

ANNOTAZIONE DEI GENOMI

1. Annotazione dei genomi (storia e importanza dell'annotazione)
2. Strumenti e risorse per l'annotazione dei genomi
 - Metodi per la determinazione della funzione dei geni
3. Metodi di annotazione di genomi procariotici
 - ORFfinder
 - PGAP
 - Sfide nell'annotazione dei genomi procariotici
4. Metodi di annotazione di genomi eucariotici
 - Metodi basati sull'allineamento
 - Metodi diretti (Blast e Blat)
 - Metodi indiretti (ab initio o de novo)
 - Metodi basati sulla ricerca per omologia (InterPro)
 - Importanza e sfide nell'annotazione dei genomi eucariotici
3. Modelli Markoviani
4. Annotazione dell'RNA



ANNOTAZIONE DEI GENOMI

1. Annotazione dei genomi



L'annotazione dei genomi è il processo di **identificazione** e assegnazione di **funzioni** ai geni all'interno di un genoma.

Questo processo coinvolge l'identificazione di regioni di interesse all'interno del genoma, come i geni codificanti per proteine, i geni non codificanti, i siti di regolazione e le regioni ripetute, e l'assegnazione di funzioni a queste regioni in base all'analisi delle sequenze e alle informazioni disponibili.

L'annotazione dei genomi è un passo fondamentale nella comprensione della biologia di un organismo e nella realizzazione di studi di biologia computazionale e di genomica funzionale.

ANNOTAZIONE DEI GENOMI

1. Annotazione dei genomi



Perché è importante ?

L'annotazione dei genomi è fondamentale per la ricerca biologica in quanto fornisce informazioni sulla struttura e sulla funzione dei geni all'interno di un organismo.

Queste informazioni sono utili per la comprensione di diversi aspetti della biologia dell'organismo, come la regolazione dell'espressione genica, la sintesi proteica e la funzione biologica delle proteine.

Inoltre, l'annotazione dei genomi è uno strumento essenziale per la genomica comparativa, che permette di confrontare e analizzare i genomi di differenti organismi per individuare differenze e somiglianze tra di essi.

La genomica comparativa è fondamentale per lo studio dell'evoluzione biologica e per la scoperta di nuovi geni e proteine, che possono avere un ruolo fondamentale in processi biologici come lo sviluppo, la riproduzione e la risposta a fattori di stress ambientali.

ANNOTAZIONE DEI GENOMI

1. Annotazione dei genomi

L'annotazione dei genomi è un campo relativamente giovane della biologia molecolare. Il primo genoma ad essere completamente sequenziato fu quello del batterio *Haemophilus influenzae*, nel 1995. Dopo di esso, furono sequenziati altri genomi batterici, fino ad arrivare alla sequenza completa del genoma umano nel 2001.

Inizialmente, l'annotazione dei genomi consisteva principalmente nella previsione delle regioni codificanti (i geni) e delle regioni non codificanti (le regioni intergeniche). Tuttavia, con l'avanzare delle tecnologie di sequenziamento e delle tecniche di analisi, l'annotazione dei genomi è diventata sempre più sofisticata e dettagliata.

batterio *Haemophilus influenzae*

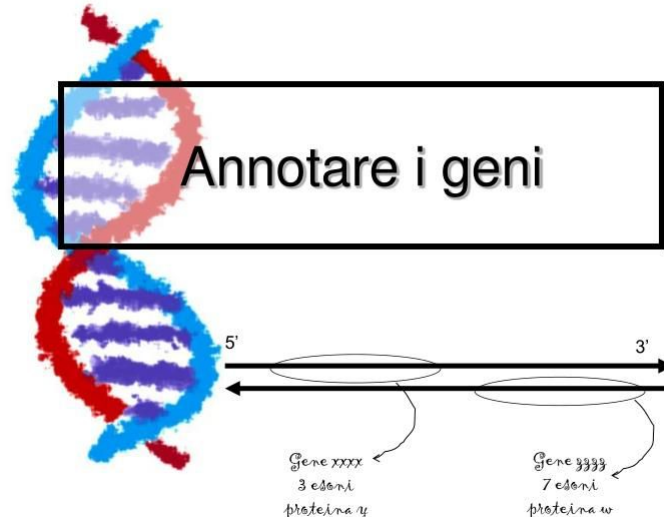


ANNOTAZIONE DEI GENOMI

1. Annotazione dei genomi



Oggi, l'annotazione dei genomi comprende anche la previsione delle variazioni genetiche, la caratterizzazione delle regioni regolatorie, la previsione dei siti di legame dei fattori di trascrizione, l'analisi delle vie metaboliche e molte altre informazioni biologiche utili per la comprensione dei processi biologici.



ANNOTAZIONE DEI GENOMI



2. Strumenti e risorse per l'annotazione dei genomi

Esistono molti strumenti e risorse utilizzati per l'annotazione dei genomi, alcuni dei quali sono:

- Programmi di previsione del gene: questi programmi utilizzano metodi di riconoscimento delle caratteristiche del gene (come la presenza di codoni di inizio e di stop) per identificare i geni all'interno del genoma.
- Database di sequenze proteiche: questi database contengono sequenze di proteine annotate da organismi correlati e possono essere utilizzati per identificare funzioni note o simili tra le proteine presenti nel genoma annotato. Ad esempio, il database di proteine UniProt è una risorsa comunemente utilizzata per l'annotazione dei genomi.

ANNOTAZIONE DEI GENOMI

2. Strumenti e risorse per l'annotazione dei genomi



- Strumenti di allineamento: gli strumenti di allineamento come BLAST possono essere utilizzati per identificare regioni del genoma con omologia con sequenze note. Questo può aiutare ad identificare i geni e le regioni funzionali all'interno del genoma.
- Strumenti di ontologia genica: questi strumenti aiutano a categorizzare i geni in base alla loro funzione biologica. Ad esempio, il Gene Ontology (GO) categorizza i geni in base alle loro funzioni molecolari, biologiche e cellulari.
- Strumenti di visualizzazione: gli strumenti di visualizzazione come IGV e Artemis possono essere utilizzati per visualizzare il genoma annotato e identificare le regioni di interesse.

