Bollettino
Unione Matematica Italiana

Sezione A – La Matematica nella Società e nella Cultura

Raffaele Giancarlo, Sabrina Mantaci

Contributi delle Scienze Matematiche ed Informatiche al sequenziamento genomico su larga scala

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **4-A**—*La Matematica nella Società e nella Cultura* (2001), n.1, p. 33–62.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=BUMI_2001_8_4A_1_33_0>

Bollettino dell'Unione Matematica Italiana, Unione Matematica Italiana, 2001.

Bollettino U. M. I.
La Matematica nella Società e nella Cultura
Serie VIII, Vol. IV-A, Aprile 2001, 33-62

Contributi delle Scienze Matematiche ed Informatiche al sequenziamento genomico su larga scala (*).

RAFFAELE GIANCARLO - SABRINA MANTACI

Abstract. – *Nel panorama della scienza contemporanea, la biologia molecolare ha recentemente assunto un ruolo di fondamentale importanza. Il bisogno crescente di conoscere intere sequenze genomiche e l'esigenza, ancora più pressante, di analizzare e confrontare tali sequenze per poter dedurre funzionalità e discendenze comuni, ha reso necessaria l'integrazione delle usuali tecniche sperimentali, proprie della ricerca biologica, con le metodologie formali della matematica e dell'informatica. Queste motivazioni hanno stimolato la nascita e lo sviluppo di un particolare settore di ricerca matematica, la* biologia computazionale, *che ha l'obbiettivo di sviluppare opportuni metodi e strumenti per problemi computazionali derivanti da questioni poste dalla ricerca genomica. Senza pretesa di essere esaustivi, questo articolo presenta alcuni contributi della ricerca matematica ed informatica al sequenziamento genomico, cioè il processo di ottenere la stringa corrispondente ad un genoma complesso, a partire dalla sua versione biochimica. L'impiego, a diversi livelli, delle idee e tecniche qui descritte ha avuto come risultato fondamentale quella che può essere considerata una delle più importanti conquiste della scienza moderna: una prima versione della stringa di DNA corrispondente al genoma umano.*

1. – Introduzione.

Nel febbraio di quest'anno, le autorevolissime riviste *Nature* e *Science* hanno consegnato alla letteratura scientifica un evento, ab-

(*) La ricerca degli autori è parzialmente sostenuta dal Progetto MURST di Rilevanza Nazionale «Bioinformatica e Ricerca Genomica».

bondantentemente anticipato dai media, di portata storica, sia per i suoi risvolti tecnici e che per le sue implicazioni future: le due versioni iniziali della sequenza corrispondente al genoma umano, ottenute in maniera indipendente dal consorzio internazionale Human Genome Project [6] e dalla compagnia privata Celera Genomics [24].

Il termine *genoma* è stato coniato dal botanico tedesco H. Winkler nel 1920 [28] per definire l'insieme dei cromosomi di una cellula aploide di organismo eucariota. In un articolo del 1995, G. Bernardi [4] osserva che tale ormai popolarissimo termine viene il più delle volte definito (quando lo si fa) solo in maniera operativa come la somma totale dei geni e delle sequenze intergeniche di una cellula aploide. Bernardi mette in evidenza, tuttavia, che un genoma è molto più che la somma delle sue parti, poiché ci sono interazioni di tipo evolutivo, funzionale e strutturale tra diverse regioni del genoma, la cui conoscenza è parte essenziale della sua caratterizzazione.

Con questa breve divagazione abbiamo voluto sottolineare che la conoscenza della sequenza corrispondente al genoma umano è solo un primo passo fondamentale verso una sua caratterizzazione *statica*, che comprende anche l'identificazione di geni e di regioni non codificanti. Tuttavia, siamo ancora lontani dalla reale comprensione di quali siano, all'interno del DNA, i meccanismi che regolano la vita. La grandissima sfida che oggi si pone per le scienze biomediche, biologiche e, per molti versi, le scienze matematiche è quella di cercare una caratterizzazione *dinamica* del genoma che permetta di descrivere funzionalità e relazioni tra le sue parti. La conoscenza della sequenza del genoma umano permette di affrontare tale sfida avendo a disposizione tutta l'informazione di base.

È noto che l'informazione genetica è immagazzinata in una cellula per mezzo di molecole nucleiche, le quali possono essere pensate come stringhe di elementi più piccoli chiamati nucleotidi. Molti dei più importanti problemi della moderna biologia molecolare (compresa la caratterizzazione sia statica che dinamica di un genoma) corrispondono a questioni di tipo combinatorico, statistico ed algoritmico su queste stringhe. Parafrasando il titolo di un libro edito da Lander e Waterman [14], le discipline matematiche

che affrontano tali questioni contribuiscono alla costruzione della teoria del «Calcolo del Segreto della Vita».

Un importante capitolo in tale teoria è costituito dal problema dell'analisi e del confronto di sequenze biologiche. Proprio in tale ambito, troviamo molti degli argomenti che per primi hanno suggerito una possibile interazione fra la biologia molecolare e la matematica. Fonti autorevoli (vedi [11], [22]) attribuiscono a Stanislaw Ulam una serie di formalizzazioni matematiche rivelatesi fondamentali, sia come contributo tecnico ma ancor più come indicazione della strada da seguire per lo sviluppo di un nuovo tipo di ricerca matematica che avesse importanti riscontri nel campo di una scienza tradizionalmente considerata di natura sperimentale. All'inizio degli anni Settanta, la disponibilità di (allora poche) sequenze biologiche aveva già reso evidente che il comportamento di una cellula era codificato nel DNA e che, per comprenderlo, occorreva stabilire relazioni tra sequenze lunghe centinaia o addirittura migliaia di basi. Ulam formalizzò questo problema definendo una nozione di *distanza* tra sequenze biologiche. Tale nozione si basa su tre tipi di operazioni che si possono definire sulle sequenze: sostituzione di un carattere con un altro; inserzione di un carattere; cancellazione di un carattere. Date due sequenze, la loro distanza è definita come il numero minimo di operazioni dei tre tipi indicati che occorre applicare sulla prima sequenza per ottenere la seconda. L'intuizione, oggi comunemente accettata come principio guida nell'analisi di sequenze biologiche, è che due sequenze molto vicine hanno (con grande probabilità) storie evolutive comuni e funzioni analoghe. Tale nozione di distanza, riformulata in seguito da altri ricercatori, può essere considerata per molti versi come il primo strumento fondamentale fornito dalla matematica e dall'informatica agli studi sulle relazioni tra parti diverse dello stesso genoma o di genomi diversi. Negli ultimi trent'anni, man mano che la disponibilità di sequenze sempre più lunghe rendeva sempre più articolati i problemi sulle relazioni fra diverse sequenze di DNA, i matematici e gli informatici sono stati sollecitati a produrre risultati scientifici sempre più specifici e algoritmi sempre più efficienti, mirati a coadiuvare i biologi nelle ricerche sulla struttura del genoma. Vogliamo sottolineare che i matematici e gli

informatici che si sono impegnati in questo tipo di ricerche, non si sono limitati ad una mera applicazione di risultati già noti, ma hanno sviluppato nuove metodologie e prodotto risultati originali che sono ormai talmente numerosi e profondi dal punto di vista teorico ed applicativo, da generare delle discipline scientifiche autonome: la Biologia Computazionale e la Bioinformatica.

In questo contesto, il Progetto Genoma [6], istituito nel 1989, ha avuto una grandissima forza propulsiva nel creare un'interazione più stretta tra biologia molecolare, matematica ed informatica. Infatti, l'ampia agenda dello Human Genome Project, che aveva come scopo finale quello di produrre l'intera sequenza del genoma umano, comprendeva fra le tappe intermedie la creazione di mappe fisiche, ossia di schemi che permettessero di conoscere la posizione in cui certi demarcatori appaiono nel DNA. Questa informazione permetteva infatti di stabilire in maniera molto precisa la posizione di certi frammenti di DNA nella molecola di provenienza, contribuendo alla ricomposizione, su una scala più ampia, delle informazioni contenute in questi piccoli frammenti.

