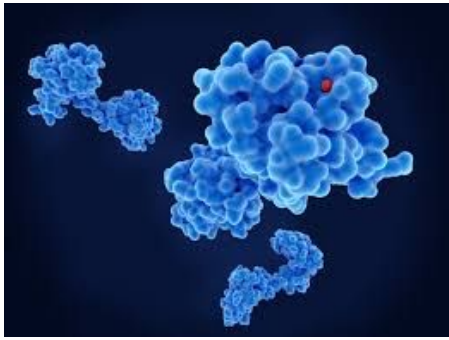
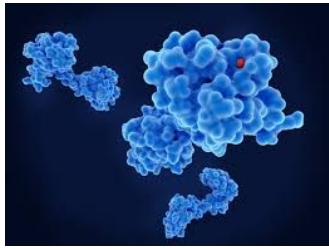
Metodi per l'analisi di sequenze proteiche



1. Cosa sono i descrittori
2. Diversi tipi di descrittori
   - Espressioni regolari
   - profili di sequenza
   - HMM
   - L'algoritmo forward-backward
   - L'algoritmo di Viterbi
   - Il modello OOPS (One Observation Per Sequence)
   - Lo ZOOPS (Zero or One Occurrence Per Sequence)
3. L'algoritmo MEME (Multiple EM for Motif Elicitation)
4. Applicazione – pattern e motivi funzionali

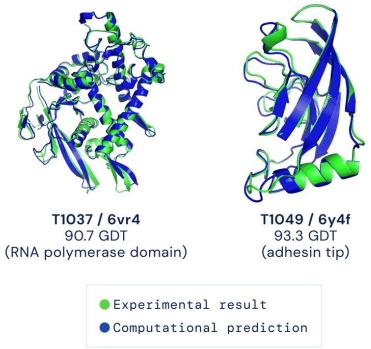
Metodi per l'analisi di sequenze proteiche



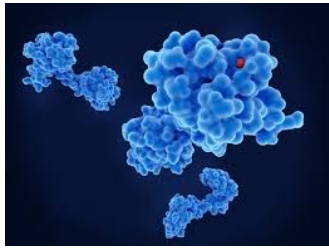
1. Cosa sono i descrittori

In generale, un descrittore è una quantità numerica o una rappresentazione matematica che descrive una particolare caratteristica di un oggetto o di un sistema.

Nell'ambito della bioinformatica, i descrittori sono spesso utilizzati per rappresentare le **proprietà** delle sequenze biologiche, come le **sequenze** di DNA o di proteine, in modo da poterle **confrontare**, classificare e analizzare in maniera efficiente.



Metodi per l'analisi di sequenze proteiche

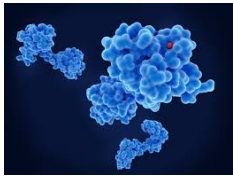


1. Cosa sono i descrittori

In particolare, i d**escrittori di elementi funzionali** nelle **sequenze proteiche** sono un tipo di descrittori utilizzati per identificare e rappresentare le caratteristiche **funzionali** delle proteine.

Questi descrittori possono essere basati su diverse proprietà delle proteine, come la composizione in amminoacidi, la presenza di **motivi funzionali** specifici, la struttura tridimensionale, la conservazione evolutiva, e così via.

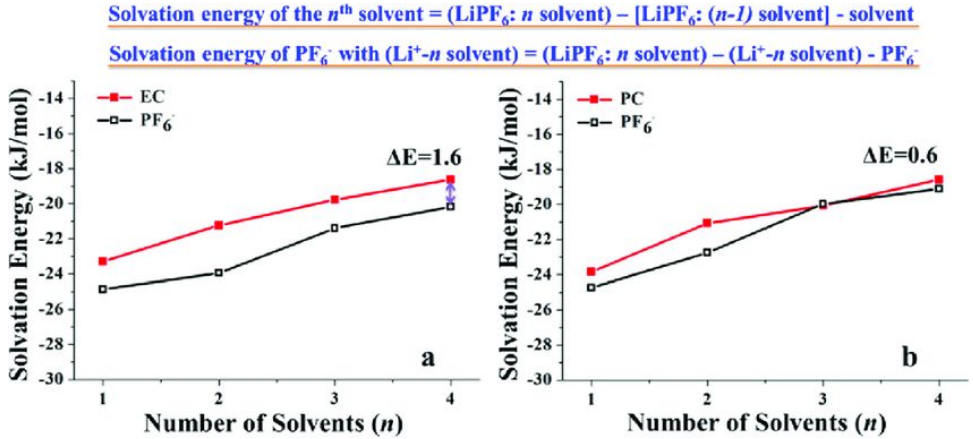
Metodi per l'analisi di sequenze proteiche



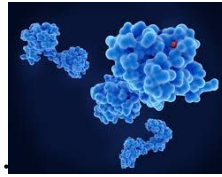
1. Cosa sono i descrittori

Ad esempio, un descrittore comune utilizzato per identificare le regioni di interazione proteina-proteina è il profilo di energia solvata (Solvation Energy Density, SED), che fornisce informazioni sulle proprietà di solvatazione di una particolare regione della proteina.

Altri descrittori utilizzati per identificare i motivi funzionali delle proteine possono includere la presenza di domini proteici specifici, la localizzazione subcellulare, la presenza di siti di fosforilazione, e così via.



Metodi per l'analisi di sequenze proteiche



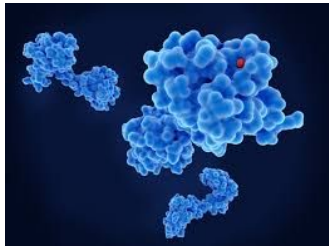
1. Cosa sono i descrittori

Molto importanti sono i descrittori basati sull'analisi dei motivi funzionali delle proteine.

Un motivo funzionale, noto anche come dominio proteico, è una regione della sequenza di una proteina che ha una funzione specifica e indipendente dalla funzione delle altre regioni della proteina.

I motivi funzionali tendono ad essere più **conservati** rispetto alle regioni non funzionali della proteina, poiché una mutazione in queste regioni potrebbe influenzare negativamente la funzione della proteina. Pertanto, la conservazione di un motivo funzionale attraverso diverse specie suggerisce una sua importanza funzionale.

