lungo periodo di tempo—in media in un tempo maggiore della metà del tempo di

coalescenza che separa le n sequenze dal loro *most recent common ancestor*—durante il quale l'albero possiede solamente due rami;

(ii) l'elevata varianza dell'altezza totale dell'albero—la standard deviation tipicamente è pari all'incirca al 60% della media.

2.3 Modello per Single-Nucleotide Polymorphism data.

Un locus *Single-Nucleotide Polymorphism* (SNP) è una singola posizione nel DNA che presenta variazioni tra differenti individui appartenenti tutti alla medesima popolazione. Tutti i locus SNPs presentano due sole varianti e questa è l'assunzione che facciamo.

Molte posizioni nel genoma si presentano identiche per i diversi individui della popolazione. L'utilizzo di dati di tipo SNPs permette di determinare, attraverso una varietà di metodi sperimentali, i siti che si presentano polimorfici all'interno della popolazione. Questo permette ai successivi studi di valutare la variabilità genetica solamente nei siti che a priori è noto esibiscono polimorfismo. L'utilità di questo approccio legato ad i dati di tipo SNPs è determinata dalla riduzione dei costi sperimentali per la valutazione della variabilità genetica.

In generale il pattern di variazione ad un particolare locus dipende dagli eventi demografici che hanno interessato la popolazione di appartenenza, dal processo mutazionale in quel particolare locus e dagli effetti della selezione.

Per i dati SNPs è stato introdotto un modello gerarchico (Nicholson et al., 2002), motivato da considerazioni di genetica di popolazione. Tale modello presenta la stessa struttura probabilistica per ogni locus SNP e caratterizza le popolazioni studiate introducendo un parametro specifico che ne cattura le differenze demografiche. Si rimanda alle appendici una descrizione completa di tale modello gerarchico.

Noi modificheremo in maniera appropriata il modello gerarchico di (NICHOLSON et al., 2002) introducendo un processo mutazionale ed un processo demografico che ci permetteranno di fare inferenza sulle grandezze di interesse per la valutazione della *conditional match probability* ai fini forensi così come abbiamo ampiamente discusso nel capitolo precedente.

Supponiamo di avere un data set $\mathcal{D}$ composto da L SNPs loci relativi ad una determinata popolazione $\mathcal{P}$ moderna e indichiamo con n la numerosità del campione. Per ogni SNP, arbitrariamente fissiamo una delle due varianti. D'ora in avanti indicheremo con 0 la variante più comune all'interno della popolazione $\mathcal{P}$ e con 1 quella più rara. Sia quindi y_j il numero di copie nel campione della variante più rara osservata nel j-esimo locus.

Introduciamo le quantità non osservate π_j, $j = 1, 2, \ldots, L$, che rappresentano la frequenza relativa della variante più comune in una popolazione ancestrale dalla quale la popolazione P moderna discende. Allora

$$\beta_j = (1 - \exp(-\mu_j t))\pi_j \qquad j = 1, 2, \ldots L \tag{2.3.1}$$

è la probabilità che nel j-esimo locus, trascorso un tempo t, si sia verificata una mutazione e la variante comune 0 si sia modificata nella variante 1, mentre

$$\alpha_j = ((1 - \exp(-\mu_j t))\pi_j + (1 - \pi_j)\exp(-\mu_j t) \qquad j = 1, 2, \ldots L \tag{2.3.2}$$

è la frequenza relativa della variante complementare a quella comune e che abbiamo indicato con 1.

Osserviamo che in assenza di un processo mutazionale, che si verifica quando il tasso di mutazione μ è pari a 0, dalle (2.3.1) e (2.3.2) si ha:

$$\beta_j = 0 \quad \text{e} \quad \alpha_j = 1 - \pi_j. \tag{2.3.3}$$

Ovvero α_j è la frequenza relativa della variante più rara nella popolazione relativa, nel nostro caso, al j-esimo locus e coincide con quella della popolazione ancestrale, in accordo con il modello per SNPs definito da (Nicholson et al., 2002).

Nel modello per SNPs (Nicholson et al., 2002), i dati vengono modellizzati in maniera binomiale:

$$Y_j \sim \text{Binomial}(n, \alpha_j) \tag{2.3.4}$$

dove Y_j è il numero di copie della variante scelta nel j-esimo locus, α_j è la frequenza relativa della medesima variante ed n è il numero delle sequenze campionate.

Nella nostra notazione, α_j rappresenta la frequenza della variante mutata , quindi il modello (2.3.4) rappresenta la distribuzione del numero di mutazioni nel j-esimo sito.

Poiché una variabile aleatoria binomiale può essere interpretata come somma di variabili aleatorie di Bernoulli indipendenti

$$Y_j = \sum_{i=1}^{n} X_{ij}, \qquad X_{ij} \sim \text{Bernoulli}(\alpha_j) \quad \forall i = 1, 2, \ldots, n \tag{2.3.5}$$

possiamo interpretare il j-esimo locus di ogni sequenza di DNA mitocondriale come una variabile aleatoria di Bernoulli di parametro α_j.

Dalla (2.3.5) segue quindi

$$f_{Y_j}(y_j) \propto \beta_j^{x_{\cdot j}}(1-\beta_j)^{n-x_{\cdot j}} = \prod_{i=1}^{n} f_{X_{ij}}(x_{ij}) \tag{2.3.6}$$

con $x_{\cdot j} = \sum_{i=1}^{n} x_{ij}$.

Inizialmente, per semplicità, supponiamo che le frequenze relative delle due varianti nella popolazione moderna non abbiano risentito di fenomeni di tipo *bottleneck*—relativi a variazioni nella numerosità della popolazione—e corrispondano quindi a quelle della popolazione ancestrale.

Definiamo una distribuzione sulle frequenze ancestrali con una distribuzione Beta simmetrica. La scelta della distribuzione sulle frequenze ancestrali richiama le seguenti

argomentazioni di genetica di popolazione. Consideriamo lo *standard neutral model*. Per questo modello è noto che in una popolazione di numerosità N, in un sito segregante che muta lentamente, la probabilità che la variante mutata sia presente in y copie, $y = 1, 2, \ldots N-1$, è proporzionale a $1/y$ (Ewens, 1979). Poiché non si conosce quale delle due varianti è quella mutata, si considera la versione simmetrizzata:

$$f_p(p) \propto \frac{1}{p(1-p)} \qquad p \in (0,1). \tag{2.3.7}$$

Nel modello mutazionale (2.1.1) il tempo T è espresso in numero di generazioni. Con riferimento al modello della coalescenza, considereremo invece il tempo come una variabile aleatoria continua. In tale modello, l'intervallo di tempo W_k durante il quale un campione di n individui ha k distinti progenitori, $2 \leq k \leq n$, è distribuito esponenzialmente con parametro di scala $\lambda = k(k-1)/2$. Inoltre gli intervalli W_k sono mutuamente indipendenti al variare di k.

Una tale descrizione fornisce una buona approssimazione per molti modelli della genetica di popolazione nei quali il tempo è espresso in numero di generazioni, con la condizione che l'unità del tempo di coalescenza sia interpretata come N generazioni dove N rappresenta la numerosità della popolazione campionata.

Definiamo la grandezza

$$T_n = \sum_{k=2}^{n} W_k \tag{2.3.8}$$

che rappresenta il tempo che separa gli n individui dal loro progenitore comune (*most recent common ancestor o MRCA*). Poiché le variabili aleatore W_k hanno speranza matematica $2/k(k-1)$, la media di T_n è la seguente

$$E(T_n) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right). \tag{2.3.9}$$

Osserviamo dalla (2.3.9) che quando il campione di numerosità n è grande, allora $E(T_n)$ tende a 2 unità del tempo di coalescenza, equivalente a $2N$ generazioni.

Sulla base di queste osservazioni, e supponendo inoltre che la popolazione ancestrale sia sufficientemente distante nel tempo rispetto alla popolazione moderna $\mathcal{P}$, il parametro α_j definito dalla (2.3.1) pu essere riscritto come:

$$\alpha_j = ((1 - \exp(-\mu_j t N))\pi_j + (1 - \pi_j)\exp(-\mu_j t N) \tag{2.3.10}$$

dove t è il tempo di coalescenza ed è distribuito come una variabile aleatoria esponenziale di media 2.

2.4 Metodi di simulazione.

La soluzione al problema di fare inferenza sulle grandezze di interesse, definite nella (2.3.2), π μ, t ed N da un campione di n sequenze è data dalla distribuzione a posteriori condizionatamente all'osservazione dei dati D. In questa sezione descriviamo alcuni algoritmi per ottenere tale soluzione.

Dalla definizione di probabilità condizionata, segue

$$f_\Pi(\pi_j|D) \propto f_\Pi(\pi_j)P(D|\pi_j) \tag{2.4.1}$$

in cui $P(D|\pi_j)$ indica il modello descritto nella sezione precedente.

L'equazione (2.4.1) non è sempre utilizzabile direttamente perché non conosciamo, in generale, un'espressione esplicita di $P(D|\pi_j)$, dipendendo il modello dal parametro α. Però si può osservare che

$$f_\Pi(\pi_j|D) \propto f_\Pi(\pi_j)P(D|\pi_j) = \int_o^1 f_{\Pi,T}(\pi_j, t)P(D|\pi_j, t)\ dt =$$

$$\int_0^1 f_{\Pi,T}(\pi_j, t)P(D|\alpha)\ dt. \tag{2.4.2}$$

Inizialmente non introduciamo incertezza nel processo mutazionale e quindi non definiamo una distribuzione sulla grandezza μ che consideriamo invece una costante;

assumiamo inoltre che anche N sia una grandezza fissa. Sia quindi

$$\alpha_j = (1 - 2\exp(-\mu t N))\pi_j + \exp(-\mu t N) = \alpha_1 \pi_j + \exp(-\mu t N) \tag{2.4.3}$$

in cui il parametro di interesse è π_j.

La (2.4.2) può essere valutata ricorrendo ad un algoritmo di tipo *Accept/Reject* (Ripley, 1987), come il seguente.

Algoritmo 2.4.1.

1. *Simulazione di p_j e t distribuiti rispettivamente come una Beta simmetrica di parametro p ed un'Esponenziale di media 2;*
2. *si calcola α_j secondo la definizione (2.4.3):*

$$\alpha_j = (1 - 2\exp(-\mu t N))\pi_j + \exp(-\mu t N)$$

3. *si accetta la coppia (π_j, t) e con probabilità definita da:*

$$u = \frac{\alpha_j^y (1-\alpha_j)^{n-y}}{\left(\frac{y}{n}\right)^y \left(1 - \frac{y}{n}\right)^{n-y}} = \frac{\text{Binomial}(y, \alpha_j)}{\text{Binomial}\left(y, \frac{y}{n}\right)} \tag{2.4.4}$$

altrimenti la si rifiuta e si torna al passo 1.

È possibile sostituire il passo 3. con un passo che sia invece basato sui metodi MCMC come il seguente.

3'. *Si accetta la coppia $(\pi_j, t)^{(k)}$ all'k-esima realizzazione della catena MCMC con probabilità definita da:*

$$\min(u^{(k)}, u^{(k-1)}), \quad \text{con} \quad u^{(k)} \sim \text{Binomial}(y, \alpha_j). \tag{2.4.5}$$

Se $(\pi_j, t)^{(k)}$ viene rifiutato, allora si accetta il valore $(\pi_j, t)^{(k-1)}$ quale nuovo $(\pi_j, t)^{(k)}$.

Infatti consideriamo una variabile aleatoria uniforme U sull'intervallo $(0,1)$. Tenendo conto della (2.4.2), segue

$$f\left(\Pi \leq \pi, U \leq \frac{1}{M}P(D|\alpha)\right) = \int_0^\infty f\left(\Pi \leq \pi, T \leq t, U \leq \frac{1}{M}P(D|\alpha)\right)\ dt =$$

$$\int_0^\infty \left(\int_0^{\frac{1}{M}P(D|\alpha)} du\right) f(\pi,t)\ dt = \frac{1}{M}\int_0^\infty f(\pi,t)P(D|\alpha)\ dt = \frac{1}{M}f(\pi|D) \qquad (2.4.6)$$

Poiché la (2.4.6) vale per ogni π, per $\pi = 1$ si ottiene

$$\frac{1}{M} = f\left(U \leq \frac{1}{M}P(D|\alpha)\right) \qquad (2.4.7)$$

e dunque dalla (2.4.6) segue

$$f(\pi_j|D) = \frac{f\left(\Pi \leq \pi, U \leq \frac{1}{M}P(D|\alpha)\right)}{f\left(U \leq \frac{1}{M}P(D|\alpha)\right)} = f\left(\Pi \leq \pi | U \leq \frac{1}{M}P(D|\alpha)\right) \qquad (2.4.8)$$

Dalla (2.4.8) risulta quindi che la scelta ottimale della costante M per l'efficienza dell'algoritmo è

$$M = \max_\alpha P(D|\alpha) \qquad (2.4.9)$$

anche se l'algoritmo funziona per ogni $M \geq \max_\alpha P(D|\alpha)$. A questo proposito osserviamo che il denominatore della (2.4.4) soddisfa la (2.4.9), essendo

$$\text{Binomial}\left(y, \frac{y}{n}\right) = \max_\alpha \text{Binomial}(y,\alpha). \qquad (2.4.10)$$

e $y \neq 0$ in quanto si considerano siti polimorfici.

Nell'approccio basato sui metodi MCMC, invece, i valori simulati possono ripetersi e una realizzazione tipica della catena sarà del seguente tipo:

$$(\pi_j,t)^{(1)}, (\pi_j,t)^{(1)}, \ldots, (\pi_j,t)^{(2)}, (\pi_j,t)^{(2)}, \ldots \qquad (2.4.11)$$

Si seleziona quindi un *burnin period* che rappresenta la fase di "assestamento"dopo la quale il vettore aleatorio $(\pi_j,t)^{(k)}$ può essere considerato come generato dalla distribuzione di interesse.

L'analisi precedente è condizionata al valore del parametro mutazionale $\theta = 2N\mu$, oltre che dal tempo di coalescenza t, ed è quindi applicabile quando la numerosità effettiva N della popolazione ed il tasso di mutazione μ sono noti. In pratica, invece, si ha una sostanziale incertezza su questi parametri. Spesso il valore di μ viene stimato confrontando sequenze omologhe, ma rimane comunque un'incertezza sui valori di tali stime. Per quanto riguarda invece la numerosità effettiva della popolazione N, tipicamente si dispone di poca informazione.

