

Laboratorio di Bioinformatica

Dispense del corso - Modulo 2

Chiara Solito

Corso di Laurea in Bioinformatica
Università degli studi di Verona
A.A. 2021/22

La presente è una dispensa riguardante il corso di **Laboratorio di Bioinformatica** del CdS in Bioinformatica (Università degli Studi di Verona). Per la stesura di questa dispensa si è fatta fede al materiale didattico fornito direttamente dal professore nell'Anno Accademico 2021/2022. Eventuali variazioni al programma successive al suddetto anno non saranno quindi incluse.

Insieme a questo documento in formato PDF viene fornito anche il codice \LaTeX con cui è stato generato.

Contents

1 Introduzione

La Bioinformatica è una disciplina molto recente, il nostro corso è il primo in laurea triennale in Italia.

Wikipedia: La bioinformatica è una disciplina scientifica dedicata alla risoluzione di problemi biologici a livello molecolare con metodi informatici.

A cosa pensiamo quando parliamo di Bioinformatica:

- Microarraay
- Genetica Forense
- Basi di Dati
- 3D Modeling
- Interazione Molecolare
- Antropologia Evolutiva
- Disegno di Molecole
- Antropologia Evolutiva

1.1 Cos'è quindi la Bioinformatica?

La bioinformatica è la disciplina scientifica che cerca di risolvere problemi biologici mediante l'elaborazione informatica dell'informazione proveniente diretta o indirettamente da essere viventi. È nata per permettere all'informatica di organizzare e rendere disponibili i dati biologici. La Bioinformatica è una scienza che risolve problemi biologici attraverso l'elaborazione di materiale proveniente da esseri viventi. Questo materiale è in genere sono **sequenze di DNA**. L'utilizzo dell'informatica è necessario per elaborare la grande quantità dei dati. La quantità di dati biologici a disposizione, è paragonabile solo all'astrofisica. Sono troppi per non affrontarli in maniera informatica.

I "fields" della bioinformatica:

- Genome Analysis
- Sequence Analysis
- Phylogenetics
- Gene Expression
- Systems Biology
- Data and Text Mining (algoritmi che leggono abstract o articoli)

La bioinformatica copre tutti questi campi.

Chiamata alle varianti: cosa cambia nella proteina o nel gene rispetto a quello di riferimento, e dopo un'analisi statistica riescono a collegarlo ad una malattia (qual è l'effetto della variante sulla proteina?)

Per questo è bene ricordare che l'informazione di partenza sono le sequenze proteiche o nucleotidiche. L'obiettivo ultimo è quello di estrarre informazioni dalle sequenze. Tali sequenze rappresentano l'informazione disponibile e possono essere:

- Sequenze genomiche
(DNA genomico: genomi, esomi o alcune regioni particolari del genoma).

- Sequenze proteiche
(cDNA cioè DNA retrotrascritto a partire da un mRNA)
- Immagini
(RX, TAC, MRI, US, ecc)
- Strutture 3D di proteine
(NMR, Cristallografia), biologia strutturale
- Informazioni provenienti da System Biology
Informazione di interazione tra molecole
- Informazioni di carattere evolutivo
- Pulsazioni, respiri, battiti cardiaci,...
- Concentrazioni di particelle nel sangue.

Una cosa molto recente è la medicina di precisione: una volta si parlava di medicina personale, conoscendo il sequenziamento genomico di una persona, si può conoscere tutta la sua storia clinica? Capiamo le malattie a livello personalizzato.

In questo modulo ci concentreremo sull'annotazione genomica e sulla bioinformatica strutturale e poi anche il risequenziamento del genoma.

Il Covid è stato un campo di prova molto importante. In questo modulo ci concentreremo sull'annotazione genomica e sulla bioinformatica strutturale e poi anche il risequenziamento del genoma.

1.2 Cos'è la Genomica?

La genomica è una branca della biologia molecolare che si occupa dello studio del genoma degli organismi viventi. In particolare si occupa della struttura, contenuto, funzione ed evoluzione del genoma. È una scienza che si basa sulla bioinformatica per l'elaborazione e la visualizzazione dell'enorme quantità di dati che produce.

Cosa studia la genomica?

