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 21 novembre 2017

Omissis

8. Ricerca

8.2. Relazione I anno di attività di ricerca RTD-B, dott. Filippo Disanto: approvazione

Il Consiglio,
VISTA: la L. 9 maggio 1989 n. 168 "Istituzione del Ministero dell'università e della ricerca scientifica e tecnologica" e in particolare il TITOLO II - Autonomia delle università e degli enti di ricerca, l'art. 6 "Autonomia delle Università" commi 1 e 2, l'art. 7 "Autonomia finanziaria e contabile delle Università" e l'art. 16 "Università";
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";
VISTO: lo Statuto dell'Università di Pisa approvato con D.R. 27 febbraio 2012 n. 2711;
VISTO: il Regolamento per l'assunzione di ricercatori a tempo determinato, emanato con D.R. prot. n. 8444 del 29 giugno 201, e successive modifiche e integrazioni;
VISTO: il 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;
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 presentata dal Dott. Disanto **(all 9);**

DELIBERA

Il Dipartimento esprime un giudizio pienamente positivo sull'attività didattica e di ricerca svolta dal dott. Filippo Disanto.

La presente delibera, contrassegnata dal **numero 65,** è approvata all'unanimità ed è immediatamente esecutiva.

Il Segretario
Dott.ssa Cristina Lossi

Il Presidente
Prof. Carlo Petronio

Relazione attività RTDb di Filippo Disanto

Ho preso servizio come RTDb presso il Dipartimento di Matematica il 15 Febbraio 2017. Le attività da me svolte sino ad oggi (15 Novembre 2017) sono elencate qui di seguito.

(1) Didattica svolta nel periodo in esame

- Corso di Geometria per Fisica (con Mario Salvetti).
- Corso di Matematica per Scienze Naturali e Geologia (con Marco Abate).

(2) Ricerca svolta nel periodo in esame

La mia attività di ricerca si è concentrata sullo studio di strutture combinatorie usate nella descrizione quantitativa di fenomeni biologici. Gli argomenti trattati sono brevemente riassunti qui di seguito. Maggiori dettagli e referenze si trovano nei manoscritti allegati.

- **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 usate 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.d**], 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.

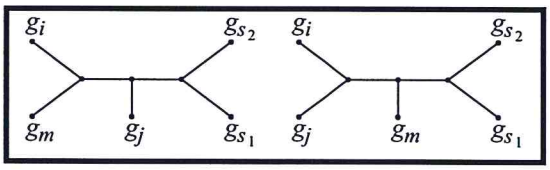

Figura 1: Alberi di taglia $n = 5$ non accessibili dall'algoritmo NJ se g_m è una sequenza mista ottenuta dalle sequenze sorgente g_{s_1} e g_{s_2}. Prese quattro arbitrarie sequenze $g_{s_1}, g_{s_2}, g_i, g_j$ ed un generico $\alpha \in (0,1)$ con $d_{m,q} = \alpha d_{s_1,q} + (1-\alpha)d_{s_2,q}$ $(q \in \{i, j, s_1, s_2\})$, 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 T_g e T_s due alberi binari con foglie $g_1, ...g_n$ ed $s_1, ..., s_n$, rispettivamente. L'albero T_g è detto albero genetico, perchè riflette una possibile storia evolutiva per le sequenze genetiche (g_i). Analogamente, l'albero T_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 T_g e T_s, la crescita del numero di configurazioni discrete tramite cui T_g può disporsi all'interno di T_s. In Fig. 2B,C, l'albero $T_g = t$ dato in A si dispone dentro l'albero T_s (quello con i rami più larghi) secondo due configurazioni differenti, C_1 e C_2. Due configurazioni di T_g in T_s sono differenti quando uno stesso nodo di T_g si presenta in due rami diversi di T_s.