Si richiedeva quindi lo sviluppo di metodi, tecnologie e professionalità di supporto per il raggiungimento di tali obbiettivi. Queste esigenze hanno avuto come risultato l'identificazione e formalizzazione di una serie di nuovi problemi di tipo matematico ed algoritmico, le cui soluzioni hanno giocato un ruolo talmente importante da essere stati messi in primo piano nei resoconti di Science e Nature sul genoma umano. Al matematico alcuni di questi problemi potranno sembrare a prima vista poco più che dei «puzzle», ma basta uno sguardo più attento per scoprire che la soluzione di tali puzzle è equivalente ad affrontare e risolvere uno dei sette più impegnativi problemi matematici aperti del Millennio [31]. Inoltre, la necessità che queste soluzioni vengano trasformate in algoritmi efficienti, in grado di gestire gigantesche quantità di dati, e che producano risultati soddisfacenti per i biologi, rende piuttosto delicato il processo di formalizzazione matematica del problema biologico.

In questo articolo vogliamo dare un'idea generale delle problematiche legate al sequenziamento genomico, facendo particolare riferimento a quei problemi in cui la ricerca matematica ha svolto un

ruolo determinante. Quindi, nel panorama brevemente delineato fino ad ora, si tratta di un'area estremamente specifica e di origine piuttosto recente.

L'articolo è organizzato come segue. Cominceremo col dare una definizione generale del problema del sequenziamento e faremo qualche cenno riguardo alle problematiche relative alla decodifica di lunghe sequenze di DNA. Daremo quindi un'idea generale di quali sono i protocolli di laboratorio che vengono utilizzati per il sequenziamento e dei problemi matematici legati all'*assembly* — il problema di ricostruire una stringa di DNA a partire da alcuni frammenti in sovrapposizione — evidenziando le difficoltà computazionali legate a certe formalizzazioni. In seguito illustreremo le due principali strategie, top-down e bottom-up, per il sequenziamento del genoma umano, applicate rispettivamente da parte dell'Human Genome Project e della Celera Genomics. Faremo inoltre qualche breve cenno riguardo ai contributi del calcolo delle probabilità e della statistica al problema del campionamento delle librerie dei cloni e dei frammenti di DNA. Descriveremo un particolare problema combinatorio fra quelli derivanti dalla costruzione di mappe fisiche e metteremo in evidenza alcuni problemi relativi alla possibilità della presenza di errori di laboratorio nei dati. Infine, descriveremo con maggiore dettaglio alcuni dei contributi dell'algoritmica e della combinatoria al problema del fragment assembly.

2. – L'Utilità del sequenziamento su larga scala.

Il grande clamore che si è sviluppato intorno alle recenti scoperte sulla struttura del DNA, ha reso ormai di dominio pubblico il ruolo fondamentale che tali scoperte potrebbero avere nel campo della ricerca biomedica. La conoscenza della sequenza corrispondente al genoma e certe altre proprietà come, ad esempio, la stabilità funzionale del genoma umano rispetto a possibili mutazioni genetiche, permetterebbero di conoscere in anticipo non solo le malattie genetiche ma anche la predisposizione genetica ad alcune patologie per certi soggetti. Se ne avrebbe come conseguenza

la possibilità di studiare delle cure *ad hoc* per ogni soggetto riguardo ad ognuna delle patologie.

Come abbiamo precedentemente accennato, il genoma di un organismo è un'entità dinamica, nel senso che svolge il suo compito di sintetizzare proteine grazie a certe interazioni fra le sue parti. Tali interazioni non avvengono necessariamente localmente, ma possono coinvolgere parti del genoma fisicamente molto distanti. Un'analisi approfondita delle funzioni del genoma deve basarsi su una visione globale della sua struttura. L'analisi di un genoma è essenziale per acquisire informazioni sulla storia evolutiva che ha portato nel tempo a certi mutamenti genetici e riguardo ai meccanismi che accomunano i diversi organismi viventi. La disponibilità delle sequenze corrispondenti a genomi di diverse specie (fra cui quello dell'uomo) e un'analisi comparativa di queste sequenze, hanno reso possibile indagini più approfondite a questo riguardo. Ad esempio, un'importante osservazione è che la lunghezza di un genoma in termini di basi non è indicativo della complessità biologica di un organismo: a prova di ciò si può osservare per esempio che il genoma umano è circa 200 volte più piccolo di quello dell'*amoeba dubia*. La complessità di un organismo sembra piuttosto dipendere dalla quantità di geni contenuti nel suo DNA o addirittura il numero di proteine che tali geni riescono a sintetizzare.

Viene naturale chiedersi «quanto» il genoma umano sia più complesso rispetto a quello di altri organismi. A questo proposito diamo alcuni dati comparativi fra il genoma umano e quello della *drosophila melanogaster* (il moscerino della frutta). Il genoma umano è lungo circa tre miliardi di basi mentre quello della drosophila è di circa 120 milioni di basi. Noi abbiamo un numero di geni stimato tra i 30 ed i 40 mila mentre la drosophila ne ha circa la metà. Circa 3 mila di tali geni sono *ortologhi*, ovvero possono essere visti come discendenti da un antenato comune: in un certo senso, tali geni rappresentano il corredo genetico «minimo» che un organismo deve avere per poter sopravvivere. Le differenze più spiccate tra il nostro patrimonio genetico e quello della drosophila si osservano riguardo ai geni che svolgono le loro funzioni sul sistema nervoso, su quello immunitario e su quello legato allo sviluppo (cf. [6], [24]).

Un'altro fatto sorprendente è quello espresso nel «postulato di Ohno»: il genoma dei mammiferi consiste di oasi di geni tra deserti vuoti. La conoscenza della sequenza del genoma umano ha consentito di verificare sperimentalmente il postulato: in realtà circa il 20% del nostro genoma è fatto di «deserti», ovvero regioni povere di geni. Inoltre, sembra che meno del 5% del nostro genoma è costituito da regioni codificanti, ossia quelle che effettivamente contribuiscono a definire i geni. Inoltre, per la parte rimanente, comunemente detta *junk DNA* (DNA spazzatura), circa il 50% è costituito da regioni ripetitive. Tuttavia, a dispetto del nome loro attribuito, queste zone non codificanti sembrano avere un ruolo importante, in quanto molte informazioni relative alla nostra evoluzione sono contenute proprio in queste zone. Inoltre tali regioni sembrano avere un ruolo fondamentale nella determinazione di nuovi geni e nella modifica di vecchi.

Infine, un altro dato interessante che si è potuto osservare mediante l'analisi della struttura del genoma è che la frequenza di mutazioni nel maschio e nella femmina non è lo stesso: sembra infatti che la maggior parte delle mutazioni avvenga nei maschi, e che quindi il genoma maschile sia maggiormente responsabile dell'evoluzione genetica.

3. – La tecnologia del sequenziamento ed una prima astrazione matematica.

Una molecola di DNA è costituita da due filamenti legati in una doppia elica. Ciascuno dei due filamenti è una catena di molecole più piccole, dette *nucleotidi*. Ogni nucleotide è composto da uno zucchero, un fosfato e una base. Quattro tipi diversi di *basi* compongono il DNA: Adenina, Citosina, Guanina e Timina (denotate, nella notazione standard, con le lettere A, C, G, T). Adenina e Timina sono complementari, nel senso che tendono a legarsi chimicamente una all'altra, e lo stesso dicasi per Guanina e Citosina. I due filamenti che costituiscono il DNA sono legati uno all'altro secondo questo principio di complementarietà. Conoscere la molecola del DNA significa quindi conoscere come si succedono, una dopo l'altra, le basi in uno dei

due filamenti (l'altro filamento si ottiene come complemento del primo). Il processo di decodifica della sequenza biochimica delle basi del DNA nella corrispondente parola nell'alfabeto $\{A, C, G, T\}$ è chiamato *sequenziamento*. Si noti che, in base alla posizione relativa di alcuni atomi di carbonio, è possibile anche stabilire l'orientamento del filamento di DNA che stiamo esaminando.

Il sequenziamento è un processo di decodifica estremamente sofisticato. Già negli anni Settanta erano state messe a punto alcune tecniche di laboratorio (cf. [19, 20]) che permettevano di individuare la sequenza iniziale (le prime 300-900 basi) di un dato frammento di DNA. Purtroppo da allora non ci sono stati sostanziali miglioramenti riguardo alla lunghezza del frammento iniziale sequenziato mediante questi protocolli di laboratorio. Ci riferiremo d'ora in poi a tali protocolli come al *sequenziamento base*, in quanto costituiscono il passo iniziale di tutte le tecniche di sequenziamento note.