- Estrazione e/o cattura di DNA da essere viventi.
- Sequenziamento del DNA con tecniche all'avanguardia come NGS (Next Generation Sequencing).
- Assemblaggio di genomi a partire da milioni di frammenti di DNA.
- Ri-sequenziamento di genomi.
- Allineamento Allineamento di frammenti frammenti di DNA a un genoma di riferimento riferimento.
- Annotazione di genomi.
- Annotazione funzionale di geni all'interno di un genoma.
- Analisi di espressione genica mediante sequenziamento dei trascritti (RNA-Seq).
- GWAS (Genome Wide Association Studies).
- Analisi di varianti tra genomi (Variant calling o Chiamata delle varianti)

2 Ottimizzazione del protocollo bioinformatico per l'annotazione di geni codificanti proteine in genomi complessi

Un genoma sequenziato ma non ben annotato non serve a molto.

Con l'avvento del sequenziamento con costi sempre più contenuti, il numero di genomi sequenziati sta incrementando considerevolmente. Lo scopo di conoscere la sequenza genomica è principalmente indirizzato a capire la funzionalità dei geni.

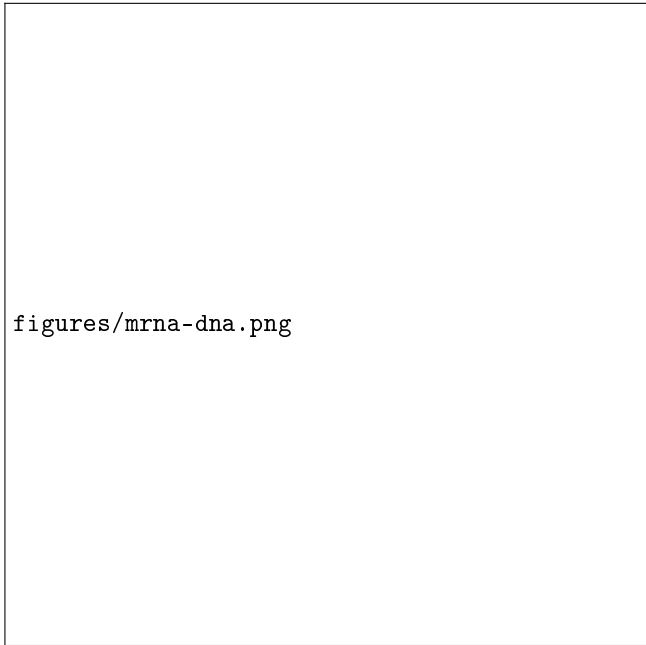
Il progetto "Encode" ha dato il via all'annotazione come la conosciamo, tramite la produzione di protocolli. Annotare un genoma significa conoscere la localizzazione, la struttura, la funzionalità di tutti gli elementi che compongono l'intero genoma. In pratica l'annotazione è quello che si deve fare dopo aver sequenziato un certo genoma. Gli elementi che vengono annotati sono:

1. Geni codificanti proteine
2. Geni non codificanti proteine
3. Elementi regolatori
4. Elementi ripetuti
5. Pseudogeni = geni che hanno perso la funzione codifica
6. Altri elementi

Come si cerca un gene all'interno del genoma? Faccio un allineamento: è la prima cosa che devo fare! Faccio un allineamento tramite BLAST (P o N o quello che è). Incollando la sequenza troverò delle sequenze simili nelle banche dati (**voglio mappare la regione**). Questo è uno dei metodi, il più semplice.

Nel corso del tempo si è capito che non basta annotare solo le zone esoniche ma bisogna considerare anche il resto come gli elementi regolatori. Con ri-sequenziamento del genoma di un paziente si intende mappare il DNA su un genoma di riferimento per confrontare le varianti manifestate. L'annotazione dei geni codificanti proteine, viene suddivisa in:


- Annotazione funzionale
Consiste nel caratterizzare ogni singolo gene, assegnando una funzione biologica a ogni proteina codificata dal gene stesso.
- Annotazione genica o semplicemente annotazione
consiste nel definire all'interno del genoma:
 - La localizzazione di ciascun gene.
- La struttura di ciascun gene (esoni, CDS, UTR).
 - Gli eventuali trascritti alternativi.



