

# ANNOTAZIONE DI GENOMI \*

**Definizione:** conoscere la localizzazione, la struttura e la funzionalità di ogni elemento che compone il genome

base: modello genico

formato: GFF3

## annotazione funzionale

caratterizzare ogni singolo gene assegnando una funzione biologica ad ogni proteina codificata

## annotazione genica

definire all'interno del genome: localizzazione e struttura dei geni e trascritti alternativi

### 1. allineamento di evidenze sperimentali

elaborate ed allineate al genome (cDNA, EST, proteine omologhe, ...)

### metodi

3. metodi basati sulla predizione genica ab initio, guidati da evidenze sperimentali mix

### 2. metodi ab initio

NO EVIDENZE SPERIMENTALI  
uso algoritmi e modelli

sensori segnale  
giunzioni esone-introne

sensori contenuto  
regioni codificanti di lunghezza variabile

### Predittori

→ dati di esempio - training  
→ dati di valutazione - testing  
Augustus

$$\text{Accuratezza} = (\text{SN} + \delta\text{P})/2$$

Sensibilità → quanto sono in grado di predire

Specificità → quanto sono corretto

### 3 lvl d'indagine

- locus genico
- regioni esoniche
- giunzioni esone-introne

### Risultato

Creazione di un CONSENSUS: sequenza che rappresenta sequenze relative e allineate

## \*GENOME BROWSERS

Sono web server con tool di annotazione automatica

### → ENSEMBL

genome browser per vertebrati  
Annota i geni; calcola gli allineamenti BLAST, VEP, e fa Validazione

### → GENS CAN

esegue una predizione genica tramite HMM, e poi raffina per dati sperimentali

### → EST

seq. veloci di tessuti interi

# GENE PREDICTION

Definizione: trovare i geni di un genoma appena sequenziato

- È alla base dell'annotazione
- È un'ipotesi
- Esplica il concetto di bioinformatica
- Come "estrapolazione" di informazioni del genoma

Cosa?

Come?

metodi indiretti

ibridi

ab initio

metodi diretti

omologia

(mappa i geni che già conosco)

procarionti

ORF  
Glimmer  
algorithm

eucarioti

mi baso su diverse  
Classi di Informazioni

ESTRINSECA

INTRINSECA

oggi allo  
stato dell'arte  
troviamo

sensor:  
segnalet

ab  
initio

de novo

non sono veri  
ab initio

HMM

SVM, neural  
networks

Gene cattori  
Sequenze

sensor  
contenuto

TRAINING

N-CROSS  
FOLDING  
EVALUATION

VALUTA  
ZIONE

DISCUSSIONE

dataset

CONCLUSIONE

1. basic tools: GENOME BROWSERS
2. valutazione sistematica: SPECIFICITÀ e SENSITIVITÀ (in data set controllati)
3. community experiment: es. CASP → GASP

discorso

NGS SEQUENCING



MEDICINA PERSONALIZZATA



## VARIANT CALLING

parte dalle read che sono pronte all'analisi perché hanno passato il  
data processing



Scoperta delle varianti e genotyping

SNP, Indel, SV, Raw variants



Analisi integrativa

Confronto con dati esterni (genoma di riferimento)

→ calibrazione della qualità

delle varianti e raffinamento del genotipo



Varianti pronte all'analisi



## VARIANT ANALYSIS

parte dalle varianti trovate

↓  
filtraggio

- tolgo le varianti comuni
- probabilità di non sinonimia → non hanno correlazione con la proteina

interpretazione biologica iterativa

· mutazioni homo o eterozigoti

l'interpretazione biologica si basa sulla probabilità che una mutazione in un  
sito altamente conservato sia dannosa → molto probabile

Come diretta conseguenza di questa probabilità, abbiamo la possibilità di:

## PREDIRE GLI EFFETTI DELLE VARIANTI

↓  
basato  
sulla sequenza

↓  
basato sulla  
struttura

↓  
basato  
sull'annotazione

→ ORTOLOGHI > PARALOGHI

→ ENTROPIA DI SHANNON

L'informazione è importante tanto quanto è bassa la probabilità  
di riceverla

# PROTEIN STRUCTURE

Obiettivo: annotazione strutturale e funzionale delle proteine, per farlo devi passare attraverso la risoluzione della loro struttura tridimensionale

Proteina

eteropolimero lineare di residui amminoacidici che hanno una struttura tale da determinare una funzione precisa e sono frutto dell'EVOLUZIONE

Struttura:  
1°, 2°, 3°, 4°  
→ angoli diedri  
→ alpha-eliche  
→ beta-foglietti

RAMACHANDRAN PLOT:

metodo di visualizzazione delle combinazioni degli angoli diedri → identifica quelle ammissibili

loops  
molto difficili da catturare (TAGLIATI IN PDB)

Elementi di struttura  
SuperSecondaria

riarrangiamento di due o più strutture secondarie → favoriti energeticamente e base di molte strutture tridimensionali

Partendo dalla famiglia:  
1. Pulisco dal PDB  
2. Calcolo gli angoli  
3. Creo i plot

## PROTEIN FOLDING

Definizione:

il riarrangiamento globulare e compatto della catena polipeptidica. Il modo in cui si ripiega la proteina è determinato dalla sequenza primaria. Tale orientamento punta a impacchettare i residui idrofobici all'interno del core della proteina

SE IN UNA STRUTTURA RISOLTA UN RESIDUO RISULTA ESSERE FUORI POSTO ALLORA È DOVUTO AD UN ERRORE O AD UNA MUTAZIONE

energia libera di Gibbs

$$\Delta G = \Delta H - T\Delta S$$

voglio trovare il minimo assoluto

Effetto idrofobico:

considero il grado di disordine  
residui idrofobici + acqua = diminuisce  
l'entropia dell'acqua