L'esigenza di conoscere lunghe e complesse sequenze di DNA, se non addirittura l'intero genoma di un organismo, ha posto la necessità di introdurre una nuova strategia di laboratorio, chiamata *Shotgun Sequencing* (cf. [16]), che, insieme ad opportuni metodi computazionali, ha la capacità di amplificare il potere di decodifica del sequenziamento base. Tale strategia è talmente importante da essere alla base di tutti gli approcci per il sequenziamento genomico. Inoltre, i sempre più potenti strumenti computazionali che sono stati affiancati allo Shotgun Sequencing ne hanno ulteriormente migliorato l'efficienza e la versatilità. Se, ad esempio, negli anni Ottanta i mezzi tecnici a disposizione permettevano di ottenere, mediante shotgun sequencing, la traduzione di sequenze composte da migliaia di basi, negli ultimi due anni il ricorso a nuove metodologie ha permesso di ottenere sequenze genomiche composte da centinaia di milioni (genoma della Drosophila — vedi [1, 17]) e miliardi di basi (genoma Umano — vedi [6, 24]). Sottolineamo che i problemi computazionali che nascono dall'applicazione dello Shotgun Sequencing a una sequenza lunga miliardi di basi sono ben diversi da quelli incontrati per sequenze lunghe solo migliaia di basi. Per il momento desideriamo mettere in luce il denominatore comune, evidenziando che gli

strumenti che permettono di gestire sequenze così lunghe sono basati su approcci di tipo matematico ed informatico.

3.1. *Shotgun Sequencing - Protocollo di laboratorio.*

In questo paragrafo diamo un'idea generale, omettendo i dettagli, delle tecniche di laboratorio che vengono utilizzate nello Shotgun Sequencing. Supponiamo di voler ottenere la sequenza di un segmento di DNA che chiameremo *sorgente*:

(a) La sorgente viene riprodotta in un abbondante numero di copie. Ogni copia viene suddivisa in *frammenti*. Essendo il processo di suddivisione casuale, ciascuna delle copie di DNA sarà tagliata in punti diversi, come rappresentato in Figura 1. Dalla Figura si può dedurre che tale suddivisione gode di due proprietà: (a.1) frammenti provenienti da copie diverse possono essere in sovrapposizione; (a.2) l'intera collezione di frammenti costituisce un ricoprimento della sorgente: il procedimento è tale da poter assumere che il ricoprimento sia uniforme.

(b) Ogni frammento viene inserito in un virus, chiamato *vettore*, in un punto predeterminato del suo DNA. Questa operazione ha lo scopo di mantenere questi frammenti in laboratorio, oltre che quello di riprodurli in diverse copie. Il frammento ora viene chiamato *insert* e la collezione di *inserts* è la *libreria* (vedi Figura 2).

(c) Si estrae un campione dalla libreria in maniera tale che, con alta probabilità, esso costituisca un ricoprimento uniforme della sorgente.

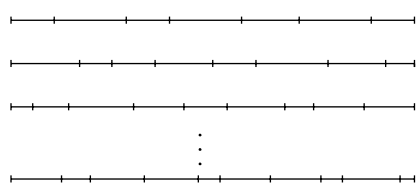

Figura 1. – Il segmento di DNA viene riprodotto in diverse copie e suddiviso in frammenti. In ognuna delle copie in Figura, i punti di taglio sono indicati da barre verticali. Si noti che in questa fase si perde l'informazione relativa alla posizione dei frammenti nella copia di provenienza.

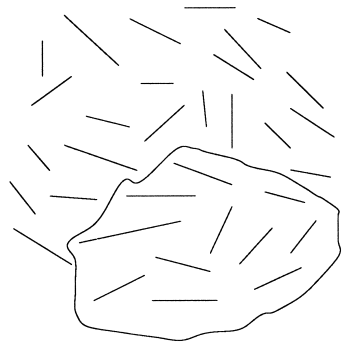

Figura 2. – I frammenti della Figura 1 formano una libreria.

(d) Un batterio viene poi infettato da un singolo vettore estratto dal campione; il batterio riproducendosi crea molte copie dell'insert.

(e) Le diverse copie di ogni singolo insert possono essere utilizzate per ricavare la sequenza iniziale del frammento corrispondente all'insert e la stringa così ottenuta viene detta *read* (vedi Figura 3). La lunghezza di una read può variare dalle 300 alle 900 basi.

Si ottiene così una collezione di stringhe sull'alfabeto $\{A, C, G, T\}$, corrispondenti ai reads del campione, a partire dalle quali vogliamo ricostruire la sorgente. Tale problema computazionale, che si presenta come un classico esempio di problema inverso, è noto come *fragment assembly* (si veda Fig. 4). Se la posizione dei frammenti nella sorgente fosse nota, il problema dell'assembly sa-

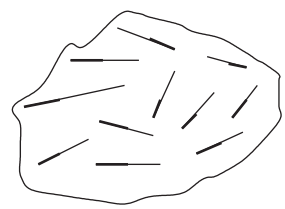

Figura 3. – Dalla libreria in Figura 2 si estrae un campione. Le parti iniziali dei frammenti nel campione vengono tradotte mediante il sequenziamento base. Per ogni frammento in figura, la parte in grassetto rappresenta la parte sequenziata.

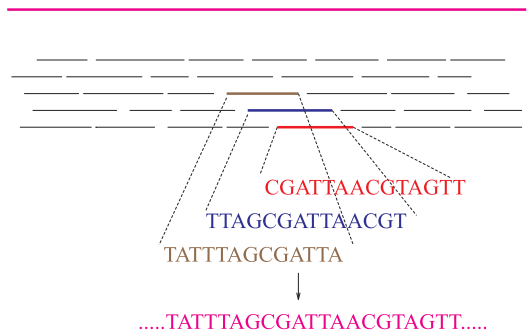

Figura 4. – A partire dalla collezione di reads si determinano le loro sovrapposizioni. Mettendo assieme questa informazioni si detemina la sequenza corrispondente alla sorgente.

rebbe banale; sfortunatamente, l'unica tecnica nota per estrarre frammenti dalla sorgente non ci permette di conoscere questa informazione essenziale. Tuttavia, il fatto che i reads siano in sovrapposizione, come osservato al punto (a), permette di stabilire la posizione relativa dei reads e quindi un ordinamento dei frammenti corrispondenti. Grazie a questa informazione, attraverso alcune sofisticate tecniche combinatorie, di cui faremo qualche cenno nella Sezione 7, siamo in grado di ricostruire la sorgente (vedi Fig. 4) con un ottimo margine di precisione.

Con riferimento al punto (c), si noti che un ricoprimento della sorgente è assicurato certamente da una campionatura abbondante. D'altra parte, la scelta di un campione di dimensioni eccessive può avere effetti disastrosi sulle finanze del progetto, in quanto ottenere i reads dai frammenti è un processo piuttosto costoso. Un problema matematico di fondamentale importanza legato allo shotgun sequencing è quindi quello di stabilire un livello di campionatura che garantisca ricoprimento mantenendo i costi del progetto sufficientemente bassi. Tale problema verrà brevemente discusso nella Sezione 5.

3.2. *Shotgun Sequencing - Il problema computazionale.*

Come già detto nel Paragrafo 3.1, il fragment assembly consiste nell'ottenere la sorgente a partire dalla collezione di reads. A livello

intuitivo, l'obbiettivo è quello di rimettere assieme i vari pezzi del mosaico in maniera tale che, con alta probabilità, ne risulti la stringa sorgente. Una formalizzazione matematica, vicina all'intuizione data e che ne caratterizzi tutti gli aspetti biologici, non si presta ad una semplice trattazione. Infatti, esistono diverse formalizzazioni del fragment assembly [13, 21], ognuna delle quali risulta soddisfacente sotto certi aspetti, mentre lascia a desiderare per altri. Per il momento, ci limitiamo a descrivere una formalizzazione di facile comprensione, ed, in alcuni casi, utile in pratica. A partire da questo esempio particolare, si metteranno in evidenza i vincoli che in generale ogni formalizzazione del problema del fragment assembly deve rispettare.