L'algoritmo 2.4.1 può essere facilmente modificato in maniera tale da incorporare incertezza sulle grandezze μ ed N. Assumiamo che prima di osservare i dati, N e μ siano mutuamente indipendenti e siano indipendenti anche da t e π_j. Allora l'equazione (2.4.2) può essere riscritta nel seguente modo:

$$f(\pi_j|D) \propto f(\pi_j)P(D|\pi_j) =$$
$$\int_0^\infty \int_0^\infty \int_0^\infty f_{\Pi,T}(\pi_j, t) f_N(n) f_M(\mu) P(D|\alpha_j)\ dt\ dN\ d\mu. \tag{2.4.10}$$

Motivazioni analoghe a quelle già esposte permettono di scrivere il seguente algoritmo.

Algoritmo 2.4.2

1. *Simulazione di N da f_N;*
2. *simulazione di μ da f_M;*
3. *simulazione di π_j e t distribuiti rispettivamente come una Beta simmetrica di parametro p ed un'Esponenziale di media 2;*
4. *calcolo di α_j secondo la definizione (2.4.3);*

$$\alpha_j = (1 - 2\exp(-\mu t N))\pi_j + \exp(-\mu t N)$$

5. *si accetta π_j, t, N e μ con probabilità definita da*

$$u = \frac{\alpha_j^y (1-\alpha_j)^{n-y}}{\left(\frac{y}{n}\right)^y \left(1-\frac{y}{n}\right)^{n-y}} = \frac{\text{Binomial}(y, \alpha_j)}{\text{Binomial}\left(y, \frac{y}{n}\right)} \tag{2.4.11}$$

altrimenti si ritorna al passo 1.

Come nell'algoritmo 2.4.1, il passo 5. può essere sostituito da un passo basato sui metodi MCMC.

Finora abbiamo considerato $T \sim \text{Exp}(2)$, ovvero abbiamo considerato una popolazione ancestrale, dalla quale la popolazione moderna discende, sufficientemente distante nel tempo.

Nel modello della coalescenza, il tempo T che separa il campione dal *most recent common ancestor* (MRCA) è stato definito nella (2.3.8) come $T = \sum_{k=2}^{n} W_k$, dove i tempi W_k hanno una distribuzione esponenziale di parametro di scala $k(k-1)/2$, e sono mutuamente indipendenti tra loro. Questa osservazione ci permette di modificare l'algoritmo 2 tenendo conto della definizione del tempo di coalescenza T nel seguente nuovo algoritmo.

Algoritmo 2.4.3.

1. *Simulazione di N da f_N;*
2. *simulazione di μ da f_M;*
3. *simulazione di π_j da una distribuzione Beta simmetrica di parametro p e W_k indipendenti e distribuiti esponenzialmente con parametro $k(k-1)/2$, $k = 2, \ldots, n$;*
4. *calcolo di t ed α_j secondo le definizioni (2.2.8) e (2.3.3) rispettivamente;*

$$t = \sum_{k=2} W_k, \quad \alpha_j = (1 - 2\exp(-\mu t N))\pi_j + \exp(-\mu t N)$$

5. *si accetta π_j, t, N e μ con probabilità definita da*

$$u = \frac{\alpha_j^y (1-\alpha_j)^{n-y}}{\left(\frac{y}{n}\right)^y \left(1-\frac{y}{n}\right)^{n-y}} = \frac{\text{Binomial}(y, \alpha_j)}{\text{Binomial}\left(y, \frac{y}{n}\right)} \tag{2.4.12}$$

altrimenti si ritorna al passo 1.

2.5 Modello demografico: modello di coalescenza con dimensione della popolazione variabile.

Il modello standard della coalescenza è un caso speciale con $\lambda(s) \equiv 1$ del modello in cui l'equazione (2.2.2) è sostituita con la seguente:

$$P(t_k > t | t_{k-1} = t') = \mathrm{Exp}(\gamma_k(\Lambda(t') - \Lambda(t))) \tag{2.5.1}$$

dove $\Lambda(\cdot)$ è una funzione differenziabile non decrescente così definita:

$$\Lambda(t) \equiv \int_0^t \frac{1}{\lambda(s)} ds. \tag{2.5.2}$$

Un modello così definito approssima la genealogia di un campione relativo ad una *random mating population*—popolazione ad accoppiamento casuale, secondo le assunzioni del modello di Wright-Fisher (vedi appendice A)—di numerosità $N\lambda(t)$ al tempo $N \int_0^t \lambda(s) ds$ generazioni fa (Hudson, 1991).

Intuitivamente un incremento del tempo di coalescenza corrisponde ad avere un maggior numero di generazioni durante le quali la numerosità della popolazione è più grande rispetto a quelle durante le quali invece è più piccola.

K-size model.

Un semplice modello che tenga conto della variabilità della numerosità della popolazione è il cosiddetto "*k-size model*"o "*k-step model*". Nel caso particolare in cui $k = 2$, il *k-size model* è definito dalle seguenti equazioni:

$$\lambda(t) = \delta, \quad \Lambda(t) = t/\delta \qquad 0 < t < t_g \tag{2.5.3a}$$

$$\lambda(t) = 1, \quad \Lambda(t) = t_g/\delta + t - t_g \qquad t > t_g \tag{2.5.3b}$$

Lo scenario demografico descritto dalle equazioni (2.5.3) è quello di una popolazione di numerosità effettiva N costante fino al tempo t_g. Al tempo t_g si verifica istantaneamente una variazione nella dimensione della popolazione che da una numerosità N passa ad una numerosità $N\delta$.

Modello a crescita esponenziale.

Un altro modello demografico che descrive una variazione nella numerosità della popolazione è quello che prevede una crescita esponenziale al tasso r. Nei termini delle funzioni $\lambda(t)$ e $\Lambda(t)$, tale modello è specificato dalle seguenti equazioni:

$$\lambda(t) = c\exp(-Rt) \tag{2.5.4a}$$

$$\Lambda(t) = c\exp(Rt)/R \tag{2.5.4b}$$

nelle quali $R = Nr$ e c è una costante scelta arbitrariamente.

Quando $R > 0$, ci sono meno eventi di coalescenza recenti rispetto al modello standard della coalescenza, a parità di lunghezza totale attesa dei rami. Osserviamo anche che il modello standard della coalescenza può essere ottenuto dal modello a crescita esponenziale con un'appropriata scelta della costante c.

Quando invece $R < 0$, ogni evento di coalescenza ha una probabilità positiva di non occorrere in un tempo finito, portando a gruppi composti da sequenze separate da un numero infinito di mutazioni.

Una crescita esponenziale della popolazione ha l'effetto di accorciare i rami più lunghi dell'albero di coalescenza, rendedo più brevi i tempi di coalescenza tra gli individui della popolazione. Questo fatto conferisce all'albero una struttura a forma di stella come nella seguente figura.

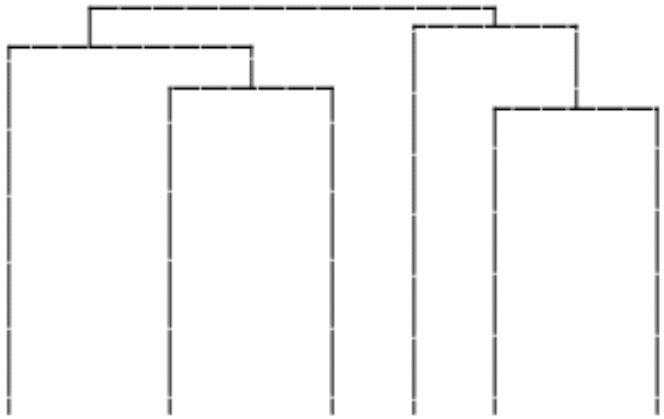

FIGURA 2.5.1. Esempio di albero di coalescenza con una popolazione a crescita esponenziale.

Modello a crescita esponenziale a due parametri o modello di coalescenza con crescita.

Una crescita puramente esponenziale, come quella descritta nelle equazioni (2.5.4) non fornisce un buon modello per la popolazione umana considerata globalmente. In un tale modello infatti alti tassi di crescita recenti avrebbero come implicazione la scomparsa di piccole popolazioni avvenuta alcune migliaia di anni fa.

Partendo da questa osservazione, (Marjoram e Donnelly, 1994) , è stato considerato un modello a crescita esponenziale a due parametri, descritto dalla seguente funzione:

$$\lambda(t) = \begin{cases} \exp(R(t_g - t)) & \text{se } 0 < t < t_g \\ 1 & \text{se } t > t_g \end{cases} \tag{2.5.5}$$

Lo scenario demografico di questo modello corrisponde a quello di una popolazione di numerosità costante N fino a Nt_g generazioni fa, tempo dopo il quale la popolazione cresce al tasso r per generazione fino a raggiungere la numerosità N_c:

$$N_c = N(1+r)^{Nt_g} \simeq N\exp(Rt_g). \tag{2.5.6}$$

Nelle ipotesi di questo modello, che seguendo (Wilson et al., 2002) chiamiamo *modello di coalescenza con crescita*, la funzione $\Lambda(\cdot)$ definita nella (1.5.2) ha la seguente

espressione:

$$\Lambda(t) = \begin{cases} \frac{1}{R}\exp(-Rt)[\exp(Rt) - 1] & \text{se } 0 < t < t_g \\ t - t_g + \frac{1}{R}[1 - \exp(-Rt_g)] & \text{se } t > t_g \end{cases} \tag{2.5.7}$$

Dalla (2.5.7) deriva che la distribuzione dei tempi di coalescenza è la seguente:

$$P(t_{n-k+1} > t | t_{n-k} = t') = \begin{cases} \exp\{\frac{\gamma_k}{R}[\exp(Rt') - \exp(Rt)]\exp(-Rt_g)\} & \text{se } t' < t < t_g \\ \exp\{\gamma_k(t_g - t + \frac{1}{R}(\exp[R(t' - t_g)] - 1])\} & \text{se } t' < t_g < t \\ \exp[\gamma_k(t' - t)] + t_{n-k} & \text{se } t_g < t' < t \end{cases} \tag{2.5.8}$$

Il *modello di coalescenza con crescita* si riduce al modello standard della coalescenza quando $t_g = 0$ e al limite per $R \to 0$.

2.6 Modello per SNPs II.

Nella sezione 2.3, lo scenario demografico sottostante al modello per SNPs che abbiamo proposto è quello della coalescenza standard, descritto brevemente nella sezione 2.2 (vedi appendice per maggiori dettagli). Questo modello prevede una numerosità effettiva N della popolazione che rimane costante nel tempo. Nella sezione 2.5 abbiamo descritto altri modelli demografici che prevedono, invece, una variazione nella numerosità della popolazione. In questa sezione vedremo come sia possibile incorporare questi scenari demografici nel modello per SNPs proposto.

È verosimile supporre che molte popolazioni abbiamo subito nel corso del tempo un incremento non indifferente nella loro numerosità effettiva N a causa o di un precedente rapido declino oppure a causa di un'espansione geografica.

Per molte di queste popolazioni abbiamo visto nella sezione 2.5 che un modello di crescita esponenziale è ragionevole oltreché facilmente trattabile dal punto di vista matematico.

Nella sezione 2.3, nella descrizione del modello per SNPs, abbiamo introdotto le variabili aleatorie W_k, con $k = 2, \ldots, n$, che rappresentano gli intervalli di coalescenza, ovvero gli intervalli di tempo durante i quali il campione di n individui ha k distinti progenitori. Nella notazione utilizzata nella sezione 2.5 abbiamo indicato

$$W_k = t_{n-k+1} - t_{n-k}, \quad t_0 = 0 \qquad k = 2, \ldots, n \tag{2.6.1}$$

dove t_k è il tempo del k-esimo evento di coalescenza.

Nel modello della coalescenza standard, gli intervalli di coalescenza W_k, $k = 2, \ldots, n$, hanno la distribuzione di una variabile aleatoria esponenziale di parametro di scala $\gamma_k = k(k-1)/2$ e sono mutuamente indipendenti. L'effetto di una variazione nella numerosità della popolazione comporta una differente distribuzione degli intervalli W_k. In particolare questi intervalli non sono più mutuamente indipendenti.

Supponiamo che la numerosità effettiva della popolazione al tempo del campionamente sia N e misuriamo il tempo in unità di generazioni. Indichiamo poi con $Np(t)$ la numerosità della popolazione al tempo t e definiamo la funzione $\lambda(t)$ nella seguente maniera:

$$\lambda(t) = \frac{1}{p(t)}, \qquad t \in (0, \infty) \tag{2.6.2}$$

Sotto un'ampia classe di modelli demografici (richiamati nella sezione 2.5), la distribuzione dell'intervallo di tempo W_k durante il quale esistono k progenitori del campione condizionata al tempo $t_{n-k} = W_{k+1} + \ldots + W_n$, ha la seguente forma:

$$P(W_k > t - t' | W_{k+1} + \ldots + W_n = t') = P(t_{n-k+1} > t | t_{n-k} = t') =$$
$$\exp\left[-\binom{k}{2} \int_{t'}^{s+t'} \lambda(t)dt\right] = \exp\left[-\gamma_k \left(\Lambda(s + t') - \Lambda(t')\right)\right], \tag{2.6.3}$$

dove abbiamo indicato con γ_k il coefficiente binomiale ed $s = t - t'$.

La distribuzione (2.6.3) fornisce una maniera diretta di simulare gli intervalli di tempo $W_2, \ldots, W_n$. Infatti gli algoritmi descritti nella sezione 2.4 sono ancora validi, con

le opportune modifiche, fintanto che è possibile simulare le variabili aleatorie W_k dalla distribuzione (2.6.3).

2.7 Metodi di simulazione II.

In questa sezione vedremo che negli scenari demografici di una crescita esponenziale della numerosità della popolazione, gli algoritmi precedentemente descritti rimangono ancora validi con le opportune modifiche.

Modello a crescita esponenziale.