ANNOTAZIONE DEI GENOMI



- Metodi per la determinazione della funzione dei geni

La determinazione della funzione dei geni è un aspetto fondamentale dell'annotazione dei genomi, poiché fornisce informazioni sulle funzioni biologiche dei geni e dei loro prodotti proteici. Esistono diversi approcci utilizzati per determinare la funzione dei geni attraverso l'annotazione.

1. Uno degli approcci più comuni è **l'omologia**, che si basa sulla ricerca di sequenze simili a quella del gene di interesse in database di sequenze note. Se una sequenza simile è stata precedentemente annotata con una funzione nota, si può presumere che il gene di interesse svolga una funzione simile o correlata.

Questo approccio viene utilizzato anche per identificare i geni **ortologhi**, che sono omologhi in diversi organismi e svolgono funzioni simili o identiche.

ANNOTAZIONE DEI GENOMI



- Metodi per la determinazione della funzione dei geni

2. Un altro approccio è la **predizione** della **struttura** proteica. La struttura tridimensionale delle proteine è spesso correlata alla loro funzione biologica, quindi la predizione della struttura proteica può fornire informazioni sulla funzione del gene.
Questo approccio viene utilizzato congiuntamente con l'omologia per determinare la funzione di un gene.
3. Infine, la caratterizzazione dei **domini** proteici può fornire informazioni sulla funzione del gene. I domini proteici sono regioni funzionali di una proteina che possono svolgere una specifica attività biologica.
La presenza di un determinato dominio proteico può fornire informazioni sulla funzione del gene.

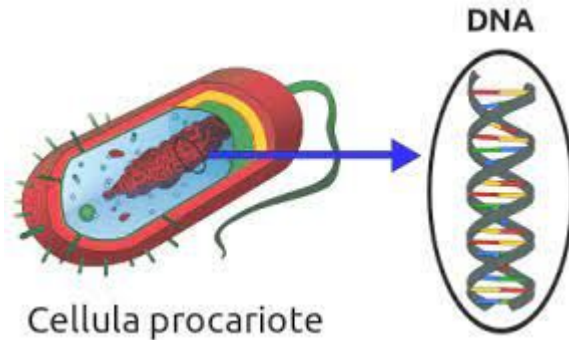
ANNOTAZIONE DEI GENOMI

2. Metodi di annotazione di genomi procariotici



L'annotazione dei **genomi procariotici** è molto più semplice di quella degli eucarioti .
Nell'annotazione di genomi procariotici, il primo passo consiste nella ricerca di regioni del genoma contenenti i codici per le proteine, cioè gli open reading frame (**ORF**).

Questo può essere fatto utilizzando una combinazione di tecniche: di **predizione** degli ORF e di **ricerca di omologie**.



ANNOTAZIONE DEI GENOMI



2. Metodi di annotazione di genomi procariotici

La **predizione degli ORF** prevede l'individuazione di sequenze di DNA che contengono un codice di avvio (start codon) seguito da una sequenza di nucleotidi che codifica per una lunga sequenza di amminoacidi, concludendosi con un codice di stop (stop codon). Questi ORF sono generalmente di lunghezza superiore ai 100 nucleotidi, anche se la lunghezza può variare in base alla specie considerata.

Il genoma viene scansionato in tutti i possibili 6 frame di lettura, 3 per ciascun filamento, e tradotto, sulla base del codice genetico, nelle corrispondenti sequenze proteiche.

Le reading frame delimitate da un codone di start ed uno di stop vengono dette aperte, da cui ORF.

La probabilità di incontrare un codone di stop nel frame di lettura è pari a $64(\text{num. tot. codoni})/3 \text{ numero di frame} = 21$.

ANNOTAZIONE DEI GENOMI

2. Metodi di annotazione di genomi procariotici



La probabilità di incontrare un codone di stop nel frame di lettura è pari a $64(\text{num. tot. codoni})/3 \text{ numero di frame} = 21$.

La lunghezza attesa di un ORF casuale è circa 20 aminoacidi. Essendo di solito le sequenze proteiche più lunghe di 20, per individuare i geni che codificano proteine in un genoma batterico si imposta una lunghezza di sequenza di aminoacidi maggiore di una data soglia (di solito 50 a.a.).

Vi sono parecchi programmi per l'identificazione di ORF . Vedremo alle esercitazioni l'utilizzo di ORFfinder fornito dall'ncbi

ANNOTAZIONE DEI GENOMI

2. Metodi di annotazione di genomi procariotici



La **ricerca di omologie** prevede invece l'utilizzo di tecniche di comparazione tra il genoma in esame e genomi di organismi correlati già annotati, al fine di individuare le regioni del genoma che codificano per proteine simili o omologhe.

Questo permette di identificare nuovi ORF e di assegnare una funzione predittiva ai geni identificati.

In genere, entrambi i metodi vengono utilizzati in combinazione per ottenere una predizione più accurata degli ORF e delle loro funzioni.

Una volta individuati gli ORF, è possibile procedere con la loro annotazione funzionale attraverso l'uso di diversi strumenti di annotazione automatica o manuale, come descritto in precedenza.

ANNOTAZIONE DEI GENOMI

- ORFfinder



ORFfinder è uno strumento di bioinformatica utilizzato per la ricerca di open reading frame (ORF) all'interno di sequenze nucleotidiche.

ORFfinder cerca gli ORF nella sequenza di input, identificando i codoni di inizio e di stop. L'output di ORFfinder comprende la posizione degli ORF all'interno della sequenza, la loro lunghezza e l'orientamento, insieme alla sequenza nucleotidica degli ORF e la loro traduzione in sequenze di amminoacidi.

ORFfinder è uno strumento ampiamente utilizzato per la predizione degli ORF in sequenze di genomi procariotici, ma può essere utilizzato anche per sequenze di genomi eucariotici e per sequenze di cDNA. È un metodo rapido ed efficace per l'identificazione degli ORF e può essere utilizzato in combinazione con altri strumenti di annotazione per una completa annotazione del genoma.