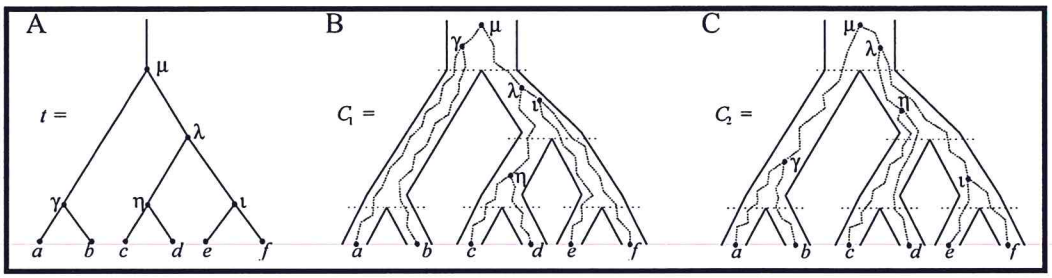

Figura 2: Diverse configurazioni per un albero genetico all'interno di un albero di specie. **(A)** Un albero genetico $T_g = t$ su $n = 6$ sequenze $(g_1, g_2, g_3, g_4, g_5, g_6) = (a, b, c, d, e, f)$. Nodi interni sono identificati da lettere greche. **(B,C)** Due differenti configurazioni dell'albero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

Per uno stesso albero di specie T_s, due diversi alberi genetici $T_g^{(a)}$ e $T_g^{(b)}$ possono avere un numero diverso di configurazioni in T_s. Un numero maggiore di configurazioni per $T_g^{(a)}$ tende ad identificare in $T_g^{(a)}$ uno scenario evolutivo per le sequenze (g_i) più probabile rispetto a quello descritto da $T_g^{(b)}$. 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.

Diversi tipi di strutture combinatorie sono state introdotte per codificare le diverse configurazioni che un albero genetico T_g può assumere in un albero di specie T_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 T_g in T_s, una CH è una funzione che associa i nodi interni di T_g ai rami di T_s, dove l'immagine di un nodo di T_g identifica il ramo di T_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 T_g nel ramo radice di T_s. Presa una configurazione di T_g in T_s ed un dato nodo di T_s, una AC corrisponde invece all'insieme dei nodi di T_g che si trovano sotto il nodo di T_s, secondo la configurazione in esame. Ad esempio, per la configurazione di Fig. 2B ed il nodo radice di T_s, la AC associata è l'insieme di nodi $\{a, b, \lambda, \iota, \eta, c, d, e, f\}$ di T_g. Le strutture CCH e neAC sono classi di equivalenza delle CH e delle AC, rispettivamente.

Negli articoli [**3.b**, **3.c**], si studia il numero di AC, neAC, e CCH per alberi genetici T_g ed alberi di specie T_s che sono isomorfi a meno della lunghezza dei loro rami, $T_g \simeq T_s \simeq t$. 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, si determina la crescita asintotica del numero medio di strutture per un albero random t considerato sotto differenti distribuzioni di probabilità.

- **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 $p_2(C) = 4$.

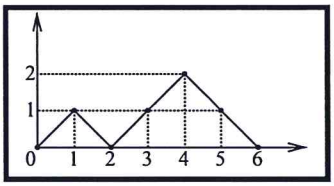

Figura 3: Un cammino di Dyck di semi-lunghezza $n = 3$. Se h_i denota l'ordinata finale dell'i-esìmo passo u, allora $(h_1, h_2, h_3) = (1, 1, 2)$.

Nel manoscritto [**3.a**], 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 descritto in [**3.a**], 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, $T_g \simeq T_s \simeq t$. Quando, ad esempio, t è un albero "caterpillar" con $n + 1$ foglie (Fig. 4, sinistra), le CH che codificano le configurazioni di T_g