Shortest Common Superstring.

Si consideri il seguente problema, noto come *Shortest Common Superstring (SCS)* (cf. [13]): date n stringhe $x_1, \ldots, x_n$, costruire una stringa X di lunghezza minima che contenga tutte le x_i come sottostringhe. Se, ad esempio, si considera l'insieme delle stringhe $\{ACAC, CTAC, AC, CTCT\}$, una loro SCS è $CTCTACAC$. Supponiamo ora che le stringhe x_i siano i reads relativi allo shotgun sequencing di una sequenza di DNA. La SCS X corrisponde ad una possibile ricostruzione del «mosaico». Possiamo quindi prenderla come soluzione del problema di fragment assembly, senza però avere alcuna garanzia che tale ricostruzione sia corretta. In alcune circostanze, tale modo di procedere può anche fornire dei buoni risultati pratici [13]. Tuttavia ogni formalizzazione del fragment assembly deve tenere conto di alcuni vincoli che appaiono concretamente nel sequenziamento del DNA:

- *Presenza di errori di laboratorio nel sequenziamento:* ottenere un read da un frammento è un processo soggetto ad errori legati alla tecnologia delle macchine utilizzate per il sequenziamento base. Un problema fondamentale è quello di assegnare una probabilità ad ogni lettera che compare in una specifica posizione del read. Ad esempio, se la stringa risultante dal sequenziamento base di un insert è $ACACTGC\ldots$, qual'è la probabilità che la quinta lettera sia

effettivamente una T? Tali stime sono di fondamentale importanza per progetti di sequenziamento su larga scala. Il problema di stabilire delle misure accurate circa l'affidabilità dei dati in ingresso è stato fra i punti più controversi negli studi di fattibilità di *Whole-Genome Shotgun Sequencing* [12, 27], un metodo mediante il quale si è recentemente realizzato sequenziamento del genoma umano. A questo proposito si pongono importanti quesiti ai quali studi di tipo probabilistico permettono di dare una risposta soddisfacente. Per una formulazione matematica di tali problemi e per lo stato dell'arte, si veda [8].

- *Presenza di sequenze ripetitive:* il DNA può presentare delle zone ripetitive: esistono cioè delle particolari sequenze che si ripetono più volte in certe aree della molecola in questione. Tali aree sono estremamente difficili da ricostruire e costituiscono quindi una grossa complicazione per il problema dell'assembly.

Consideriamo ad esempio la scelta di una SCS come soluzione per il fragment assembly: cerchiamo una stringa di lunghezza minima, e quindi la più economica possibile in termini di lunghezza, che contenga tutti i reads. In altre parole utilizziamo implicitamente il principio di *parsimonia* come fondamento per la nostra funzione obbiettivo. Questo principio si rivela non adatto come funzione obbiettivo di fragment assembly per stringhe ripetitive. Infatti consideriamo per esempio, le stringhe in input $\{ACAC, ACAG, ACAC, ACAT, GCCTGAC\}$. Da questa collezione di reads è possibile ottenere due diverse shortest common superstrings: *ACACAGCCTGACAT* e *ACAGCCTGACACAT*. Supponiamo, invece, che la stringa iniziale da cui abbiamo ricavato i reads sia *ACACAGCTGACACACAT*. Si noti che le due superstringhe ottime soffrono di *super-compressione* cioè una parte della stringa sorgente è stata compressa. Come anticipato, ciò è dovuto alla natura ripetitiva della stringa sorgente utilizzata nell'esempio. Considerato che i genomi di eucarioti (fra cui anche il genoma umano) presentano di fatto vaste aree ripetitive, un buon algoritmo di fragment assembly per tali genomi deve essere in grado di ricostruire queste aree in maniera corretta quindi non deve dar luogo a problemi di super-compressione. Come affrontare e ri-

solvere in maniera soddisfacente tale problema? Questo è stato un altro punto molto controverso per progetti di sequenziamento su larga scala. Diversi contributi di matematica ed informatica si sono rivelati fondamentali in quest'area. Alcuni di questi verranno presentati in dettaglio nella Sezione 4.2.

Si noti infine che, sebbene la sua formalizzazione sia di facile comprensione, SCS è un problema NP-hard [10]. Ciò significa che stabilire l'esistenza di un algoritmo efficiente per la sua soluzione è equivalente a risolvere uno dei più difficili problemi aperti della matematica moderna [31]. Stabilita la sua difficoltà computazionale, la teoria dell'NP-Completezza ci autorizza a cercare soluzioni non ottime per SCS (attraverso algoritmi approssimati o euristiche). Si noti che la presenza di problemi NP-hard è un tema ricorrente per quasi tutte le formalizzazioni dei problemi legati al sequenziamento genomico. Questo è chiaramente legato alla difficoltà intrinseca dei problemi in questione.

4. – Principi di progetto per sequenziamento genomico.

In questa sezione illustreremo i due più importanti principi di progetto per il sequenziamento su larga scala, che sono stati applicati in particolare al sequenziamento del genoma umano. Il primo, di tipo top-down, è stato adottato dal consorzio pubblico Human Genome Project [29], mentre l'altro, di tipo bottom-up, è quello utilizzato dalla compagnia privata Celera Genomics [30].

Human Genome Project è un consorzio multinazionale di diversi centri di ricerca, istituito nel 1988 [7]. Sin dall'inizio, il Progetto Genoma fu strutturato in maniera tale da soddisfare diverse esigenze: una era quella di dividerlo in sotto-progetti, affinchè i diversi laboratori del consorzio, sparsi in diverse nazioni, potessero lavorare in parallelo; un'altra era quella di superare le limitazioni inerenti al processo di shotgun sequencing, cui si è accennato nella Sezione 3.2; un'altra necessità era quella di ottenere una soluzione finale quanto più accurata possibile (per i dettagli vedere [6]). Queste motivazioni hanno portato il consorzio a scegliere un approccio in cui la fase di

sequenziamento viene preceduta da una fase preliminare, che consiste nella costruzione di *mappe fisiche*. Le mappe permettono infatti di conoscere le informazioni necessarie a ricostruire il genoma a partire dai frammenti sequenziati in maniera indipendente. Questa strategia ha permesso di affidare alle diverse sedi il sequenziamento di pezzi diversi del genoma umano, ad esempio i cromosomi. Tuttavia, i processi di costruzione di mappe fisiche si sono rivelati molto lenti, tanto da risultare il collo di bottiglia dell'intero progetto [5, 15, 23, 27].

A metà degli anni Novanta, il sequenziamento dei piccoli genomi di alcuni organismi attraverso tecniche che non utilizzavano esplicitamente il concetto di mappa [1, 17], ha messo in discussione l'utilità della costruzione di mappe fisiche, facendo anche intravedere la possibilità di una notevole accellerazione nel sequenziamento del genoma umano. Da allora una tecnica di sequenziamento di interi genomi non vincolato alla costruzione di mappe è diventato argomento di studio di diversi ricercatori [23, 27]. L'idea intuitiva alla base di questa tecnica consiste nel dividere l'intero genoma in frammenti di piccola dimensione, sequenziare tutti i frammenti e poi assemblare le stringhe così ottenute in un'unica stringa. È chiaro che, se da un lato una strategia del genere permette di evitare la costosa e lunga operazione di costruzione di mappe fisiche, dall'altro porta ad un problema di fragment assembly di enormi dimensioni e di non facile soluzione. La realizzabilità pratica di tale strategia di sequenziamento è stata a lungo discussa, in modo anche controverso [12, 23, 25, 27]. Tale strategia è stata infine applicata dalla compagnia privata *Celera Genomics* per il sequenziamento del genoma umano, con ricorsi minimi alla costruzione di mappe fisiche [24].