<https://www.ncbi.nlm.nih.gov/orffinder/>

ANNOTAZIONE DEI GENOMI

- ORFfinder

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for [Linux x64](#).



ANNOTAZIONE DEI GENOMI

- PGAP



PGAP (Prokaryotic Genome Annotation Pipeline) è uno strumento di annotazione automatica sviluppato dal National Center for Biotechnology Information (NCBI) per i genomi procariotici.

PGAP utilizza una combinazione di approcci di annotazione basati su omologia, predizione di geni, analisi di interruzioni di frame di lettura aperti (ORF), predizione di funzione e caratterizzazione di elementi di inserzione.

PGAP fornisce anche la possibilità di revisionare manualmente l'annotazione automatica. Il risultato finale dell'annotazione è un file di testo in formato GenBank pronto per essere caricato nei database pubblici.

ncbi/pgap

NCBI Prokaryotic Genome Annotation Pipeline



ANNOTAZIONE DEI GENOMI

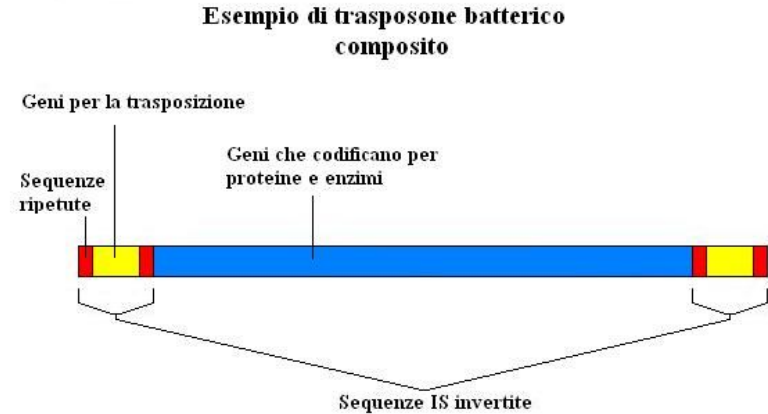
- Sfide nell'annotazione dei genomi procariotici



L'annotazione dei genomi procariotici può presentare diverse sfide che possono influenzare la qualità dell'annotazione e la comprensione della biologia.

Una delle principali sfide è rappresentata dalla presenza di elementi di trasposizione, come i **trasposoni**, che possono causare la frammentazione del genoma in piccole sequenze o la loro ripetizione all'interno del genoma.

Questi elementi possono anche contenere sequenze di geni che possono essere erroneamente annotati, portando a errori nell'identificazione dei geni funzionali.



ANNOTAZIONE DEI GENOMI

2. Metodi di annotazione di genomi eucariotici



La annotazione di genomi **eucariotici** è un processo più complesso rispetto alla annotazione di genomi procariotici a causa delle dimensioni maggiori dei genomi e della complessità della struttura del genoma eucariotico.

Tra i metodi più usati:

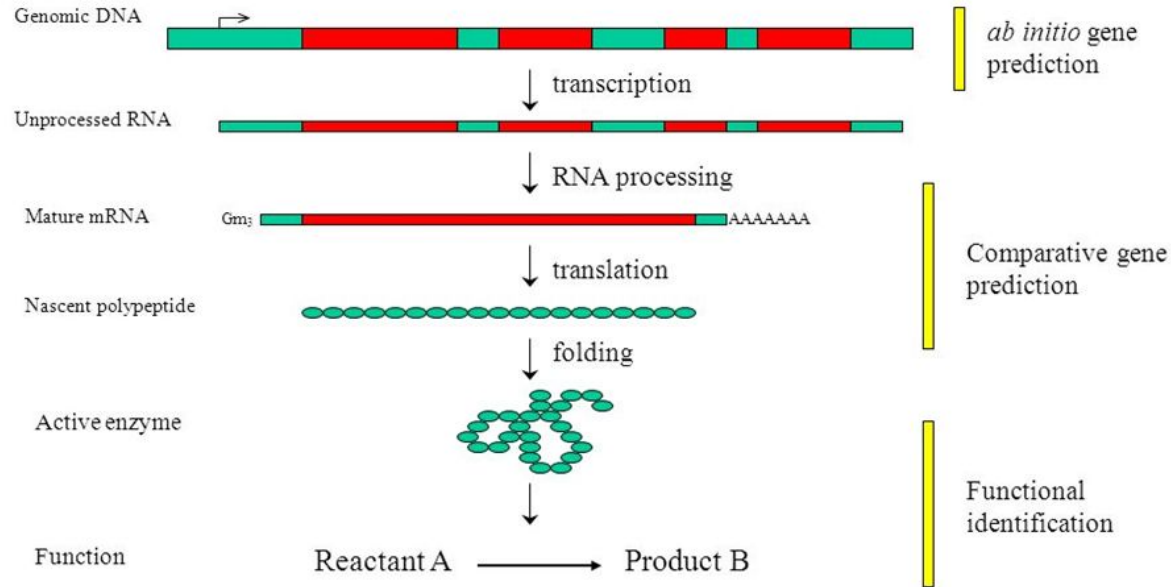
Allineamento di sequenze: L'allineamento di sequenze viene utilizzato per identificare sequenze simili a quelle già note. Questo metodo è utile per identificare i geni che hanno omologia con sequenze già caratterizzate, come ad esempio quelli presenti nei database pubblici.

ANNOTAZIONE DEI GENOMI

2. Metodi di annotazione di genomi eucariotici



Annotation of eukaryotic genomes



ANNOTAZIONE DEI GENOMI

2. Metodi di annotazione di genomi eucariotici



Il modo più semplice e affidabile di annotare i geni in un genoma eucariotico consiste nell'allineare alla sequenza genomica le sequenze di proteine o di trascritti (generalmente nella forma di cDNA) per identificare i probabili siti di trascrizione. Quando le sequenze di queste proteine o cDNA provengono da campioni della specie il cui genoma si vuole annotare si parla di **metodi diretti** o di allineamento in **cis**.