figures/mrna-dna.png

2.1 Il modello genico

In questo modulo ci occuperemo anche di studiare algoritmi di predizione di struttura di geni dalla loro sequenza. Quello che bisogna fare è utilizzare HMM per costruire un modello che descriva probabilisticamente le conformazioni possibili. Occorre operare sempre una divisione in training e test set. In ogni caso bisogna sempre partire dalla definizione di gene riassunta nell'immagine seguente:



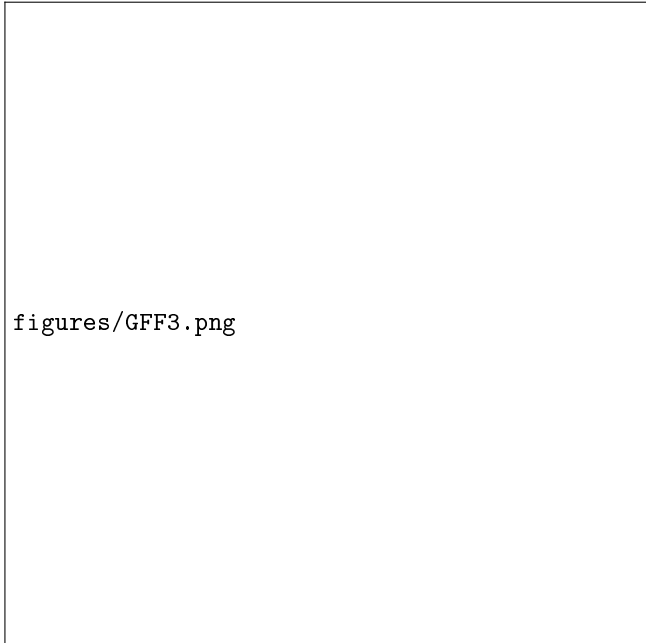
figures/model.png

Un gene codificante proteine è composto da diversi elementi:

- Esone: regione che viene mantenuta dopo la maturazione.
- Introne: regione che viene eliminata durante la maturazione.
- mRNA: RNA maturo, composto da esoni.

- CDS: regione codificante dell'mRNA.
- UTR: regione non tradotta dell'mRNA.

2.2 Formato di File di Annotazione GFF3



figures/GFF3.png

2.3 Metodi per l'identificazione dei geni codificanti proteine

- Metodi basati sull'allineamento delle evidenze sperimentali.
- Metodi basati sulla predizione genica ab initio.
- Metodi basati sulla predizione genica ab initio guidata da evidenze sperimentali.
- Metodi basati sul confronto tra genomi.

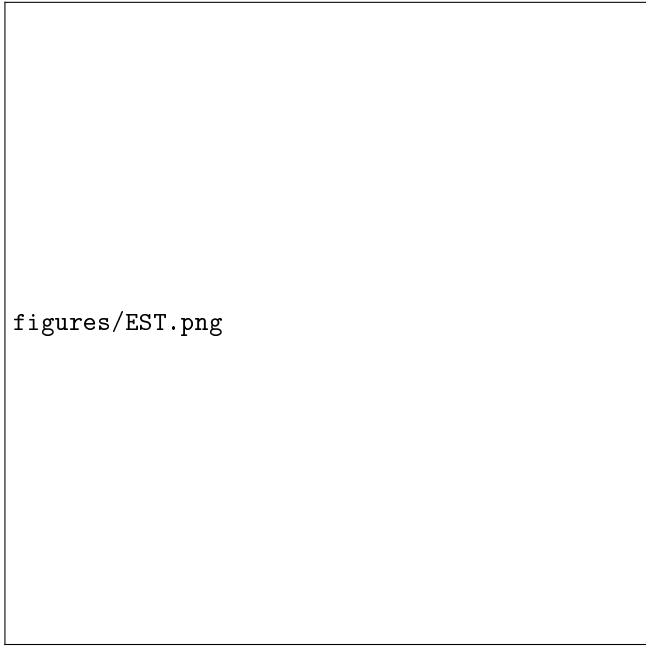
2.3.1 Metodi basati sull'allineamento delle evidenze sperimentali

Si possono utilizzare diverse evidenze sperimentali, che opportunamente elaborate e allineate al genoma permettono di identificare le regioni codificanti proteine:

- cDNA full-length: sequenze di RNA maturi (mRNA) retrotrascritti a cDNA, quindi completo di UTR e CDS.
- EST (Expressed Sequence Tags): brevi frammenti parziali, tra 400-800 bp, di mRNA retrotrascritti retrotrascritti a cDNA.
- Proteine omologhe: sequenze aminoacidiche corrispondenti a proteine omologhe di organismi evolutivamente vicini.
- Tiling arrays: microarray con sonde equamente spaziate su tutto il genoma, permettono l'identificazione di regione espresse mediante l'ibridazione di campione marcati.
- MPSS: Massively Parallel Signature Sequencing, piattaforma che analizza il livello di espressione e identifica una regione di 17-20 bp degli mRNA tramite sequenziamento.
- RNA-seq: frammenti di cDNA di lunghezza tra 50-150 bp che derivano dal sequenziamento shotgun di un intero trascrittoma.

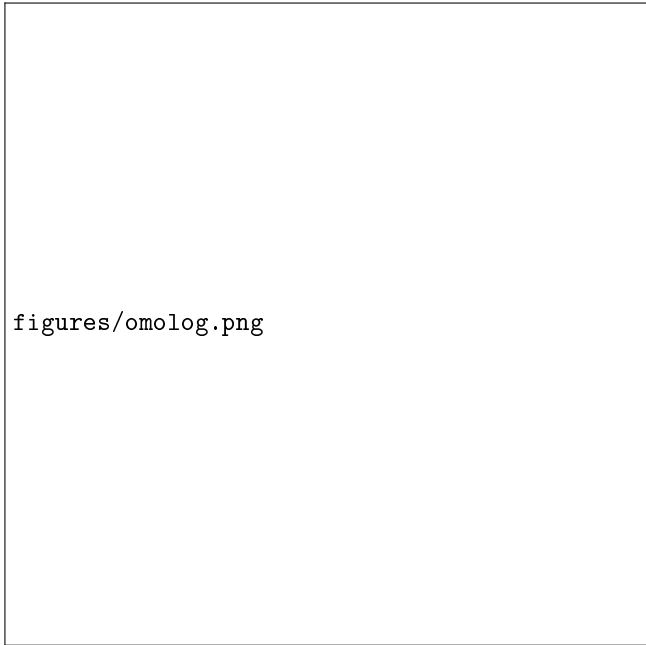
EST - Express Sequence Tag Sequenziamenti veloci di tessuti interi, in cui le etichette dei primi e degli ultimi nucleotidi sono ben sequenziati, in mezzo invece hanno errori. Però per annotare velocemente sono molto importanti.

Sono dei brevi frammenti di lunghezza tra 400-800 bp di cDNA ottenuto dalla retrotrascrizione di un frammento di RNA maturo.



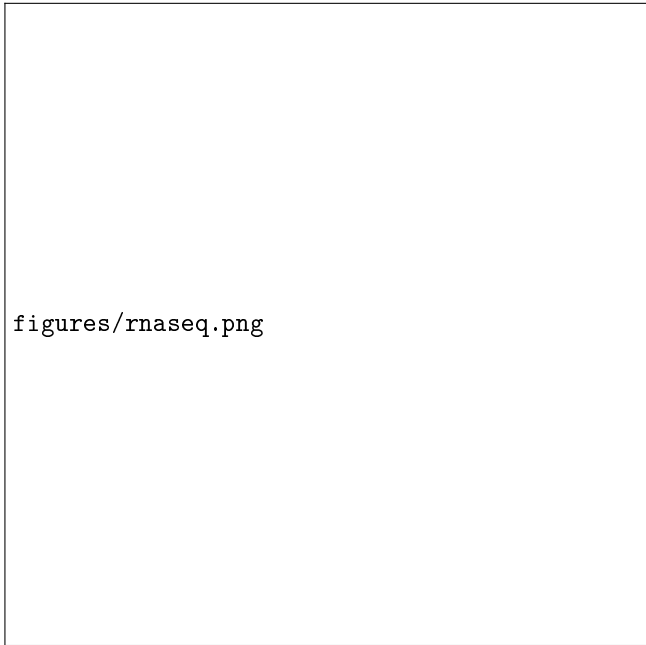
figures/EST.png

Proteine Omologhe Dalla sequenza proteica delle proteine si può risalire alla sequenza nucleotidica e quindi alla zona codificante (CDS) del gene che l'ha codificata.