Metodi per l'analisi di sequenze proteiche



1. Cosa sono i descrittori

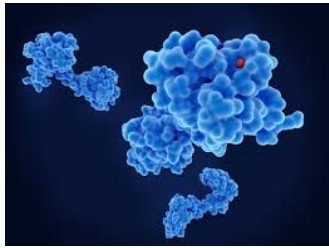
In generale, l'uso di descrittori di elementi funzionali può aiutare a identificare le regioni di una proteina che sono importanti per la sua funzione, a confrontare le sequenze proteiche e a predire le interazioni proteina-proteina o proteina-ligando.

Esistono **diversi tipi** di descrittori, come le espressioni regolari, le matrici posizionali di peso, i profili PSSM e i modelli nascosti di Markov.

Si differenziano per la loro complessità e la loro capacità di catturare informazioni specifiche sulla sequenza proteica.

Metodi per l'analisi di sequenze proteiche

- Espressioni regolari

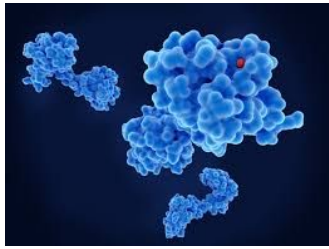


Un'espressione regolare, in informatica, è una sequenza di caratteri che definisce un pattern o modello di ricerca.

Sono spesso utilizzate per la ricerca di pattern specifici nelle sequenze di DNA e proteine, ad esempio per identificare regioni di sequenze conservate tra diverse specie o per riconoscere sequenze di segnale e di localizzazione subcellulare.

Le espressioni regolari sono formate da una combinazione di caratteri speciali e normali. Ad esempio, il carattere "a" rappresenta semplicemente la lettera "a", mentre il carattere "." o “X” (***wild card***) rappresenta qualsiasi carattere. Le espressioni regolari sono utilizzate in diversi linguaggi di programmazione, come Perl, Python, Java, JavaScript e molti altri.

Metodi per l'analisi di sequenze proteiche

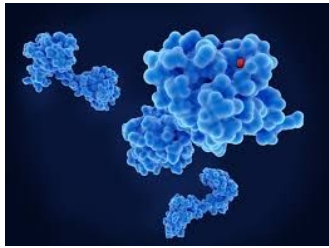


2. Diversi tipi di descrittori

Le espressioni regolari sono un metodo relativamente semplice per codificare un motivo funzionale perché consentono di descrivere in modo conciso una serie di caratteri che possono essere presenti in diverse posizioni all'interno della sequenza.

In altre parole, le espressioni regolari consentono di identificare pattern di sequenza conservati che possono essere associati a funzioni specifiche.

Metodi per l'analisi di sequenze proteiche



2. Diversi tipi di descrittori

Un esempio di espressione regolare per la ricerca del motivo funzionale del dominio SH3 (Src Homology 3) potrebbe essere:

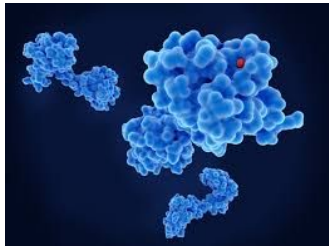
PXXP

In questo caso, "P" rappresenta l'aminoacido prolina e "X" rappresenta qualsiasi altro aminoacido. Quindi, questa espressione regolare identificherebbe qualsiasi sequenza di quattro aminoacidi che abbia una prolina alla prima e alla quarta posizione, mentre la seconda e la terza posizione possono essere qualsiasi altro aminoacido.

Questo pattern è comune per il dominio SH3, che si lega a sequenze contenenti il motivo PXXP.

La prolina è un amminoacido particolarmente rigido e la presenza di due proline in questo pattern permette di creare una struttura adatta all'interazione con il dominio SH3.

Metodi per l'analisi di sequenze proteiche



2. Diversi tipi di descrittori

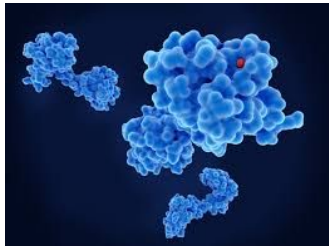
I **docking site** sono regioni specifiche delle proteine che interagiscono con altre proteine o molecole, come substrati o ligandi.

Queste regioni sono generalmente caratterizzate da **motivi** di legame altamente **conservati** e sono importanti per la funzione della proteina in quanto permettono l'interazione specifica con altre molecole.

Ad esempio la fosfatasi PP1 ha un docking site denominato SILK molto conservato nelle fosfatasi ortologhe, con alcuni residui invariati. L'analisi di tutte le sequenze di SILK depositate in banca dati permette di estrarne l'espressione regolare:

[GS] IL [KR] [^DE]

Metodi per l'analisi di sequenze proteiche



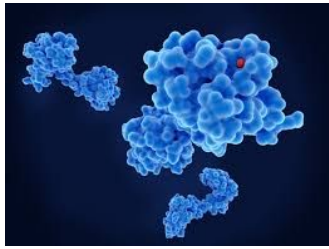
2. Diversi tipi di descrittori

[GS] IL [KR] [^DE]

Questa espressione regolare rappresenta una sequenza di amminoacidi di lunghezza 5. I simboli tra parentesi quadre indicano che in quella posizione possono esserci solo specifici amminoacidi:

[GS]: può essere G o S
IL: deve essere I ed L
[KR]: può essere K o R
[ˆDE]: il carattere ˆ indica i residue che sicuramente non sono ammessi nella posizione

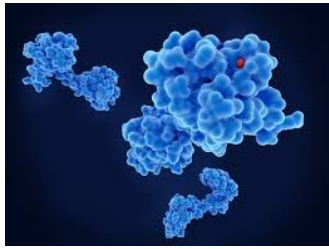
Metodi per l'analisi di sequenze proteiche



2. Diversi tipi di descrittori

I loghi dei motivi funzionali delle proteine, spesso chiamati "sequence logos" o "motif logos", sono rappresentazioni grafiche dei motivi amminoacidici conservati all'interno di una famiglia di proteine o di un dominio proteico. I loghi vengono generati sulla base di un allineamento multiplo di sequenze di proteine e forniscono un modo per visualizzare l'informazione conservata in queste sequenze.