4.1. *Costruzione Top-Down: Il progetto genoma umano.*

Supponiamo di voler sequenziare una certa molecola di DNA *target*. Una *libreria di cloni* è un insieme di frammenti di DNA (*cloni*) ottenuti suddividendo in pezzi diverse copie del target. In generale, una volta che la molecola target viene frammentata, l'informazione relativa alla posizione dei singoli cloni rispetto alla sequenza di pro-

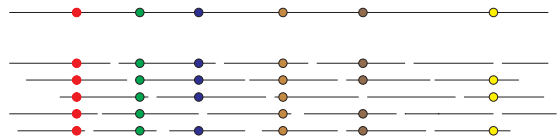

Figura 5. – I puntini colorati sulla linea continua in alto rappresentano (l'occorrenza di) stringhe caratteristiche uniche nel frammento di DNA. I cloni sono rappresentati da linee più piccole in basso. I puntini colorati nei cloni rappresentano le stesse stringhe caratteristiche del frammento. Si noti che, grazie all'unicità di tali stringhe caratteristiche, bisogna allineare puntini con colori uguali. Ciò dà un ordinamento dei cloni, ovvero una mappa fisica della stringa originale.

venienza viene perduta. Una possibile strategia per superare questa difficoltà può essere quella di costruire una *mappa fisica* della molecola di DNA in questione. Essa infatti stabilisce il punto in cui determinati demarcatori (stringhe caratteristiche) compaiono nella sequenza di DNA esaminata.

La presenza o meno di questi demarcatori sui frammenti del campione estratto dalla libreria, permette di stabilire un ordinamento relativo dei cloni, nonché la posizione di ogni singolo clone nella sequenza del DNA di provenienza (vedi Figura 5). Ci riferiremo a tale problema come all'*ordinamento della libreria dei cloni* [13]. Si noti che, a differenza dell'assembly, in questo caso non conosciamo le sequenze alfanumeriche corrispondenti ai cloni. L'unica informazione che possiamo sfruttare è che, se due cloni contengono lo stesso demarcatore, allora andranno sovrapposti in maniera tale che i demarcatori uguali si trovino nella stessa posizione (Figura 5).

È evidente che la costruzione di mappe fisiche e l'ordinamento di librerie di cloni sono problemi intimamente connessi. Seguendo la letteratura, ci riferiremo ad entrambi come *costruzione di mappe fisiche* [13].

Illustriamo ora il procedimento seguito per il sequenziamento genomico mediante l'approccio *top-down*, che consiste essenzialmente nella costruzione di mappe fisiche a diversi livelli. In Figura 6, si riportano tre livelli diversi di mappe. Al primo livello, la molecola di DNA viene riprodotta in diverse copie, ciascuna delle quali viene frammentata in cloni di grandi dimensioni (tipicamente intorno al

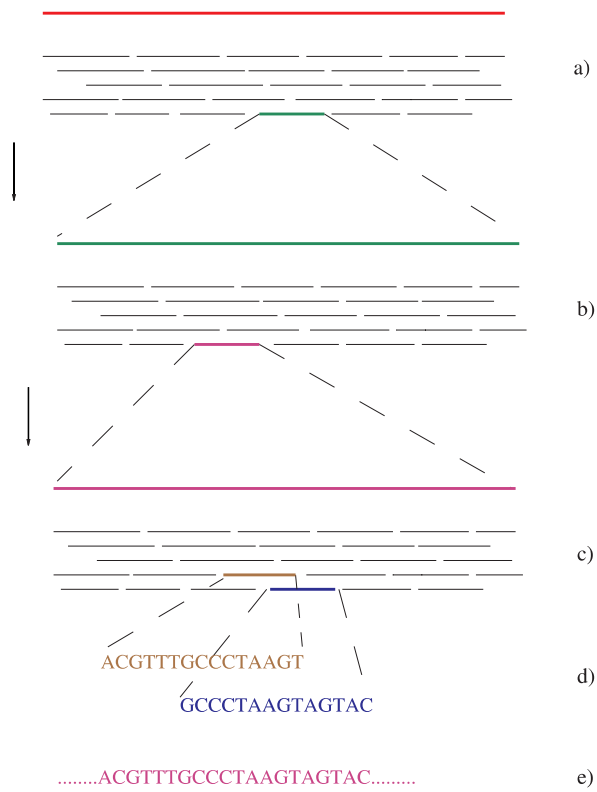

Figura 6. – Strategia di sequenziamento top-down. a) Creazione e ordinamento degli YAC; b) Creazione e ordinamento dei cosmids c) Creazione e ordinamento dei plasmids d) Sequenziamento dei plasmids mediante shotgun sequencing; e) ricostruzione a ritroso delle sequenza al livello superiore, mediante l'uso delle mappe.

milione di basi). Tali cloni assumono il nome dei particolari vettori di clonazione in cui vengono inseriti che possono essere *YAC* (Yeast Artificial Chromosome) o *BAC* (Bacterial Artificial Chromosome). Tale operazione ha sia lo scopo di mettere i cloni in condizione di essere mantenuti in laboratorio, sia quello di poterli riprodurre in diverse copie. Per fissare le idee, supponiamo che si tratti di YAC. In base ai demarcatori, è possibile stabilire un ordinamento dei YAC, da cui si deduce una mappa fisica della sequenza iniziale (Figura 6, a)). Passando al secondo livello (Figura 6, *b*)), ogni YAC viene a sua

volta duplicato e suddiviso in parti più piccole (di solito intorno alle 40000 basi) che prendono il nome di *cosmids*. Come si è fatto per la sequenza iniziale, per ogni YAC si costruisce una mappa fisica. Siamo in grado quindi di conoscere la posizione di ogni cosmid nello YAC di provenienza. A loro volta i cosmids vengono duplicati e frammentati in *plasmids* (Figura 6, *c*)), e per ogni cosmid si costruisce una mappa fisica. Finalmente ogni plasmid viene sequenziato mediante la tecnica dello shotgun sequencing (Figura 6, *d*)). Una volta sequenziate le parti, procedendo a ritroso, i vari livelli di mappe fisiche permettono di ricostruire facilmente la sequenza iniziale (Figura 6, *e*)).

I problemi matematici che derivano da tale tipo di approccio sono numerosi e di varia natura. Infatti, ai problemi riguardanti il campionamento dei cloni e l'assembly, di cui si è già discusso nella Sezione 3.2, si aggiungono interessanti problemi combinatorici legati alle diverse tecniche per la costruzione di mappe fisiche. Descriveremo in particolare una di queste tecniche nella Sezione 6.

4.2. *Costruzione Bottom-Up: La Celera Genomics.*

La Celera Genomics è una compagnia privata che, contemporaneamente al consorzio pubblico Human Genome Project, si è occupata del sequenziamento del genoma umano.

La strategia utilizzata dalla Celera, denominata *Whole-Genome Shotgun Sequencing*, si differenzia da quella utilizzata dall'Human Genome Project in quanto fa un uso minimo delle mappe fisiche, mentre applica delle procedure di fragment assembly su larghissima scala. In questo caso si parlerà di una stategia di tipo *bottom-up*.

In questo tipo di approccio, l'intero genoma viene direttamente suddiviso in piccoli frammenti. A differenza della procedura di sequenziamento classica, si applica qui il cosiddetto *pairwise-end sequencing*: il sequenziamento base viene cioè applicato alle due estremità di ogni frammento, ottenendo così un certo numero di coppie di reads, corrispondenti ai due estremi di uno stesso frammento. Queste coppie verranno chiamate *mate-pairs*.

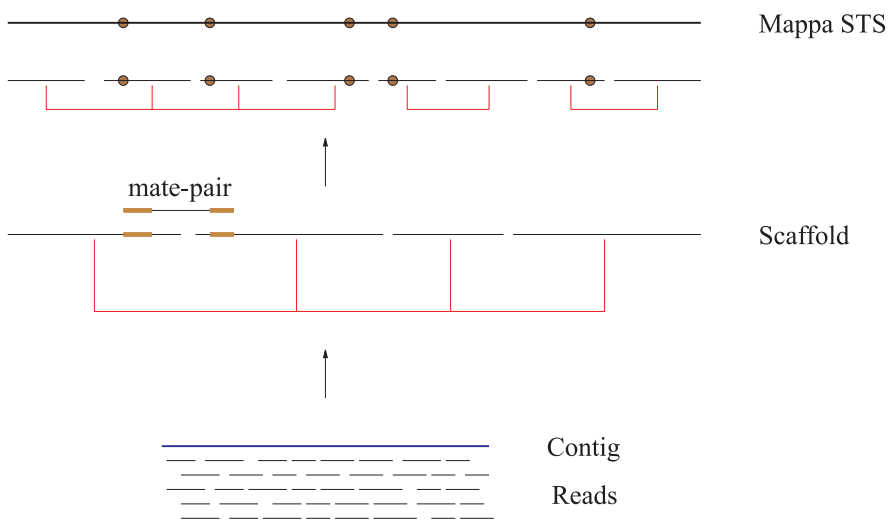

Figura 7. – Strategia di sequenziamento bottom-up.

Il primo passo della costruzione bottom-up consiste nel raggruppare in un'unica classe i reads che, con alta probabilità, si assemblano in un'unica stringa. Per ogni classe siamo in grado di determinare tale stringa, corrispondente ad una regione di basi contigue nel genoma (*contig*) (vedi Fig. 7). Il secondo passo consiste nel raggruppare i contigs in sottoinsiemi ordinati, chiamati *scaffolds*, utilizzando l'informazione relativa ai mate-pairs. Infatti, se un read è contenuto in un contig e il suo mate appare in un altro contig, allora siamo in grado di determinare il loro orientamento relativo. Con questa tecnica riusciamo a raggruppare gli scaffolds in classi, tuttavia ciò non è sufficiente a determinare un ordine totale fra tutti i contigs. Quindi, per sapere in che ordine e con quale orientamento vanno disposti gli scaffolds sul DNA di origine si fa ricorso a una mappa STS, un tipo particolare di mappa di cui discuteremo più diffusamente nella Sezione 6. In ogni caso, anche qui vale il principio di utilizzare demarcatori unici per poter posizionare gli scaffolds. Con riferimento alla Figura 7, i demarcatori sono i punti messi in evidenza sulla mappa ed ognuno è allineato al suo corrispondente nello scaffold. A questo punto il sequenziamento è quasi completo, a meno di alcune aree non sequenziate (*gaps*) che restano fra i vari cosmids. Queste parti

mancanti possono essere completate mediante il ricorso a tecniche alternative [24].

Osserviamo infine che la nozione di mate-pair ha anche un ruolo fondamentale nel risolvere efficientemente il problema delle aree ripetitive, quando la taglia delle sequenze che si ripetono è più piccola della taglia dei frammenti da cui vengono estratti i mate-pairs.

5. – Probabilità e Statistica.

Il Calcolo delle Probabilità e la Statistica hanno fornito fondamentali contributi alle metodologie usate per il sequenziamento genomico (per una introduzione, si può consultare [26]). Qui ci limiteremo a presentare una trattazione estremamente semplice di alcuni problemi statistici legati al problema del ricoprimento del genoma da parte di un campione di cloni estratti da una libreria. Tale trattazione, descritta qui assumendo delle condizioni ideali, può essere estesa a casi più realistici e specializzata al caso di campionatura di frammenti per shotgun sequencing. Per tutto lo spettro di applicazioni che ci interessano, le stime che ne derivano sono abbastanza accurate [26]: in particolare questo tipo di trattazione assume un ruolo fondamentale nel progetto di sequenziamento genomico su larga scala [9].

Presentiamo qui un modello molto semplice che ci suggerisce un metodo per l'estrazione di un campione da una libreria di cloni, che ricopra con alta probabilità l'intero genoma. Al fine di schematizzare il problema, supponiamo che tutti i cloni abbiano lunghezza L. Inoltre diremo che due cloni sono in sovrapposizione solo nel caso in cui essi si sovrappongono almeno per una frazione θ della loro lunghezza. Intuitivamente, il parametro θ dice «quanta informazione» occorre per potere affermare con certezza che due cloni si sovrappongono.

Man mano che i cloni vengono selezionati e che le sovrapposizioni vengono determinate, osserveremo la formazioni di agglomerati di cloni in sovrapposizione tra loro, che chiameremo *isole apparenti* (in quanto alcune sovrapposizioni non saranno determinate). Inoltre tali agglomerati risulteranno divisi da spazi, che chiameremo *ocea-*

ni. La Figura 7 illustra molto bene il processo: ogni contig è un'isola apparente. Lo scaffold in Figura è costituito da quattro isole apparenti e tre oceani.

In funzione del numero di cloni N selezionati dalla libreria, siamo interessati ad avere informazioni statistiche circa (a) il numero e la lunghezza delle isole apparenti; (b) la lunghezza degli oceani. Naturalmente, affinché sia garantito il ricoprimento, vogliamo che il numero delle isole apparenti si riduca ad 1 e che la lunghezza degli oceani si riduca a 0. La conoscenza di tali informazioni, come funzione della taglia del campione, ci permette di stabilire quando possiamo essere sicuri, con un piccolissimo margine di errore, di avere un ricoprimento del genoma.

Denotiamo con G la lunghezza del genoma in questione. L'estrazione di un clone dalla libreria può essere rappresentato da una coppia $([0, r],[0, l])$ di sottointervalli di $[0, G]$. Nel primo sottointervallo, r detto *punto destro* rappresenta punto nel genoma in cui il clone termina mentre nel secondo l (*punto sinistro*) rappresenta il punto d'inizio del clone nel genoma. Inoltre, sia $c = NL/G$ il *fattore di ricoprimento* o *di ridondanza*, cioè il numero atteso di cloni che ricoprono un particolare punto del genoma. Il fattore di ricoprimento ci dice, in altre parole, di quanto abbiamo amplificato la lunghezza originaria del genoma, selezionando N cloni dalla libreria. La sequenza punti destri può essere modellata come un processo di Poisson, nell'intervallo $[0, g]$, $g = G/L$, con numero medio di arrivi uguale a c. Lo stesso vale per i punti sinistri. Da quanto detto, si possono ricavare diverse informazioni statistiche, di cui ci limitiamo a segnalare le seguenti:

1. Il numero atteso di isole apparenti è $Ne^{-c(1-\theta)}$.

2. Il numero atteso di cloni in un'isola apparente è $e^{c(1-\theta)}$.

3. La lunghezza attesa di un'isola apparente, in basi, è $L\lambda$, dove $\lambda = \theta + (e^{c(1-\theta)} + 1)/c$.

4. La probabilità di avere un oceano di lunghezza xL alla fine di un'isola apparente è $e^{-c(x+\theta)}$.

Si noti come il fattore di ricoprimento c e il parametro θ giochino un ruolo fondamentale nella scelta della dimensione del campione da estrarre. Ad esempio, per problemi di assembly, si desidera un numero atteso di isole uguale ad uno. Dati c e θ, si può determinare la dimensione del campione.

6. – Combinatorica ed Algoritmica: Le mappe fisiche.

La costruzione di mappe fisiche su larga scala è un'area di ricerca relativamente nuova, la cui importanza è legata alla necessità di sequenziare genomi sempre più lunghi e complessi, fra cui, in particolare, il genoma umano. Diversi modelli e metodi sono stati proposti per questo tipo di problema: quelli più frequentemente utilizzati sono *double digest* mapping, *partial digest* mapping, *Hibridization* mapping [21]. Qui ci soffermeremo in particolare su di un approccio noto come *STS content mapping* (vedi [13]). Molte delle idee di base e degli algoritmi che vengono utilizzati in questa particolare metodologia sono comuni a diversi altri approcci di *Hibridization mapping*, mentre i metodi di mapping basati su *digest* richiedono altri tipi di tecniche che qui non verranno descritte. Per consultazione si può vedere [21].

Dal punto di vista matematico, l'aspetto comune a tutti i problemi menzionati è che la loro modellizzazione ed il loro studio si riduce spesso alla ricerca di soluzioni per problemi combinatorici ed algoritmici su grafi. Si noti inoltre che spesso queste formalizzazioni danno luogo a problemi algoritmici NP-Hard. Ancora una volta, questo ci autorizza ad utilizzare metodi e risultati noti nel campo della teoria della Complessità Computazionale.

Qui abbiamo scelto di discutere un particolare modello utile per la soluzione del problema del mapping, ossia il *problema degli 1 consecutivi*. Vedremo però che nel caso concreto, ossia ammettendo la possibilità di errori di laboratorio, questo modello teorico non risulta efficace e si deve ricorrere a delle euristiche che conducono ad un classico problema NP-hard, ossia il *Il Problema del Commesso Viaggiatore*.