figures/omolog.png

RNA Seq Sono sequenze di lunghezza tra 50-150 bp che derivano dal sequenziamento shotgun di un intero trascrittoma, cioè dalla retro-trascrizione di tutto l'RNA in cDNA di un particolare momento cellulare, poi spezzato e sequenziato con tecnologie NGS.



figures/rnaseq.png

2.3.2 Metodi basati sulla predizione genica ab initio

Per identificare le regioni codificanti i predittori utilizzano algoritmi e modelli matematici specifici che utilizzando informazione intrinseca dell'organismo analizzato cercano di identificare la localizzazione e la struttura dei geni.

- Sensori di segnale (*signal sensors*): permettono di identificare le giunzioni esone-introne e le estremità delle regioni codificanti.
- Sensori di contenuto (*content sensors*): permettono di identificare le regioni codificanti di lunghezza variabile.

I predittori hanno bisogno di dati di esempio per imparare le caratteristiche dell'organismo analizzato (dati di training) e dei dati di prova per valutare l'accuratezza delle predizioni (dati di test).

Predittore	Predizione ab initio	Predizione di geni eucarioti	Training in locale per nuovi genomi	Utilizzo di EST e Proteine per la predi- zione	Utilizzo di RNA-Seq per la predi- zione	Predizione degli UTR	Predizione dei trascritti alterna- tivi
Augustus	SI	SI	SI	SI	SI	SI	SI
Snap	SI	SI	SI	NO	NO	NO	NO
GeneMark-ES	SI	SI	NO	NO	NO	NO	NO
GeneID	SI	SI	SI	SI	SI	SI	SI
FGenesh	SI	SI	SI	NO	NO	NO	NO
Genescan	SI	SI	NO	SI	SI	SI	NO
MZEF	SI	SI	NO	NO	NO	NO	NO
mGene.NGS	SI	SI	SI	SI	SI	SI	NO
Contrast	SI	SI	SI	SI	NO	SI	NO
GrailExp	SI	SI	NO	SI	NO	SI	NO
TwinScan/N-Scan	SI	SI	SI	SI	NO	NO	SI

2.3.3 Metodi basati sulla predizione genica ab initio guidata da evidenze sperimentali

Predizione genica ab initio: utilizza dati di training che potrebbero non essere rappresentativi di tutti i geni del genoma.

Evidenze sperimentali: non coprono mai tutto il genoma, quindi non permettono l'annotazione completa di tutti i geni codificanti proteine.

I migliori metodi di predizione genica utilizzano una metodologia ibrida tra predizione genica ab initio e l'utilizzo degli allineamenti delle evidenze sperimentali:

- cDNA
- Proteine
- EST
- RNA-seq

2.4 Annotazione finale

Creazione di un consensus utilizzando le evidenze sperimentali e le predizioni geniche. Ciascuna evidenza viene pesata dando un peso maggiore ai dati sperimentali rispetto alle predizioni. Principali programmi di integrazione:

- Evidence Modeller
- JIGSAW
- GAZE

3 Pipeline automatizzata di annotazione genica

Basate su automazione di programmi di predizione e allineamento esistenti.

- Vantaggio: relativamente semplici da utilizzare.
- Svantaggio: consentono un controllo limitato dei passaggi intermedi dell'annotazione.

Pipeline di annotazione più utilizzate:

- PASA
- MAKER

3.1 Obiettivo

L'ottimizzazione del protocollo bioinformatico per l'annotazione dei geni codificanti proteine in genomi complessi. A questo scopo non verrà utilizzata una pipeline automatica di annotazione ma, attraverso la scelta di metriche adeguate, verrà valutato ogni singolo passaggio intermedio dell'annotazione in modo da fornire una procedura ottimizzata sulla base delle evidenze sperimentali a disposizione.

3.2 Genoma di riferimento

Genoma dell'organismo eucariote *Vitis vinifera*, versione V1 PN40024 12X del consorzio French-Italian Public Consortium for Grapevine Genome, con una dimensione di 487 Mb.