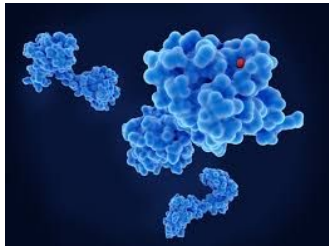
Metodi per l'analisi di sequenze proteiche



2. Diversi tipi di descrittori

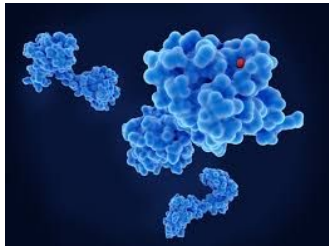
Un logo del motivo consiste di colonne di lettere che rappresentano gli amminoacidi, con l'altezza di ogni lettera proporzionale alla frequenza di quell'amminoacido in quella posizione specifica del motivo. In altre parole, un amminoacido altamente conservato in una posizione avrà una lettera più alta, mentre un amminoacido meno conservato avrà una lettera più bassa. L'altezza totale di una colonna indica l'informazione conservata nella posizione, misurata in bit.

Metodi per l'analisi di sequenze proteiche



I loghi dei motivi funzionali delle proteine sono utili per evidenziare regioni conservate e importanti per la funzione di una proteina, per identificare siti di legame o per prevedere nuovi membri di una famiglia di proteine sulla base della conservazione del motivo. Gli strumenti bioinformatici, come WebLogo, MEME Suite e Skylign, consentono di creare loghi di motivi funzionali delle proteine a partire da allineamenti di sequenze.

Metodi per l'analisi di sequenze proteiche



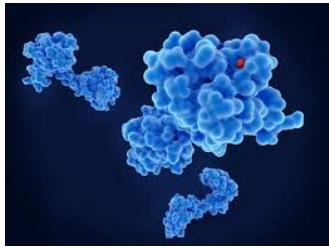
2. Diversi tipi di descrittori

Un programma che si basa sull'utilizzo di PSSM (Position-Specific Scoring Matrix) è PSI-blast, che confronta il profilo con tutte le sequenze di una banca dati per identificare proteine omologhe.

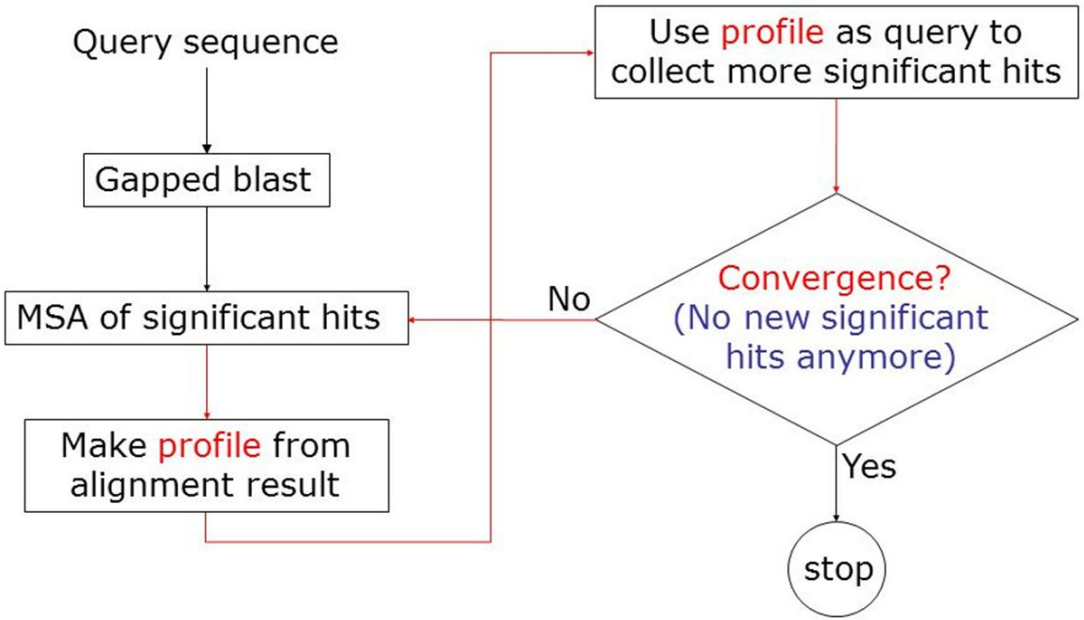
Questo tipo di ricerca è implementato nella pagina del BLASTp dell'NCBI

Metodi per l'analisi di sequenze proteiche

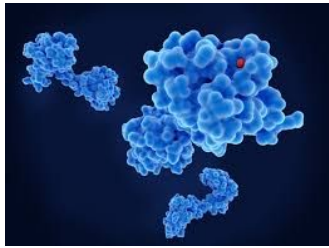
2. Diversi tipi di descrittori



PSI-BLAST – step di analisi



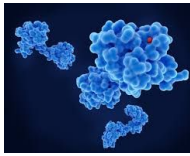
Metodi per l'analisi di sequenze proteiche



Ecco come funziona PSI-BLAST in termini generali:

a. Inizia con una sequenza di query e utilizza l'algoritmo BLAST per trovare sequenze simili in un database di sequenze proteiche.
b. A partire dalle sequenze simili trovate, crea un allineamento multiplo delle sequenze e genera una PSSM.
c. Utilizza la PSSM (il profilo) per cercare nuovamente nel database di sequenze proteiche, questa volta con una maggiore sensibilità e specificità rispetto al BLAST standard.
d. Ripete iterativamente i passaggi b e c per migliorare ulteriormente la PSSM e la ricerca di sequenze simili.

Metodi per l'analisi di sequenze proteiche

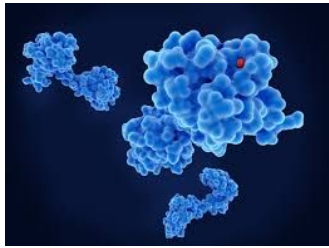


2. Diversi tipi di descrittori

L'algoritmo PSI-BLAST è particolarmente utile per identificare proteine simili che condividono una struttura o una funzione conservata ma hanno una bassa identità di sequenza. Questo lo rende uno strumento importante per l'analisi delle famiglie di proteine, la predizione della funzione delle proteine e la ricerca di omologhi evolutivamente correlati.