Nel caso di una popolazione con crescita esponenziale la funzione $\Lambda(\cdot)$ ha la seguente forma:

$$\Lambda(t) = \frac{1}{R}\exp(Rt). \tag{2.7.1}$$

Sostituendo la (2.7.1) nella (2.6.3), segue che la distribuzione degli intervalli di tempo W_k, per $k = 2, \ldots, n$ è definita dalle seguenti probabilità:

$$P(W_k > s | W_{k+1} + \ldots W_n = t') = \exp\left\{-\gamma_k \frac{\mathrm{e}^{Rt'}}{R}[\mathrm{e}^{Rs} - 1]\right\} \tag{2.7.2}$$

Poiché la distribuzione è di tipo esponenziale, possiamo generare la variabile aleatoria di interesse cercando una funzione $g(u)$ tale che $Y = g(U)$ abbia la distribuzione richiesta, essendo U distribuita uniformemente sull'intervallo $(0, 1)$.

Sia $R > 0$ il tasso di crescita della popolazione. Dalla (2.7.2) segue

$$RW_k = \ln\left[1 - \frac{R}{\gamma_k}\exp\left(-R\sum_{j=1}^{n-k} W_{k+j}\right)\ln U\right] \qquad U \sim \mathrm{Unif}(0,1) \tag{2.7.3}$$

da cui, essendo $W_k = t_{n-k+1} - t_{n-k}$ e $t_{n-k} = W_{k+1} + \ldots + W_n$, risulta

$$t_{n-k+1} = \frac{1}{R}\ln\left[\exp(Rt_{n-k} - \frac{R}{\gamma_k}\ln U_k\right]. \tag{2.7.4}$$

Algoritmo 2.7.1.

1. *Per* $k = n, n-1, \ldots, 2$ *si simula* $t_n - k + 1$ *definito dalla (2.7.4)*

$$t_{n-k+1} = \frac{1}{R} \ln \left[\exp(R t_{n-k} - \frac{R}{\gamma_k} \ln U_k \right], \qquad t_0 = 0$$

   e si calcola $T = \sum_{k=2}^{n} W_k = t_{n-1}$*;*

2. *si calcola* N *definito da* $N(t) = N_0 \exp(-RT)$*, dove* N_0 *è la numerosità della popolazione attuale;*

3. *si simula di* μ *da* f_M*;*

4. *si simula di* π_j *da una distribuzione Beta simmetrica di parametro* p

5. *si accettano i valori* π_j*,* t*,* N*,* μ *con probabilità definita da*

$$u = \frac{\alpha_j^y (1-\alpha_j)^{n-y}}{\left(\frac{y}{n}\right)^y \left(1-\frac{y}{n}\right)^{n-y}} = \frac{\text{Binomial}(y, \alpha_j)}{\text{Binomial}\left(y, \frac{y}{n}\right)}$$

   con $\alpha_j = (1 - 2\exp(-\mu t N))\pi_j + \exp(-\mu t N)$ *altrimenti si ritorna al passo 1.*

Modello a crescita esponenziale a due parametri o medello di coealescenza con crescita.

Nel caso di una popolazione con crescita esponenziale a due parametri, la distribuzione degli intervalli dei tempi di coalescenza è la seguente:

$$P(W_k > s | W_{k+1} + \ldots + W_n = t') =$$

$$\begin{cases} \exp\{\frac{\gamma_k}{R}[\exp(Rt') - \exp(R(t'+s))]\exp(-Rt_g)\} & \text{se } t' < t'+s < t_g \\ \exp\{\gamma_k(t_g - (t'+s) + \frac{1}{R}(\exp[R(t'-t_g)] - 1])\} & \text{se } t' < t_g < t'+s \\ \exp[-\gamma_k s] & \text{se } t_g < t' < t'+s \end{cases} \tag{2.7.5}$$

Applicando le sostituzioni $W_k = t_{n-k+1} - t_{n-k}$ e $t_{n-k} = W_{k+1} + \ldots + W_n$, otteniamo il seguente algoritmo.

Algoritmo 2.7.2.

1. *Si simula il tempo t_g al quale ha inizio la crescita della popolazione;*
2. *Per $k = n, n-1, \ldots, 2$ si simula $t_n - k + 1$ come*

$$t_{n-k+1} = \begin{cases} \frac{1}{R} \ln \left[\exp(Rt_{n-k} - \frac{R}{\gamma_k} \exp(Rt_g) \ln U \right] & \text{se } t_{n-k} < t_{n-k+1} < t_g \\ t_g + \frac{1}{R} \left[\exp(R(t_{n-k} - t_g)) - 1 \right] - \frac{1}{\gamma_k} \ln U & \text{se } t_{n-k} < t_g < t_{n-k+1} \\ \sim \text{Exponential di media } 1/\gamma_k & \text{se } t_g < t_{n-k} < t_{n-k+1} \end{cases} \tag{2.7.6}$$

   e si calcola $T = \sum_{k=2}^{n} W_k = t_{n-1}$;

3. *si simula N definita dalla relazione*

$$\ln(N_c/N) = Rt_g N \tag{2.7.7}$$

   nell'ipotesi che $\ln(N_c/N)$, N ed R siano mutuamente indipendenti ed essendo N_c la numerosità della popolazione attuale;

4. *si simula di μ da f_M;*
5. *si simula di π_j da una distribuzione Beta simmetrica di parametro p*
6. *si accettano i valori π_j, t, N, N_c, μ con probabilità definita da*

$$u = \frac{\alpha_j^y (1-\alpha_j)^{n-y}}{\left(\frac{y}{n}\right)^y \left(1 - \frac{y}{n}\right)^{n-y}} = \frac{\text{Binomial}(y, \alpha_j)}{\text{Binomial}\left(y, \frac{y}{n}\right)}$$

   con $\alpha_j = (1 - 2\exp(-\mu t N))\pi_j + \exp(-\mu t N)$ altrimenti si ritorna al passo 1.

Capitolo 3.

3.1 Descrizione dei dati.

In questo capitolo descriviamo i risultati relativi all'analisi del modello statistico che abbiamo proposto nel capitolo precedente applicato ad un database composto da $n = 49$ individui, ciascuno dei quali sequenziato nella regione compresa tra le basi 16,024–16,383 del *D-loop* mitocondriale, ed appartenenti tutti ad una medesima popolazione $\mathcal{P}$. La popolazione $\mathcal{P}$ in questione è la popolazione toscana—sottopopolazione della popolazione caucasica—e le 49 sequenze provengono da un database curato a livello internazionale per fini forensi dal Federal Bureau of Investigation (FBI).

Le 49 sequenze di DNA mitocondriale considerate presentano 55 siti polimorfici. La variabilità, all'interno della popolazione $\mathcal{P}$, di ciascun sito è caratterizzata dalla presenza di due sole basi. Pertanto ogni sito polimorfico è caratterizzato da una delle due varianti presenti. La cornice teorica nella quale applichiamo il modello statistico proposto al database a disposizione è dunque quella di un *K-allele model* con $K = 2$, secondo la definizione 1.5.5.

Molti dei 55 siti presi in considerazione sono scarsamente polimorfici e presentano la variante più rara in un solo individuo. I siti di questo tipo sono scarsamente informativi e pertanto ne abbiamo considerato uno solo, arrivando così ad analizzare un totale di 28 loci polimorfici.

3.2 Dimensione della popolazione e parametri di crescita.

In questo paragrafo discutiamo la scelta delle prior relative alla dimensione N della popolazione effettiva e dei parametri di crescita.

Nel caso di un modello con crescita della popolazione, in (Wilson *et al.*, 2003) è stata adottata come prior per il tasso di crescita r una distribuzione gamma di media 0.5% con parametro di forma $\alpha = 2$ e rate $\beta = 400$. Una tale scelta è stata basata sui dati di crescita della popolazione umana riportati da (Cavalli-Sforza *et al.*, 1994)—dati che si riferiscono alla dimensione della popolazione intesa come censo—e sull'osservazione che per quanto riguarda la popolazione effettiva è verosimile supporre che i tassi di crescita siano un po' più bassi.

Per il modello che prevede una popolazione che si mantiene costante nel tempo, in molti studi, tra cui per esempio (Wilson e Balding, 1998), (Forster *et al.*, 1996), (Tavaré *et al.*, 1996), (Takahata, 1993), (Fullerton, 1994), (Hammer, 1995), (Harding *et al.*, 1997), la numerosità effettiva è stata spesso stimata nell'ordine di 5000 individui. Questa è l'ipotesi che assumiamo anche noi per il modello con popolazione costante della coalescenza standard.

Al fine di introdurre incertezza sul parametro relativo alla dimensione della popolazione, consideriamo quindi nel caso di una popolazione la cui numerosità si mantiene costante nel tempo una distribuzione gamma di media 5000 e parametro di forma $\alpha = 5$, ed in alternativa una lognormale di parametri, media e standard deviation, 9 ed 1 rispettivamente.

Nel caso del modello che prevede una crescita esponenziale, è verosimile che la numerosità effettiva N della popolazione ancestrale, quindi la dimensione della popolazione nel tempo che precede la crescita, fosse un po' più piccola (Forster *et al.*, 1996). Per questo motivo scegliamo come prior una distribuzione gamma di media 3000, con parametro di forma $\alpha = 3$ e rate $\beta = 10^{-3}$.

Per specificare una prior per la numerosità effettiva N_c della popolazione moderna—la quale ha subito un incremento nella sua dimensione—assegniamo invece una distribuzione gamma con parametro di forma $\alpha = 5$ e rate $\beta = 1$ a $\ln(N_c/N)$ in accordo

con (Wilson *et al.*, 2003).

Dalla prior precedentemente definita per N_c deriva anche la prior per il tempo t_g al quale ha inizio la crescita della popolazione per il modello di crescita esponenziale a due parametri, tenendo conto della relazione

$$\ln \frac{N_c}{N} = r t_g N \tag{3.2.1}$$

e nell'ipotesi in cui $\ln(N_c/N)$, N, ed r siano mutuamente indipendenti.

3.3 Tasso di mutazione e tempi di coalescenza.

In questo paragrafo descriviamo le prior che utilizziamo per modellizzare l'incertezza relativa al tasso di mutazione.

In (Harding *et al.*, 1997) viene considerato un tasso di mutazione per sito per generazione dell'ordine di 10^{-8}. Sulla base di considerazioni analoghe a quelle di (Tavaré *et al.*, 1997), useremo come prior una distribuzione gamma con parametro di forma $\alpha = 2$ e rate $\beta = 4.94 \times 10^{-8}$.

Osserviamo che nonostante le distribuzioni di probabilità con le quali possiamo modellizzare l'incertezza del tasso di mutazione μ e della numerosità effettiva N della popolazione non siano uniche, in molti casi le distribuzioni a posteriori non sono sensibili a ragionevoli specificazioni.

Per le prior relative agli intervalli dei tempi di coalescenza, rimandiamo alla descrizione dei singoli algoritmi effettuata nel capitolo 2.

Per quanto riguarda l'intervallo di tempo con cui specificare la lunghezza delle generazioni, seguendo l'impostazione di (Thomson *et al.*, 2000), abbiamo considerato un valore pari a $G = 25$ anni. In alcuni studi viene modellizzata anche l'incertezza su tale valore G. Osserviamo però che, almeno per ciò che concerne il nostro caso, i dati non sono informativi sul valore di G.

3.4 Risultati.

Nella seguente tabella sono riassunti i risultati relativi al modello della coalescenza standard nell'implementazione degli algoritmi 2.4.2 e 2.4.3.

Tabella 3.4.1. Risultati degli algoritmi 2.4.2 e 2.4.3

Parametri	*Min*	*1stQu*	*Median*	*Mean*	*3rdQu*	*Max*
			Prior			
(a) coalescenza standard-algoritmo 2.4.2						
dimensione effettiva N della popolazione	263	3408	4677	5017	6320	18010
TMRCA (× anni × generazioni)	6	6.05e+4	1.51e+5	2.48e+5	3.27e+5	3.88e+6
(b) coalescenza standard-algoritmo 2.4.3						
dimensione effettiva N della popolazione	443	3354	4644	4980	6246	20660
TMRCA (× anni × generazioni)	1.6e+4	1.23e+5	1.99e+5	2.47e+5	3.15e+5	1.97e+6

Parametri	*Min*	*1stQu*	*Median*	*Mean*	*3rdQu*	*Max*
			Posterior			
(a) coalescenza standard-algoritmo 2.4.2						
dimensione effettiva N della popolazione	411	3405	4733	5095	6370	18570
TMRCA (× anni × generazioni)	18	6.57e+4	1.64e+5	2.74e+5	3.51e+5	4.46e+6
(b) coalescenza standard-algoritmo 2.4.3						
dimensione effettiva N della popolazione	249	3257	4562	4866	6101	15520
TMRCA (× anni × generazioni)	3.95e+5	1.3e+7	1.99e+7	2.23e+7	2.97e+7	1.22e+8

Osserviamo che il *time since the most recent common ancestor* (TMRCA), nell'implementazione dell'algoritmo 2.4.3 del modello della coalescenza standard, è maggiore rispetto a quello che risulta dall'implementazione dell'algoritmo 2.4.2. È naturale aspettarsi un risultato simile in quanto le mutazioni tendono a dilatare il tempo di coalescenza e lo fanno in misura tanto maggiore quanto sono distribuite su intervalli di tempo più brevi, come succede nel caso dell'algoritmo 2.4.3 rispetto all'algoritmo 2.4.2.

Nelle seguenti figure riportiamo le posteriori delle distribuzioni degli alleli. La prior utilizzata per la distribuzione degli alleli nella popolazione ancestrale (vedi appendice C) è una distribuzione Beta simmetrica di parametro p. Nel nostro modello abbiamo considerato i casi $p = 0.1$ e $p = 1$ come suggerito da (NICHOLSON *et al.*, 2002). Il primo caso è un'approssimazione della distribuzione suggerita dal modello neutrale standard (vedi appendice C), mentre nel secondo si ottiene una distribuzione uniforme. Un'analisi di sensitività dimostra che le conclusioni non dipendono dalla particolare scelta di p; i risultati che presentiano fanno riferimento al caso $p = 0.1$.

Per la numerosità N della popolazione, abbiamo utilizzato due distribuzioni a priori: una distribuzione lognormale(9,1) ed una distribuzione gamma(5, 10^{-3}). Le distribuzioni a posteriori relative a queste due prior sono molto simili ed i risultati che presentiamo si riferiscono al caso della distribuzione gamma(5, 10^{-3}).

Come naturale aspettarsi per il modello della coalescenza, i dati non sono molto informativi sulla dimensione N effettiva della popolazione e il tasso di mutazione μ, però lo sono sul loro prodotto $\theta = 2N\mu$.

Osserviamo infine che la dimensione effettiva N della popolazione è una variabile aleatoria discreta, che però abbiamo descritto come variabile continua. Si tratta di un modo di procedere usuale in ambito inferenziale, per ciò che riguarda i modelli di coalescenza, che comporta un errore in pratica trascurabile.

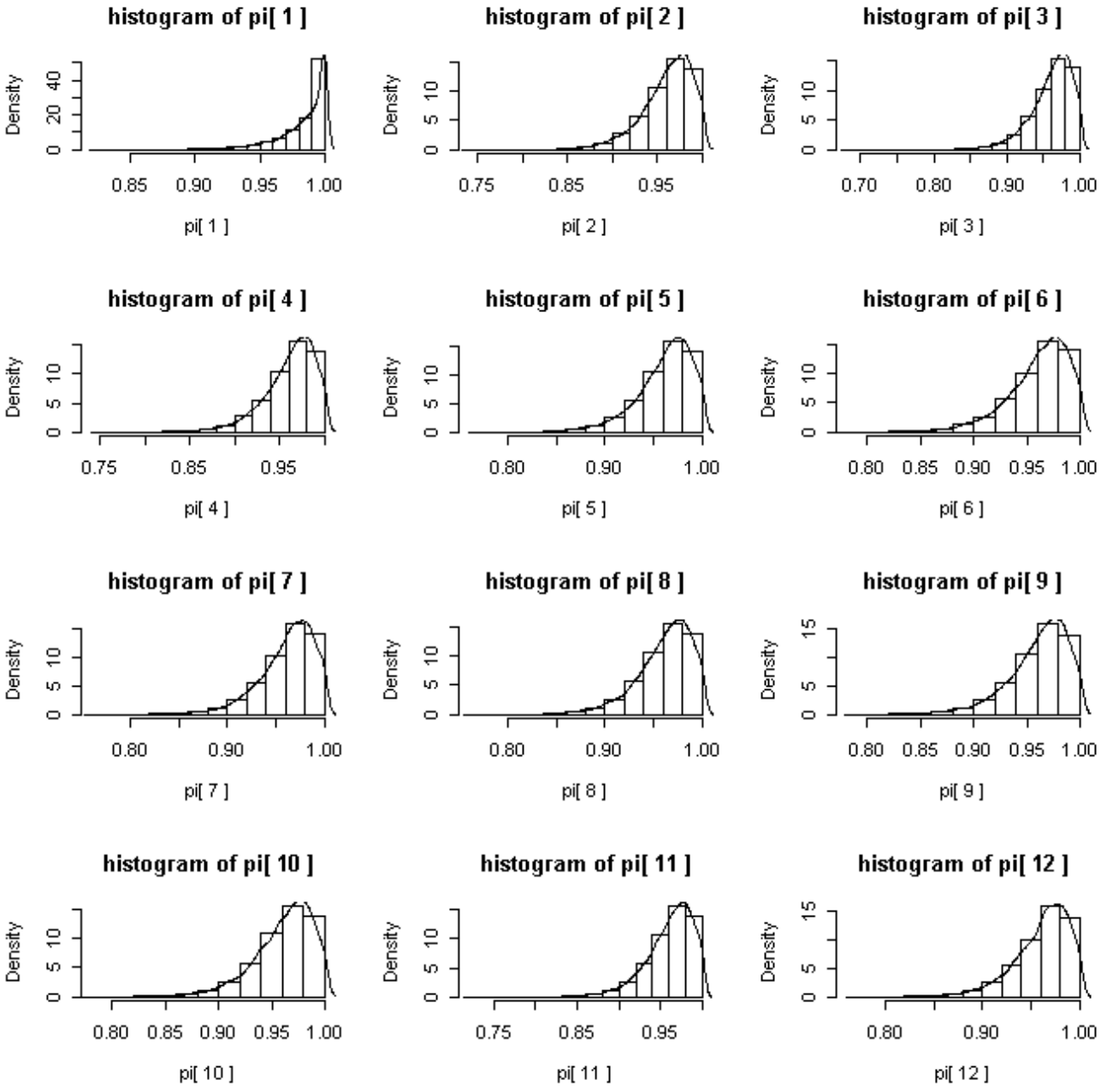

FIGURA 3.4.1 Distribuzione a posteriori delle frequenze degli alleli per l'algoritmo 2.4.2.

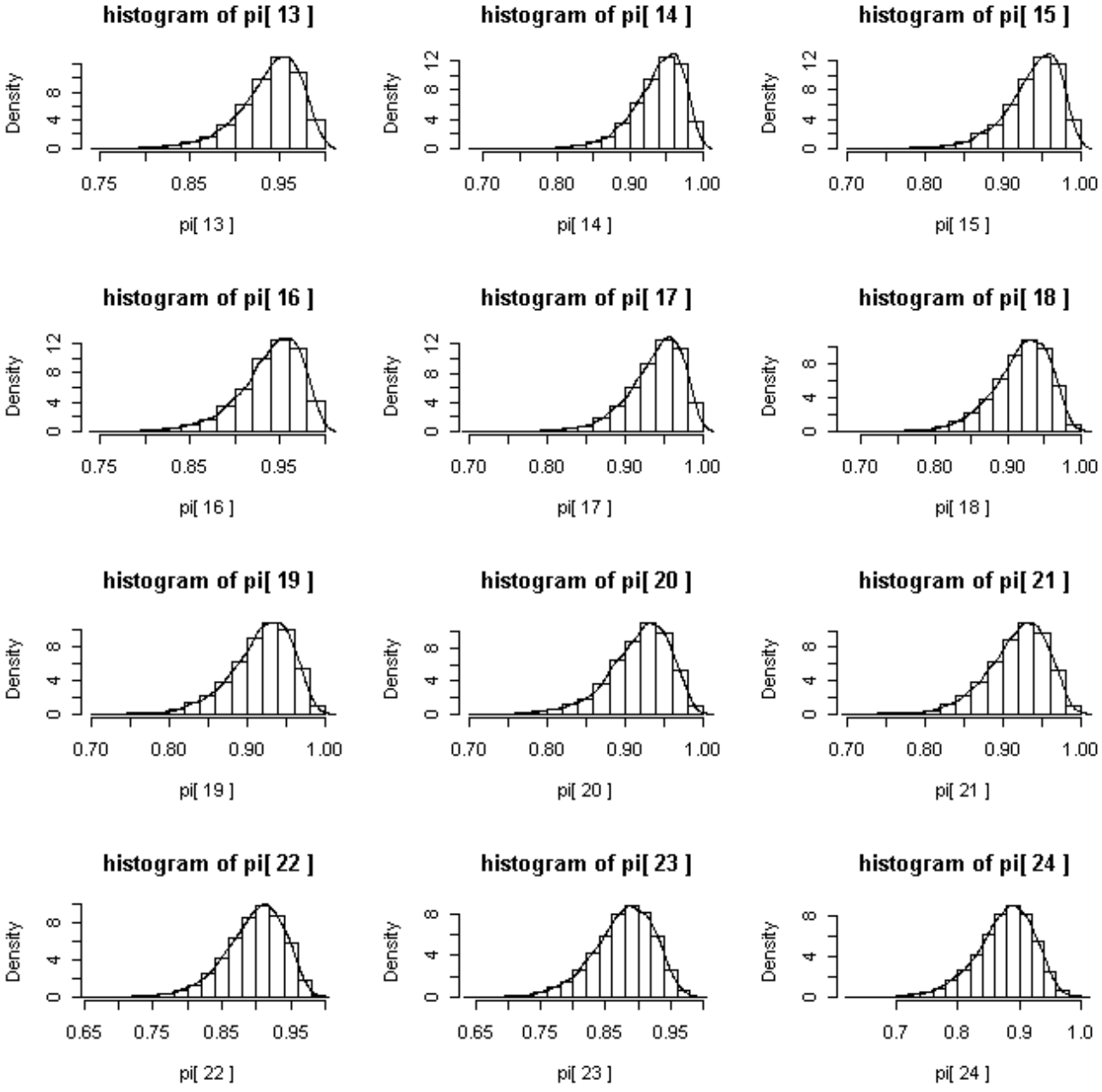

FIGURA 3.4.2 Distribuzione a posteriori delle frequenze degli alleli per l'algoritmo 2.4.2.

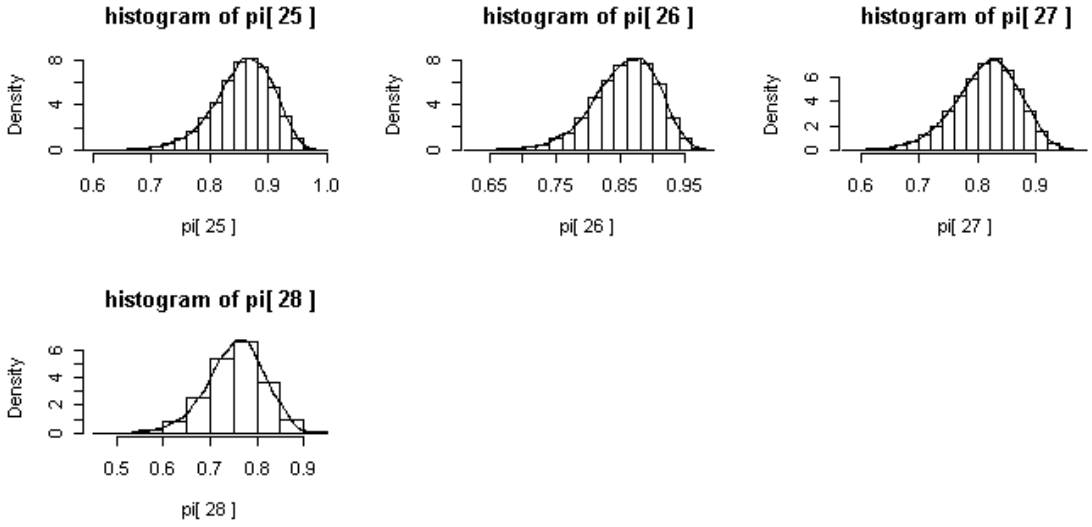

FIGURA 3.4.3 Distribuzione a posteriori delle frequenze degli alleli per l'algoritmo 2.4.2.

Nelle seguenti figure riportiamo inevece le posteriori delle distribuzioni del tasso di mutazione per generazione $\theta = 2N\mu$ per i differenti loci.

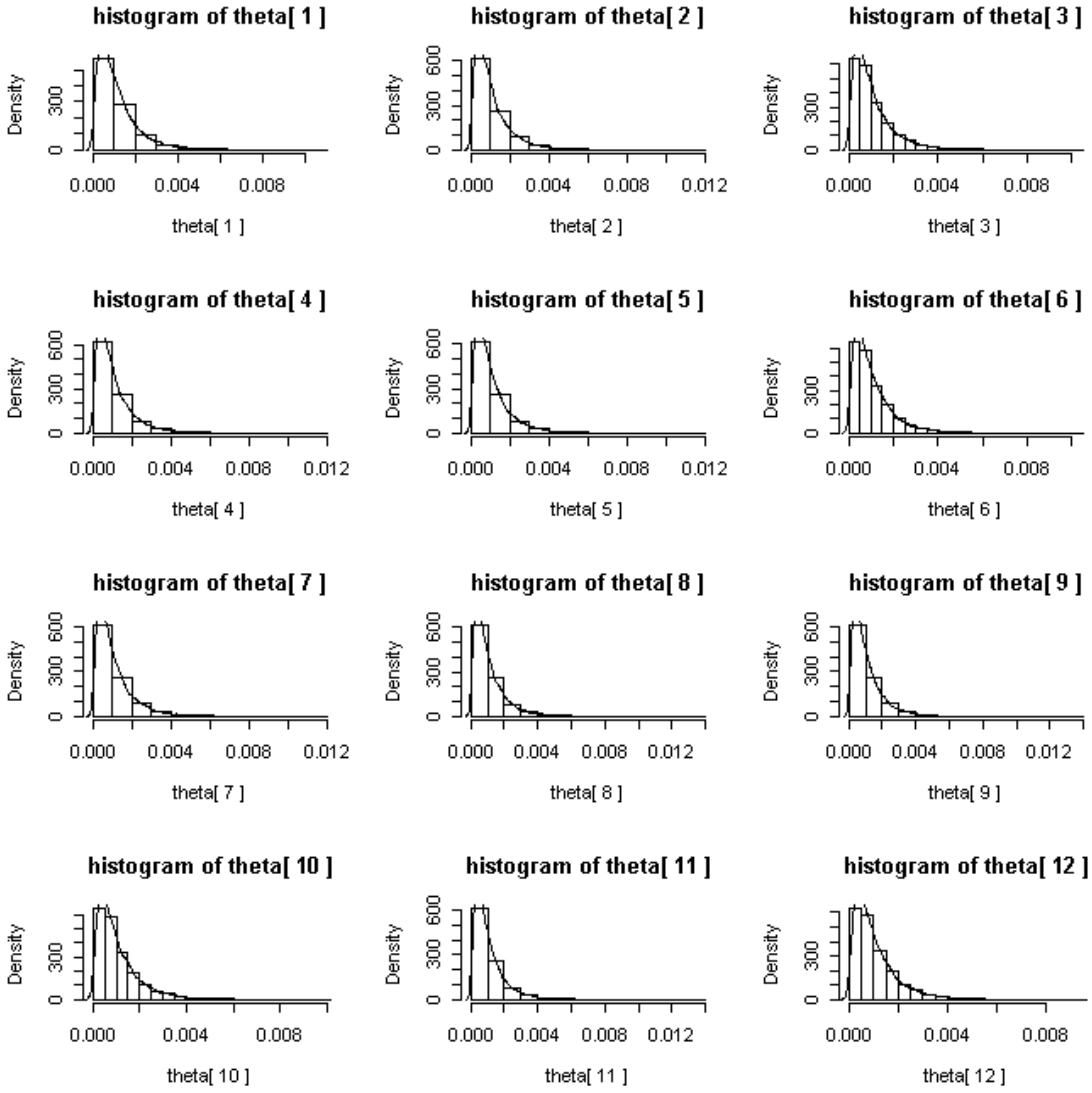

FIGURA 3.4.4 Distribuzione a posteriori del parametro di mutazione per generazione.