Motivi di questa scelta:

- Il genoma è disponibile
- Ci sono dati sperimentali disponibili (EST, 454, RNA-Seq, cDNA full-length)

3.3 Preparazione del dataset di riferimento

- 16.054 contig di cDNA full-length prodotte dal consorzio FrenchItalian Public Consortium for Grapevine Genome → 3752 cDNA non ridondanti.
- Rimozione delle sequenze con ORF non completa 3.436 → sequenze.
- Le 3.436 sequenze sono state suddivise in due gruppi in maniera casuale:
 - 936 sequenze di cDNA full-length → training.
 - 2.500 sequenze di cDNA full-length → test.



3.4 Preparazione ed allineamento delle evidenze sperimentali

- **EST:**
2.713.343 sequenze EST pubbliche (NCBI, Sequenziamento 454 + banca dati del consorzio).
Allineamento e generazione modelli genici con Gmap.
→ 1.649.082 trascritti putativi ridondanti (56.630 non ridondanti).
- **Proteine omologhe:**
Allineamento al genoma delle sequenze proteiche di tutto il database SWISSPROT utilizzando Blat, Blast e Genewise.
→ 22.355 trascritti putativi ridondanti (5.808 non ridondanti).
- **RNA-seq:**
114.726.580 reads RNA-seq sequenziati dal laboratorio di genomica dell'Università di Verona (pool di 45 campioni provenienti da 15 tessuti e organi a diversi stadi di sviluppo).
Allineamento e generazione modelli genici con suite Bowtie + Tophat + Cufflinks.
→ 40.324 trascritti putativi ridondanti (17.444 non ridondanti).

3.5 Statistiche generali degli allineamenti delle evidenze sperimentali

Statistiche generali	EST	Proteine omologhe	RNA-seq
Numero di modelli genici allineati	56.630	5.808	17.444
Numero di modelli genici multi esonici	19.485	3.175	17.366
Media della lunghezza dei modelli genici	1.034,12	874,42	2.236,89
N50 della lunghezza dei modelli genici	2.257	1.563	2.751
Media del numero di esoni per modello genico	3,30	4,39	6,75

Distribuzione della percentuale di sovrapposizione di nucleotidi tra allineamenti e riferimento.

figures/stats.png

4 Il covid

Gli articoli scientifici durante il covid Le tempistiche di review degli articoli non sono compatibili con quelle di lotta al covid, le ricerche portano via tanto tempo, una volta mandate a un giornale sono analizzate da un editor: decide se è d'accordo con gli obiettivi del giornale. Allora viene mandato ai reviewers, anonimi esperti che analizzano il lavoro, commentano, approvano, non approvano, ecc. In questo caso l'editor può bocciare, chiedere minor or major revisions oppure accettarlo. Si può fare avanti e indietro tra queste cose per mesi. Durante la lotta al covid si sono creati dei giornali di pre-print: appena finita la ricerca la si può depositare su queste banche dati, a cui viene assegnato un DOI (protegge la proprietà intellettuale). In quel caso è già a disposizione della comunità scientifica:

- Molti articoli venuti fuori nelle notizie erano in pre print, quindi non erano articoli revisionati manualmente
- L'informazione negli articoli di pre print è stata categorizzata da algoritmi di machine learning, e questo ha aiutato a comprendere articoli di aiuto alla lotta al covid.

Cosa c'è ancora da fare in merito al covid? Trovare un vaccino che sia pan-coronavirus. Trovare le regioni della proteina spike più conservate, che permettano di neutralizzare la proteina di tutte le varianti. Lo sviluppo di antivirali in grado di abbattere il virus, al momento ce n'è uno Pfizer (marzo 2022), che non colpiscono le proteine spike ma quelle di replicazione e funzionamento.

5 Predizione Genica

5.1 Informazioni di Base

Ogni predizione genica è un'ipotesi che aspetta di essere testata, i risultati dei test informano il nuovo set di ipotesi.