☑	macin [Scapharca broughtonii]	163	163	100%	4e-51	100%	AFQ02694.1
☑	Mytimacin-6 [Mytilus galloprovincialis]	82.8	82.8	94%	3e-19	54%	AHG59339.1
☑	hydramacin [Ruditapes philippinarum]	75.5	75.5	94%	4e-16	52%	AGM14601.1
☑	mytimacin-2 [Mytilus galloprovincialis]	73.2	73.2	67%	2e-15	61%	CCC15016.1
☑	macin [Ruditapes philippinarum]	72.8	72.8	91%	3e-15	46%	APY18889.1
☐	antibacterial peptide [Cyclina sinensis]	72.8	72.8	72%	3e-15	57%	AFI24614.1
☑	mytimacin-3 [Mytilus galloprovincialis]	71.2	71.2	75%	2e-14	58%	CCC15017.1
☑	macin [Ruditapes philippinarum]	69.7	69.7	94%	5e-14	47%	APY18888.1
☐	RecName: Full=Hydramacin-1; Short=Hm-1; Flags: Precursor	67.4	67.4	97%	4e-13	45%	B3RFR8.1

Metodi per l'analisi di sequenze proteiche

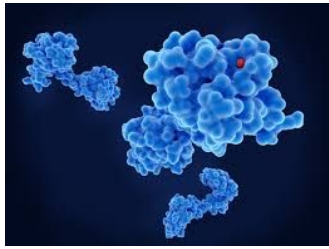


2. Diversi tipi di descrittori

Le espressioni regolari però non tengono conto della frequenza dei residui nelle varie posizioni del motivo. Per esempio un'espressione regolare che contiene in una posizione [K,R] non tiene conto della frequenza delle lisine e arginine nelle istanze note del motivo e sarebbero valutate egualmente.

Se nelle instanze note del motivo la lisina fosse più frequente dell'arginina vorremmo dare un punteggio maggiore.

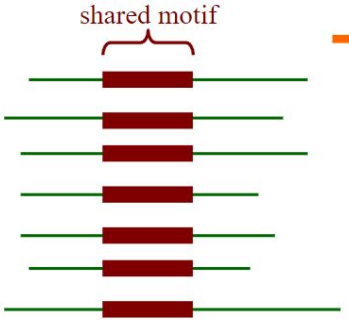
Metodi per l'analisi di sequenze proteiche

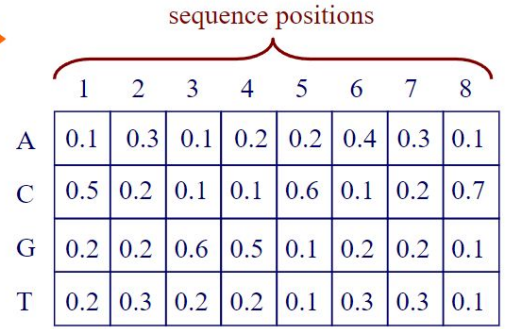


2. Diversi tipi di descrittori

Per risolvere questo problema, un motivo funzionale può essere rappresentato come una matrice, le cui colonne rappresentano le diverse posizioni dell'allineamento delle istanze del motivo e le cui righe rappresentano tutti i possibili residui (nucleotidi o aminoacidi).

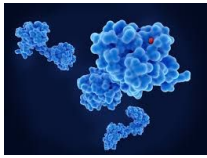
Ogni cella di questa matrice deve riportare un punteggio relativo un particolare residuo in una particolare posizione del motivo.





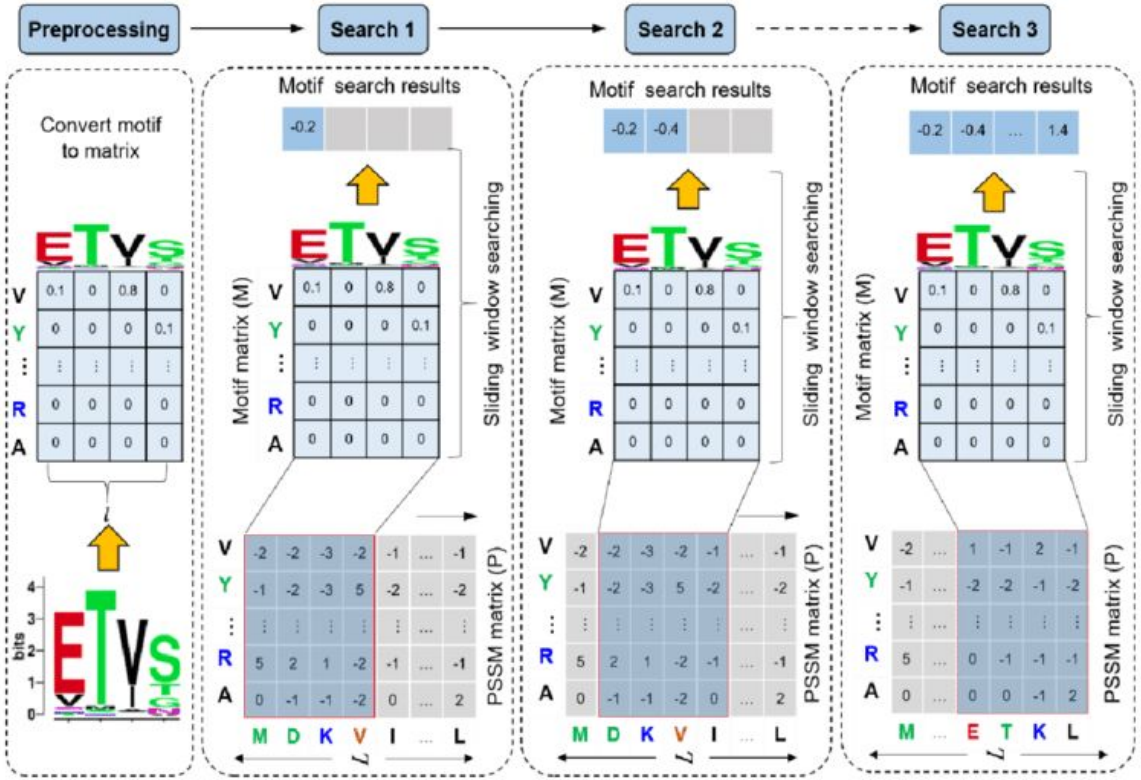
	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

