BIOTECNOLOGIE CELLULARI, MOLECOLARI E COMPUTAZIONALI

A.A. 2019/2020

LONGOBARDI LORETA

Le **biotecnologie** sono scienze sperimentali che cioè partono da un problema , per passare poi alla formulazione di un'ipotesi seguita da un esperimento che nel caso la verifica porta alla formulazione di un modello .

es. formulazione del modello semi conservativo della replicazione di Mendelson e Stahl.

BIOINFORMATICA

La **bioinformatica** è una disciplina scientifica che studia la biologia servendosi di strumenti informatici coinvolgendo, oltre alla biologia e all'informatica, altri campi tra cui matematica, statistica, biochimica.

si occupa di:

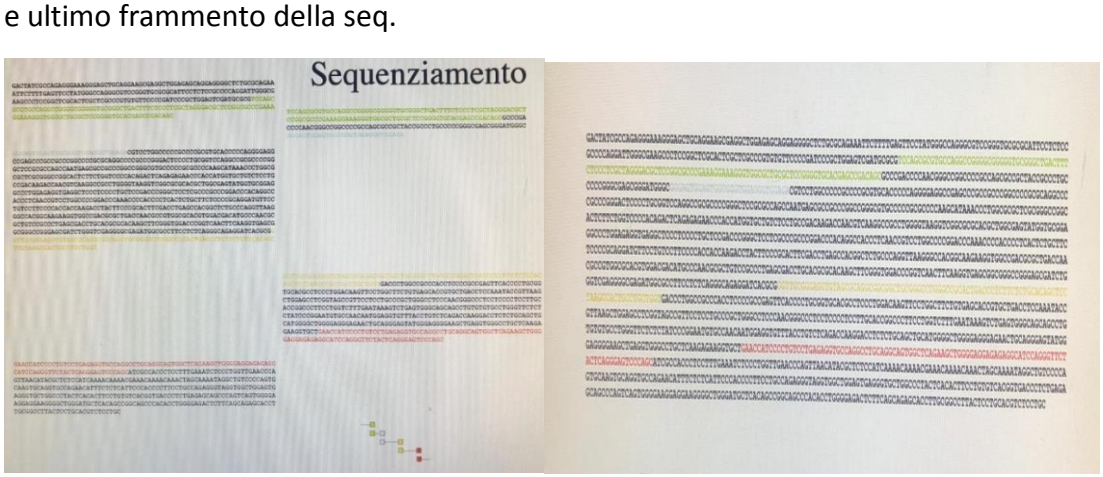
-fornire modelli per l'interpretazione e la manipolazione dei dat**i** provenienti da esperimenti di biologia molecolare e biochimica

-organizzare le conoscenze acquisite su genoma e proteoma in dati da rendere accessibili a tutti

-usare strumenti informatici per la gestione di stringhe di caratteri (sequenze di DNA, RNA e proteine) e per la loro analisi al fine di conoscerne sequenze rilevanti, la loro evoluzione e funzione,

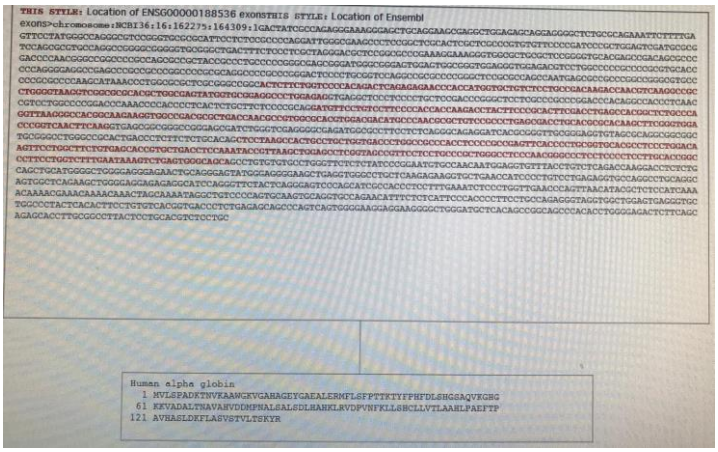
SEQUENZIAMENTO

La tecnica di sequenziamento consiste nel connettere fra loro e dunque allineare delle sequenze brevi ottenute sperimentalmente al fine di produrre un'unica sequenza genomica finale.i frammenti che si possiedono sono frammenti random dei quali cioè non si sa l'effettiva posizione all'interno del genoma finale, per cui la tecnica che si adotta è quella di osservare il tratto iniziale e finale dei vari frammenti in modo da individuare quelli che sono consecutivi, infatti il tratto iniziale di un frammento sarà quello finale di un altro e viceversa, ciò vuol dire che i due frammenti sono consecutivi pertanto possiamo fonderli tenendo la parte comune una sola volta. I due frammenti che hanno uno il tratto iniziale e l'altro finale non in comune con altri sono rispettivamente il primo e ultimo frammento della seq.