L'estrapolazione dei geni è un processo di:

- Identificare i fenomeni comuni nei geni noti
- Costruire un modello accurato che descriva il fenomeno
- Scannerizzare il genoma per identificare le regioni che matchano il mio modello
- Abbiamo una predizione, che dobbiamo validare e testare

5.1.1 Appocchi alla Gene Finding

- Metodi Diretti
Vado direttamente a cercare dove posso mappare le cose conosciute: cerco match esatti o simili di EST, cDNA o proteine
- Metodi Indiretti
 1. Cerco per omologia qualche gene particolare (homology)
 2. Cercare qualcosa che sia come tutti i geni (ab initio)
 3. Metodi che combinano i due precedenti

5.1.2 Procarioti

Identificare Open Reading Frames (ORF) lunghe, che possono codificare per proteine.

- Identificazione del codone di start ATG che massimizza la lunghezza dell'ORF
- Codoni di end (UAA, UAG, UGA)
- TATA box (TATAAT), la sequenza a -35 o siti di binding ribosomiale: siti di inizio trascrizionale o di traduzione,
- Bias di codone (codon usage)

GLIMMER, GeneMark. 90% sia per la sensibilità che la specificità.

Cosa possiamo misurare nei geni (e modellare)? Molta della conoscenza è prevenuta nei confronti delle caratteristiche protein-coding.

- ORF (Open Reading Frames): una sequenza definita da AUG inframe e il codone di stop, che a turno definiscono una sequenza aminoacidica putativa.
- Codon Usage: più frequentemente misurato tramite CAI (Codon Adaptation Index)
- Frequenze nucleotidiche e correlazioni (valore e struttura)
- Siti funzionali: siti di splicing, promotori, UTRs, siti di poliadenilazione

5.1.3 Codon Adaptation Index

I parametri sono determinati empiricamente dall'esame di un set "grande" di geni d'esempio. Questo ovviamente lo rende imperfetto:

- I geni di solito hanno codoni inusuali per un motivo
- La potenza predittiva è dipendente dalla lunghezza della sequenza.

5.1.4 Informazioni Generali sui Software di Predizione

- È, in generale, organismo-specifico
- Funziona meglio su geni che sono "ragionevolmente" simili a qualcosa visto in precedenza
- Trova regioni codificanti per proteine meglio di quelle non codificanti
- In assenza di informazioni esterne (dirette), forme alternative non verranno identificate
- È imperfetto!

5.2 Classi di Informazioni

5.2.1 Informazioni Estrinseche

formata da tutto ciò che la letteratura scientifica ci fornisce: EST, cDNA e prodotti proteici. Non è costituita soltanto dall'informazione circa la sequenza genica. Le informazioni estrinseche sono prelevate da programmi tipo BLAST, BLAT ecc. e da file in formato FASTA.

- BLAST family, FASTA, ecc.
Pros: veloci, statisticamente ben fondato
Cons: non c'è comprensione/modelli di struttura dei geni
- BLAST, Sim4, EST_GENOME, ecc.
Pros: le strutture geniche sono incorporate
Cons: splicing non-canonico, più lento di blast

5.2.2 Informazioni Intrinseche

Predittori genici. Possiamo avere:

- *de novo* che usano sequenze di uno o più genomi come unico input.
- *Ab initio*: non usano genomi informativi. Usano solo l'informazione della sequenza genica senza informazioni esterne.

Per quanto riguarda i metodi *ab initio*, abbiamo che essi simulano la trascrizione e lo splicing di un trascritto per ricavare informazioni circa i geni presenti. Per fare questo usano **sensori segnale** e **sensori contenuto**.

Sensori Segnale sono studiati per riconoscere codoni di inizio e di stop, enhancer, silencer, siti di splicing alternativo ecc. Per perseguire tale scopo utilizzano una PSSM ovvero un profilo che è una matrice 20xn (n lunghezza della query) che associa per ogni casella la frequenza di un certo AA in una certa posizione. Per addestrare quindi l'algoritmo *ab initio* a riconoscere ad esempio i promotori, gli fornisco una PSSM creata allineando sequenze annotate di promotori. Dopo averlo addestrato, lo applico sul genoma in questione. Di fatto il profilo non è altro che un HMM di livello zero cioè un modello a stati che mi identifica la probabilità di osservare un certo amminoacido in una certa posizione. Creare un profilo non fa altro che catturare il pattern di conservazione per le sequenze allineate.

Sensori Contenuto servono per riconoscere il contenuto di tale sequenza codificante.

Quando sono disponibili più genomi da confrontare è possibile stimare il grado di conservazione nelle sequenze.