Gli strumenti più usati per effettuare l'annotazione sono il **Blast** o il **Blat**, quest'ultimo fornito dalla Università della California UCSC.

Questi algoritmi devono trovare il migliore allineamento della sequenza query, una proteina o un cDNA, traducendo la sequenza genomica o retro-traducendo la sequenza proteica e considerando che la sequenza query darà luogo ad allineamenti locali corrispondenti agli esoni che l'hanno codificata. Nel caso di allineamento in-cis ci si aspetta una identità perfetta o quasi perfetta.

ANNOTAZIONE DEI GENOMI

2. Metodi di annotazione di genomi eucariotici



Nel caso in cui non siano disponibili le sequenze note di proteine o cDNA per alcuni organismi, si può cercare di mappare sul genoma di una specie le sequenze di proteine o trascritti ottenute per una specie diversa ma sufficientemente vicina evolutivamente per cui ci si possa aspettare di trovare chiare relazioni di omologia.

In questo caso si parla di “allineamenti in **trans**”.

ANNOTAZIONE DEI GENOMI



- Metodi basati sull'allineamento

Quindi ricapitolando...

Metodi Diretti:

- Mediante ricerca di match perfetti o quasi perfetti con EST, cDNA o sequenze proteiche dello stesso organismo (allineamento **cis**)

Metodi Indiretti

- Mediante ricerca di somiglianze con un gene noto (allineamento **trans**);
- Mediante ricerca di strutture simili a un gene ideale (**ab initio** o de novo)
 - Metodi basati sul contesto (contenuto in **GC**, frequenze dei codoni, etc.)
 - Metodi basati su segnali (siti accettori e donatori di **splicing**, promotori, codoni di start/stop, segnali di poliadenilazione, etc.)

ANNOTAZIONE DEI GENOMI



- Metodi diretti

Supponiamo di avere una regione del genoma eucariotico in cui è presente una sequenza di trascrizione nota, come un promotore o un enhancer.

Utilizzando l'allineamento in cis, si può cercare di identificare eventuali esoni nella stessa regione che siano omologhi a sequenze note di geni.

Una volta identificati questi esoni, è possibile utilizzare ulteriori informazioni, come la presenza di introni o di domini proteici specifici, per confermare la loro identità e annotarli come geni.

BLAST può essere utilizzato anche come metodo diretto per l'annotazione di un genoma eucariotico. In questo caso, le sequenze dei geni noti o delle proteine vengono utilizzate come query in BLAST per cercare omologie nei database target. Le sequenze omologhe trovate vengono quindi annotate come possibili geni o regioni funzionali.

ANNOTAZIONE DEI GENOMI



- Metodi diretti

Tuttavia, l'uso di BLAST come metodo diretto presenta alcune limitazioni, tra cui la possibilità di ottenere risultati non specifici a causa della presenza di famiglie di geni altamente omologhe, l'incapacità di trovare regioni intergeniche e la dipendenza dalla qualità delle sequenze di riferimento utilizzate come query.

Per superare queste limitazioni, spesso è necessario combinare l'approccio BLAST con altri metodi di annotazione.

BLAT (BLAST-like alignment tool), un algoritmo di allineamento che permette di confrontare sequenze di DNA o RNA di lunghezza variabile. BLAT è stato sviluppato specificamente per l'allineamento di sequenze di genomi eucariotici, e può essere utilizzato per annotare i geni attraverso l'allineamento di trascritti noti o di proteine omologhe.

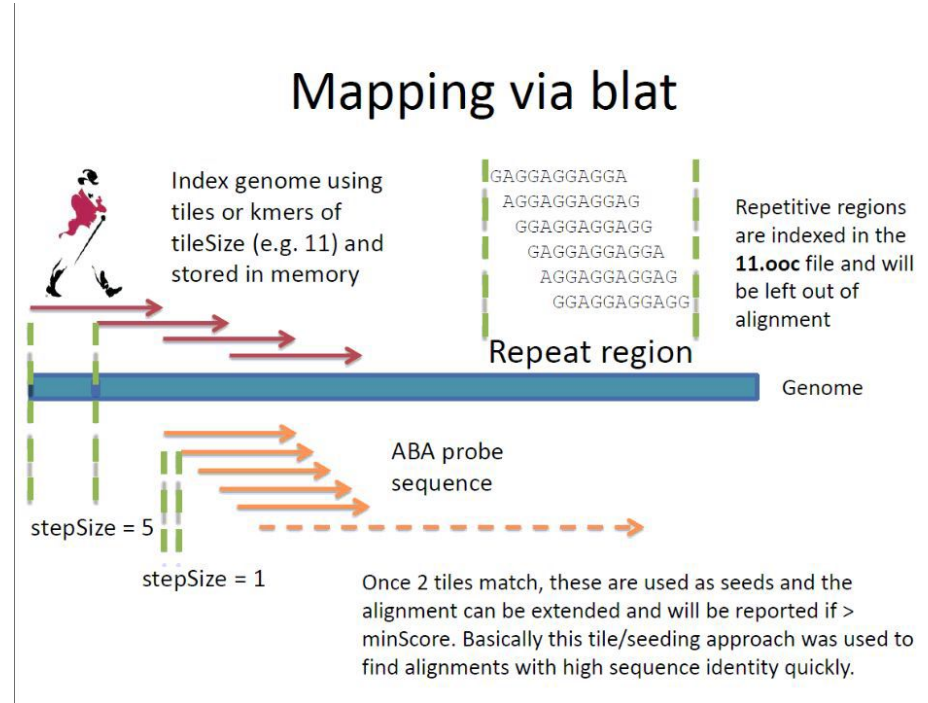
ANNOTAZIONE DEI GENOMI

- Metodi diretti



In pratica, BLAT utilizza una combinazione di allineamento globale e locale per cercare di identificare sequenze omologhe tra il genoma di riferimento e la sequenza di confronto.

Una volta identificati i punti di omologia, BLAT utilizza un modello statistico per valutare la significatività dell'allineamento e identificare eventuali esoni.