Longobardi Loreta

Analizzando poi la seq finale ottenuta è possibile individuare dei tratti codificanti che una volta assemblati producono la seq genica codificante per l'alfa globina umana



Un genoma umano però è tipicamente lungo 3-4 miliardi di basi, per cui il processo di sequenziamento è complicato. Se ad es si vuole sequenziare un gene lungo 1000 basi e si ottengono 1000 frammenti lunghi ognuno 1000 basi allora vuol dire che sullo stesso nucleotide siamo passati 1000 volte(Exp depth) quindi ogni frammento corrisponde esattamente all'intero gene.se si intende sequenziare un genoma virale lungo 10kb usando gli stessi frammenti allora sullo stesso nucleotide ci siamo passati 100 volte (1000000/10kb) e si copre circa il tot del genoma.se si intende sequenziare un genoma mammifero lungo un miliardo di basi usando gli stessi frammenti si ha una depth di 0,001 che ci dice che siamo passati su una base ogni mille (su ogni base sei passato un millesimo di volte) e quindi che il miliardo di basi copre circa un millesimo dell'intero genoma . se si vuole infine seq un genoma batterico lungo 1mln di basi, si ha che su ogni base siamo passati una volta ma la percentuale di genoma sequenziato è di circa il 67%. Per capire il perché di tale risultato si prende come es due persone in una stanza e si ricorda la regola della moltiplicazione che si applica a una combinazione di eventi che stabilisce che la probabilità che si verifichino contemporaneamente l'evento A e l'evento B equivale al prodotto delle probabilità di ciascun evento, per cui la probabilità che il loro compleanno si lo stesso giorno è 1/365 mentre quella di non avere il compleanno lo stesso giorno è 364/365, se le persone sono 3 la probabilità che la terza persona non sia nata in nessuno dei giorni degli altri due è 363/365:

Target	Target size	nSeqs (1 kb)	Exp depth	Target
gene	1 kb	1,000	1000x	100%
virus genome	10 kb	1.000	100x	~100%
bacterial genome	1,000 kb	1.000	1x	~67 %
mammalian genome	1,000,000 kb	1.000	0.001x	<0.1%

The probability changes with the number of people (D=365)

N	P(No Match)	P(No Match)
1	365/365	D/D
2	365/365*364/365	D/D*(D-1)/D
3	365/365*364/365*363/365	D/D*(D-1)/D*(D-2)/D

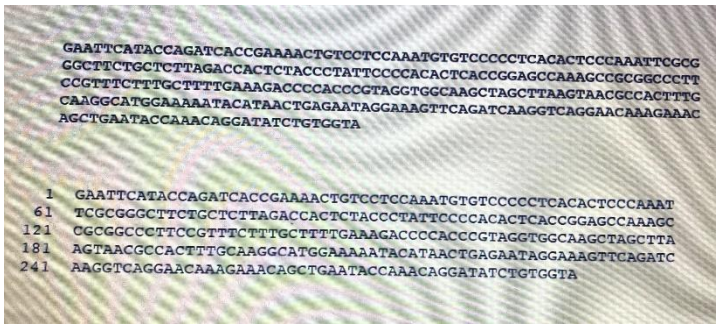
Longobardi Loreta

Più aumenta il numero di persone e più aumenta la percentuale di match addirittura costruendo una tabella ci si accorge che in alcuni gg capita che in uno stesso giorno è il compleanno di più di due persone, inoltre se arriviamo a 350 persone si nota che comunque il calendario di 365 gg ha dei buchi cioè giorni in cui nessuno compie il compleanno , quindi ne abbiamo coperto solo i 2/3 (67%) che corrisponde e giustifica quel 67% del genoma batterico che possiamo paragonare al calendario di 365 gg .