6.1. *Sequence Tagged Sites.*

Fissiamo una sequenza di DNA (*target*) che può essere un intero genoma oppure un singolo cromosoma. Un *Sequence Tagged Site* (STS), è una sequenza di DNA lunga tra le 300 e le 500 basi che presenta alle sue estremità due piccole sottosequenze, di taglia intorno a 20 nucleotidi, che compaiono solo una volta all'interno del target: di conseguenza, anche un STS compare una sola volta nel target. Sebbene sia possibile riconoscere quali sequenze sono degli STS, in genere non si conosce a priori la posizione in cui un dato STS compare nel DNA di origine.

Data una collezione di STS, ci poniamo il problema di trovare un metodo generale che ci permetta di identificare i punti in cui ogni STS compare nel target. Una soluzione a questo tipo di problema ci permette di realizzare un metodo di costruzione per le mappe fisiche detto *STS content mapping*. STS content mapping viene realizzato utilizzando una libreria di cloni che ricoprono la sequenza iniziale e che sono in sovrapposizione fra loro. La fase iniziale consiste nel determinare in quali cloni della libreria compare ogni determinato STS. Particolari tecniche di laboratorio sono in grado di fornirci questa informazione sotto forma di una matrice binaria L, di m righe, una per ogni clone, ed n colonne, una per STS. Avremo quindi che $L[i, j] = 1$ se e soltanto se l'STS j compare nel clone i. Si noti che, essendo i cloni in sovrapposizione, lo stesso STS può trovarsi in cloni diversi. Il problema matematico che ci accingiamo a formalizzare permette di determinare, a partire dalla matrice L, un ordinamento della libreria dei cloni, e quindi di determinare una mappa fisica del target.

Il caso ideale.

Cominciamo con il considerare lo scenario ideale, cioè quello in cui non si presenta nessun tipo di errore sperimentale e quindi sappiamo con certezza se un STS compare in un clone. In questo caso, ordinare la libreria di cloni, e quindi stabilire le posizioni degli STS nel target, consiste nel trovare un'opportuna permutazione delle colonne di L tale che la risultante matrice L' gode della proprietà *de-*

gli 1 consecutivi. Tale proprietà è verificata se, in ogni riga di L', tutti gli 1 compaiono in colonne consecutive. Questo è un problema classico di Ottimizzazione Combinatoria per il quale esiste un elegante algoritmo di complessità polinomiale [21].

Il caso reale.

I dati che vengono dal laboratorio, purtroppo, sono soggetti a tre tipi di errori:

- si ha un *falso positivo* quando si registra la presenza di un STS in un clone che in realtà non lo contiene;
- abbiamo al contrario un *falso negativo* quando l'esperimento non individua la presenza di un STS in un clone che in realtà lo contiene;
- Un *clone chimerico* è un clone contenuto nella libreria e che non compare nella stringa target. La formazione di cloni chimerici avviene nel momento in cui i cloni vengono inseriti nei vettori; in questa fase infatti può capitare che due cloni non contigui nel target, si uniscano in un unico pezzo, dando luogo così ad un clone (chimerico appunto) che in realtà non deriva direttamente dalla sequenza originaria.

Per chiarirci le idee circa la migliore formalizzazione del problema in esame, supponiamo che ci venga data (da un oracolo infallibile) una matrice L' che dia l'ordinamento corretto dei cloni. Potremo osservare che se una riga i corrisponde ad un clone chimerico (ovvero a due cloni separati), ci saranno due sequenze di uno, separate da una sequenza di zero. Se una riga corrisponde a un clone con un falso negativo, troveremo due sequenze di uno, separate da uno zero. Infine se il clone presenta un falso positivo, nella riga corrispondente avremo due sequenze di zero separate da un uno. Chiameremo *gap* ogni cambiamento lungo una riga della matrice da zero a uno o da uno a zero. Si può notare che il numero di gaps presenti nella matrice è indicativo di quanti errori sono stati effettuati.

Il problema dell'ordinamento dei cloni in presenza di errori può essere quindi formalizzato come segue: data una matrice L, trovare

una permutazione delle sue colonne tale che la risultante matrice L' abbia un numero minimo di gaps. Tale problema può essere risolto in maniera molto elegante trasformandolo in un ben noto problema su grafi. Si introduce una colonna addizionale alla fine di L i cui valori sono tutti 0. Il grafo corrispondente alla nuova matrice ha $n + 1$ vertici, uno per ogni colonna di L. Dati i vertici i e j, l'arco (i, j) ha peso uguale alla *distanza di Hamming* (cf. [13]) tra il vettore corrispondente alla colonna i e quello corrispondente alla colonna j. Si può provare che la permutazione cercata corrisponde a un percorso su questo grafo che tocchi tutti i vertici e che abbia peso minimo. In altre parole consiste nel trovare la soluzione ad un'istanza del ben noto *Problema del Commesso Viaggiatore*.

Il Problema del Commesso Viaggiatore è noto essere un problema NP-Hard. Questo significa che in realtà non si conosce nessun algoritmo efficiente per la formalizzazione di STS content mapping in presenza di errori. Fortunatamente, data la sua importanza pratica in molte aree e il fascino della sua struttura combinatoria, il Problema del Commesso Viaggiatore è stato abbondantemente studiato in letteratura. Programmi efficienti per calcolare soluzioni approssimate sono disponibili e tenuti in grandissima considerazione nella comunità matematica internazionale [3].

In questa sezione abbiamo dato un semplice esempio di come un particolare problema legato al sequenziamento genomico possa essere tradotto in un problema di natura matematica e computazionale. Vogliamo sottolineare che molti problemi di ricerca relativi al sequenziamento restano ancora aperti (ad esempio *Radiation Hybrid Mapping* [2]). Per molti di questi problemi si cerca ancora un'opportuna formalizzazione matematica che permetta di affrontarli e risolverli in maniera efficiente utilizzando metodi di natura algoritmica.

7. – Combinatorica ed Algoritmica: Fragment assembly.

Nella Sezione 3.2 avevamo già enunciato il problema della Shortest Common Superstring come un possibile metodo per la soluzione del fragment assembly. Tuttavia abbiamo osservato che tale tipo di soluzione si rivela poco soddisfacente, visto che SCS non fornisce

una soluzione corretta per l'assembly nel caso in cui si verifichino degli errori di sequenziamento nei reads e nel caso in cui la sequenza iniziale sia ripetitiva. In realtà si può notare che qualunque tipo di formalizzazione matematica del fragment assembly che non tenga conto degli errori nei dati di ingresso e delle ripetizioni, incorre in problemi dello stesso tipo. Per questo motivo spesso si ricorre a particolari euristiche, mediante le quali si tenta di trovare delle soluzioni approssimate, tollerando, entro un limite fissato, un certo numero di errori di laboratorio.

La strategia utilizzata nella pratica consiste nel dividere il problema dell'assembly in tre fasi:

- La prima fase, quella di *overlap*, consiste nel determinare, per ogni coppia di reads nel campione scelto, degli overlap approssimati. Si cerca cioè, per ogni coppia di reads (u, v), un prefisso di u che si approssimi a un suffisso di v, entro un certo tasso prefissato di tolleranza dell'errore.

- Una volta note tutte le sovrapposizioni fra i reads, nella fase di *layout*, si vuole trovare un ordinamento dei pezzi compatibile con gli overlap trovati. Per la soluzione di questo problema sono stati proposti diversi metodi: quello di cui accenneremo in questa sede consiste nella ricerca di certi cammini nel *grafo degli overlap*.

Nel caso esatto, cioè ammettendo che non ci siano errori nel sequenziamento dei reads, i nodi di questo grafo rappresentano i reads del campione, ed esiste un arco fra il nodo i e il nodo j di peso p se esiste un prefisso di i di taglia p che si sovrappone a un suffisso di j. In questo grafo cercheremo il cammino che massimizza gli overlap fra tutti i frammenti: l'ordinamento dei reads si riduce ancora una volta al Problema del Commesso Viaggiatore.

Nel caso reale bisogna tenere conto del fatto che gli overlap sono approssimati. In questo caso un arco fra i e j indicherà che un suffisso di i di taglia p si approssima a un prefisso di j. Il peso p' dell'arco (i, j) sarà dato quindi da p meno una certa «penalità» legata alle differenze nella sovrapposizione dei reads. In questo caso una soluzione al Problema del Commesso Viaggiatore darà un ordinamento che da un lato massimizza le

taglie degli overlap e dall'altro minimizza il numero degli errori.

- La terza fase, quella del *consenso* mira a stabilire, in base ai reads in sovrapposizione, quale delle quattro basi è più probabile che appaia in una certa posizione del DNA. Questo viene realizzato mediante i classici metodi di *multiple sequence alignment*.

La metodologia di assembly qui presentata deve essere in realtà interpretata come una descrizione sintetica di una famiglia di metodi di assembly. Infatti, la scelta delle euristiche utilizzate per ciascuna delle tre fasi sopra descritte, dipenderà dalle sequenze particolari a cui l'assembly deve essere applicato. Per esempio, la fase di layout è stata affrontata mediante diversi tipi di euristiche che permettono di determinarne una soluzione mediante tecniche di *branch and bound*, e che si avvalgono di una serie di trasformazioni su grafi. Inoltre, le fasi di overlap e consenso, comunemente considerate «semplici», hanno un diverso livello di difficoltà secondo che le sequenze in oggetto generiano centinaia di migliaia o milioni di reads. La scelta dell'euristica dipenderà quindi dalla taglia dell'input. Un esempio particolarmente istruttivo in questo senso è dato dall'analisi comparativa di queste due fasi per l'assembly del genoma di Drosophila e del genoma umano, rispettivamente, realizzati dalla Celera Genomics.

8. – Conclusioni.

In questo articolo abbiamo presentato alcuni dei metodi matematici ed algoritmici utilizzati del sequenziamento genomico su larga scala. I soddisfacenti risultati ottenuti nella determinazione della sequenza del genoma umano hanno messo in evidenza l'importanza del ruolo svolto dai matematici e dagli informatici nella soluzione di un antico problema della biologia molecolare, cui difficilmente si sarebbe pervenuti con il semplice uso della sperimentazione di laboratorio e senza il ricorso a metodi formali e computazionali. Tuttavia i rapporti di Nature e Science, non si sono limitati ad evidenziare i contributi delle cosiddette scienze esatte per il sequenziamento, ma hanno

anche indicato quanti e quali strumenti matematici e computazionali sono già stati utilizzati nei primi passi verso la nuova sfida lanciata dalla ricerca genomica, ossia l'*analisi* del genoma umano.

A tale proposito possiamo citare la sorprendente scoperta, annunciata recentemente dai media, che il numero di geni nel genoma umano è molto inferiore alle stime attese. La determinazione dei geni viene realizzata mediante l'uso di sofisticati algoritmi, basati su ancor più sofisticati metodi matematici. Ci si può quindi aspettare che la realizzazione di algoritmi più efficienti e l'applicazione di metodi matematici più adeguati, permetta anche di migliorare capacità di identificare geni. L'analisi dell'informazione contenuta nel genoma è la sfida che terrà impegnati nei prossimi anni i biologi molecolari e anche, come ormai dato per scontato dalla comunità scientifica internazionale, i matematici e gli informatici.

Ringraziamenti.

Vogliamo ringraziare il Prof. Antonio Restivo e il Prof. Umberto Bottazzini per gli utili commenti che ci hanno dato in seguito alla lettura di una versione preliminare del presente articolo. Ringraziamo inoltre la Prof. Marie-France Sagot per la cortese ospitalità presso l'Institut Pasteur di Parigi durante la stesura iniziale del lavoro. Ringraziamo infine l'anonimo revisore dell'articolo per i preziosi consigli che ci hanno consentito di migliorarne la presentazione.

REFERENCES

[1] M. D. Adams et al, *The genome sequence of drosophila melanogaster*, Science, **287** (March 2000), 2185-2195.

[2] R. Agarwala et al. *A fast and scalable radiation hybrid map construction and integration strategy*, Genome Research, **10** (2000), 350-364.

[3] D. Applegate - R. Bixby - V. Chvátal - W. Cook, *On the solution of Traveling Salesman Problems*, Documenta Mathematica, extra volume **ICM (III)** (1998), 645-656.

[4] G. Bernardi, *The human genome: Organization and evolutionary history*, Ann. Rev. Genetics, **29** (1995), 445-476.

[5] F. COLLINS - D. GALAS, *A new five-years plan for the U. S. Human Genome Project*, Science, **262** (1993), 43-46.

[6] International Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome*, Nature, **409** (February 2001), 860-912.

[7] National Research Council, *Mapping and sequencing the human genome*, National Academy Press, Washington D. C., 1988.

[8] B. EWING - L. HILLER - M. C. WENDL - P. GREEN, *Base-calling of automate sequencer traces using phred, I, accuracy assessment*, Genome Research (1998), 175-185.

[9] R. D. FLEISCHMAN et al., *Whole-Genome random sequencing and assembly of haemophilus influenzae rd*, Science (1995), 496-512.

[10] M. R. GAREY - D. S. JOHNSON, *Computers and Intractability - A Guide to the theory of NP-Completenss*, W. H. Freeman and Company, 1979.

[11] W. B. GOAD, *Sequence analysis: Contributions by Ulam to molecular genetics*, in N. G. Cooper, editor, From Cardinals to Chaos. Reflections on the life and legacy of Stanislaw Ulam, pp. 288-291. Cambridge University Press, 1989.

[12] P. GREEN, *Against a Whole-Genome Shotgun*, Genome Research, **7** (1997), 410-417.

[13] D. GUSFIELD, *Algorithms on Strings, Trees and Sequences-Computer Science and Computational Biology*, Cambridge University Press, 1997.

[14] E. LANDER - M. WATERMAN, editors, *Calculating the secrets of life: Contributions of the Mathematical Sciences to Molecular Biology*, National Academy Press, 1995.

[15] E. MARSHALL - E. PENNISI, *NIH launches the final push to sequence the genome*, Science, **272** (1996), 188-189.

[16] A. M. MAXAM - W. GILBERT, *A new method for sequencing DNA*, Proc. Natl. Acad. Sci. USA, **74** (2) (1997), 560-564.

[17] E. MYERS et al, *A whole-genome assembly of drosophila*, Science, **287** (March 2000), 2196-2204.

[18] G. MYERS, *Whole-genome DNA sequencing*, IEEE Computational Biology (1999), 33-43.

[19] F. SANGER - S. NICKLEN - A. R. COULSON, *DNA sequencing with chain-terminating inhibitors*, **74** (12) (1977), 5463-5467.

[20] F. SANGER et al, *Nucleotide sequence of bacteriophage λ DNA,* J. Molecular Biology, **162** (4) (1982), 729-773.

[21] J. SETUBAL - J. MEIDANIS, *Introduction to Computational Molecular Biology*, PWS Publishing Company, Boston, 1997.

[22] S. M. ULAM, *Some ideas and prospects on biomathematics*, Annual Review of Biophisics and Bioengineering, 1972.

[23] J. C. VENTER et al, *Shotgun sequencing of the human genome*, Science, **280** (June 1998), 1540-1542.
[24] J. C. VENTER et al, *The sequence of the human genome*, Science, **291** (February 2001), 1304-1351.
[25] J. C. VENTER - H. O. SMITH - L. HOOD, *A new strategy for genome sequencing*, Nature, **381** (May 1996), 364-366.
[26] M. S. WATERMAN, *Introduction to Computational Biology (Maps, Sequencies and Genomics) - Interdisciplinary Statistics*, London: Chapman & Hall, 1995.
[27] J. L. WEBER - E. W. MYERS, *Human whole-genome sequencing*, Genome Research, **7** (1997), 401-409.
[28] H. WINKLER, *Verbreitung und ursache der parthenogenesis im pflanzenund tierreich*, Jena: Fischer, 1920.
[29] http://www.ornl.gov/hgmis.
[30] http://www.celera.com.
[31] http://www.claymat.org.
[29] http://www.ensembl.org.

Raffaele Giancarlo - Sabrina Mantaci, Dipartimento di Matematica e Applicazioni
Università di Palermo, via Archirafi 34 - 90123 Palermo, Italy
(raffaele, sabrina@altair.math.unipa.it)