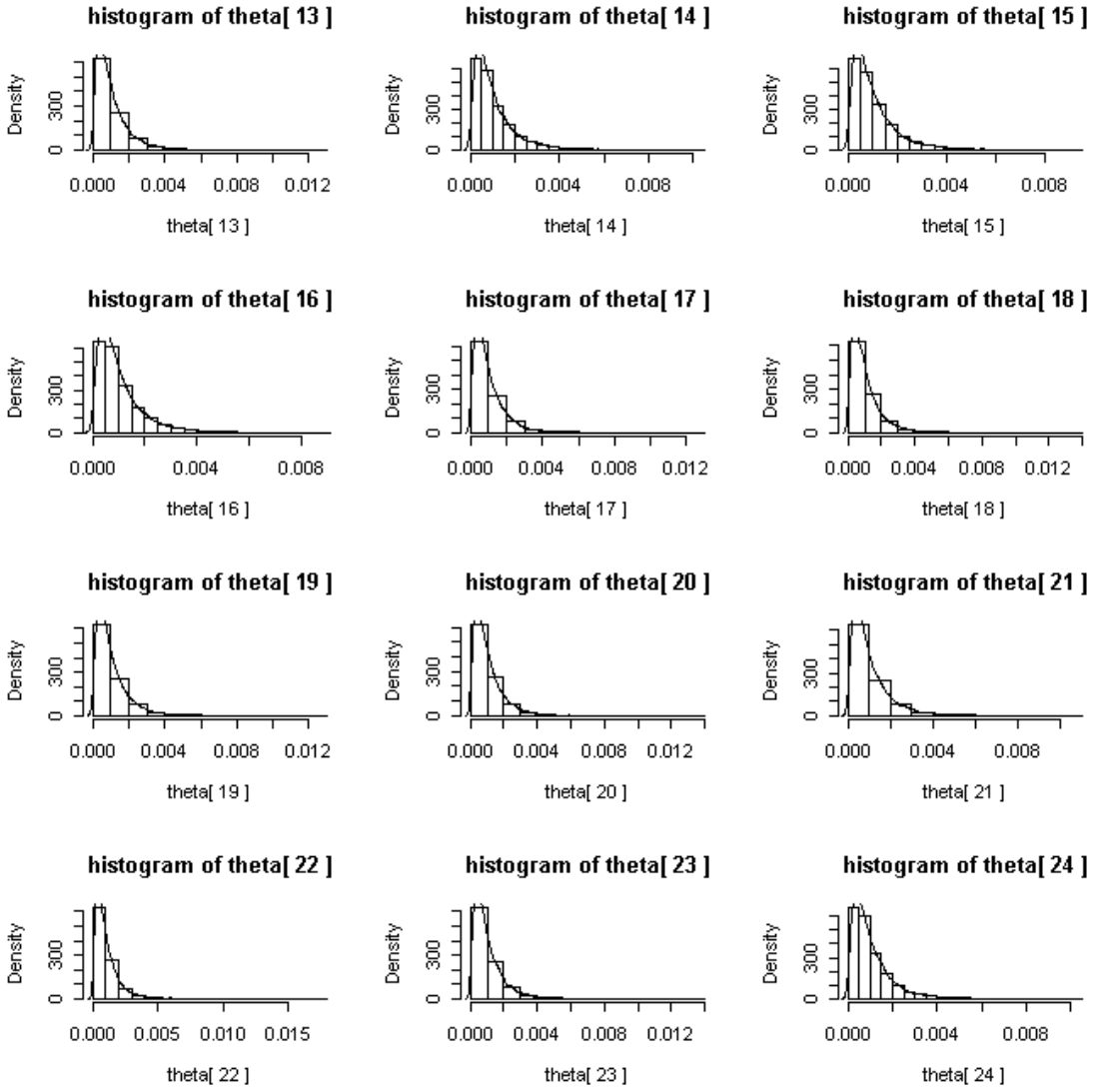

FIGURA 3.4.5 Distribuzione a posteriori del parametro di mutazione per generazione.

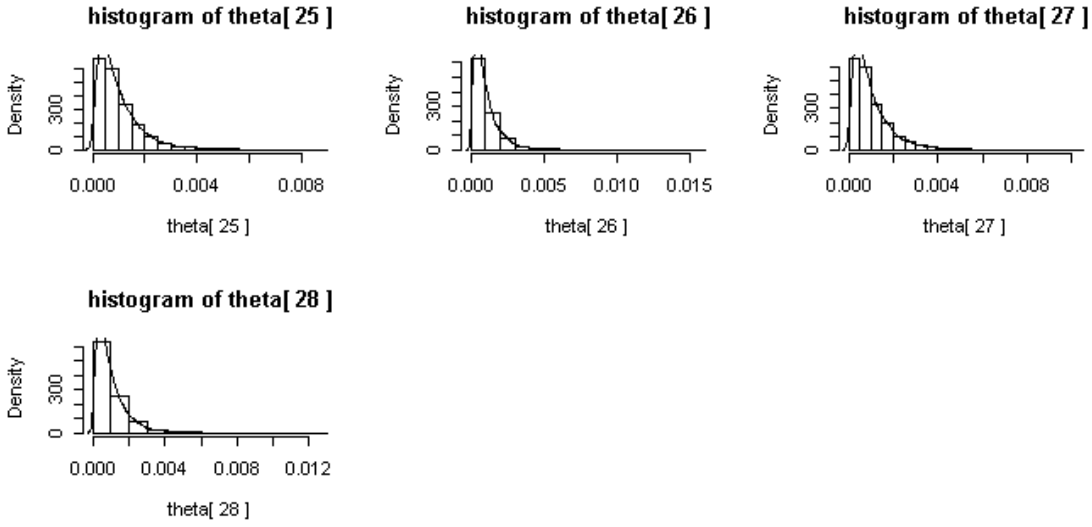

FIGURA 3.4.6 Distribuzione a posteriori del parametro di mutazione per generazione.

Qui di seguito riportiamo i risultati relativi ai modelli implementati con gli algoritmi 2.7.1 e 2.7.2 rispettivamente. Per essi valgono osservazioni analoghe a quelli già effettuate a proposito del modello della coalescenza standard implementato con gli algoritmi 2.4.2 e 2.4.3.

Tabella 3.4.2. Risultati dell'algoritmo 2.7.1

Parametri	*Min*	*1stQu*	*Median*	*Mean*	*3rdQu*	*Max*
			Prior			
(c) crescita esponenziale						
dimensione popolazione ancestrale	155	1676	2702	2983	3889	9995
dimensione effettiva N della popolazione	758	2.52e+4	6.97e+4	153e+3	166.4e+3	2.365e+6
tasso di crescita r (%)	0.0077	0.24	0.41	0.49	0.67	2.16
TMRCA (× anni × generazioni)	6102	202.8e+3	560.9e+3	123.1e+4	133.8e+4	190.2e+5

Parametri	*Min*	*1stQu*	*Median*	*Mean*	*3rdQu*	*Max*
			Posterior			
(a) crescita esponenziale						
dimensione N della popolazione ancestrale	146	1823	2750	3088	3996	12580
dimensione N della popolazione moderna	509	18970	54670	143000	146700	4.143e+6
tasso di crescita r (%)	0.0023	0.24	0.43	0.51	0.70	2.2
TMRCA (× anni × generazioni)	10.72e+3	174.7e+3	380.7e+3	607.8e+3	736.4e+3	1.225e+7

Tabella 3.4.3. Risultati dall'algoritmo 2.7.2

Parametri	*Min*	*1stQu*	*Median*	*Mean*	*3rdQu*	*Max*
			Prior			
(d) coalescenza con crescita						
dimensione popolazione ancestrale	155	1676	2702	2983	3889	9995
dimensione effettiva N della popolazione	1529	1.66e+4	2.67e+4	2.95e+4	3.85	9.89e+4
tasso di crescita r (%)	0.0077	0.24	0.41	0.49	0.67	2.16
tempo al quale inizia la crescita	2660	8519	1.39e+4	2.25e+4	2.35e+4	7.42e+5
TMRCA (× anni × generazioni)	1.06e+5	1.04e+6	1.75e+6	2.34e+6	2.93e+6	2.31e+7

Parametri	*Min*	*1stQu*	*Median*	*Mean*	*3rdQu*	*Max*
			Posterior			
(d) coalescenza con crescita						
dimensione popolazione ancestrale	113	429	576	747	763	8374
dimensione effettiva N della popolazione	431	2484	4212	6348	6764	87260
tasso di crescita r (%)	0.0035	0.223	0.398	0.478	0.623	2.92
tempo al quale inizia la crescita	3.82e-1	6.71e+3	1.19e+4	2.61e+4	2.86e+4	1.37e+6
TMRCA (× anni × generazioni)	2.28e+4	1.40e+5	2.47e+5	5.11e+5	5.33e+5	1.11e+7

3.5 Validità del modello.

Per controllare l'accuratezza delle distribuzioni a posteriori, abbiamo simulato 50 dataset, ciascuno dei quali composto da 70 individui, dal modello della coalescenza standard implementato dall'algoritmo 2.4.2. I parametri (dimensione effettiva della popolazione, tempo di coalescenza, i tassi di mutazione e le frequenze degli alleli) sono stati ottenuti campionando dalle distribuzioni a priori discusse in §3.2.

Per ogni parametro, sia H_D la funzione indicatrice che assume il valore 1 se il $p\%$ intervallo della posteriori, noto il dataset D, comprende il valore corretto della simulazione. La media $\bar{H}$ osservata sui dataset va quindi confrontata con una Binomiale di parametri $n = 50$ e probabilità p.

Nella seguente tavola abbiamo riportato i valori di $\bar{H}$ e tra parentesi i valori esatti della *standard deviation*.

Tabella 3.5.1. Analisi del modello.

$p\%$	*Parametri*			
	N	mu	T	pi
	media di H_D *(SD(%))*			
10	3 (2.12)	4 (2.12)	3 (2.12)	2 (2.12)
30	12 (3.2)	17 (3.2)	16 (3.2)	10 (3.2)
50	27 (3.54)	23 (3.54)	19 (3.54)	19 (3.54)
70	35 (3.2)	29 (3.2)	31 (3.2)	29 (3.2)
90	40 (2.12)	46 (2.12)	38 (2.12)	39 (2.12)

Osserviamo che nella maggior parte dei casi $\bar{H}$ è più piccolo della media p, però la differenza è sempre minore di 3σ. Le differenze maggiori riguardano le frequenze degli alleli e la spiegazione può risiedere nel fatto che nella distribuzione a priori della popolazione ancestrale si può scegliere con uguale probabilità sia la variante comune che quella più rara presenti nella popolazione moderna.

3.6 Esempio di calcolo della conditional match probability.

Consideriamo due individui X_1 e X_2 che possiedono gli stessi aplotipi mitocondriali. Poiché il genoma mitocondriale viene trasmesso in blocco per via materna e non presenta eventi di ricombinazione, vogliamo calcolare la probabilità che i due individui siano fratelli.

Una formulazione più precisa del problema in termini di ipotesi statistiche è la seguente. X_1 e X_2 sono due individui—sequenziati nella regione del Dloop mitocondriale compresa tra le basi 16024–16383. Consideriamo le seguenti ipotesi mutuamente esclusive:

H: X_1 e X_2 sono fratelli;

$\bar{H}$: X_1 e X_2 non sono fratelli.

Le due ipotesi vengono valutate con un approccio bayesiano e in particolare:

$$\frac{P(H|X_1,X_2)}{P(\bar{H}|X_1,X_2)} = \frac{P(X_1,X_2|H)}{P(X_1,X_2|\bar{H})} \times \frac{P(H)}{P(\bar{H})}. \tag{3.6.1}$$

Il *likelihood ratio* ha dunque la seguente espressione

$$\frac{P(X_1,X_2|H)}{P(X_1,X_2|\bar{H})} = \frac{1}{P(X_1|X_2,\bar{H})} \tag{3.6.2}$$

(vedi §1.3 per le condizioni sotto le quali la (3.6.2) è valida.)

Indichiamo con $A_1A_2\ldots A_L$ la sequenza mitocondriale comune ai due individui. La (3.6.2)—la probabilità che due individui non imparentati tra loro abbiano tale medesima

sequenza—è la *conditional match probability* di cui abbiamo ampiamente discusso nel paragrafo 1.5.

Sia $\mathcal{D}$ il database della popolazione a cui appartengono i due individui X_1 e X_2 e sia $\mathcal{D}_1 = \mathcal{D} \cup \{X_1, X_2\}$, ovvero un nuovo database composto da $\mathcal{D}$ aggiornato con le due sequenze X_1 e X_2. Supponiamo poi che anche nel nuovo database aggiornato $\mathcal{D}_1$, tutti i locus polimorfici presentino ancora due possibili alleli e che la cornice teorica sia ancora quella di *K-allele model* con $K = 2$.

Se i locus polimorfici appartengono a regioni della molecola mitocondriale distanti tra di loro, è verosimile l'ipotesi di indipendenza tra i vari siti. Questo fatto ci permette di modellizzare ciascun locus come un *one locus finite-sites model* con due alleli e la *conditional match probability* ha la seguente espressione:

$$P(X_1 = A_1A_2 \dots A_L | X_2 = A_1A_2 \dots A_L, \bar{H}, \mathcal{D}_1) =$$

$$\prod_{j=1}^{L} \int_{\theta_j, \pi_j t} P(X_{1j} = A_j | X_{2j} = A_j, \bar{H}, \theta_j, \pi_j, t) P(\theta_j, \pi_j, t | \bar{H}, \mathcal{D}_1) \; d\theta_j d\pi_j dt =$$

$$\prod_{j=1}^{L} \int_{\theta_j, \pi_j t} \left(1 - \frac{\theta_j t}{2}\right) \left(\pi_j + (1 - \pi_j)\frac{\theta_j t}{2}\right) P(\theta_j, \pi_j, t | \bar{H}, \mathcal{D}_1) \; d\theta_j d\pi_j dt \qquad (3.6.3)$$

dove t è il tempo che separa due individui non imparentati con il loro *most recent common ancestor*, $\theta_j = 2N\mu_j$ dove N è la dimensione effettiva della popolazione di appartenenza dei due individui X_1 e X_2, μ_j il tasso di mutazione del locus j-esimo ed infine X_{1j} è il j-esimo locus della sequenza X_1.