ANNOTAZIONE DEI GENOMI

- Metodi diretti



BLAT è uno strumento molto utilizzato nell'annotazione di genomi eucariotici, in particolare quando si ha a che fare con specie poco studiate o con genomi molto grandi.

“BLAT”

Selezionate il genoma che vi interessa

Incollate la sequenza da cercare

Human BLAT Search

BLAT Search Genome

Genome: Assembly: Query type: Sort output: Output type:

Human Mar 2006 BLAT's guess query.score hyperlink

submit I'm feeling lucky clear

ANNOTAZIONE DEI GENOMI



- Metodi indiretti

–Mediante ricerca di somiglianze con un gene noto (allineamento **trans**);

Questo metodo si basa sull'ipotesi che i **geni** di un organismo siano **conservati** all'interno di una data famiglia genica e che quindi la sequenza nucleotidica di un gene conservato in un organismo possa essere utilizzata per identificare geni simili in un altro organismo.

Il metodo più comune per l'allineamento trans è l'uso del programma **BLAST** (Basic Local Alignment Search Tool), che confronta la sequenza nucleotidica o proteica di un gene sconosciuto con le sequenze di geni noti contenuti in database pubblici come **Nucleotide (GenBank)** o **UniProt**.

Il risultato dell'allineamento viene espresso come un punteggio di similarità (**score**) che indica quanto la sequenza del gene sconosciuto è simile a quella del gene noto.

ANNOTAZIONE DEI GENOMI



- Metodi indiretti

-Mediante ricerca di strutture simili a un gene ideale (**ab initio** o de novo)

Metodi di annotazione mediante allineamenti in cis o trans possono non avere a disposizione dati sufficienti per annotare su un genoma l'insieme di geni in esso contenuti.

Si possono utilizzare in questi casi i metodi predittivi, detti metodi di annotazione de-novo o ab initio che ricercano lungo la sequenza genomica regioni geniche in base a modelli statistici.

L'approccio ab initio si basa su algoritmi di predizione delle caratteristiche dei geni, come la lunghezza, la presenza di siti di splicing e l'identificazione di regioni codificanti, che vengono utilizzati per identificare le regioni che contengono i geni. Questi algoritmi utilizzano modelli di Markov nascosti o altri metodi di apprendimento automatico per identificare i confini dei geni.

ANNOTAZIONE DEI GENOMI

- Metodi indiretti



Il problema dell'identificazione di geni in una sequenza genomica può essere ridotto all'identificazione gli intervalli che delimitano gli esoni putativi e le altre regioni della struttura del gene.

Metodi di predizione di geni ab inizio si basano generalmente sui segnali che permettono l'identificazione di un gene e della sua struttura (ossia dei siti di splicing, dei codoni di inizio fine traduzione, ecc.) e su modelli statistici che incorporano questi segnali.

ANNOTAZIONE DEI GENOMI



- Metodi indiretti

I segnali utilizzati per costruire un modello di gene possono essere di varia natura.

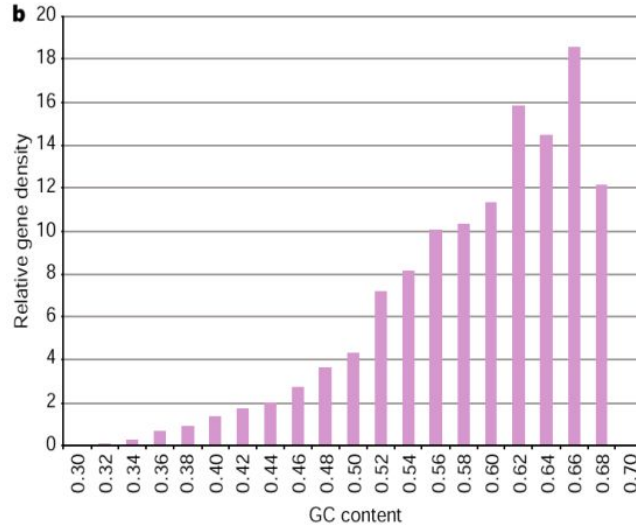
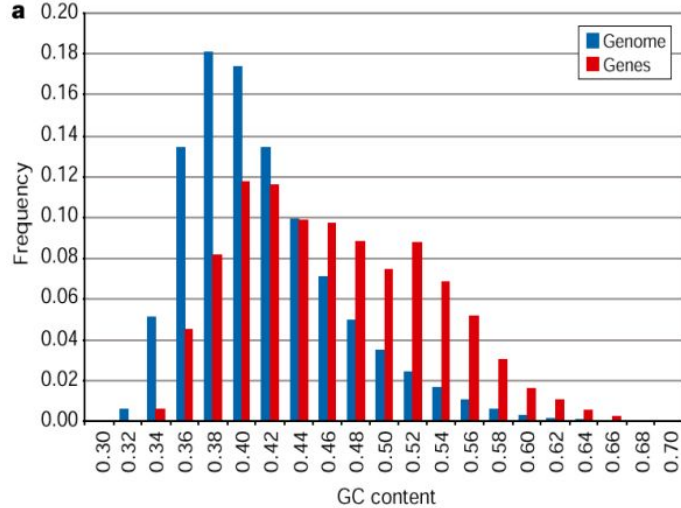
Alcuni sono descrivibili mediante sequenze consenso o matrici che riportano frequenze nucleotidiche in posizioni specifiche di sequenze associate ad una funzione. Altri segnali possono essere usati per identificare un gene nella sequenza del genoma e riguardano le lunghezze di esoni e introni per i quali sono note distribuzioni caratteristiche che possono essere differenti da specie a specie.

Lunghezze di esoni e introni sono anche dipendenti da caratteristiche di composizione della sequenza: per esempio introni di geni trovati regioni ricche di G+C sono mediamente più corti di quelli regioni povere in G+C.

ANNOTAZIONE DEI GENOMI

- Metodi indiretti

Contenuto in G+C



ANNOTAZIONE DEI GENOMI

- Metodi basati sulla ricerca per omologia



InterPro è un database di famiglie di proteine e domini proteici omologhi, che possono essere utilizzati per l'annotazione di genomi eucariotici. InterPro integra le informazioni provenienti da diverse fonti di dati, tra cui Pfam, PRINTS, ProDom, SMART e PROSITE, per identificare i domini proteici e le famiglie proteiche presenti nelle sequenze proteiche.

Per utilizzare InterPro nell'annotazione dei genomi eucariotici, si possono utilizzare programmi come InterProScan, che analizza le sequenze proteiche e le confronta con le informazioni contenute in InterPro, per identificare i domini proteici e le famiglie proteiche presenti nelle sequenze.



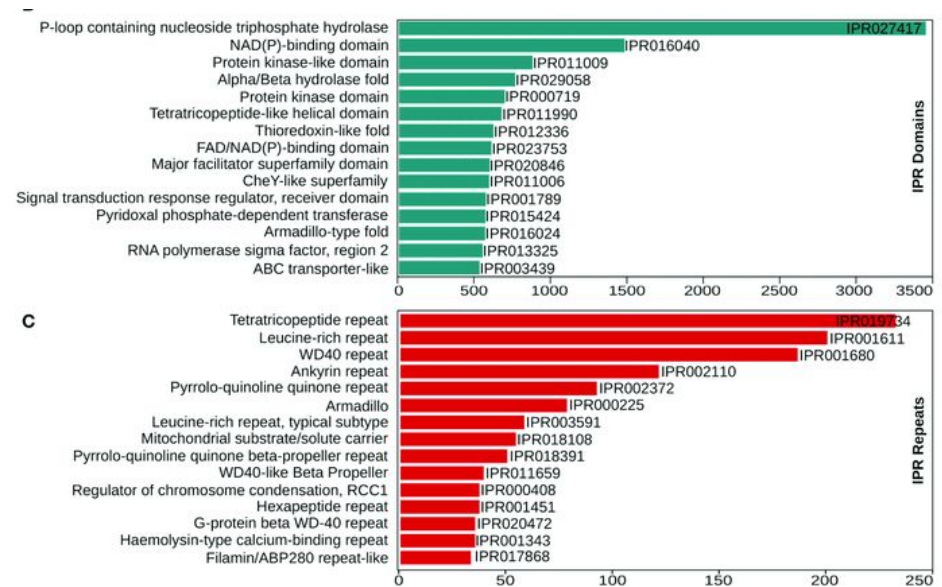
ANNOTAZIONE DEI GENOMI

- Metodi basati sulla ricerca per omologia



L'utilizzo di **InterPro** nella ricerca di famiglie di geni e proteine omologhe è particolarmente utile perché molte famiglie di proteine hanno funzioni conservate tra diverse specie, consentendo di identificare rapidamente potenziali funzioni di nuovi geni e proteine in un genoma eucariotico.

Inoltre, l'utilizzo di InterPro consente di identificare i domini proteici presenti in una proteina, che possono essere utili per predire le funzioni e le interazioni proteiche.



ANNOTAZIONE DEI GENOMI



- Metodi basati sulla ricerca per omologia

In generale, si può dire che l'annotazione basata sull'allineamento di sequenze di nucleotidi o proteine noti sia una forma di **annotazione per omologia**, poiché si cerca di identificare una corrispondenza tra le sequenze annotando i geni in modo simile ai geni noti.

L'annotazione basata sulla ricerca di domini o di motivi proteici, invece, è una forma di **annotazione funzionale**, poiché si cerca di identificare la funzione del gene sulla base della presenza di sequenze o strutture noti associati a una determinata funzione.

- Tuttavia, è importante notare che i due metodi non sono necessariamente esclusivi l'uno dall'altro e spesso vengono utilizzati in combinazione per ottenere un'annotazione più completa e accurata dei genomi.
- Inoltre, l'annotazione funzionale per omologia può anche essere basata sull'analisi di altre caratteristiche, come la localizzazione cellulare o la regolazione dell'espressione genica, piuttosto che solo sui domini proteici o motivi.

ANNOTAZIONE DEI GENOMI



- Importanza e sfide nell'annotazione dei genomi eucariotici

Ci sono molti esempi di successo nell'annotazione dei genomi eucariotici che hanno portato a una migliore comprensione della biologia degli organismi.

Primo fra tutti l'annotazione del **genoma umano**: L'annotazione del genoma umano ha portato a una migliore comprensione dei meccanismi di regolazione genica e dei processi patologici come il cancro e le malattie genetiche.

Annotazione del genoma di **Drosophila melanogaster**: L'annotazione del genoma di *Drosophila melanogaster*, una mosca della frutta, ha portato alla scoperta di nuovi geni coinvolti nella regolazione dello sviluppo degli animali e nella risposta a fattori ambientali come la temperatura.

ANNOTAZIONE DEI GENOMI



- Importanza e sfide nell'annotazione dei genomi eucariotici

L'annotazione dei genomi eucariotici presenta diverse sfide rispetto all'annotazione dei genomi procariotici. Alcune delle principali sfide sono:

- Presenza di **introni**: gli introni sono regioni non codificanti presenti nei geni eucariotici che vengono rimossi durante la trascrizione per generare la sequenza di mRNA matura. L'annotazione dei genomi eucariotici deve quindi considerare la presenza degli introni e identificare correttamente i confini esatti tra esoni e introni.
- **Splicing Alternativo**: in molti geni eucariotici, gli esoni possono essere assemblati in diverse combinazioni attraverso il processo di alternative splicing, generando molteplici varianti di mRNA e di proteine. L'annotazione dei genomi eucariotici deve quindi considerare anche la presenza di alternative splicing e cercare di identificare tutte le possibili varianti di trascrizione e traduzione.

ANNOTAZIONE DEI GENOMI



- Importanza e sfide nell'annotazione dei genomi eucariotici
- Presenza di **regioni** altamente **ripetitive**: i genomi eucariotici possono contenere regioni altamente ripetitive, come le sequenze di DNA satellite e le sequenze ribosomali, che possono essere difficili da annotare correttamente a causa della loro alta omologia e della loro struttura altamente ripetitiva. L'annotazione dei genomi eucariotici deve quindi utilizzare approcci specifici per gestire queste regioni ripetitive e cercare di identificare correttamente i confini tra i geni all'interno di queste regioni.
- Variazione del genoma tra le diverse specie: i genomi eucariotici possono variare notevolmente tra le diverse specie, anche all'interno dello stesso regno (ad esempio tra i mammiferi). L'annotazione dei genomi eucariotici deve quindi tener conto di queste variazioni e utilizzare approcci basati su omologia con attenzione, in modo da non sovrastimare o sottostimare la presenza di geni.

ANNOTAZIONE DEI GENOMI

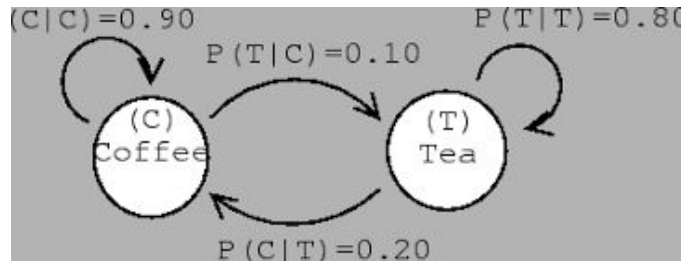
3. Modelli Markoviani



I modelli markoviani sono una classe di **modelli probabilistici** utilizzati in bioinformatica per analizzare e predire sequenze di dati, tra cui sequenze di nucleotidi e sequenze di proteine.

In particolare, i modelli markoviani assumono che la probabilità di un evento dipenda solo dall'evento immediatamente precedente, o da un numero limitato di eventi precedenti.

In altre parole, la probabilità di un evento successivo dipende solo dallo stato corrente del sistema e non dai suoi stati passati.



ANNOTAZIONE DEI GENOMI

3. Modelli Markoviani



In bioinformatica, i modelli markoviani sono spesso utilizzati per l'annotazione genica, ovvero per identificare i geni all'interno di una sequenza di DNA.

Ciò è possibile poiché i modelli markoviani possono essere **addestrati** su sequenze di DNA note per identificare i pattern comuni associati ai geni e utilizzarli per identificare i geni in sequenze di DNA sconosciute.

Una classe di modelli markoviani sono i **modelli markoviani nascosti (HMM**, acronimo di Hidden Markov Model): permettono di analizzare sequenze di dati quando l'informazione sullo stato del sistema non è completamente osservabile, ma piuttosto **parzialmente osservabile** o nascosta.

ANNOTAZIONE DEI GENOMI

3. Modelli Markoviani



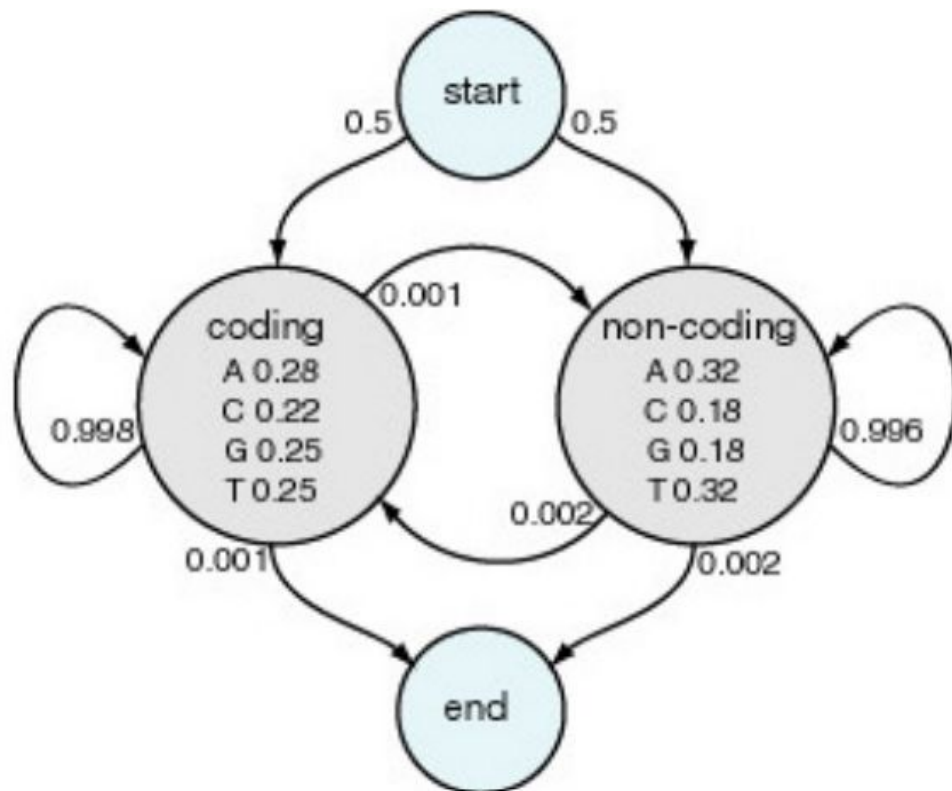
Per fare un esempio, immagina di avere una sequenza di lettere, dove alcune lettere sono "nascoste", ovvero non sono visibili ma influenzano le lettere che sono visibili. Gli HMM sono in grado di modellare questi tipi di sequenze, prendendo in considerazione sia le lettere visibili che quelle nascoste.

Gli HMM vengono utilizzati per identificare i geni all'interno di una sequenza di DNA o di proteine.

Questo è possibile perché gli HMM possono essere **addestrati** su sequenze di DNA o proteine note per **identificare** i pattern comuni associati ai geni e utilizzarli per identificare i geni in sequenze di DNA o proteine sconosciute.

ANNOTAZIONE DEI GENOMI

3. Modelli Markoviani



Questo è un modello semplice di HMM in cui sono presenti due stati, uno codificante e uno non codificante, e delle probabilità di transizione da uno stato all'altro.