Metodi per l'analisi di sequenze proteiche

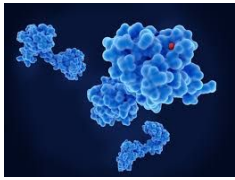


2. Diversi tipi di descrittori

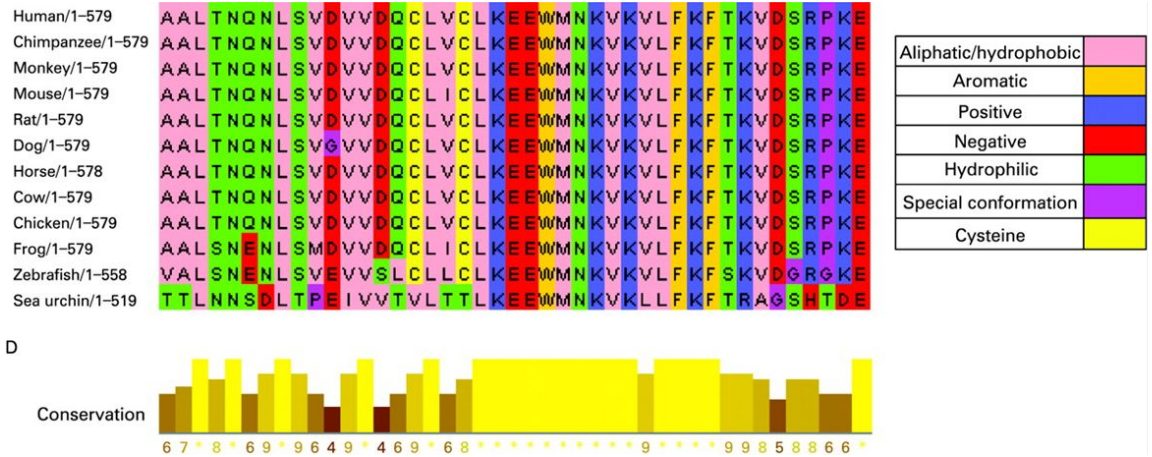
Il processo di generazione della funzione Motif-PSSM basata sulla matrice PSSM.



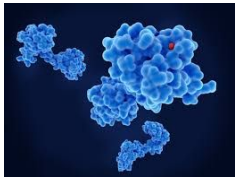
Metodi per l'analisi di sequenze proteiche



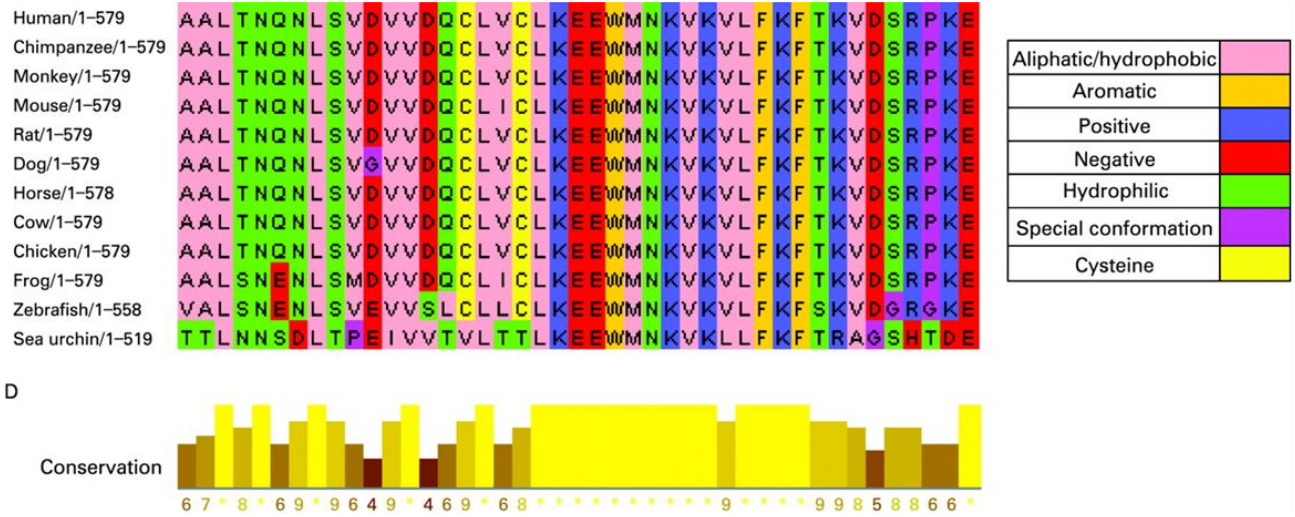
PSSM sta per Position-Specific Scoring Matrix (Matrice di Punteggio Specifica per Posizione). È un'importante struttura dati utilizzata nelle analisi bioinformatiche per rappresentare le preferenze amminoacidiche all'interno di un motivo o di un dominio proteico conservato. La PSSM viene generata sulla base di un allineamento multiplo di sequenze di proteine e fornisce un punteggio per ogni possibile amminoacido in ogni posizione del motivo o del dominio.



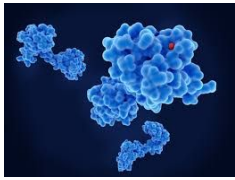
Metodi per l'analisi di sequenze proteiche



Una PSSM è una matrice di dimensioni L x 20, dove L è la lunghezza del motivo o del dominio, e 20 è il numero di amminoacidi standard. Ogni elemento della matrice rappresenta il punteggio per un determinato amminoacido in una specifica posizione. Un punteggio elevato indica una maggiore preferenza per un amminoacido in quella posizione, mentre un punteggio basso indica una minore preferenza.



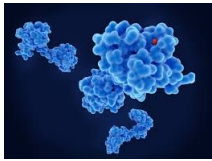
Metodi per l'analisi di sequenze proteiche



Le PSSM vengono utilizzate in diversi contesti bioinformatici, tra cui:

1. **Ricerca di sequenze simili**: la PSSM può essere utilizzata per cercare proteine con sequenze simili in un database di sequenze, utilizzando algoritmi come PSI-BLAST (Position-Specific Iterated BLAST).
2. **Predizione di motivi o domini**: la PSSM può essere utilizzata per identificare regioni conservate all'interno di una sequenza di proteine e prevedere la presenza di motivi o domini funzionali.
3. **Valutazione della conservazione delle sequenze**: la PSSM può essere utilizzata per quantificare la conservazione di una posizione specifica all'interno di un allineamento di sequenze e identificare posizioni importanti per la funzione della proteina.

Metodi per l'analisi di sequenze proteiche



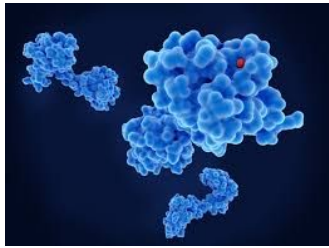
2. Diversi tipi di descrittori

Ecco un esempio di tabella di profilo di sequenza per un motivo di 7 residui:

	Posizione 1	Posizione 2	Posizione 3	Posizione 4	Posizione 5	Posizione 6	Posizione 7
A	0.2	0.1	0.05	0.05	0.9	0.05	0.05
C	0.1	0.1	0.1	0.7	0.1	0.0	0.0
G	0.7	0.0	0.0	0.0	0.0	0.0	0.3
T	0.0	0.8	0.8	0.2	0.0	0.2	0.2

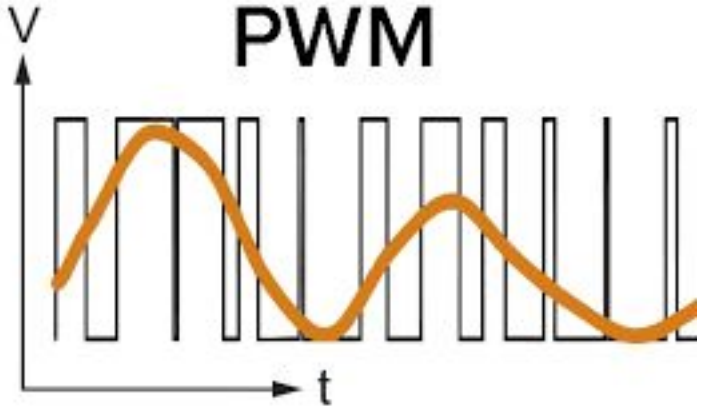
Nella tabella, le righe rappresentano i quattro possibili nucleotidi, mentre le colonne rappresentano le posizioni del motivo. Ogni cella contiene un punteggio che rappresenta la probabilità di trovare un dato nucleotide in una data posizione del motivo, basata sull'allineamento di molte istanze di quel motivo.

Metodi per l'analisi di sequenze proteiche

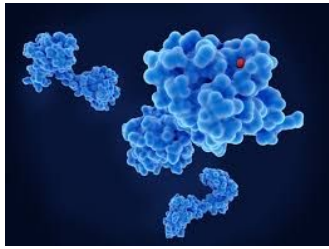


2. Diversi tipi di descrittori

La Position Weight Matrix (**PWM**), o Matrice dei Pesi per Posizione, è un tipo di matrice utilizzata nelle analisi bioinformatiche per rappresentare le preferenze di nucleotidi o aminoacidi in un motivo conservato, come un sito di legame per un fattore di trascrizione o un dominio proteico specifico. Le PWM sono generate a partire da un allineamento multiplo di sequenze e forniscono un punteggio per ogni possibile nucleotide o amminoacido in ogni posizione del motivo.

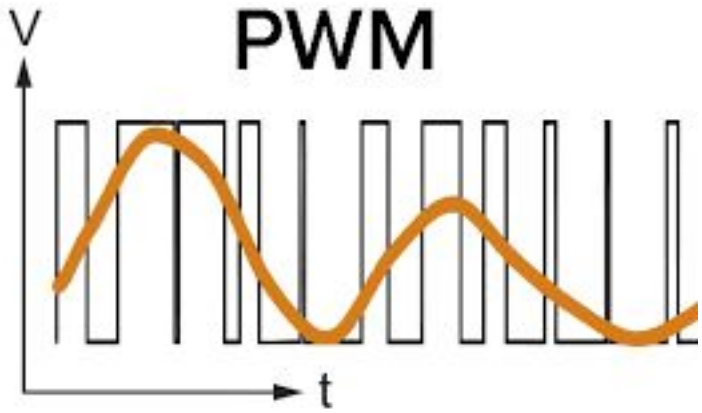


Metodi per l'analisi di sequenze proteiche



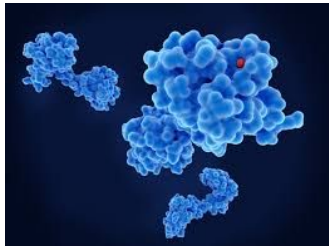
2. Diversi tipi di descrittori

Una **PWM** è una matrice posizionale di peso che riporta i dati relativi alle sequenze che sono già state identificate come omologhe e che vogliamo descrivere. Uno degli scopi è definire dei descrittori capaci di incorporare le informazioni relative a un insieme di proteine omologhe sulla base delle sequenze che sono già state allineate.



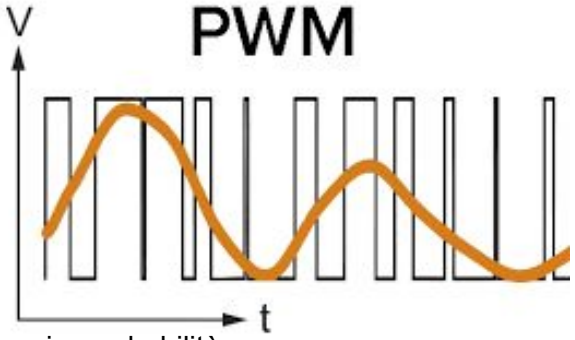
Una PWM è una matrice di dimensioni L x N, dove L è la lunghezza del motivo e N è il numero di possibili nucleotidi (4 per il DNA) o amminoacidi (20 per le proteine). Ogni elemento della matrice rappresenta il peso (o la probabilità) di un determinato nucleotide o amminoacido in una specifica posizione del motivo.

Metodi per l'analisi di sequenze proteiche



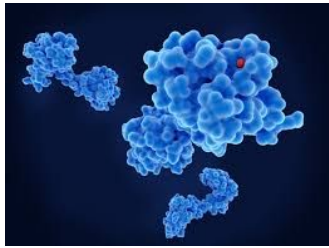
Le PWM sono spesso utilizzate per:

1. Identificare i siti di legame di fattori di trascrizione nel DNA: una PWM può essere utilizzata per trovare sequenze simili a un motivo noto di legame di un fattore di trascrizione nel genoma e prevedere nuovi siti di legame.
2. Predizione di motivi o domini proteici: una PWM può essere utilizzata per identificare regioni conservate all'interno di una sequenza di proteine e prevedere la presenza di motivi o domini funzionali.
3. Analisi di sequenze simili: la PWM può essere utilizzata per confrontare sequenze e valutare la loro similarità sulla base del punteggio del motivo conservato.



