Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2002/05

Spazi campionari e polimorfismo nelle regioni HV1 e HV2 del DNA mitocondriale

Federico Mattia Stefanini

Università degli Studi
di Firenze

Biostatistics

Spazi campionari e polimorfismo nelle regioni HV1 ed HV2 del DNA mitocondriale

Federico M. Stefanini

Dipartimento di Statistica 'G.Parenti', Università degli Studi di Firenze,

Viale Morgagni 59, I-50134, Firenze, Italia

Tel.: +39 055 4237266 Fax: +39 055 4223560

stefanin@ds.unifi.it http://www.ds.unifi.it/∼stefanin/

Riassunto. L'impiego del DNA mitocondriale in ambito forense è in costante aumento. L'uso efficiente dei dati molecolari richiede la formulazione di modelli statistici in cui la struttura gerarchica della popolazione di molecole è esplicitamente riconosciuta.

In questo lavoro viene costruita una gerarchia di spazi campionari relativi ai diversi livelli di osservazione del polimorfismo mitocondriale per quanto riguarda le regioni HV1 e HV2 del *D-Loop*. La formalizzazione è immediatamente utile nell'elicitazione dell'informazione a priori, nella costruzione di modelli statistici e nella implementazione al calcolatore di algoritmi Monte Carlo per l'adattamento dei modelli.

1 Introduzione

Le popolazioni umane mostrano livelli elevati di polimorfismo nel DNA mitocondriale (mtDNA). L'elevato grado di polimorfismo unito alla bassa propensione alla degradazione tipica di tale molecola promettono un'autentica rivoluzione nella scienza forense [3] principalmente nei problemi di attribuzione di identità sulla base di reperti biologici e per valutare quantitativamente l'evidenza a carico di una persona sospettata impiegando tracce di materiale biologico proveniente dalla scena di un crimine [5, 6, 7].

A differenza di quanto accade per il DNA nucleare, l'impiego dei dati molecolari mitocondriali in ambito forense non è ancora largamente diffuso

principalmente per le problematiche di metodo e valutazione quantitativa dell'evidenza. Quale che siano le assunzioni e le semplificazioni operate nella formulazione di un modello, è importante stabilire un quadro di riferimento concettuale riguardo la struttura gerarchica del sistema genetico.

Nei paragrafi che seguono viene proposta una formalizzazione che cattura gli aspetti salienti del polimorfismo mtDNA nelle regioni HV1 e HV2 ai diversi livelli di osservazione possibili in un sistema biologico, a partire dalla singola molecola di mtDNA sino a popolazioni di individui umani.

2 Il DNA mitocondriale

Il DNA mitocondriale è organizzato in una molecola circolare (anello) costituito da due catene appaiate (*strands*), dette H ed L, ognuna di lunghezza complessiva pari a 16569 paia di basi [1], etichettate con i numeri da 1 a 16569. La regione di tale anello nota come *D-loop* contiene due sottoregioni ipervariabili, note come HV1 e HV2, rispettivamente costituite dalle coppie di basi tra 16000 e 16430 la prima e tra 40 e 440 la seconda.

Il polimorfismo delle regioni HV1 e HV2 si sostanzia nella diversità delle sequenze presenti in una popolazione. In ogni posizione la coppia di basi, una per *strand*, è un elemento dell'insieme:

$$\{(A,T),(T,A),(C,G),(G,C)\}\,, \tag{1}$$

in cui la scrittura (A,T) indica, per convenzione, che sulla catena H è situata la base A mentre nella medesima posizione ma sulla catena L si trova la base T. Una coppia di basi appartiene al sottoinsieme ottenuto dal prodotto cartesiano $\mathcal{A}^2$ dell'alfabeto $\mathcal{A} = \{A,T,C,G\}$, ma non tutte le coppie sono possibili. La specificità degli appaiamenti descritti nella (1) ha ragioni chimico-fisiche. In assenza di evidenze sperimentali contrarie alla regola di appaiamento espressa dalla (1), vale la seguente:

Assunzione 1: In condizioni fisiologiche l'mtDNA non è presente in forma heteroduplex. Pertanto la caratterizzazione della sequenza relativa ad una sola catena non causa la perdita di informazione (per convenzione si adotterà la catena H), cioé di polimorfismo.

Restringendo l'interesse alle sottoregioni ad alto polimorfismo, HV1 ed HV2, conviene sostituire alla sequenza intera dell'mtDNA quella regione ottenuta concatenando le sequenze delle sole sottoregioni HV1 ed HV2. Pertanto, ogni elemento di tale insieme ristretto indica anche una classe di equivalenza, quella di tutte le sequenze che hanno quella data successione di basi in HV1 ed HV2 ma differiscono (eventualmente) per una o piú basi in posizioni fuori da tale regione. La restrizione è motivata dal fatto che le due regioni sono altamente polimorfe, perciò di prevalente interesse in ambito forense. Inoltre, il sequenziamento dell'intera molecola di mtDNA risulta costoso, non sempre fattibile e promette contenuti vantaggi aggiuntivi derivati dal sequenziamento di regioni poco variabili.

Attualmente si ritiene che gli eventi di ricombinazione nell'mtDNA tra le regioni HV1 ed HV2 siano praticamente assenti. Pertanto, conviene considerare le due regioni come completamente associate (*linked*):

Assunzione 2: Le regioni HV1 ed HV2 sono situate su di una molecola non soggetta a ricombinazione. Indichiamo come HV12 la regione ottenuta giustapponendo le sequenze da 16000 a 16430 e quella da 40 a 440. La lunghezza complessiva è pari a $431 + 401 = 832$ basi. La perdita di polimorfismo indotta dalla restrizione a HV12 non è rilevante.

Dall'Assunzione 2, lo spazio campionario Ω_0 per il livello di osservazione costituito da una sola molecola HV12 si ottiene ricorrendo al prodotto cartesiano dell'alfabeto $\mathcal{A}$, dunque $\Omega_0 = \mathcal{A}^{832}$.

Tuttavia lo spazio Ω_0 non cattura tutto il polimorfismo di HV12, perché rimangono escluse le molecole che recano inserzioni e delezioni in uno o più posizioni della sequenza, ovvero le sequenze di lunghezza diversa da 832 basi.

Assunzione 3: La lunghezza minima di una sequenza HV12 deleta è di 782 basi, quella massima di una sequenza con inserzioni è di 882 basi.

L'Assunzione 3, consente di definire lo spazio Ω_1 in cui plausibilmente è situata ogni possibile sequenza per la regione HV12, incluse le sequenze che recano delezioni ed inserzioni:

$$\Omega_1 = \bigcup_{j=782}^{882} \mathcal{A}^j. \tag{2}$$

2.1 Spazi ottimizzati con tipologia di mutazione fissata

Lo spazio Ω_1 è adeguato per rappresentare il polimorfismo del DNA mitocondriale, ma la sua cardinalità è assai grande. Vi sono almeno due validi motivi per ricercare spazi di cardinalità contenuta. Il primo riguarda la scelta di una funzione di probabilità per gli eventi semplici di tipo $\{\omega\} \subset \Omega_1$. Il modello saturo richiede un numero di parametri tanto elevato da comportare l'intrattabilità dei calcoli. In secondo luogo, il polimorfismo effettivamente osservato nelle popolazioni umane, benché in termini relativi assai ricco, risulta estremamente inferiore a quello teoricamente definito da Ω_1. Una sequenza HV12 puó contenere una delezione in una certa posizione ma non è ragionevole che la delezione includa l'intera regione del *D-loop*. L'identificazione di trascritti codificati in HV12, seppure con funzione ignota, suggerisce che essi esercitino un ruolo di una qualche utilità che verrebbe meno in caso di delezione. Inoltre, è ragionevole assumere che HV12 eserciti un ruolo stabilizzatore meccanico e/o termodinamico, oltre al noto ed importantissimo ruolo di controllo della espressione genica nella parte tradotta fuori dalla regione HV12.

In considerazione del numero non trascurabile di studi effettuato su mtDNA di umani appartenenti a diverse popolazioni e razze, si potrebbe assumere che il sottoinsieme di mutazioni osservate in HV12 [4] esaurisca gran parte delle tipologie ammesse dal regolare funzionamento dell'mtDNA.

Assunzione 4: Il sequenziamento della regione HV12 è stato effettuato su di un campione grande rispetto al polimorfismo delle popolazioni umane. Pertanto, i tipi di mutazione ad oggi non osservati costituiscono una parte negligibile del polimorfismo totale.

Uno spazio campionario 'ottimizzato' per la cardinalità può essere definito in base all'Assunzione 4. Sia α la sequenza della catena H nella regione HV12 osservata da Anderson [1]. In accordo con le convenzioni correnti per l'indicazione delle sequenze [4], si può indicare una generica sequenza HV12 (stato della sequenza in una data molecola) ricorrendo all'elenco dei modi e delle posizioni in cui essa differisce da α.

In Tabella 1 è riportato l'insieme delle posizioni e delle modificazioni di sequenza osservate nella regione HV12 umana (MITOMAP, 2001).

Sia $\mathcal{M}_2$ la collezione delle coppie posizione-mutazione. Allora, lo spazio campionario Ω_2 puó essere formulato a partire dal generatore $\mathcal{G}_2$ che è sottoinsieme dell'insieme delle parti di $\mathcal{M}_2$:

$$\mathcal{G}_2 \subset \mathcal{P}(\mathcal{M}_2). \tag{3}$$

Lo spazio Ω_2 è generato da $\mathcal{G}_2$, applicando ad α ognuna delle trasformazione in g per ogni g in $\mathcal{G}_2$, ovvero: $\Omega_2 = \{\omega : g \mapsto \omega \wedge g \in \mathcal{G}_2\}$,

Assunzione 5: Il numero di mutazioni, rispetto α, contemporaneamente presenti in una sequenza non supera le 50.

In base all'Assunzione 5, lo spazio Ω_2 deve essere modificato ponendo il vincolo $card(g) < 50$.

2.2 Spazi ottimizzati con tipologia di mutazione non prefissata: una scelta operativa

In questa sede è conveniente rappresentare una generica sequenza ω come elenco g di trasformazioni da operare sulla sequenza α di Anderson per ottenere ω.

L'attuale stato di conoscenza dei meccanismi molecolari alla base del funzionamento mitocondriale non consente di impiegare le Assunzioni 4 e 5 acriticamente. Il campione dei sequenziamenti disponibili, seppur non trascurabile, non sembra così grande da rassicurare il ricercatore nei confronti dell'esistenza di sequenze rare sì, ma con frequenza non nulla nella popolazione. La presenza di campionamento mitocondriale (*drift*) durante la formazione dello zigote potrebbe comportare un numero di sequenze di HV12 diverse mantenute nella popolazione che è sensibilmente inferiore al numero massimo di sequenze teoricamente differenti. Un'analogo effetto potrebbe derivare da fenomeni di selezione zigotica contro sequenze fisiologicamente non efficaci e dalla trasmissione di un numero di molecole estremamente piccolo dalla madre ai discendenti (*bottleneck effect*).

pos	mut	pos	mut	pos	mut	pos	mut	pos	mut
7	A-G	9	G-A	40	T-C	41	C-ins	43	C-ins
56	C-ins	58	T-C	63	T-C	64	C-T	66	G-T
72	T-A	72	T-C	73	A-G	89	T-C	93	A-G
95	A-C	97	G-A	98	C-del	103	G-A	106	G-del
107	G-del	108	A-del	109	G-del	110	C-del	111	A-del
114	C-T	125	T-C	127	T-C	132	C-G	143	G-A
146	T-C	150	C-T	151	C-T	152	T-C	153	A-G
159	T-C	182	C-T	185	G-T	185	G-C	185	G-A
186	C-A	186	C-T	188	A-G	189	A-C	189	A-G
190	A-ins	191	A-ins	192	T-C	193	A-G	194	C-T
195	T-C	196	T-C	198	C-T	199	T-C	200	A-G
202	A-G	203	G-A	204	T-C	207	G-A	208	G-A
210	A-G	214	A-G	215	A-G	222	C-T	222	C-G
225	G-A	226	T-C	227	A-G	228	G-A	234	A-G
235	A-G	236	T-C	239	T-C	241	A-G	242	C-T
247	G-A	248	A-G	249	A-del	250	T-C	257	A-G
258	C-T	263	A-G	264	C-T	271	C-T	273	C-G
282	T-C	290	A-del	291	A-del	294	T-ins	295	C-T
297	A-G	302	CCC-ins	303	C-del	303	C-CC	303	C-CCC
309	C-T	309	C-del	309	CCC-ins	310	C-ins	310	T-C
311	C-CC	315	C-CC	315	C-CCC	315	C-del	316	G-A
318	T-C	319	T-C	325	C-T	353	C-CC	357	A-G
373	A-G	374	A-G	375	C-T	385	A-G	390	A-G
420	C-A	444	A-G	456	C-T	462	C-T	471	T-C
475	A-G	477	T-C	489	T-C	493	A-G	497	C-T
499	G-A	508	A-G	512	A-C	513	G-A	514	C-CAC
514	C-del	515	A-del	520	C-$(CA)_3$C	520	C-del	520	C-CAC
521	A-del	522	C-del	523	A-del	533	A-G	568	C-CCC
568	C-$(C)_4$	568	C-$(C)_5$	568	C-$(C)_6$	568	C-$(C)_7$	575	C-T
16025	T-A	16025	T-G	16037	A-G	16039	G-A	16041	A-G
16042	G-A	16048	G-A	16051	A-G	16059	A-G	16067	C-T
16069	C-T	16070	A-G	16071	C-T	16075	T-C	16080	A-G
16082	T-C	16083	C-A	16086	T-C	16089	C-T	16092	T-C
16093	T-C	16094	T-G	16095	C-T	16097	T-C	16102	T-C
16104	C-A	16104	C-T	16107	C-T	16108	C-T	16111	C-T
16113	A-T	16114	C-A	16114	C-T	16124	T-C	16126	T-C
16127	A-G	16129	G-A	16131	T-C	16132	A-T	16133	C-T
16134	C-T	16136	T-C	16139	A-del	16140	T-C	16142	C-T
16144	T-A	16145	G-A	16146	A-G	16147	C-G	16147	C-T
16148	C-T	16150	C-T	16153	G-A	16154	T-C	16155	A-G
16158	A-T	16160	A-G	16161	T-C	16162	A-G	16163	A-G
16164	A-G	16166	A-C	16167	C-T	16168	C-T	16169	C-T
16171	A-G	16172	T-C	16173	C-T	16174	C-T	16176	C-G
16176	C-T	16179	C-T	16180	A-G	16181	A-G	16182	A-G
16182	A-del	16183	A-C	16183	A-del	16183	A-AC	16183	A-ACC
16184	C-A	16184	C-T	16184	C-del	16184	C-CC	16184	C-CCC
16184	C-$(C)_4$	16184	C-$(C)_5$	16185	C-T	16186	C-A	16186	C-T
16187	C-T	16188	C-A	16188	C-G	16188	C-T	16189	T-C
16190	C-ins	16191	C-T	16192	C-T	16193	C-CC	16193	C-ins
16193	C-T	16194	A-G	16195	T-C	16195	T-G	16196	G-A
16197	C-G	16198	T-C	16203	A-G	16206	A-C	16207	A-G
16209	T-C	16212	A-G	16213	G-A	16214	C-T	16215	A-C
16215	A-G	16217	T-C	16218	C-T	16219	A-G	16220	A-C
16220	A-G	16220	A-del	16221	C-T	16222	C-T	16222	C-del
16223	C-T	16224	T-C	16227	A-G	16230	A-G	16231	T-C
16232	C-A	16232	C-T	16233	A-G	16234	C-T	16235	A-G
16236	C-T	16239	C-G	16239	C-T	16240	A-G	16240	A-T
16241	A-G	16242	C-T	16243	T-C	16245	C-T	16246	A-C
16246	A-T	16247	A-G	16248	C-T	16249	T-C	16252	A-G
16254	A-G	16255	G-A	16256	C-T	16257	C-A	16257	C-T
16258	A-G	16259	C-T	16259	C-del	16260	C-T	16261	C-T
16263	T-C	16264	C-G	16264	C-T	16265	A-C	16265	A-G
16265	A-T	16266	C-A	16266	C-G	16266	C-T	16268	C-T
16268	C-del	16269	A-G	16270	C-A	16270	C-T	16271	T-C
16272	A-G	16273	G-A	16274	G-A	16275	A-G	16275	A-del
16277	A-T	16278	C-T	16280	A-G	16281	A-G	16284	A-G
16286	C-T	16287	C-T	16288	T-C	16289	A-G	16290	C-T
16291	C-T	16292	C-T	16293	A-C	16293	A-G	16294	C-T
16295	C-T	16296	C-T	16297	T-C	16298	T-C	16299	A-G
16300	A-G	16301	C-T	16302	A-G	16303	G-C	16304	T-C
16304	T-G	16305	A-G	16308	A-G	16309	A-G	16310	G-A
16311	T-C	16312	A-G	16313	T-C	16314	A-G	16316	A-G
16318	A-T	16319	G-A	16319	G-C	16319	G-del	16320	C-T
16323	T-C	16324	T-C	16325	T-C	16326	C-T	16327	C-T
16331	A-G	16335	A-G	16336	G-A	16342	T-C	16343	A-C
16343	A-T	16343	A-G	16344	C-T	16349	A-T	16352	T-C
16353	C-T	16354	C-T	16355	C-A	16355	C-T	16356	T-C
16357	T-C	16359	T-C	16360	C-T	16361	G-A	16362	T-C
16365	C-T	16366	C-T	16367	A-G	16368	T-C	16381	T-C
16389	G-A	16390	G-A	16391	G-A	16395	C-T	16398	G-A
16399	A-G	16400	C-T	16422	T-C	16424	T-C	16429	C-T
16440	T-C	16456	G-A	16463	A-G	16482	A-G	16483	G-A
16497	A-G	16519	T-C	16540	C-T				

Tabella 1: Posizioni e tipo di mutazioni del *D-loop* mitocondriale. La scrittuta $(C)_4$ equivale alla sequenza CCCC (MITOMAP, 2001).

Nel rilassare l'Assunzione 4 lo spazio campionario Ω_2 deve essere arricchito di un certo numero di elementi ma senza giungere ad una cardinalità uguale a quella di Ω_1.

Assunzione 6: La sequenza α di HV12 può presentarsi mutata in una qualsiasi posizione delle 832: tutte le posizioni sono mutazionalmente calde.

La costruzione dell'insieme dei tipi di mutazione è effettuata in passi successivi riconoscendo le tipologie di mutazione genetica rilevanti in questa sede: sostituzione, inserzione e delezione.

Sia ω una generica sequenza HV12. Sia $\omega[j]$ la base azotata localizzata in posizione j della sequenza ω. Sia $\mathcal{A} = \{A, T, C, G\}$ l'alfabeto relativo alle basi azotate e $\mathcal{A}\backslash\omega[j]$ il sottoinsieme ottenuto escludendo la base che compare in $\omega[j]$. Per sostituzione puntiforme g_S di una base in posizione j nella sequenza ω si intende la trasformazione che sostituisce la base $\omega[j]$ con la base $B \in \mathcal{A}\backslash\omega[j]$, formalmente indicata con $g_S(\omega, j, B)$.

Per delezione g_D di ampiezza l in posizione j nella sequenza ω si intende una trasformazione che accorcia la sequenza da $len(\omega)$ a $len(\omega) - l$ eliminando l basi azzotate a partire da quella in posizione j, quindi la sequenza deleta è $g_D(\omega, j, l)$.

Per inserzione g_I di ampiezza l in posizione j nella sequenza ω si intende una trasformazione che allunga la sequenza da $len(\omega)$ a $len(\omega) + l$ inserendo l basi azzotate a partire da quella in posizione j, formalmente $g_I(\omega, j, l)$.

Evidentemente una sostituzione puntiforme è equivalente ad una delezione puntiforme seguita da una inserzione puntiforme (o viceversa), dunque impiegando queste due ultime classi di mutazioni si potrà formalmente trattare anche le sostituzioni.

Nella formalizzazione proposta in seguito, è utile impiegare "l'intorno" della sequenza di Andreson in HV12, cioé l'insieme di sequenze che differiscono per un numero di trasformazioni di α inferiore ad un assegnato valore d. Un intorno $\mathcal{L}_{\omega,d}$ della sequenza ω è definito come

$$\mathcal{L}_{\omega,d} = \{\omega_j : (g_j \mapsto \omega_j) \wedge (card(g_j) \leq d) \wedge (g_j \in \mathcal{G}_2)\}\,.$$

In base alle informazioni genetico-molecolari disponibili, un valore ragionevole di d può essere definito incrementando i valori sperimentalmente osservati entro ogni tipologia di mutazione. Intuitivamente, il procedimento corrisponde all'inclusione di sequenze che differiscono 'di poco' da quelle osservate. Ad esempio, inserzioni e delezioni sono tipicamente di una sola base, ma in Ω_2 sono incluse sequenze con delezione di due basi. Analogamente, inserzioni di poliC fino a 7 basi sono state osservate, ma in Ω_2 risultano definite anche sequenze con inserzioni pari a 10 basi, e in qualsiasi posizione della sequenza.

Si procede alla costruzione di Ω_3 allargando Ω_2 a sequenze plausibili ma non osservate in base al fatto che piccole variazioni di sequenza possono essere fisiologiche come le forme più frequenti nella popolazione.

La costruzione effettuata deve essere messa in relazione con i dati sperimentali: sono state osservate 47 differenti sequenze ed un totale di 49 posizioni in cui la sequenza originale di Anderson è mutata. Inserzioni e delezioni sono tipologie di mutazione non trascurabili, dato che ammontano a circa il 10.8% delle mutazioni totali [3].

Assunzione 7: La sequenza α di HV12 può presentarsi deleta od inserta in più posizioni ma la lunghezza della sequenza risultante non differisce più di 50 basi rispetto alla lunghezza della sequenza $len(\alpha)$.

Dalla Assunzione 7, siano definiti i due insiemi di trasformazioni $\mathcal{T}_I$ e $\mathcal{T}_D$ in cui la locazione l è assegnata ed i parametri non esplicitati possono assumere uno dei possibili valori ammessi:

$$\mathcal{T}_{I,l} = \{g_I : |len(g_I(\omega)) - len(\omega)| \leq 50\}$$

$$\mathcal{T}_{D,k} = \{g_D : |len(g_D(\omega)) - len(\omega)| \leq 50\}$$

Le trasformazioni in $\mathcal{T}_D$ sono ovviamente modificate quando la posizione dista dal termine della sequenza meno dell'ampiezza massima della delezione.

Per includere le sostituzioni di una base tra le possibili trasformazioni si deve considerare il cartesiano $\mathcal{T}_{I,l} \times \mathcal{T}_{I,k}$, in cui $l = k = 1$ indica la sostituzione, $l = 0$ indica una delezione e $k = 0$ un'inserzione.

In conclusione una singola mutazione puntiforme è un punto del cartesiano

$$\mathcal{M}_3 = \mathcal{T}_{I,l} \times \mathcal{T}_{I,k}.$$

Mutazioni in più posizioni sono rappresentate da un sottoinsieme dell'insieme potenza $\mathcal{G}_3 = \mathcal{P}(\mathcal{M}_3)$.

Assunzione 8: Il numero massimo di mutazioni che possono essere contenute in una sequenza di HV12 è pari a 50, nel senso che la sequenza mutata deve giacere nell'intorno $\mathcal{L}_{50}$ di α, e il loro effetto congiunto produce differenze di lunghezza (rispetto α) inferiori a 50 basi.

Lo spazio campionario risultante è composto da tutte le sequenze che sono generabili a partire da α applicando meno di 51 trasformazioni tra quelle indicate da $\mathcal{M}_3$:

$$\Omega_3 = \{\omega : \omega \wedge g \in \mathcal{G}_3 \wedge card(g) \leq 50 \wedge |len(\alpha) - len(\omega)| \leq 50\}.$$

L'insieme di tutte le sequenze ad oggi ottenute, e basate sulle mutazioni descritte in Tabella 1 è un sottoinsieme di Ω_3.

3 Il mitocondrio

Il mitocondrio è l'organello cellulare in cui è situata la molecola di mtDNA. A livello di metabolismo cellulare, il genoma (molecola-circolare) mitocondriale viene duplicato nella fase che precede l'aumento del numero di mitocondri.

Evidenze recenti hanno mostrato come il genoma mitocondriale all'interno di un mitocondrio possa raggiungere 2.5 copie per organello.

Assunzione 9 Il numero massimo di molecole circolari, e più precisamente di regioni HV12, dentro un organello mitocondriale è pari a 5.

In un mitocondrio possono esserci sequenze HV12 differenti perché il processo di replicazione del mtDNA non è a prova di errore (*error prone*). Pertanto, lo spazio campionario relativo al livello di osservazione del singolo mitocondrio è

$$\Omega_4 = \bigcup_{j=1}^{5} \{\Omega_3\}^j, \tag{4}$$

E' importante rammentare che il livello di osservazione Ω_4 non è facilmente accessibile, sia per i costi che per le difficoltà tecniche sperimentali.

Una possibile semplificazione si ottiene ponendo $\Omega_4 = \Omega_3$.

Assunzione 10: Un qualsiasi mitocondrio è osservato sempre e solo nella fase in cui contiene una sola molecola di mtDNA.

L'Assunzione 10 può essere motivata in presenza di amplificazione PCR, dato che una delle sequenze eventualmente differenti puó in tal circostanza essere selettivamente amplificata attraverso un meccanismo di campionamento e reazione a catena. Ovvero, la prima sequenza amplificata domina sull'amplificazione di tutte le altre sequenze.

4 Dalle cellule alle popolazioni: una gerarchia di livelli di osservazione

In una cellula sono contenuti un numero di mitocondri che può variare da 1 a circa 10000. La cellula costituisce il tipico livello di osservazione perché il materiale biologico costituito dall'mtDNA raggiunge una quantità che consente pratiche di laboratorio abbastanza ripetibili. In alcuni casi è anche possibile valutare la ripetibilità dei dati ottenuti, una circostanza di massimo rilievo nelle applicazioni forensi.

Dato l'elevato numero di mitocondri, dunque di sequenze HV12, contenute in una cellula, si ripropone la circostanza già affrontata per il livello di osservazione di un solo mitocondrio. Pertanto, un adeguato spazio campionario Ω_5 può essere definito da:

$$\Omega_5 = \bigcup_{j=1}^{10000} \{\Omega_4\}^j \,, \tag{5}$$

in cui j indica il numero dei fattori nel prodotto cartesiano.

E' importante sottolineare che il tasso di mutazione per l'mtDNA è da 5 a 10 volte quello nucleare, dunque relativamente alto [3]. In generale esso è ritenuto sufficientemente basso da far sì che la probabilità di avere sequenze diverse in un mitocondrio con genomi multipli per un numero non piccolo di mitocondri sia negligibile.

Assunzione 11: Il tasso di mutazione per HV12 è nell'ordine di $1 \cdot 10^{-5}$ per sequenza per replicazione, un valore sufficientemente piccolo perché la probabilità di avere in una cellula un numero rilevante di sequenze mutate sia trascurabile rispetto alla sequenza ereditata che è rappresentata con frequenza maggiore.

A questo livello di osservazione poche copie di mtDNA possono essere mutate rispetto al grande numero di quelle identiche, pertanto la strumentazione di laboratorio non riconoscerà il debole segnale molecolare di polimorfismo. Tale segnale è risulta nell'ordine di grandezza del rumore di fondo, dunque esso è coperto dal forte segnale dovuto al grande numero di copie non mutate.

Si consideri ora un tessuto biologico composto da K cellule. Lo spazio campionario relativo alle sequenze provenienti da un certo numero di cellule (espianto di tessuto-organo) può essere definito a partire dallo spazio Ω_5. Se l'espianto è costituito da k cellule, allora lo spazio campionario è

$$\Omega_6 = \Omega_5^k, \tag{6}$$

in cui $k = K$ corrisponde all'espianto dell'intero tessuto.

Il valore k non è in genere sotto lo stretto controllo del ricercatore, ad esempio puó dipendere dalle caratteristiche del campione biologico rinvenuto sulla scena del crimine.

A livello di osservazione tissutale, ed ancora di più in organi diversi, il numero di divisioni cellulari che separano due cellule casualmente scelte è relativamente elevato. Pertanto il numero di repliche del mtDNA che separano due sequenze HV12 ottenute da organi-tessuti diversi è altrettanto grande. In queste circostanze, la probabilità di avere molteplici copie di una sequenza mutata durante la replicazione dell'mtDNA originario diventa non trascurabile. Questo fenomeno è indicato come eteroplasmia (ad esempio, [2]), ed indica che lo stesso individuo umano può presentare sequenze HV12 diverse in dipendenza dall'organo o tessuto da cui il campione biologico proviene. Le sequenze mutate, in questa circostanza, hanno raggiunto una numerosità tale da consentirne l'osservazione sperimentale.

Per questo motivo strutture biologiche quali gli organi, oppure un individuo, sono rappresentabili come una T-pla di tessuti di numerosità $\underline{N}_T =$

$(N_1, \ldots, N_T)$. Tracce rinvenute sul luogo di un crimine contengono un numero di tessuti $1 \leq t \leq T$, con il vettore $\underline{n}_t = (n_1, \ldots, n_t)$ che indica il numero di cellule per tipologia tissutale. L'individuo, od una sua traccia, è rappresentabile come punto nello spazio campionario

$$\Omega_7 = \Omega_6^{n_1} \times \ldots \times \Omega_6^{n_t}, \tag{7}$$

in cui eventualmente per ogni i vale $n_i = N_i$.

In completa analogia con l'astrazione operata per un individuo, una popolazione di M individui è un punto nello spazio

$$\Omega_8 = \Omega_7^M, \tag{8}$$

in cui i prodotti sono cartesiani e per individui rappresentati dallo stesso numero di cellule per tessuto.

Una collezione di popolazioni di numerosità $M_1, M_2 \ldots$ è rappresentabile come un punto nello spazio $\Omega_9 = \prod \Omega_8^{M_i}$.

Ovvie modificazioni valgono qualora siano considerati campioni biologici estratti da individui differenti di popolazioni diverse, ovvero $m_i < M_i$ ed $n_i < N_i$.

Il livello di osservazione Ω_7, insieme a quello del singolo individuo, è particolarmente importante per le applicazioni in ambito forense, nelle quali l'interesse è diretto verso il contenuto informativo discriminante. I dati molecolari rappresentano le evidenze impiegate per stabilire un legame probabilistico tra un individuo ed il campione rinvenuto sulla scena del crimine.

5 Discussione

La formalizzazione degli spazi campionari proposta in questo lavoro può essere impiegata come ausilio nell'elicitazione dell'informazione posseduta dagli specialisti di queste tecniche molecolari. Le semplificazioni operate per definire spazi e modelli da impiegare nell'analisi dati possono essere discusse con supporto di una formalizzazione che diminuisce la possibilità di fraintendimenti.

Un ulteriore impiego riguarda lo sviluppo di modelli empirici e teorici per la stima del polimorfismo. Il ricorso ad algoritmi di tipo Monte

Carlo richiede l'implementazione software di spazi di sequenze, oltre che di parametri, in cui la formalizzazione sviluppata può costituire un rilevante ausilio. La formulazione di processi stocastici per la dinamica delle sequenze deve tenere conto dei differenti livelli di strutturazione gerarchica delle sequenze mitocondriali, dal mitocondrio alle popolazioni.

Dal punto di vista operativo è importante stabilire quanti e quali siano i mitotipi presenti in un individuo (sequenze mitocondriali differenti). Per rispondere a questa domanda bisogna rammentare che il risultato di un'analisi molecolare si basa su metodi indiretti che sono soggetti ad errore, ad esempio nella rilevazione della fluorescenza, e che possiedono certe caratteristiche di sensibilità e specificità. Al campionamento biologico dei mitocondri a diversi livelli della gerarchia illustrata in precedenza, si aggiungono gli effetti dovuti al laboratorio, al protocollo, al momento temporale dell'analsi, alla variabilità della singola misura. Essi non possono essere ignorati se si desidera usare i dati di tipo molecolare in maniera efficientemente in ambito forense, ma anche qualora la certificazione di qualità delle analisi mtDNA faccia parte della pratica di laboratorio.

Ringraziamenti

Ringrazio Giampietro Lago ed il Raggruppamento Carabinieri Investigazioni Scientifiche, Roma, che ha parzialmente finanziato questo lavoro. Parte di questa ricerca è stata svolta entro un progetto del COFIN 2001, coordinato da Fabio Corradi.

Riferimenti bibliografici

[1] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Douin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, and I. G. Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–465, 1981.

[2] A. Berti, V. Manzari, and G. Lago. *Il DNA mitocondriale nell'analisi forense*. Unpublished seminar notes, 1998.

[3] J. M. Butler and B. C. Levin. Forensic applications of mitochondrial DNA. *TIBTECH*, 16:158–162, 1998.

[4] Center for Molecular Medicine. *MITOMAP: A Human Mitochondrial Genome Database.* Emory University, Atlanta, GA, USA, http://www.gen.emory.edu/mitomap.html, 2001.

[5] P. Gill, P.L. Ivanov, C. Kimpton, R. Piercy, L. Benson, G. Tully, I. Evett, K. Sullivan, and E. Hagelberg. Identification of the remain of the Romanov family by DNA analysis. *Nature Genetics*, 6:130–135, 1994.

[6] M. Holland and W. Parsons (eds). *Mitochondrial DNA sequence analysis in forensic casework: Methods and Current issues.* Unpublished Congress Acta, 1998.

[7] P.L. Ivanov, M. J. Wadhams, R.K. Roby, M. M. Holland, V. W. Weedn, and T. J. Parsons. Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia George Romanov established the authenticity of the remains of Tsar Nicholas II. *Nature Genetics*, 12:417–420, 1996.