Poiché il tempo che separa due individui non imparentati non dipende da θ_j e da π_j e dai dati ma, nel modello della coalescenza, ha la distribuzione di un'esponenziale di media 2, la (3.6.3) possiamo riscriverla nella seguente maniera:

$$\prod_{j=1}^{L} \int_{\theta_j, \pi_j t} \left(1 - \frac{\theta_j t}{2}\right) \left(\pi_j + (1 - \pi_j)\frac{\theta t}{2}\right) P(\theta_j, \pi_j | \bar{H}, \mathcal{D}_1) P(t | \bar{H}) \; d\theta_j d\pi_j dt \qquad (3.6.4)$$

Possiamo calcolare la (3.6.2) come media ergodica campionando i parametri (θ_j, π_j) con gli algoritmi descritti nel capitolo 2.

Nel caso più generale in cui nel database $\mathcal{D}_1$ i locus polimorfici presentino più di due possibili alleli, alla (3.6.3) andrà sostituita la formula (1.5.10).

Nel problema pratico analizzato, nel database $\mathcal{D}_1$, aggiornato con la sequenza mitocondriale X comune ai due individui X_1 ed X_2, la variabilità per ogni sito è caratterizzata dalla presenza di due sole basi. Pertanto per ogni sito siamo nel contesto di un *one locus finite-sites model* con due alleli. La sequenza di mtDNA, comune ad i due individui X_1 e X_2, in corrispodenza dei 28 siti polimorfici studiati presenta, inoltre, in soli due loci la variante più rara.

È ragionevole ipotizzare che il processo mutazionale operi in maniera indipendente sui 28 loci, trovandosi, questi, in regioni sufficientemente distanti l'una dall'altra.

Nella seguente tabella sono esposti i risultati relativi alla computazione della *conditional match probability* per i quattro modelli descritti nel capitolo 2.

Tavola 3.6.1. Conditional match probability.

Modelli	*Conditional match probability* $P(X_1 = X \mid X_2 = X, \bar{H}, \mathcal{D}_1)$
Modello coalescenza standard (a)	7.05e-4
Modello coalescenza standard (b)	1.79e-4
Modello con crescita popolazione (c)	2.07e-4
Modello coalescenza con crescita (d)	6.90e-4

Osserviamo che la *conditional match probability* così calcolata presenta una proprietà di robustezza rispetto ai differenti modelli mutazionali e demografici considerati.

Conclusioni

In questo lavoro abbiamo affrontato il problema relativo alla valutazione della capacità identificativa del DNA mitocondriale, evidenziando gli aspetti principali legati all'utilizzo di questo tipo di genoma ai fini dell'identificazione personale.

I termini del quesito che ci siamo posti sono relativamente semplici e possono essere così sintetizzati. Attesa la completa sovrapponibilità di due aplotipi mitocondriali, in che misura è possibile affermare che tale dato deriva dal fatto che i soggetti, di cui sono stati sequenziati i mitocondri, verificano la circostanza che si tratti di fratelli? In termini diversi, il problema consiste nel comprendere quale sia la probabilità che due individui, presi a caso nella popolazione e pertanto non geneticamente correlati, condividano la medesima sequenza mitocondriale.

La risposta a tale quesito non è banale. Allo stato dell'arte, l'approccio consiste nell'osservare, tra tutte le sequenze disponibili in database di riferimento, la frequenza della sequenza oggetto dell'analisi. Il dato di compatibilità basato su sequenze mtDNA, però, non può avere una stima statisticamente rigorosa in quanto mancano database di riferimento sufficientemente estesi. Non è quindi disponibile la distribuzione degli aplotipi mitocondriali caratterizzanti le popolazioni. A maggior ragione tale considerazione vale se si tratta di una popolazione antica e non chiaramente definibile.

Alla luce di questi problemi, abbiamo formalizzato un metodo che consente un'indicazione numerica della potenzialità identificativa del DNA mitocondriale in alternativa all'approccio tipicamente descrittivo a cui attualmente si fa ricorso. La metodologia presentata permette contemporaneamente di fare inferenza su molti parametri genealogici e demografici e bene si presta ad essere modificata implementando scenari evolutivi eventualmente più realistici di quelli considerati.

Rimangono alcuni problemi aperti legati ai fenomeni dell'eteroplasmia ed alla inserzione e delezione di nucleotidi. In particolare, per quanto riguarda la possibilità di prendere in considerazione la presenza di delezioni o inserzioni, disponendo di dati opportuni che permettano un'analisi appropriata, non dovrebbero esserci difficoltà concettuali nel generalizzare in tale direzione il modello proposto.

APPENDICE A

Modelli della genetica di popolazione.

A.1 Modello di Wright-Fisher.

Il modello di Wright-Fisher è alla base di molti modelli che descrivono l'evoluzione delle frequenze dei geni in presenza di deriva genetica, mutazioni e selezione. Le assunzioni per una popolazione diploide sono le seguenti: (i) l'accoppiamento è casuale (*random mating population*); (ii) la dimensione N della popolazione è costante nel tempo; (iii) le generazioni non si sovrappongono; (iv) il locus non è legato al sesso (locus autosoma); (v) la popolazione è nell'equilibrio di Hardy-Weinberg.

Consideriamo quindi un singolo locus autosoma con due alleli A_1 ed A_2, con genotipi A_1A_1, A_1A_2 e A_2A_2 con rispettive *fitness* w_{11}, w_{12} e w_{22}. Il parametro wij di fitness, presente in caso di selezione, riflette la probabilità di sopravvivenza dello zigote di genotipo A_iA_j, $i, j = 1, 2$.

L'ipotesi di una popolazione in equilibrio di Hardy-Weinberg stabilisce che se gli alleli A_1 ed A_2 hanno rispettivamente frequenza p e $q = 1 - p$, allora le frequenze genotipiche saranno p^2, $2pq$ e q^2 per rispettivamente A_1A_1, A_1A_2 e A_2A_2.

Supponiamo che alla generazione n-esima ci siano j geni di tipo A_1 e $2N - j$ geni di tipo A_2. La frequenza dell'allele A_1 alla generazione n è quindi $p(n) = j/2N$ e dopo selezione diviene:

$$\phi(n) = \frac{p(n)[p(n)w_{11} + q(n)w_{12}]}{\bar{w}(n)} \tag{A.1.1}$$

dove $\bar{w}(n) = p(n)^2w_{11} + 2p(n)q(n)w_{12} + p(n)^2w_{22}$ rappresenta la *fitness* media.

Se alla selezione segue anche un processo di mutazione, assumendo mutazioni simmetriche con probabilità u (A_1 muta in A_2 ed A_2 muta in A_1 con probabilità u), la frequenza dell'allele A_1 diviene la seguente:

$$\psi(n) = \phi(n)(1-u) + (1-\phi(n))u. \tag{A.1.2}$$

Per formare la generazione successiva $n+1$, costituta da N individui, si campionano $2N$ gameti indipendenti alla generazione n-esima in accordo ad uno schema binomiale. Ovvero, se ψ è la frequenza dell'allele A_1 nella generazione n-esima dopo la selezione e la mutazione, allora la probabilità che che ci siano k geni di tipo A_1 nella generazione successiva è

$$\binom{2N}{k}\psi^k(1-\psi)^{2N-k} \tag{A.1.3}$$

e la frequenza dell'allele A_1 alla $n+1$-esima generazione è appunto $k/2N$.

Il modello di Wright-Fisher è un esempio di catena di Markov le cui probabilità di transizione sono espresse dalla (A.1.3). Essendo queste indipendenti dal tempo, la catena di Markov è omogenea. Quando la dimensione N della popolazione è grande, si può approssimare la catena di Markov con un processo di diffusione che risulta in genere più semplice da studiare rispetto a quello originale.

A.2 Modello della coalescenza standard.

Le ipotesi di lavoro sono quelle del modello di Wright-Fisher neutrale, in assenza di selezione. Supponiamo di avere un campione di n individui da una popolazione la cui dimensione N si mantiene costante nel tempo. Alcuni degli n individui possono condividere un genitore nella generazione precedente. Consideriamo il numero di distinti progenitori τ generazioni fa, $\tau = 0, 1, 2, \ldots$. Man mano che si risale indietro nel tempo, il numero di distinti progenitori del campione di individui decresce fino a ridursi ad uno solo, il *most recent common ancestor* (MRCA).

Quando risalendo indietro nel tempo si trova il progenitore comune di due individui, diciamo che questi coalescono. Si assume che gli eventi di coalescenza avvengano solo tra coppie di individui. Questa ipotesi è motivata dal fatto che la probabilità che k individui non abbiamo in comune, nella generazione precedente, il genitore ha la seguente espressione:

$$\prod_{i=1}^{k-1} \frac{N-i}{N} = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right) \tag{A.2.1}$$

dove il termine $O\left(\frac{1}{N^2}\right)$ include gli eventi di coalescenza tra tre o più individui e quegli eventi che si presentano multipli nella medesima generazione.

Dalla (A.2.1) per l'indipendenza degli eventi di coalescenza nelle differenti generazioni, la probabilità che k individui non condividano un progenitore comune nelle precedenti τ generazioni è la seguente:

$$\left[\prod_{i=1}^{k-1} 1 - \frac{i}{N}\right]^{\tau}. \tag{A.2.2}$$

Se esprimiamo il tempo di coalescenza in unità di N generazioni, ovvero se operiamo la sostituzione $t = \tau/N$, allora

$$\left[\prod_{i=1}^{k-1} 1 - \frac{i}{N}\right]^{Nt} = \left[\left(1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right)\right)^{N}\right] \tag{A.2.3}$$

ed al limite quando la dimensione N della popolazione tende ad infinito—diffusion approximation—otteniamo:

$$\left[\left(1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right)\right)^{N}\right] \longrightarrow \exp\left(-\binom{k}{2}\right) \tag{A.2.4}$$

Una più ampia classe di modelli può essere approssimata se il tempo di coalescenza si misura in unità di $N\sigma^{-2}$, dove σ^2 è la varianza del numero di discendenti degli individui. Per $\sigma^2 = 1$ si ottiene appunto il modello di Wright-Fisher.

Indichiamo con T_k il tempo durante il quale sono presenti k distinti individui. Dalla (A.2.4) segue che la distribuzione di T_k è quella di una variabile aleatoria esponenziale di media $k(k-1)/2$.

Poiché coalescono solamente coppie di individui e non anche differenti coppie nella medesima generazione, in un campione di n individui vi sono $n-1$ eventi e dunque tempi di coalescenza $\{T_n, T_{n-1}, \ldots, T_2\}$.

Il processo di coalescenza è completato da una topologia, ovvero da una mappa che definisce quali coppie di individui coalescono. Generalmente si rappresenta la topologia come una famiglia di relazioni di equivalenza

$$\mathcal{R} = \{\mathcal{R}_N(\tau) : \tau = 0, 1, 2, \ldots\}\,. \tag{A.2.5}$$

Per ogni coppia di individui i e j, diciamo che $i\mathcal{R}_N(\tau)j$ se e solo se i e j possiedono un progenitore comune alla τ-esima generazione. Nell'ipotesi di neutralità tutti gli individui hanno la medesima probabilità di coalescere tra di loro.

Sotto condizioni abbastanza deboli, il processo limite per grandi popolazioni (i.e. $N \to \infty$) in cui il tempo è misurato in unità di generazioni (i.e. $t = \tau/N$) è un processo di Markov a tempo continuo con generatore infinitesimale

$$Q_{\xi,\eta} = \begin{cases} -\binom{|\xi|}{2} & \text{if } \eta = \xi \\ 1 & \text{if } \xi \prec \eta \\ 0 & \text{otherwise} \end{cases} \tag{A.2.6}$$

dove ξ ed η sono relazioni di equivalenza sull'insieme $\{1, 2, \ldots, n\}$, $|\xi|$ è il numero di classi di equivalenza di ξ e $\xi \prec \eta$ se e solo se η può essere ottenuto da ξ immegendo due classi di equivalenza di ξ in una singola classe di η.

Al processo riproduttivo così descritto viene infine super-imposto un processo mutazionale in accordo ad un processo di Poisson di parametro $\theta/2$, con $\theta = 2Nu$, essendo u il tasso di mutazione.

A.3 Infinitely Many Sites Model, K-allele Model e Finite Sites Model.

L'*infinitely-many sites model* è stato introdotto da (KIMURA, 1969) e (WATTERSON, 1975) come modello per descrivere le sequenze di DNA. Ogni individuo è descritto da una stringa al più numerabile di siti non soggetti ad eventi di ricombinazione. Si assume che la probabilità di mutazione per sito sia molto piccola in maniera tale che il numero totale di mutazioni per individuo per generazione sia finito. Ogni mutazione che si verifica introduce un nuovo sito segregante, e non vengono prese in considerazione le retro-mutazioni, ovvero mutazioni che colpiscono più volte uno stesso sito. Da questo deriva che ogni sito può essere classificato come sito ancestrale oppure sito mutante. Classificando, per esempio, i siti del primo tipo come 0 e quelli mutanti come 1, la sequenza di DNA è descritta da una stringa di 0 ed 1.

L'*infinitely-many sites model* è in genere utilizzato per le sequenze di DNA mitocondriale che è aploide, si eredita per via materna ed è consistente con l'ipotesi di assenza di eventi di ricombinazione.

Un *K-allele model* è un modello in cui ciascun locus è caratterizzato da k possibili alleli. Quando si verifica una mutazione per un allele di tipo i, vi è una probabilità m_{ij} che l'allele mutato sia di tipo j. Un *K-allele model* è descritto matematicamente da una catena di Markov la cui matrice di transizione dà le probabilità che un locus di tipo i sia stato rimpiazzato da un locus di tipo j.

Il *finite sites model* è un *K-allele model* in cui $K = 2$ o $K = 4$. Nel *finite sites model* quando si verifica una mutazione, si sceglie a caso il locus a cui assegnare la mutazione stessa e la base che si trova in quel locus muta in accordo ad una matrice di transizione M.

La differenza tra l'*infinitely-many sites model* ed il *finite sites model* risiede nell'allo-

cazione delle mutazioni nella sequenza. Infatti il numero totale di mutazioni segue la stessa distribuzione nei due modelli. Però mentre nell'*infinitely-many sites model* ogni mutazione genera un nuovo sito segregante, nel *finite sites model* un sito segregante è un sito colpito da almeno una mutazione.