È importante notare che la Position Weight Matrix (PWM) e la Position-Specific Scoring Matrix (PSSM) sono concetti simili ma non identici. Entrambe rappresentano le preferenze di nucleotidi o amminoacidi in un motivo conservato, ma la PSSM utilizza punteggi logaritmici mentre la PWM utilizza pesi o probabilità.

Metodi per l'analisi di sequenze proteiche

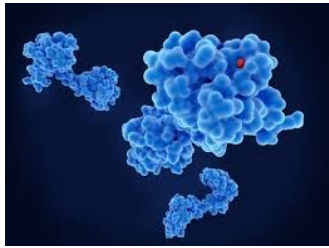


2. Diversi tipi di descrittori

Un altro scopo è riconoscere sequenze omologhe non ancora riconosciute come tali. La frequenza di un residuo in una posizione può essere normalizzata dal numero di sequenze che compongono l'allineamento e dalla frequenza assoluta di quel residuo in un dataset di background, e questo valore normalizzato e riportato come logaritmo

Si possono avere casi in cui compaiano logaritmi di zero, cioè quando un residuo non è mai trovato in una data posizione del motivo. Per ovviare a questo problema si possono sommare i logaritmi alle frequenze delle **pseudoconte** , ovvero dei valori uguali per tutti i residui.

Metodi per l'analisi di sequenze proteiche

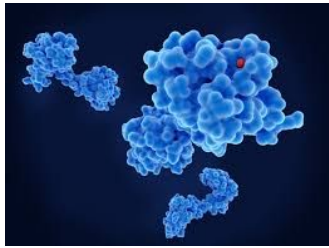


2. Diversi tipi di descrittori

I modelli nascosti di Markov (HMM, Hidden Markov Models) sono una classe di modelli statistici utilizzati per la modellizzazione di sequenze di dati, ad esempio sequenze di nucleotidi o di aminoacidi. Un HMM è utilizzato per rappresentare sistemi che presentino una struttura di dipendenza temporale e una natura nascosta o non osservabile.

L'HMM viene addestrato su un set di dati noti, e quindi può essere utilizzato per la predizione e l'analisi di nuovi dati.

Metodi per l'analisi di sequenze proteiche

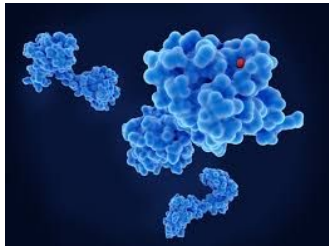


2. Diversi tipi di descrittori

Un HMM è composto da tre elementi principali:

1. Stati nascosti: una serie di stati non osservabili che rappresentano le proprietà sottostanti del sistema. Ad esempio, in un HMM per l'analisi delle sequenze proteiche, gli stati nascosti potrebbero rappresentare le diverse conformazioni secondarie delle proteine, come alfa-elica, foglietto-beta e coil.
2. Transizioni tra stati: le probabilità di passare da uno stato nascosto a un altro. Queste probabilità definiscono la dinamica temporale del sistema e le relazioni tra gli stati nascosti.
3. Emissioni: le probabilità di osservare un simbolo (ad esempio, un nucleotide o un amminoacido) in un determinato stato nascosto. Queste probabilità definiscono il legame tra gli stati nascosti e le sequenze osservabili.

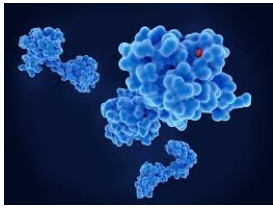
Metodi per l'analisi di sequenze proteiche



In pratica, si utilizza un HMM per "addestrare" un modello a riconoscere dei pattern specifici nelle sequenze. Il modello impara a riconoscere questi pattern durante la fase di addestramento, in cui gli viene fornito un insieme di sequenze di esempio.

Una volta addestrato, il modello può essere utilizzato per identificare pattern simili in nuove sequenze sconosciute, aiutando così a comprendere meglio la funzione e l'evoluzione delle sequenze biologiche.

Metodi per l'analisi di sequenze proteiche

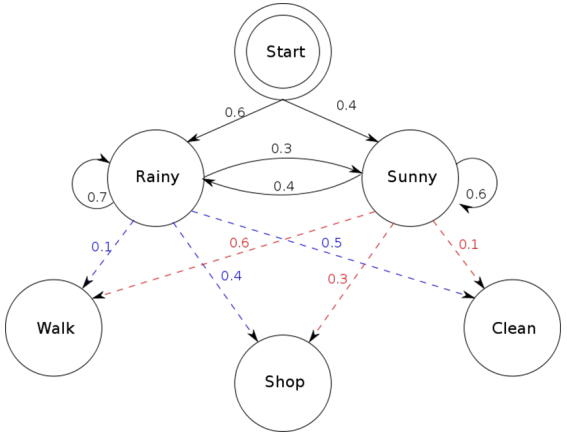


Un semplice esempio

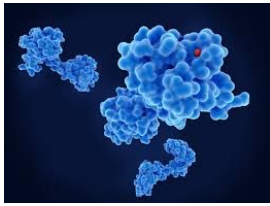
Una nostra amica vive in Australia e fa una vita estremamente monotona in cui passeggia, fa comprere e pulisce (queste sono le nostre osservazioni o simboli).

Gli stati possibili sono due «bel tempo» e «pioggia» e non sono da noi direttamente osservabili, in quanto non siamo in Australia.

Tuttavia le tre azioni che fa la nostra amica dipendono da questi stati con determinate probabilità, dette probabilità di emissione.