A ciascuno di questi due stati sono associate delle probabilità di ammissione dei quattro nucleotidi, calcolate come frequenza dei nucleotidi in regioni codificanti e sequenze non codificanti in una serie di geni noti.

ANNOTAZIONE DEI GENOMI

3. Annotazione dell'RNA



L'annotazione dell'RNA è importante perché ci permette di **identificare** e caratterizzare i **trascritti** presenti in un campione biologico.

Con l'annotazione dell'RNA possiamo identificare i **geni** che vengono trascritti, la loro struttura (ad esempio la presenza di introni ed esoni) e la loro espressione relativa. Questa informazione è fondamentale per comprendere la **regolazione** dell'espressione genica, la funzione dei geni e la loro relazione con le diverse patologie.

Inoltre, l'annotazione dell'RNA è fondamentale per poter utilizzare le informazioni generate dalle tecnologie di sequenziamento ad alta capacità, come il RNA-seq, che permettono di ottenere una visione completa del trascrittoma di un organismo.

ANNOTAZIONE DEI GENOMI

3. Annotazione dell'RNA



Il processo di annotazione dell'RNA inizia con la raccolta dei dati sperimentali sull'RNA, come ad esempio i dati di RNA-seq, che forniscono informazioni sulla presenza, l'abbondanza e la struttura dei trascritti di RNA.

Successivamente, i dati di RNA-seq vengono processati e analizzati per identificare i trascritti di RNA e le loro varianti di splicing. Questo può essere fatto utilizzando strumenti di assemblaggio trascrittomici che consentono di ricostruire i trascritti di RNA a partire dai dati di RNA-seq.

Una volta che i trascritti di RNA sono stati identificati, vengono **annotati** utilizzando informazioni provenienti da diverse fonti, come ad esempio la presenza di motivi funzionali, l'analisi della struttura secondaria e la conferma sperimentale.

ANNOTAZIONE DEI GENOMI

3. Annotazione dell'RNA



L'annotazione dell'RNA può essere effettuata utilizzando diverse strategie:

- Una delle principali strategie consiste nel cercare **similitudini** con RNA noti presenti in banche dati pubbliche come GenBank o RefSeq. Ciò può essere fatto utilizzando strumenti di allineamento come BLAST o programmi di allineamento di sequenze multiple come ClustalW.
- In alternativa, la funzione dell'RNA può essere predetta in base alla **struttura tridimensionale** della molecola. Questo può essere fatto utilizzando tecniche di modellizzazione di strutture RNA come RNAfold e Mfold.
- Inoltre, alcuni strumenti utilizzano approcci basati su **omologia funzionale**, cercando similitudini con RNA noti che svolgono funzioni simili.

ANNOTAZIONE DEI GENOMI

3. Annotazione dell'RNA



- Altri approcci includono l'utilizzo di tecniche di **predizione** di **strutture** secondarie e terziarie dell'RNA, nonché l'utilizzo di algoritmi di machine learning per predire funzioni e **interazioni** dell'RNA in base a determinati pattern di sequenza e struttura.

Una volta annotate le funzioni dell'RNA, queste informazioni possono essere utilizzate per comprendere la biologia dell'organismo di interesse.

Ad esempio, l'identificazione di RNA codificanti proteine può essere utilizzata per identificare nuovi bersagli terapeutici per le malattie.

Inoltre, l'annotazione dell'RNA può aiutare a comprendere i meccanismi molecolari coinvolti in processi biologici come lo sviluppo, la differenziazione cellulare e la risposta allo stress.

ANNOTAZIONE DEI GENOMI

3. Annotazione dell'RNA



Recenti studi sostengono che una proporzione preponderante del trascrittoma umano è costituita da trascritti non codificanti di dimensioni superiori a 200 bp, i cosiddetti long non coding RNA (lncRNA).

lncRNA (long non-coding RNA) sono una classe di RNA che non codifica per proteine ma svolgono funzioni regolatorie nell'organismo. Negli ultimi anni sono stati oggetto di crescente interesse perché coinvolti in numerosi processi biologici, tra cui la regolazione dell'espressione genica, la modulazione dell'attività cellulare e lo sviluppo di patologie umane.

L'annotazione dei lncRNA può essere fatta utilizzando gli stessi approcci usati per l'annotazione dell'RNA in generale, come la ricerca di somiglianze con sequenze di RNA noti o la predizione di funzioni sulla base della struttura dell'RNA.

ANNOTAZIONE DEI GENOMI

3. Annotazione dell'RNA



L'annotazione dei lncRNA si basa sull'utilizzo di strumenti bioinformatici dedicati che valutano se una sequenza trascritta contenga il potenziale per produrre una proteina, il cosiddetto Coding Potential Assessment (**CPA**).

Quindi nell'annotazione dei lncRNA non si tratta solo di capire se un eventuale trascritto contenga un meno una ORF, molti lncRNA infatti contengono ORF, potenzialmente di dimensioni non trascurabili, ma si differenziano dai trascritti per proteina per il fatto che:

1. queste ORF non sono conservate tra specie filogeneticamente vicine,
2. hanno una composizione in base generalmente anomala e
3. raramente contengono domini proteici che possano suggerire una qualche evidenza funzionale.

La determinazione del CPA è quindi basata su caratteristiche composizionali (contenuto in CG) o evolutive (conservazione con altre specie).

ANNOTAZIONE DEI GENOMI

3. Annotazione dell'RNA



L'annotazione dei lncRNA si articola in due fasi:

1. inizialmente trascrittoma viene assemblato per ricostruire le sequenze (possibilmente complete) dei trascritti che lo compongono. La ricostruzione dei trascritti completi può essere guidata dal corrispondente genoma di riferimento.

Se questo non fosse disponibile possono essere usati varianti di algoritmi di assemblaggio del genoma, considerando la possibilità di eventi di splicing alternativo, sono in grado di effettuare un assemblaggio de-novo dei trascritti (un programma molto usato è Trinity);

2. successivamente i trascritti completi ricostruiti vengono analizzati con uno o più strumenti per il calcolo del CPA per determinare quali trascritti siano potenzialmente codificanti le proteine e quali no