IL PROBLEMA:

Il folding di una proteina è marginalmente stabile poiché il grado di  $\Delta G$  è dell'ordine di poche Kcal/mol

Ad oggi è impossibile ripiegare le proteine, ma è possibile ricreare il modello strutturale

# Dal protein folding → COMPARATIVE MODELING

## El Principio:

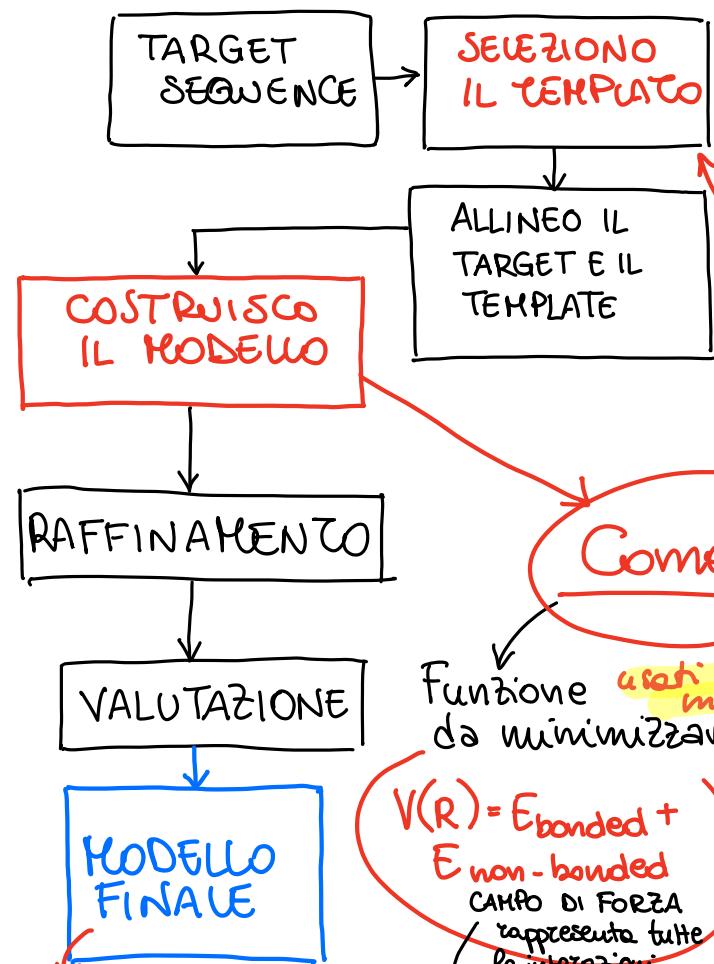
Sappiamo che proteine allineate con una buona percentuale di identità di sequenza avranno una divergenza nella conformazione del loro core di meno di 1 Å avranno una conformazione molto simile.

più sono vicine evolu-  
tivamente, più lo  
sovranno a livello di  
struttura

- Clothiā e lesk
- RMSD
- grafici

TEMPLATE + TARGET → modellare struttura 3D  
protein sequence ignota del target

# HOMOLOGY MODELING



BLAST in database per cercare proteine simili / della stessa famiglia

ALLINEAMENTI  
MULTIPLI →

dico conoscere benissimo  
il template; prelevo e  
Studio il PDB e cerco  
in letteratura

# PREDIZIONI AB INITIO

Fragfold → porzioni con stesse  
strutture supersecondarie  
Rosetta → 9 a.a. → combinazione dei  
nuovi e  
vecchi

## HMM modeling

andare a fondo alla storia  
evolutiva → **CALCOLO DEL  
PROFILO DI LINN**

- Stati: match, insert, del.
  - Osservabili: frequenze aa
  - matrice probabilità trans.
  - matrice probabilità emissione  
probabilità iniziali

**Forward:** trovare le probabilità della sequenza dato il modello e

i parametri → trova tutti i path  
**Viterbi**: trovare il path più  
probabile che abbia generato la  
mia sequenza

**AlphaFold**  
Crea matrici di interazione e contatto  
**NON HA RISOLTO IL PROBLEMA DEL FOLDING**

# DOCKING

Studio delle interazioni ligando-proteina

**GOAL:** data una struttura proteica, predire quali ligandi lega e dove

Devo trovare una conformazione ligando-proteina che minimizza l'energia totale del complesso.

Challenge: predire il binding e l'energia cercando nello spazio delle conformazioni possibili

Come?

## TROVARE I METODI PER CAMPIONARE LO SPAZIO CONFORMAZIONALE

→ metodo della griglia

→ genetic search (algoritmo di ricerca)

Devo rispettarle perché la conformazione sia possibile

## ELABORO SCORING FUNCTIONS

- metodi empirici
- metodi Knowledge-based
- metodi basati sui campi di forza

## Clustering

Devo generare un centinaio di modelli

## CALCOLARE L'INTERAZIONE È DIFFICILE

- Interazioni idrofobiche
  - Forze elettrostatiche
  - legami a idrogeno
  - Forze di van der Waals
  - Ponti salini
  - Interazioni di tipo pi-greco
- CREO UN CAMPO DI FORZA EMPIRICO  
ovvero sommo tutte le forze dei gruppi funzionali, ottenute dai dati sperimentali

Devo usarle per calcolare uno score e un minimo energetico

Il compito degli algoritmi è di trovare un gran numero di conformazioni possibili e quello delle funzioni di punteggio è classificare la qualità delle soluzioni.

1. Scelgo il target → trovo la struttura tridimensionale
2. Scelgo tra docking rigido (assumo le catene laterali rigide) o flessibile
3. Docking manuale o Docking automatico
4. Effettuo il clustering e valuto i miei docking (es. RMSD)