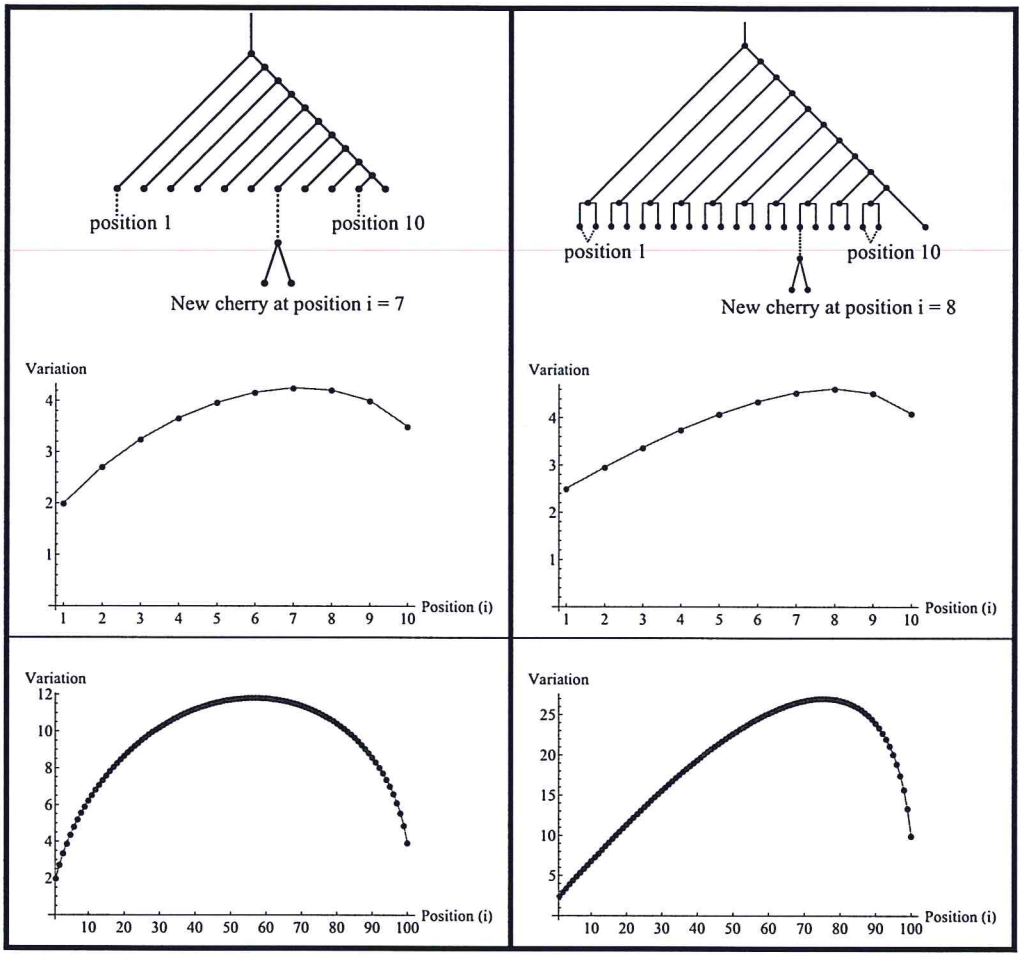

Figura 4: Variazione del numero di CH per alberi caterpillar (sinistra) e alberi lodgepole (destra) di taglia $n = 10$ (sopra) e $n = 100$ (in basso). Un nuovo nodo interno (cherry) sostituisce una foglia in posizione $i \in [1, n]$ e l'incremento è misurato come il rapporto tra il numero di CH nell'albero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

in T_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, nel manoscritto [**3.a**] 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).

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

- F. Disanto. Estimates of the heights for three types of weighted Dyck paths, in preparazione. [Allegato **3.a**]
- F. Disanto, N.A. Rosenberg (2017). On the number of non-equivalent ancestral configurations for matching gene trees and species trees, in stampa in *Bulletin of Mathematical Biology, Special issue: Algebraic Methods in Phylogenetics.* [Allegato **3.b**]
- F. Disanto, N.A. Rosenberg. Enumeration of compact coalescent histories for matching gene trees and species trees, sottomesso a *Journal of Mathematical Biology.* [Allegato **3.c**]
- J. Kim, F. Disanto, N.M. Kopelman, N.A. Rosenberg. An extended analysis of the behavior of admixed taxa in the neighbor-joining algorithm, sottomesso a *Bulletin of Mathematical Biology.* [Allegato **3.d**]

(4) Conferenze svolte nel periodo in esame

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