APPENDICE B

Approssimazione alla diffusione nella genetica di popolazione.

B.1 Approssimazione alla diffusione.

Un processo di diffusione è un processo di Markov $\{X_t,\ t \in [0,\infty)\}$—noto lo stato presente, la probabilità degli stati futuri non è alterata dalla conoscenza degli stati passati—a tempo continuo ed a stati continui.

Nell'ambito dei modelli di genetica di popolazione, $X(t)$—con cui indichiamo lo stato del processo al tempo t—può indicare per esempio la frequenza relativa di un particolare allele al tempo t.

Un processo di diffusione è caratterizzato dalla media (*drift* o parametro di deriva) e dalla varianza (diffusione) dell'incremento infinitesimale $\Delta_h X(t) = X(t+h) - X(t)$ durante l'intervallo di tempo $(t, t+h)$. Il parametro di *drift* è così definito:

$$a(x,t) = \lim_{h\to 0} \frac{1}{h} E[\Delta_h X(t) | X(t) = x]. \tag{B.1.1}$$

Il parametro di diffusione è definito nella seguente maniera:

$$b(x,t) = \lim_{h\to 0} \frac{1}{h} E[(\Delta_h X(t))^2 | X(t) = x]. \tag{B.1.2}$$

Per h sufficientemente piccolo, la quantità $a(x,t)h$ è approssimativamente la media dell'incremento infinitesimale $\Delta_h X(t)$ durante l'intervallo di tempo $(t, t+h)$. Infatti dalla (B.1.1) segue:

$$E[\Delta_h X(t) | X(t) = x] = a(x,t)h + o(h). \tag{B.1.3}$$

La quantità $b(x,t)h$ è approssimativamente la varianza dell'incremento infinitesimale $\Delta_h X(t)$ durante l'intervallo di tempo $(t, t+h)$ per h sufficientemente piccolo, poiché

$$\text{var}[\Delta_h X(t)|X(t)=x] = E[(\Delta_h X(t))^2|X(t)=x] - (E[\Delta_h X(t)|X(t)=x])^2 =$$

$$b(x,t)h - (a(x,t)h)^2 + o(h) = b(x,t)h + o(h). \tag{B.1.4}$$

B.2 Approssimazione alla diffusione nel modello neutrale di Wright-Fisher.

Consideriamo il modello neutrale di Wright-Fisher per una popolazione ad accoppiamento casuale e di numerosità N (ovvero $2N$ gameti). Consideriamo poi un one-locus model con due alleli A_1 ed A_2, con probabilità di mutazione u ed indichiamo con $Y(n)$ il numero di gameti A_1 nell'n-esima generazione.

Il numero di gameti alla $n+1$-esima generazione condizionato a quello dell'n-esima generazione ha una distribuzione binomiale:

$$P(Y(n+1)=j|Y(n)=i) = \binom{2N}{j}\psi^j(1-\psi)^{2N-j}, \tag{B.2.1}$$

dove

$$\psi = \frac{i}{2N}(1-u) + \frac{2N-i}{2N}u. \tag{B.2.2}$$

Definiamo il processo riscalato in termini di unità di $2N$ generazioni—corrispondente alla sostituzione $t = n/2N$ nel modello $Y(n)$:

$$X(t) = \frac{Y(\lfloor 2Nt \rfloor)}{2N}, \qquad t \geq 0 \tag{B.2.3}$$

dove $\lfloor 2Nt \rfloor$ è il più grande intero minore o uguale a $2Nt$. Il processo $X(t)$ è ancora un processo di Markov. Osserviamo anche che il modello così definito è omogeneo essendo

le probabilità di transizione stazionarie, e pertanto i parametri di *drift* e diffusione non dipenderanno dal tempo.

Al fine di calcolare il parametro di *drift*, osserviamo che

$$2NE\left[X\left(t+\frac{1}{2N}\right)-X(t)|X(t)=\frac{i}{2N}\right]=E[Y(\lfloor 2Nt\rfloor+1)-i|Y(\lfloor 2Nt\rfloor)=i]$$

$$2N\psi-i=i(1-u)+(2N-i)u-i=2Nu\left(1-\frac{i}{n}\right). \tag{B.2.4}$$

Poniamo $h=1/2N$ e consideriamo che la numerosità della popolazione tenda ad infinito ($N\to\infty$ o $h\to 0$). Assumiamo inoltre che esista il limite $\lim_{N\to\infty}4Nu=\theta$. Allora per $x=i/2N$, il parametro di *drift* è il seguente:

$$a(x)=\lim_{h\to 0}\frac{1}{h}E[X(t+h)-X(t)|X(t)=x]=\frac{\theta}{2}(1-2x). \tag{B.2.5}$$

Per il calcolo del parametro di diffusione, osserviamo che

$$2NE\left[\left(X\left(t+\frac{1}{2N}\right)-X(t)\right)^2|X(t)=\frac{i}{2N}\right]=E[(Y(\lfloor 2Nt\rfloor+1)-i)^2|Y(\lfloor 2Nt\rfloor)=i]$$

$$=\frac{1}{2N}2N\psi(1-\psi) \tag{B.2.6}$$

essendo $2N\psi(1-\psi)$ la varianza di una distribuzione binomiale di parametri ψ e $2N$.

Ponendo $x=i/2N$ e dall'ipotesi $\lim_{N\to\infty}4Nu=\theta$, segue che $\psi\to x$ per $N\to\infty$, da cui il seguente parametro di diffusione

$$b(x)=x(1-x). \tag{B.2.7}$$

I risultati ottenuti nelle (B.2.5) e (B.2.7) sono condizionati all'esistenza del limite $\lim_{N\to\infty}4Nu$ che abbiamo indicato con θ. Naturalmente la probabilità di mutazione u non dipende dalla numerosità N della popolazione ed è invece una costante. Pertanto in un modello realistico ci si aspetta che tale limite non esista. L'ipotesi matematica che

abbiamo fatto di esistenza del limite si presta però alle seguenti considerazioni. Innanzitutto il processo di diffusione al limite—che è più semplice da trattare rispetto al processo originario—è un'approssimazione del modello reale quando la numerosità della popolazione è fissa ed è finita. In secondo luogo la probabilità di mutazione u è generalmente molto piccola, tipicamente dell'ordine di 10^{-5} o 10^{-6}. Pertanto possiamo interpretare l'esistenza del limite $\lim_{N\to\infty} 4Nu$ come un indicatore della bontà dell'approssimazione—ci si aspetta che l'approssimazione sia buona quando la numerosità della popolazione è dell'ordine del reciproco della probabilità di mutazione.

B.3 Approssimazione alla diffusione nel modello di Wright-Fisher con mutazioni e selezione.

Come per il modello neutrale, il modello di Wright-Fisher con mutazioni e selezioni è un processo di Markov omogeneo e pertanto i parametri di *drift* e di diffusione, nella approssimazione alla diffusione, non dipendono dal tempo.

Consideriamo un one locus model con due alleli A_1 ed A_2 per una popolazione aploide di dimensione N. Le mutazioni, simmetriche, si verificano con probabilità u. Supponiamo inoltre che l'allele A_2 abbia un vantaggio selettivo, *fitness*, nella trasmissione rispetto all'allele A_1 ed indichiamo il suo parametro di selezione con w.

Sia $Y(n)$ il numero di individui con l'allele A_1 all'n-esima generazione. Le probabilità di transizione sono le seguenti:

$$P(Y(n+1) = j | Y(n) = i) = \binom{N}{j} \psi^j (1-\psi)^{N-j} \tag{B.3.1}$$

dove

$$\psi = \frac{p(1-u) + (1-p)(1+s)u}{p + (1-p)(1+s)}, \qquad p = \frac{i}{N}. \tag{B.3.2}$$

Ponendo $\theta = 2Nu$ e $\sigma = 2Ns$ ed operando la sostituzione $t = n/N$—ovvero misurando il tempo in unità di N generazioni—al limite per N tendente ad infinito si ottengono, in maniera analoga a quanto visto per il modello neutrale, per il processo $X(t) = Y(\lfloor tN \rfloor)$ i seguenti parametri di *drift* e diffusione rispettivamente:

$$a(x) = -\frac{\sigma}{2}x(1-x) + \frac{\theta}{2}(1-2x) \tag{B.3.3a}$$

$$b(x) = x(1-x) \tag{B.3.3b}$$

Come il parametro di mutazione u, anche il parametro di selezione s è una costante e dunque il limite $\lim_{N\to\infty} 2Ns = \sigma$ esiste per ipotesi matematica. In pratica questo ci dice che il modello al limite è una buona approssimazione di quello reale quando il parametro s è dell'ordine inverso della dimensione della popolazione, dunque l'approssimazione alla diffusione per grandi popolazioni è accurata solo in presenza di una debole selezione.

APPENDICE C

Modello gerarchico per SNPs loci.

C.1 Introduzione al modello gerarchico per SNPs loci.

In questa sezione descriviamo il modello gerarchico per dati di tipo SNP presentato da (Nicholson *et al.*, 2002).

Consideriamo una collezione di P popolazioni per ciascuna delle quali si dispone dei dati relativi ad L single-nucleotide polymorphisms (SNPs) loci. Un locus SNP è una singola posizione nel DNA che presenta variazioni tra differenti individui appartenenti ad una determinata popolazione. Mentre un nucleotide è identificato con una delle lettere A, C, G, T—associate rispettivamente alle basi contenenti Adenina, Citosina, Guanina e Timina—un SNP presenta due sole varianti che indicheremo con 0 ed 1.

Indichiamo con n_{ij} il numero di cromosomi tipizzati all'$i-esimo$ SNP della $j-esima$ popolazione. Per ogni SNP scegliamo in maniera arbitraria e fissiamo una delle due varianti, sia poi x_{ij} il numero di copie della variante prescelta al locus $i-esimo$ nel campione della $j-esima$ popolazione. Infine denotiamo con α_{ij}, $0 \leq \alpha \leq 1$ la frequenza (non osservata) nella popolazione $j-esima$ della variante prescelta al locus $i-esimo$.

Assumiamo per ipotesi che nella maggior parte o in tutti gli SNP analizzati, la variante osservata fosse presente anche nella popolazione ancestrale a quella studiata. L'ipotesi è consistente per gli SNPs presenti in un range di popolazioni contemporanee qualora gli stessi SNPs siano il risultato di un singolo evento mutazionale piuttosto che di mutazioni ricorrenti.

Le differenze che si riscontrano nelle frequenze degli alleli tra popolazioni differenti sono un risultato dovuto in larga parte o esclusivamente ad eventi demografici. Hanno un ruolo fondamentale infatti i processi inerenti alla segregazione mendeliana, il fatto poi che gli individui abbiano un numero differente di discendenti e lo scambia di materiale genetico attraverso migrazioni. Ignorando un fenomeno di selezione per cui certi SNPs possono avere maggiore probabilità di essere trasmessi ai discendenti, il meccanismo probabilistico lo stesso per tutti gli SNPs.

C.2 Modello gerarchico per SNPs loci.

Introduciamo in questa sezione il modello gerarchico per SNPs. Si tratta di un modello non puramente statistico ma che ha anche un'interpretazione nell'ambito della genetica di popolazione.

I dati sono modellizzati come binomiali: dati n ed α

$$x_{ij} \sim \text{Binomial}(n_{ij}, \alpha_{ij}). \tag{C.2.1}$$

Modelliamo la struttura di dipendenza tra le frequenze α degli alleli della popolazione nella seguente maniera. Introduciamo un'altra collezione di quantità non osservate, una per ogni locus: π_i, $i = 1, 2, \ldots, L$. Nell'ambito della genetica di popolazione, queste quantità corrispondono alle frequenze degli alleli in una popolazione ancestrale a quella campionata. Introduciamo, inoltre, i parametri c_j, $j = 1, 2, \ldots, P$, uno per ogni popolazione. I parametri c_j specificano quanto tendono ad essere distanti, in termini di varianza, dai loro valori tipici le frequenze degli alleli di ciascuna popolazione. Formalmente, condizionatamente a π e c:

$$\alpha_{ij} \sim \text{Normal}(\pi_i, c_j \pi_i (1 - \pi_i)) \tag{C.2.2}$$

ristretta all'intervallo $[0, 1]$, ovvero la distribuzione normale che ai punti 0 e 1 associa una probabilità pari alla massa della distribuzione normale su $(-\infty, 0)$ e $(1, +\infty)$ rispettivamente.

Per completare il modello gerarchico, consideriamo prior indipendenti su π e c ed assumiamo quindi che $\pi_1, \pi_2, \ldots, \pi_L$ e $c_1, c_2, \ldots, c_P$ siano i.i.d. con densità $p_\pi \sim Beta(a, a)$ e p_c rispettivamente.

C.3 Una motivazione al modello nell'ambito della genetica di popolazione.

Il modello per dati SNPs presentato trova una giustificazione teorica nell'ambito della genetica di popolazione.

Come ipotesi di lavoro si considerare una popolazione ancestrale relativamente recente. Come usuale in genetica di popolazione, la scala dei tempi è misurata in unità dell'ordine della numerosità della popolazione e dunque l'espressione "relativamente recente"va intesa in termini di centinaia o migliaia di generazioni.

Consideriamo un particolare locus autosomico di una particolare popolazione moderna. Con un abuso di notazione, indichiamo con π la frequenza dell'allele nel locus nella popolazione ancestrale e con α la frequenza nella popolazione discendente. Assumiamo poi che sia $\pi \neq 0,\ 1$, ovvero che il single-nucleotide polymorphism considerato esibisca variazione nella popolazione ancestrale. Supponiamo inoltre che nel locus in questione la popolazione discendente sia stata formata campionando N_0 cromosomi dalla popolazione ancestrale e, nel corso delle generazioni, abbia registrato delle fluttuazioni $N_1, N_2, \ldots N_t$ fino alla presente. Se la numerosità iniziale N_0 è relativamente piccola rispetto a quella della popolazione ancestrale, si dice che la popolazione discendente è passata attraverso un collo di bottiglia (*bottleneck*). Più in generale ogni improvviso declino della numerosità della popolazione è indicato con il termine *bottleneck*.