6 Modelli di Biologia Computazionale

Nella "ricerca" di un gene i modelli possono essere visti come "generatori di sequenze" (attraverso gli HMM) o classificatori di sequenze" (attraverso le ANN).

6.1 HMM

Gli esoni e gli introni di una sequenza da modellare e poi da generare sono identificati da uno stato. La catena di acquisizione degli stati parte dal 5' fino al 3' in cui ogni base è generata grazie ad una matrice di emissione condizionata solo dallo stato corrente.

La transizione da uno stato all'altro dipende dalla matrice di transizione che è soggetta ai vincoli biologici tali per cui un ORF è adiacente ad un esone ecc. I valori da inserire nella matrice sono stimati dagli osservabili cioè guardando le sequenze.

Basic HMM

1. Insieme di possibili stati X : codone di inizio, esone, introne, accettore, codone di stop
2. Insieme di osservabili: A, C, T, G
3. Matrice delle transizioni φ
4. Matrice delle emissioni H (contiene le frequenze di un nucleotide in uno stato particolare)
5. Distribuzione di probabilità iniziali

HMM sono coinvolti nel processo generativo e abbiamo degli ordini di catena di Markov che sono il numero di posizioni precedenti sulla quale la corrente posizione dipende.

Ordine 0: mononucleotide

$$P_0(s) = p(t) \cdot p(t) \cdot p(a) \cdot p(c) \cdot p(g) \cdots = \prod_{i=1}^N p(s_i)$$

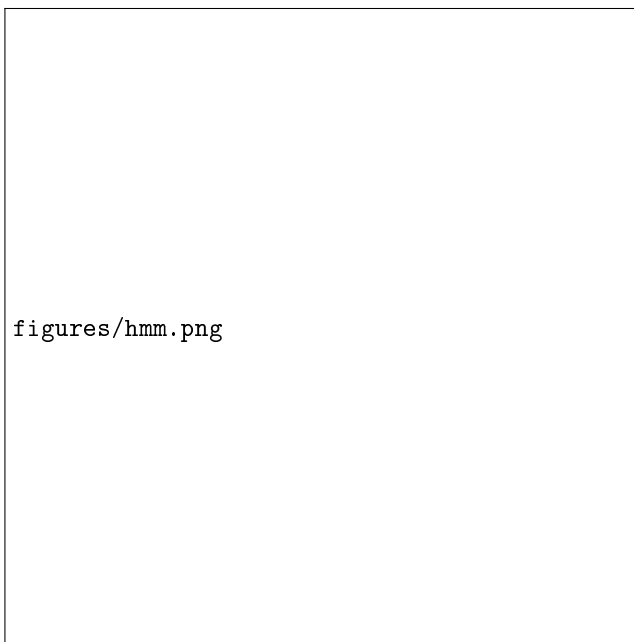
Ordine 1: dinucleotide

$$P_1(s) = p(t) \cdot p(t|t) \cdot p(a|t) \cdot p(c|a) \cdots = p(s_1) \prod_{i=2}^N p(s_i | s_{i-1})$$

Ordine 2: trinucleotide

$$P_2(s) = p(tt) \cdot p(a|tt) \cdot p(c|ta) \cdot p(g|ac) \cdots = p(s_1 s_2) \prod_{i=3}^N p(s_i | s_{i-2} s_{i-1})$$

Posso andare a costruire un modello addestrando l'HMM su sequenze non codificanti e codificanti. Ottengo ad esempio una cosa di questo tipo:



Nell'utilizzo di HMM è importante usare server che utilizzano tools che siano documentati. Se la documentazione non è presente non si utilizza quel tool.

7 Curve ROC: specificità e sensibilità

Questi due parametri servono per verificare se le performance di predizione sono state corrette o meno. La sensibilità è la frazione di geni conosciuti che sono stati correttamente predetti. La specificità è la frazione di geni predetti che corrispondono a veri geni.

BLAST in genere è poco sensibile ma molto specifico mentre PSI-BLAST è il contrario. Per il PSI-BLAST la specificità potrebbe essere condizionata dalla presenza dei falsi positivi che inducono ad una corruzione della ricerca.

Di seguito alcuni grafici che rappresentano le distribuzioni ideali e reali degli scores: