UNIVERSITÀ DI PISA
DIPARTIMENTO DI MATEMATICA
Largo Bruno Pontecorvo, 5
I - 56127 - Pisa
Tel. +39 050 2213223
Fax +39 050 2210678
matematicaprotocollo@pec.unipi.it
http://www.dm.unipi.it
C.F. 80003670504
P.I. 00286820501

Consiglio di Dipartimento del 4 novembre 2019

Omissis

7. Ricerca e Terza Missione

7.3. Relazione conclusiva sull'attività di ricerca svolta dal dott. Filippo Disanto, RtdB: parere

Il Consiglio,
PRESO ATTO: del contratto di lavoro subordinato per ricercatore a tempo determinato ex art. 24, punto 3, lettera b) della Legge 240/2010, di durata triennale, stipulato con il dott. Filippo Disanto nell'ambito del Programma per giovani ricercatori "Rita Levi Montalcini", con decorrenza 15 febbraio 2017, in scadenza il 14 febbraio 2020;
ACCERTATO: che, secondo quanto disposto dall'art. 4 del suddetto contratto, il ricercatore, non oltre 90 giorni prima della scadenza di ciascun anno di durata del contratto, è tenuto a presentare al Dip.to una dettagliata relazione sull'attività di ricerca svolta;
PRESO ATTO: della relazione conclusiva, inerente il triennio 15 febbraio 2017-14 febbraio 2020, presentata dal Dott. Disanto (**all. n. 6**);
VISTA: la L. 30 dicembre 2010 n. 240 "Norme in materia di organizzazione delle università, di personale accademico e reclutamento, nonché delega al Governo per incentivare la qualità e l'efficienza del sistema universitario";

DELIBERA

di esprimere un giudizio pienamente positivo sull'attività di didattica e di ricerca svolta dal dott. Filippo Disanto nel periodo indicato.

La presente delibera, contrassegnata dal **numero 63,** è approvata all'unanimità.

Il Segretario
Dott.ssa Cristina Lossi

Il Presidente
Prof. Matteo Novaga

Relazione attività RTDb Levi Montalcini: **Filippo Disanto**

L'attività didattica e di ricerca svolta a partire dalla mia presa di servizio come RTDb Levi Montalcini presso il Dipartimento di Matematica dell' Univeristà di Pisa è riassunta qui di seguito.

1 Didattica svolta nel periodo in esame

A.A. 2016/17: n.40 ore di lezione per il corso di Matematica a Geologia e Scienze Naturali (titolare Prof. M. Abate).

A.A. 2017/18: n.39 ore di esercitazione per il corso di Geometria 1 a Fisica (titolare Prof. M. Salvetti).

A.A. 2017/18: n.42 ore di lezione per il corso di Matematica a Geologia e Scienze Naturali (titolare Prof. M. Abate).

A.A. 2018/19: n.46 ore di esercitazione per il corso di Geometria 1 a Fisica (titolare Prof. M. Salvetti).

A.A. 2018/19: n.42 ore di lezione per il corso di Matematica a Geologia e Scienze Naturali (titolare Prof. M. Abate).

A.A. 2019/20 (in corso): n.44 ore di esercitazione per il corso di Geometria 1 a Fisica (titolare Prof. M. Salvetti).

A.A. 2019/20 (in corso): n.110 ore di lezione per il corso di Matematica a Geologia e Scienze Naturali.

2 Ricerca svolta nel periodo in esame

La mia attività di ricerca si è concentrata sullo studio di strutture combinatorie e modelli probabilistici usati nella descrizione quantitativa di fenomeni biologici. Gli argomenti trattati sono brevemente riassunti qui di seguito. Maggiori dettagli e referenze si trovano nei corrispondenti articoli pubblicati (Sezione 3).

Alberi filogenetici con sequenze miste e l'algoritmo Neighbor-Joining. Date n sequenze genetiche $g_1, ..., g_n$, tramite metodi di allineamento si può assegnare ad ogni coppia (g_i, g_j) una misura $d_{i,j}$ della distanza evolutiva tra g_i e g_j. L'algoritmo *Neighbor-Joining* (NJ) è una delle procedure computazionali più note (Saitou and Nei, Mol. Biol. Evol. 4: 406-425, 1987; Gascuel and Steel, Mol. Biol. Evol. 23: 1997-2000, 2006) per ottenere un albero filogenetico che descriva le relazioni evolutive tra le sequenze (g_i), a partire dalla matrice $D = (d_{i,j})$ delle distanze tra le sequenze. Prendendo come input D, NJ modifica D in passi successivi cercando volta per volta la trasformazione della matrice D corrente che minimizza una funzione obiettivo. Alla fine della procedura, la sequenza di matrici ottenute viene letta come una sequenza di operazioni che trasforma un albero a stella con foglie $g_1, ..., g_n$ in una albero binario sullo stesso insieme di foglie. Nell'albero finale, indicato con NJ(D), i sottogruppi geneticamente simili di sequenze tendono a formare sottoalberi, mentre la lunghezza dei rami di NJ(D) riflette la distanza genetica tra i vari sottogruppi di sequenze.

Nell'articolo (3.a), si studiano alcune proprietà degli alberi NJ(D), quando tra $g_1, ..., g_n$ esiste una sequenza mista g_m ottenuta linearmente come combinazione di due sequenze sorgente g_{s_1} ed g_{s_2}. In altri termini, quando per ogni indice $q \neq m$ si ha $d_{m,q} = \alpha d_{s_1,q} + (1-\alpha) d_{s_2,q}$, per un fissato parametro $\alpha \in (0,1)$. Nell'articolo menzionato, si dimostra come in presenza di sequenze miste ci siano categorie di alberi non accessibili tramite l'algoritmo NJ (Fig. 1). Attraverso simulazioni, si misura poi la probabilità di particolari proprietà per gli alberi NJ(D) in funzione di n ed α, confrontando i risultati empirici ottenuti con calcoli esatti svolti per il caso senza sequenze miste.

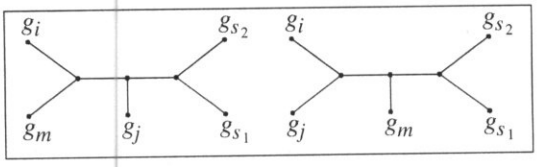

Figura 1: Alberi di taglia $n = 5$ non accessibili dall'algoritmo NJ se g_m è una sequenza mista ottenuta dalle sequenze sorgente g_{s_1} e g_{s_2}. Prese quattro arbitrarie sequenze $g_{s_1}, g_{s_2}, g_i, g_j$ ed un generico $\alpha \in (0,1)$ con $d_{m,q} = \alpha d_{s_1,q} + (1-\alpha) d_{s_2,q}$ $(q \in \{i, j, s_1, s_2\})$, l'albero NJ$(D)$ non è mai del tipo in figura.

Proprietà enumerative delle configurazioni di alberi genetici in alberi di specie. Prendiamo n sequenze genetiche $g_1, ..., g_n$ di individui appartenenti alle specie $s_1, ..., s_n$, dove g_i è la sequenza dell'individuo scelto per la specie s_i. Siano G e S due alberi binari

con foglie $g_1, ...g_n$ ed $s_1, ..., s_n$, rispettivamente. L'albero S è detto albero genetico, perchè riflette una possibile storia evolutiva per le sequenze genetiche (g_i). Analogamente, l'albero S è detto albero di specie in quanto corrisponde ad una possibile storia evolutiva per le specie (s_i). Un problema combinatorio di interesse in biologia è quello di studiare, al variare della taglia e della forma di G e S, la crescita del numero di configurazioni discrete tramite cui G può disporsi all'interno di S. In Fig. 2B,C, l'albero $G = t$ dato in A si dispone dentro l'albero S (quello con i rami più larghi) secondo due configurazioni differenti, C_1 e C_2. Due configurazioni di G in S sono differenti quando uno stesso nodo di G si presenta in due rami diversi di S.

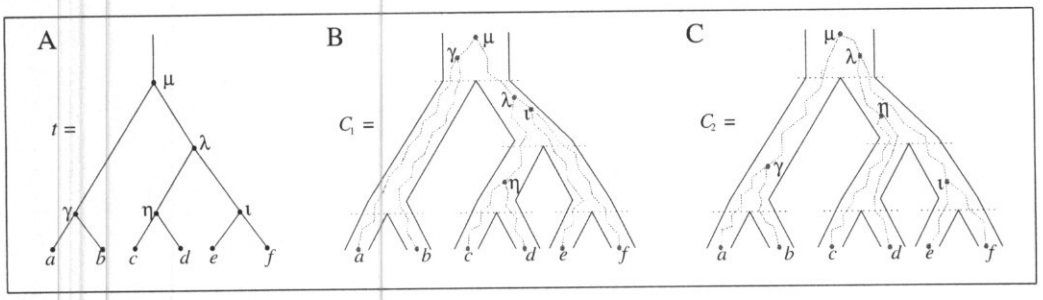

Figura 2: Diverse configurazioni per un albero genetico all'interno di un albero di specie. **(A)** Un albero genetico $G = t$ su $n = 6$ sequenze $(g_1, g_2, g_3, g_4, g_5, g_6) = (a, b, c, d, e, f)$. Nodi interni sono identificati da lettere greche. **(B,C)** Due differenti configurazioni dell'albero genetico di (A) in un albero di specie S (albero con rami più larghi). Due configurazioni di G in S sono differenti quando uno stesso nodo di G si presenta in due rami diversi di S. Nella configurazione C_1, il nodo interno λ di G si trova inserito nel ramo destro sotto la radice di S. In C_2, il nodo λ di G è inserito nel ramo sopra la radice di S. In C_1 e C_2 tutti i nodi interni di G diversi dalla radice si dispongono in rami diversi di S.

Per uno stesso albero di specie S, due diversi alberi genetici G_a e G_b possono avere un numero diverso di configurazioni in S. Un numero maggiore di configurazioni per G_a tende ad identificare in G_a uno scenario evolutivo per le sequenze (g_i) più probabile rispetto a quello descritto da G_b. Inoltre, il numero di configurazioni di un albero genetico in un albero di specie è un parametro importante nello studio della complessità di alcuni algoritmi filogenetici (Degnan and Salter, Evol. 59: 24-37, 2005 ; Degnan et al., Math. Biosci. 235: 45-55, 2012 ; Wu, Evol. 66: 763-775, 2012).

Diversi tipi di strutture combinatorie sono state introdotte per codificare le diverse configurazioni che un albero genetico G può assumere in un albero di specie S. Tra tali strutture troviamo le *Coalescent Histories* (CH), le *Compact Coalescent Histories* (CCH), le *Ancestral Configurations* (AC), e le *non-equivalent Ancestral Configurations* (neAC). Per una data configurazione di G in S, una CH è una funzione che associa i nodi interni di G ai rami di S, dove l'immagine di un nodo di G identifica il ramo di S in cui il nodo è inserito secondo la configurazione considerata. Ad esempio, la CH corrispondente alla configurazione di Fig. 2B manda il nodo γ di G nel ramo radice di S. Presa una configurazione di G in S ed un dato nodo di S, una AC corrisponde invece all'insieme dei nodi di G che si trovano sotto il nodo di S, secondo la configurazione in esame. Ad esempio, per la configurazione di Fig. 2B ed il nodo radice di S, la AC associata è l'insieme di nodi $\{a, b, \lambda, \iota, \eta, c, d, e, f\}$. Le strutture CCH e neAC sono classi di equivalenza delle CH e delle AC, rispettivamente.

Negli articoli (3.b-d), si studia il numero di AC, neAC, e CCH per alberi genetici G ed alberi di specie S che sono isomorfi a meno della lunghezza dei loro rami, $G \simeq S \simeq t$. In particolare, si studia il numero di strutture combinatorie per differenti famiglie di alberi t di taglia crescente, si caratterizzano gli alberi t che, per una fissata taglia, possiedono il massimo e minimo numero di strutture, e si determina la distribuzione asintotica del numero di strutture per un albero random t considerato sotto differenti distribuzioni di probabilità. A riguardo, nel manoscritto (3.d) si dimostra che quando t viene scelto con probabilità uniforme tra gli alberi di n foglie, la distribuzione del numero di AC alla radice di t è asintoticamente lognormale. Lo stesso tipo di distribuzione asintotica si realizza quando l'albero t viene considerato nel modello di Yule (Philos.Trans. Roy. Soc. Lond. Ser. B 213: 21-87, 1924).

Cammini nel piano discreto e variabilità del numero di Coalescent Histories. Un ulteriore problema combinatorio considerato riguarda lo studio delle altezze per alcune classi di cammini nel piano discreto. Un cammino di Dyck di semi-lunghezza n è un cammino che, partendo dall'origine $(0, 0)$ del piano, raggiunge il punto di coordinate $(2n, 0)$ compiendo passi di tipo $u \equiv (+1, +1)$ e $d \equiv (+1, -1)$, senza mai raggiungere ordinate negative. Ad esempio, nel cammino di Dyck in Fig. 3 la sequenza di passi è data da *uduudd*. Ad ogni cammino di Dyck C di semi-lunghezza n si può assegnare un peso che dipende dall'altezza dei suoi passi. Fissato un intero $w \geq 0$, il peso $p_w(C)$ di C è dato dal prodotto $p_w(C) = \prod_{i=1}^{n} y_i^w$, dove y_i è l'ordinata finale dell' i-esimo passo u di C. Ad esempio, se C è il cammino di Fig. 3, allora $p_0(C) = 1, p_1(C) = 2$, e $p_2(C) = 4$.

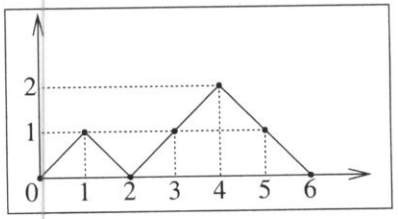

Figura 3: Un cammino di Dyck di semi-lunghezza $n = 3$. Se y_i denota l'ordinata finale dell'i-esimo passo u, allora $(y_1, y_2, y_3) = (1, 1, 2)$.

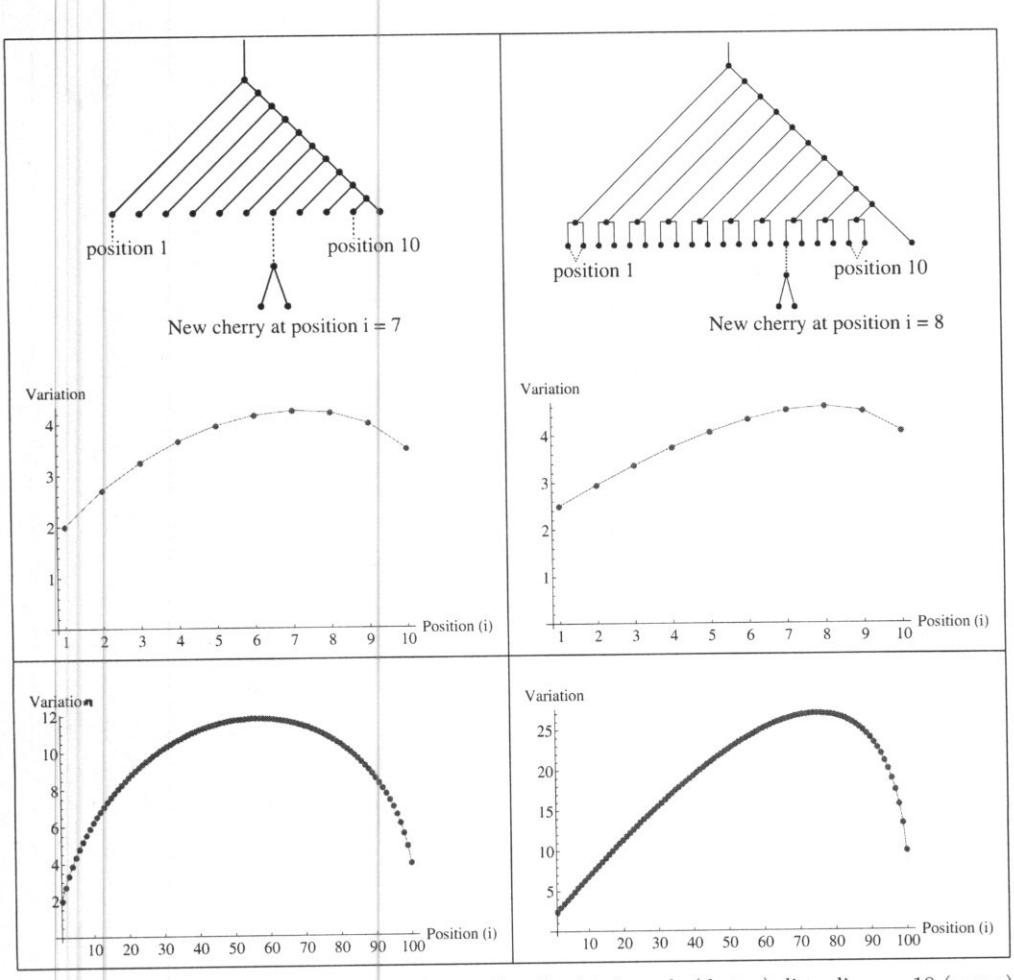

Figura 4: Variazione del numero di CH per alberi caterpillar (sinistra) e alberi lodgepole (destra) di taglia $n = 10$ (sopra) e $n = 100$ (in basso). Un nuovo nodo interno (cherry) sostituisce una foglia in posizione $i \in [1, n]$ e l'incremento è misurato come il rapporto tra il numero di CH nell'albero risultante ed il numero di CH nell'albero caterpillar o lodgepole originale.

Nell'articolo (3.e), si studia l'altezza y_i dell'i-esimo passo u di un cammino di Dyck scelto casualmente tra quelli di semi-lunghezza n con probabilità proporzionale al suo peso p_w $(w = 0, 1, 2)$. In particolare, si identificano in funzione di n e w quei valori di i che massimizzano il valore atteso $\mathbb{E}(y_i)$ dell'altezza y_i. Ad esempio, per $w = 0$, si trova che, per n sufficientemente grande, $\mathbb{E}(y_i)$ è massimo per $i \approx n/2$, mentre, quando $w = 1$, $\mathbb{E}(y_i)$ è massimo per $i \approx 3n/4$. Come mostrato nell'articolo, una corrsipondenza tra cammini di Dyck e coalescent histories permette di applicare i risultati precedenti allo studio della variabilità del numero di CH per particolari famiglie di alberi genetici e di specie isomorfi, $G \simeq S \simeq t$. Quando, ad esempio, t è un albero "caterpillar" con $n + 1$ foglie (Fig. 4, sinistra), le CH che codificano le configurazioni di G in S corrispondono a cammini di Dyck di semi-lunghezza n considerati con peso p_0. Aggiungendo un nuovo nodo interno al posto dell'i-esima foglia di t, nell'articolo (3.e) si mostra che l'incremento del numero di CH nel nuovo albero $t^{(i)}$ rispetto all'originale t è maggiore quando i massimizza il valore atteso $\mathbb{E}(y_i)$ per i cammini corrispondenti. Per n sufficientemente grande, la variazione maggiore si trova quindi nell'albero $t^{(i)}$, con $i \approx n/2$ (Fig. 4, sinistra in basso). Quando invece t è un albero "lodgepole" con $2n + 1$ foglie (Fig. 4, destra), le CH di t corrispondono a cammini di Dyck di semi-lunghezza n presi con peso p_1. Aggiungendo un nuovo nodo interno al posto di una delle due foglie di t in posizione i, l'incremento del numero di CH nel nuovo albero $t^{(i)}$ rispetto all'originale t è maggiore quando $i \approx 3n/4$, cioè quando, tra i cammini di semi-lunghezza n e peso p_1, il valore atteso $\mathbb{E}(y_i)$ è più grande (Fig. 4, destra in basso).

Alberi filogenetici con massima probabilità. Nell'articolo (3.f), si risolve in senso affermativo una congettura riguardante la struttura di alberi filogenetici di massima probabilità nel modello "multispecies colescent". Qui di seguito si riassumono brevemente alcune idee di base.

Un albero *ordinato* è un albero binario con radice in cui foglie diverse hanno etichette diverse ed i nodi interni sono ordinati linearmente in modo che ogni cammino che dalla radice dell'albero raggiunge una foglia incontri nodi interni in senso crescente (Fig. 5A). Un albero di *specie* è un albero binario con radice e foglie etichettate in cui i rami hanno una lunghezza misurata in opportune unità. I rami di un albero di specie corrispondono a popolazioni differenti e la struttura dell'albero rappresenta le relazioni di discendenza tra tali popolazioni. Un albero *genetico* è un albero ordinato realizzato internamente ad un albero di specie. Come mostrato in Fig. 5, una realizzazione dell'albero genetico G dato in A all'interno dell'albero di specie S dato in B deve rispettare

l'ordine dei nodi interni: nodi di G con etichetta crescente si dispongono dall'alto verso il basso all'interno di S determinando un ordine temporale tra le biforcazioni di G. Mentre un albero di specie rappresenta le relazioni di discendenza tra popolazoni diverse, un albero genetico interno ad un albero di specie mostra in che modo geni campionati da individui appartenenti alle popolazioni più recenti si sono evoluti nel tempo dentro l'albero di specie.

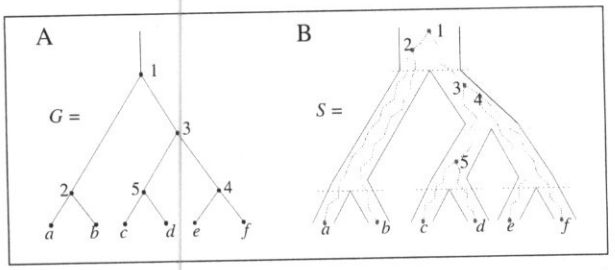

Figura 5: Un albero genetico ed un albero di specie che, a meno dell'ordinamento dei nodi interni, hanno la stessa struttura. **(A)** Un albero ordinato. **(B)** Una possibile realizzazione dell'albero ordinato dato in A all'interno di un albero di specie con la stessa struttura non ordinata.

Per un dato albero di specie S con n foglie, il modello "multispecies colescent" (Hudson, Evol. 37: 203-217, 1983 ; Degnan and Rosenberg, Trends Ecol. Evol. 24: 332-340, 2009) permette di definire e calcolare una distribuzione di probabilità sull' insieme degli alberi genetici della stessa taglia di S. In particolare, tale modello stocastico è impiegato per stimare alberi di specie a partire da alberi genetici costruiti prendendo in esame sequenze di DNA appartenenti ad individui selezionati dalle popolazioni in oggetto. In pratica, procedure computazionali—tipo l'algoritmo Neighbor-Joining menzionato sopra—creano una collezione di alberi genetici considerando molteplici aree genetiche di interesse, e successivamente si cercano quegli alberi di specie che massimizzano la probabilità degli alberi genetici osservati.

Usando un approccio esaustivo, Degnan et al. (Math. Biosci. 235: 45-55, 2012; IEEE/ACM Trans. Comput. Biol. Bioinf. 9: 1558-1568, 2012) avevano mostrato che, per un fissato albero di specie con al più 5 foglie, l'albero genetico più probabile ha la stessa struttura dell'albero di specie a meno dell'ordinamento dei nodi interni (Fig. 5 B). Gli stessi autori avevano dunque posto il problema di verificare o smentire tale proprietà per alberi di specie di taglia arbitraria. Nell'articolo (3.f) si fornisce una dimostrazione combinatoria di questo fatto, la cui veridicità ha interessanti conseguenze applicative. In particolare, il risultato ci indica come la struttura non ordinata dell'albero di specie possa essere predetta guardando la struttura non ordinata degli alberi genetici ordinati che appaiono dai dati con frequenza maggiore. Invece, come mostrato da Degnan and Rosenberg (Plos Genet. 2: 762-768, 2006), guardando alla struttura degli alberi genetici non ordinati più frequenti, si perde la corrispondenza con la struttura non ordinata dell'albero di specie.

Lunghezza dei rami esterni nel modello Kingman's coalescent. Il modello "Kingman's coalescent" (Kingman, Stoch. Proc. Appl. 13: 235-248, 1982) simula l'evoluzione in condizioni standard—in assenza di selezione naturale o cali/espansioni demografiche—di linee genetiche appartenenti ad una stessa popolazione. Il risultato di tale processo stocastico è rappresentabile come un albero binario con rami di lunghezza misurata in opportune unità di tempo. In particolare, la struttura discreta di un albero coalescent è determinata da un albero binario con nodi interni ordinati linearmente selezionato secondo la distribuzione di Yule (Philos.Trans. Roy. Soc. Lond. Ser. B 213: 21-87, 1924), mentre la lunghezza temporale dell'intervallo τ_i in cui coesistono i rami dell'albero è una variabile esponenziale di media $1/(i(i-1))$ (Fig. 6).

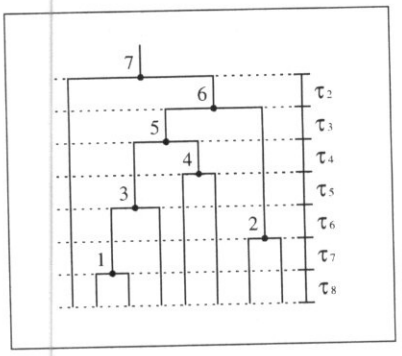

Figura 6: Un albero binario di taglia $n = 8$ con nodi interni ordinati linearmente (dal basso verso l'alto) e con lunghezza τ_i per l'intervallo temporale in cui esattamente $i \in [2, n]$ rami dell'albero coesistono. Il numero di intervalli temporali intersecato da un ramo esterno coincide con il ranking del nodo interno da cui discende.

La lunghezza di un ramo di un albero coalescent è la somma delle lunghezze assunte dagli intervalli temporali intersecati dal ramo in questione. In particolare, se il ramo considerato è esterno, cioè termina con una foglia, la densità di probabilità della sua lunghezza temporale si può ricavare dal numero di intervalli intersecati dal ramo, che corrisponde al ranking del nodo interno da cui discende

(Fig. 6). Su ogni ramo di un albero coalescent si accumula un numero di mutazioni che, secondo il modello, è una variabile di Poisson di media determinata dalla lunghezza del ramo in questione. Disponendosi su un ramo dell'albero, una mutazione modifica il corredo genetico dei soli individui sottostanti il ramo considerato. In Fig. 6, una mutazione disposta sul ramo che connette i nodi interni etichettati con 6 e 7 non differenzierà il genoma dell'individuo corrispondente alla foglia all'estrema sinistra dal genotipo originale, ma soltanto il resto della popolazione individuata dalle foglie che stanno sotto al nodo 6. Dal punto di vista delle applicazioni biologiche, dunque, una delle proprietà più importanti del Kingman's coalescent è quella di poter modelizzare la variabilità genetica di una certa poplazione in condizioni standard.

Nel manoscritto (3.g), si studia la lunghezza dei rami esterni di un albero coalescent, cioè di quei rami lungo i quali si distribuiscono mutazioni che coinvolgono singoli individui. In particolare, si usano tecniche e risultati di matematica discreta per analizzare il ranking dei nodi da cui discendono i rami esterni degli alberi binari ordinati sottostanti agli alberi coalescent. I risultati teorici ottenuti in merito alla lunghezza dei rami esterni più lunghi di un albero coalescent sono poi confrontati con dati provenienti da genoma umano di varie popolazioni (`www.1000genomes.org`).

3 Pubblicazioni ed articoli in preparazione a cui ho lavorato nel periodo in esame

(3.a) Mathematical and Simulation-Based Analysis of the behavior of admixed taxa in the Neighbor-Joining algorithm, *Bulletin of Mathematical Biology, special issue: Algebraic Methods in Phylogenetics*, 81: 452-493, 2019. [Con J. Kim, N.A. Rosenberg (Stanford University) e N.M. Kopelman (Holon Institute of Technology)]

(3.b) On the number of non-equivalent ancestral configurations for matching gene trees and species trees, *Bulletin of Mathematical Biology, special issue: Algebraic Methods in Phylogenetics*, 81: 384-407, 2019. [Con N.A. Rosenberg (Stanford University)]

(3.c) Enumeration of compact coalescent histories for matching gene trees and species trees, *Journal of Mathematical Biology*, 78: 155-188, 2019. [Con N.A. Rosenberg (Stanford University)]

(3.d) The distributions under two species-tree models of the number of root ancestral configurations for matching gene trees and species trees, manoscritto in preparazione disponibile su richiesta [Con M. Fuchs, (National Chiao Tung University, Taiwan) e N.A. Rosenberg (Stanford University)]

(3.e) Local height in weighted Dyck models of random walks and the variability of the number of coalescent histories for caterpillar-shaped gene trees and species trees, *Springer Nature Applied Sciences*, 1: article 578, 2019. [Con E. Munarini (Politecnico di Milano)]

(3.f) On the unranked topology of maximally probable ranked gene tree topologies, *Journal of Mathematical Biology*, 79: 1205-1225, 2019. [Con P. Miglionico e G. Narduzzi (studenti della Scuola Normale Superiore, Pisa)]

(3.g) A discrete approach to the external branches of a Kingman coalescent tree. Theoretical results and practical applications, manoscritto sottomesso a *Theoretical Population Biology*, preprint: bioRxiv/2019/818088. [Con T. Wiehe (Institut für Genetik, Köln)]

4 Software prodotto nel periodo in esame

`RGTProb` (`https://github.com/PasqM/RGTProb`), software che accompagna l'articolo (3.f) per il calcolo numerico e simbolico nel modello "multispecies coalescent" della probabilità di un albero genetico ordinato per un dato albero di specie.

5 Conferenze svolte nel periodo in esame

"Catalan Hypercubes", Hypergraphs, Graphs and Designs 2017, Messina 21-24/06/2017.

"Enumerative properties of gene tree configurations in matching species trees", 2*nd* International Workshop on Enumeration Problems & Applications 2018, Pisa 5-8/11/2018.

"Existence of maximally probable ranked gene tree topologies with a matching unranked topology", SIAM conference on Applied Algebraic Geometry 2019, minisymposium on Combinatorial and Algebraic Phylogenetics, Bern 9-13/07/2019

Pisa, 28 Ottobre 2019

Allegato n. 6
Pag. n. 5
Verbale del 4-11-2019