Sotto assunzioni piuttosto generali sulla demografia della popolazione, l'evoluzione della frequenza dell'allele nel locus considerato è bene approssimata dal modello di diffusione neutrale di Wright-Fisher con una frequenza iniziale π e per un periodo di separazione τ tra la popolazione ancestrale e quella discendente che soddisfa la sequente relazione:

$$\tau = \frac{1}{N_0} + \sigma^2 \sum_{k=1}^{t} \frac{1}{N_k} \tag{C.3.1}$$

(Ewens, 1979). Si tratta di un risultato limite per la numerosità della popolazione tendente ad infinito, pertanto l'approssimazione è tanto più accurata quanto N_k è grande. In pratica però l'approssimazione si rivela buona anche se alcuni N_k sono molto piccoli (Norborg, 2001). Osserviamo anche che i dettagli della demografia entrano nel modello solo attraverso la costante moltiplicativa σ^2.

La versione del processo di diffusione Wright-Fisher considerato ha media infinitesimale 0 e varianza infinitesimale $b(x) = x(1-x)$. L'incremento $\Delta_\tau X(t)$ della frequenza nell'intervallo di tempo infinitesimale τ da un valore iniziale π è in maniera approssimativa distribuita secondo una normale di media 0 e varianza $\tau\pi(1-\pi)$. Quindi approssimatamente:

$$\alpha \sim \text{Normal}(\pi, \tau\pi(1-\pi)) \tag{C.3.2}$$

che è consistente con la distribuzione marginale (C.2.2). Sia il modello naturale Wright-Fisher, che è un processo markoviano, che la sua diffusion approximation possiedono stati assorbenti corrispondenti al caso di avere o tutti o nessun cromosoma nella popolazione portatori della variante considerata.

In questa cornice teorica, nella (C.2.2) il parametro c_j relativo alla j-esima popolazione può essere interpretato in maniera naturale come il tempo, sulla scala del processo di diffusione, durante il quale la popolazione ha subito un effetto di deriva genetica cioè in cambiamento evolutivo indipendente dai meccanismi di selezione naturale ma deter-

minato dal campionamento iniziale della popolazione. Nel caso particolare in cui l'effetto maggiore sia invece quello del "collo di bottiglia", il parametro c_j può essere interpretato come l'inverso dell'ampiezza del collo di bottiglia (inverso della numerosità della popolazione determinata dai periodi di contrazione). La (C.3.2) è infatti l'approssimazione normale della distribuzione binomiale del numero di copie dell'allele considerato dopo che si è prodotto l'effetto del "collo di bottiglia".

In un'impostazione più generale, c_j^{-1} può essere visto come l'ampiezza effettiva del "collo di bottiglia". Per esempio, per una popolazione con una demografia coerente con le assunzioni del modello di Wright-Fisher che subisce un effetto *bottleneck* di ampiezza N per t generazioni, la (C.3.2) vale approssimativamente con $c = t/N$.

Supponiamo ora che i loci evolvano indipendentemente l'uno dall'altro—l'ipotesi è realistica per quei loci che non siano vicini tra loro sullo stesso cromosoma—in popolazioni che simultaneamente divergono da una comune popolazione ancestrale, in ogni locus, ignorando fenomeni di selezione naturale. Lo stesso modello descritto si applicherà, indipendentemente, a tutti i loci autosomi nella popolazione con indipendenza anche tra le differenti popolazioni.

Ci si aspetta che il modello si adatti meglio quando la divergenza dalla popolazione ancestrale origina immediatamente nuove sottopopolazioni. Delle violazioni si potrebbero verificare qualora le popolazioni divergessero dalla popolazione ancestrale in tempi differenti e le frequenze degli alleli cambiassero nel periodo che intercorre tra le divergenze. Questa eventualità non ha grandi effetti per popolazioni ancestrali molto numerose nelle quali le frequenze degli alleli cambiano molto lentamente.

Nell'impostazione descritta, la prior naturale f_π per le frequenze π_i è la distribuzione delle frequenze degli SNP nella popolazione ancestrale. In pratica però tale frequenze non è ovviamente nota e dipende dagli eventi demografici che hanno caratterizzato la popolazione ancestrale. Un naturale punto di partenza è lo standard neutral model per

il quale è noto che in una popolazione di numerostià effettiva N, in un sito segregante che presenta una bassa propensione a mutare, la probabilità che la variante mutante sia presente in y copie, $y = 1, 2, \ldots, N-1$, è proporzionale a $1/y$, (Ewens, 1979). Poiché non si conosce quale delle due varianti con cui si caratterizza il sigle-nucleotide polymorphism è quella mutante, la distribuzione sarà la versione simmetizzata:

$$f_\pi(\pi) \propto \frac{1}{\pi(1-\pi)} \qquad \pi \in (0,1) \tag{C.3.3}$$

APPENDICE D

Glossario.

Allele: ciascuno dei due componenti la coppia di geni presenti su ogni cromosoma omologo che controlla un determinato carattere. Provengono uno dal padre ed uno dalla madre.

Aploide: dotato di una sola copia di ciascun cromosoma.

Autosoma: termine con cui vengono chiamati i cromosomi ad eccezione di quelli sessuali.

Cellule somatiche: tutte le cellule che formano un individuo.

Diploide: organismo dotato di due copie omologhe di ciascun cromosoma.

Eterozigote: organismo in cui un determinato carattere è controllato da una coppia di geni diversi.

Eucariote: un organismo composto da una o più cellule dotate di un nucleo ben definito, delimitato dalla membrana nucleare.

Fenotipo: insieme delle caratteristiche visibili che distinguono un organismo. Benché resti sostanzialmente fedele a certi elementi basali tipici della specie e dell'individuo, il fenotipo di un organismo varia nel tempo in quanto è il risultato visibile dell'interazione fra genotipo ed ambiente. Il fenotipo di un uomo comprende migliaia di caratteristiche corporee e mentali.

Gamete: cellula specializzata aploide che si unisce ad un gamete del sesso opposto per formare uno zigote diploide; nei mammiferi, cellula-uovo e spermatozoo.

Gene: unità genetica fondamentale che viene trasmessa da una generazione all'altra assicurando la continuità dell'informazione genetica. È un frammento di cromosoma che controlla la sintesi di una proteina responsabile della realizzazione di un carattere ereditario.

Genoma: insieme delle informazioni genetiche presenti in un gamete. Le cellule somatiche, essenso diploidi, contentono due genomi, uno di origine paterna e uno di origine materna.

Genotipo: insieme dei geni che formano il corredo ereditario di un organismo.

Meiosi: particolare tipo di divisione cellulare che si verifica negli organismi a riproduzione sessuata, in cui, a partire da precursori diploidi, si formano gameti contenenti un corredo cromosomico aploide. Nel dimezzamento del numero dei cromosomi, per ciascuno di essi pu essere scelto a caso il cromosoma di origine paterna o materna.

Mitocondrio: organulo cellulare presente nel citoplasma degli eucarioti provvisto di un proprio cromosoma e deputato a funzioni connesse con le trasformazioni energetiche.

Mutazione genetica: modificazione del messaggio genetico che, causando un'alterazione nella sequenza dei nucleotidi durante la duplicazione del DNA, determina la formazione di una molecola di DNA differente dall'originale.

Omozigote: organismo in cui un determinato carattere è controllato da una coppia di geni diversi.

Procariote: microorganismo in cui il materiale cromosomico non organizzato in nucleo vero e proprio, delimitato da un involucro.

Ricombinaizone: scambio di pezzi, e quindi di geni, di cromosomi omologhi durante la meiosi.

Variabilità genetica: coesistenza in una popolazione di genotipi diversi.

Bibliografia

Balding, D. J., M. Bishop e C. Cannings, (eds) (2001) *Handbook of Statistical Genetics.* Chichester, Wiley.

Balding, D. J. e P. Donnelly, 1995 Inferring identity from DNA profile evidence. Proc. Natl. Acad. Sci. USA **92:** 11741–11745.

Bataille, M., K. Crainic, M. Leterreux, M. Durigon e P. deMazancourt, 1999 Multiplex amplification of mitochondrial DNA for human and species identification in forensic evaluation. Forens. Sci. Int. **99:** 165–170.

Bendall, K., e B. Sykes, 1995 Length heteroplasmy in the forst hypervariable segment of the human mitochondrial DNA control region. Am. J. Hum. Genet. **57:** 248-256.

Cavalli-Sforza, L. L., P. Menozzi, A. Piazza, 1994 *The History and Geography of Humans Genes.* Princeton University Press, Princeton.

Donnelly, P. and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

Drummond, J., G. K. Nicholls, A. G. Rodrigo e W. Solomon, 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161:** 1307–1320.

Evett, I. W. e B. S. Weir, 1998 *Interpreting DNA Evidence: Statistical Genetics for Forensic Science.* Sinauer, Sunderland, MA.

Ewens, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, Berlin.

Foreman, L. A., A. F. M. Smith e I. W. Evett, 1997 Bayesian Analysis of DNA profiling data in forensic identification applications. J. R. Statist. Soc. A **160:** 429–469.

Forster, P., R. Harding, A. Torroni e H. Bandelt, 1996 Origin and evolution of native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. **59:** 935–945.

Fullerton, S. M., R. M. Harding, A. J. Boyce e J. B. Clegg, 1994 Molecular and population genetic analysis of allelic sequence diversity at the human beta-globin locus. Proc. Natl. Acad. Sci. Usa **91:** 1805–1809.

Griffiths, R. C. e S. Tavaré, 1994a Simulating probability distributions in the coalescent. Theor. Pop. Biol. **46:** 131–159.

Griffiths, R. C. e S. Tavaré, 1994b Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

Hammer, M. F., 1995 A recent common ancestry for human Y chromosomes. Nature **378:** 376–378.

Handt, O., S. Meyer and A. von Haeseler, 1998 Compilation of human mtDNA control region sequences. Nucleic Acids Res. **26:** 126–130.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, e J. B. Clegg, 1997 A gene tree for β-globin sequences from Melanesia. J. Molec. Evol. **44(1):** S133–S138.

Hasegawa, M., A. Di Rienzo, T. D. Kocher and A. C. Wilson, 1993 Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. **37:** 347–354.

Hauswirth, W., e P. Laipis, 1985 Transmission genetics of mammalian mitochondria: a molecular model and experimental evidence in *Achievements and Perspectives of Mitochondrial Research*. Elsevier (Biomedical), vol II Biogenesis.

Helgason, A., B. Hrafnkelsson, J. R. Gulcher, et al., 2003 A population coalescent analysis of Icelandic matrilinea and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y Chromosomes. Am. J. Hum. Genet. **72:** 1370–1388.

Howell, N., S. Halvorson, I. Kubacka, D. A. McCullough, L. A. Bindoff e D. M. Tumbull, 1992 Mitochondrial gene segregation in mammals: is the bottleneck always narrow? Hum. Genet. **90:** 1 17-120.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. J. Futuyama and J. D. Antonovics. Oxford University Press, New York.

Kimura, M., 1969 The number of heterogeneous nucleotide sites maintained in a finite population due to steady flux mutation. Genetics **61:** 893–903.

Kimura, M. e J. F. Crow, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 788–798.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probability **19A:** 27–43.

Kocher, T. D. and A. C. Wilson, 1991 Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and protein coding regions, pp. 391–413 in *Evolution of Life: Fossils, Molecules and Culture* edited by S. Osawa and T. Honio. Springer Verlag, Tokyo.

Lundstrom, R., S. Tavaré, e R. H. Ward, 1992 Estimating substitution rates from molecular data using the coalescent. Proc. Natl. Acad. Sci. USA **89** 5961–5965.

Meyer, S., G. Weiss e A. Von Haeseler, 1999 Pattern of nucleotide substitution and rate heterogeneity in the hypervariable region I and II oh human mtDNA. Genetics **152:** 1103–1110.

Neuhauser, C., 2001 Mathematical models in population genetics, pp. 153–177 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Nicholson, G., A. V. Smith, F. Jónsson, O. Gústafsson, K. Stefánsson e P. Donnelly, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Statist. Soc. B **64(4):** 695–715.

Norborg, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Ripley, B. D., 1987 *Stochastic Simulation.* Wiley, New York.

Rogers, A. R., 1992 Error introduces by the infinite sites model. Mol. Biol. Evol. **9:** 1181–1184.

Rogers, A. R., A. E. Fraley, M. J. Bamshad, W. S. Watkins and L. B. Jorde, 1996 Mitochondrial mismatch analysis is insensitive to the mutational process. Mol. Biol. Evol. **13:** 895–902.

Schneider, S. e L. Excoffier, 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152:** 1079–1089.

Stephens, M., 2001 Inference under the coalescent, pp. 213–238 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Stoneking, M., 1993 DNA and recent human evolution. Evolutionary Antropology **2:** 60–73.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tajima, F., 1996 The amount of DNA polymorphism maintened in a finite population when the neutral mutation rate varies among sites. Genetics **143:** 1457–1465.

Takahata, N., 1993 Allelic genealogy and human evolution. Mol. Biol. Evol. **10:** 2–22.

Tamura, K., e M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

Tavaré, S., 1986 Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences, **17:** 57–86.

Tavaré, S., D. J. Balding, R. C. Griffiths e P. Donnelly, 1996 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner e M. W. Feldman, 2000 Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Nat. Acad. Sci. USA **97:** 7360–7365.

Torroni, A., T. G. Schurr, C. Yang *et al.*, 1992 Native American mitochondrial DNA analysis indicates the Amerind and Na-Dene populations were founded by two independent migrations. Genetics **130:** 153–162.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wilson, 1991 African populations and the evolution of mitochondrial DNA. Science **253:** 1503–1507.

Wakeley, J., 1993 Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

Wakeley, J. and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

Ward, R. H., B. L. Frazier, K. Dew-Jager e S. Pääbo, 1991 Extensive mitochondrial diversity within a single Amerindian tribe. Nat. Acad. Sci. **88:** 8720–8724.

Watterson, G. A., 1975 On the number of segretating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 256–276.

Weir, B. S., 2001 Forensics, pp. 721–739 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Weiss, G. e A. Von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539-1546.

Wilson, I. J. e D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Wilson, I. J., M. E. Weale e D. J. Balding, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Statist. Soc. A **166:** 1–33.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Yang, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10(6):** 1396–1401.

Yang, Z., 1996 Statistical Properties of a DNA Sample Under the Finite–Site Model. Genetics **144**: 1941–1950.

Yang, Z. e B. Rannala, 1997 Bayesian phylogenetic inference using DNA sequence: a Markov Chain Monte Carlo method. Mol. Biol. Evol. **14:** 717–724.