dunque abbiamo paragonato un anno al genoma e assunto che Seq un anno significa mettere 365 persone in una stanza e cercare di coprire tutti i 365 gg, li coprirai mai ? no perché capitano sempre dei match , cioè dei gg in cui due o più persone compiono gli anni lo stesso giorno, e più mach ci sono meno si copre tutto il calendario quindi nel caso del genoma più volte si passa sulla stessa base e meno stiamo seq completamente il genoma nonostante il numero di basi ottenuto con 1000 frammenti lunghi 1000 basi è pari alla lunghezza dell'intero genoma, perché vuol dire che ci saranno delle basi che non sono state proprio sequenziate.

quindi in sostanza tenderemo a coprire tutto il genoma ma coprirlo è molto complicato infatti così come per seq l'intero anno serviranno mooolte persone ,così per coprire tutto il genoma serviranno molti tentativi di sequenziamento al fine di seq tutte le basi .

una volta che abbiamo ottenuto il sequenziamento di un tratto genico, ciò che risulta sono delle read (letture di seq):



nel secondo caso a inizio di ogni seq è indicata la posizione corrispondente a ogni primo nucleotide della seq stessa. ad es 181 = A , 183 =T

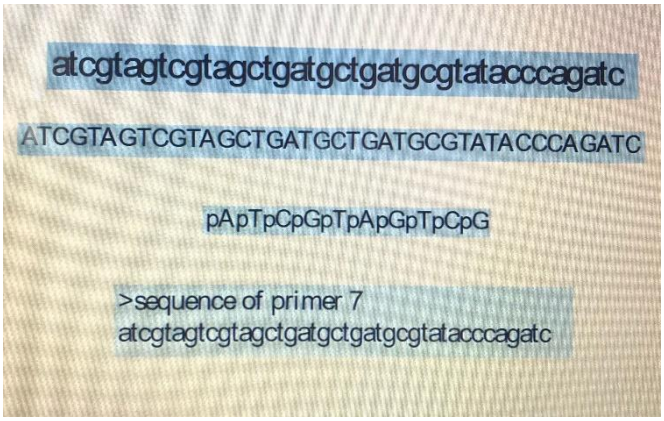
Longobardi Loreta

MANIPOLAZIONE DELLE SEQUENZE

DNA e proteine sono lunghe sequenze di monomeri quali nucleotidi e amminoacidi e in linguaggio informatico la sequenza si traduce in lunghe **stringhe** di caratteri che nel caso del dna sono A G C T nel caso della proteine i 20 a.a , pertanto la bioinformatica ci ha dato la possibilità di conoscere tali sequenze e quindi di studiarle. tali molecole biologiche hanno una loro **linearità** e **polarità** ciò un inizio e una fine:

- Per gli acidi nucleici è 5'p-> 3'OH
- Per le proteine è NH2-terminale-> COOH terminale

Pertanto allo stesso modo in cui scriviamo un testo da sx verso dx così si scrive una seq nucleotidica o amminoacidica secondo la polarità della molecola . le sequenze nucleotidiche possono essere scritte o come una lista di lettere minuscole o maiuscole che rappresentano i singoli nucleotidi, oppure utilizzando entrambi i caratteri per identificare distintamente introni ed esoni in una stessa seq o per identificare un inserto in un plasmide batterico ,o indicando i gruppi fosfato , oppure includendo una piccola descrizione della seq sottostante (**formato FASTA**)



Le operazioni che in campo bioinformatico è possibile effettuare con la molecole biologiche sono molte:

- **Editing**: modificare .così come modifichiamo un testo word così è possibile manipolare le sequenze, per es così come si può sostituire una parola in un testo , così si può cambiare una base in una seq per simulare una mutazione , o posso aggiungerla , eliminarla per simulare inserzione e delezione e come si può cercare in un lungo testo scritto in word una specifica parola, così in una lunga sequenza di dna è possibile ricercare una specifica seq genica
- **Complement**: cioè a partire da una seq di dna trovare la complementare. Il calcolo della sequenza del filamento complementare viene effettuato applicando la nota regola di appaiamento delle basi, per cui A è convertita in T, C in G, G in C, T in A. Inoltre, per rispettare la polarità dei filamenti, la sequenza è invertita in modo da risultare scritta in direzione 5'-3'.
- **Composition**: cioè avendo una seq di dna chiedere qual è la sua composizione in basi, ad es quante sono le A
- **Ricercare** un sito di restrizione
- **Translate:** cioè avendo una seq di dna tradurla

Longobardi Loreta