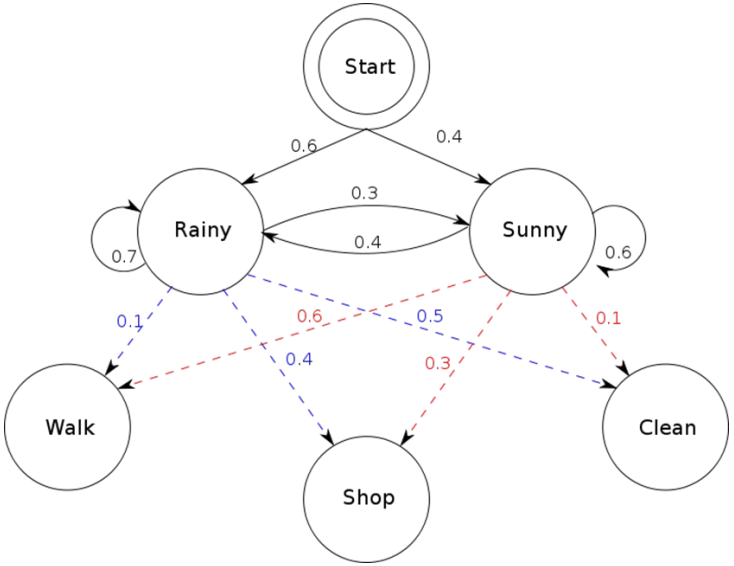


Metodi per l'analisi di sequenze proteiche



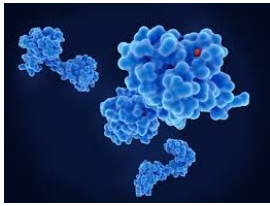
Anche gli stati però sono dipendenti tra loro, ad esempio è più probabile avere due giorni piovosi l'uno di fila all'altro (0,7), piuttosto che una transizione da un giorno piovoso ad uno di sole (0,3). Queste sono le probabilità di transizione.

Noi, pur non potendo osservare direttamente gli stati (che sono quindi **nascosti**) conosciamo bene il comportamento della nostra amica ed il trend metereologico di quella regione geografica. Siamo quindi a conoscenza dei parametri che regolano il susseguirsi degli stati e delle osservazioni dell'Hidden Markov Model

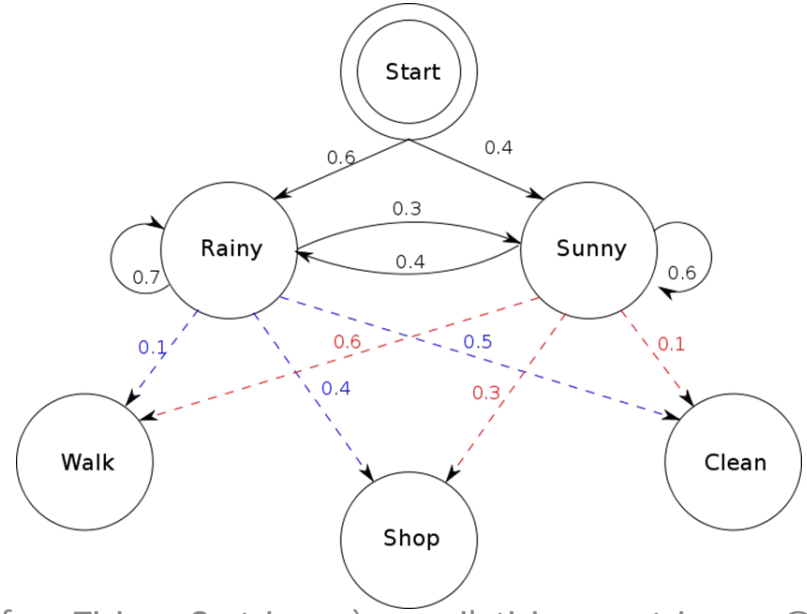


Metodi per l'analisi di sequenze proteiche

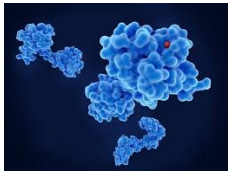
Gli stati nascosti in un modello nascosto di Markov sono quegli stati del modello che non sono direttamente osservabili, ma sono solo ipotetici.



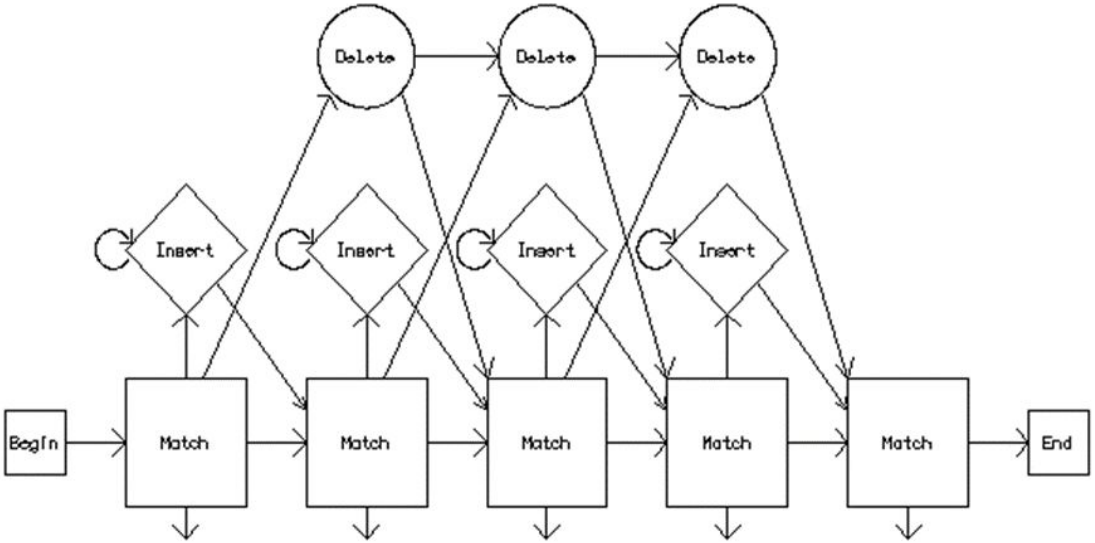
Ad esempio, in una sequenza proteica, gli stati nascosti possono rappresentare informazioni sulla struttura della proteina, come l'orientamento di particolari amminoacidi rispetto ad altri.



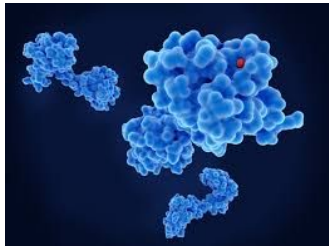
Metodi per l'analisi di sequenze proteiche



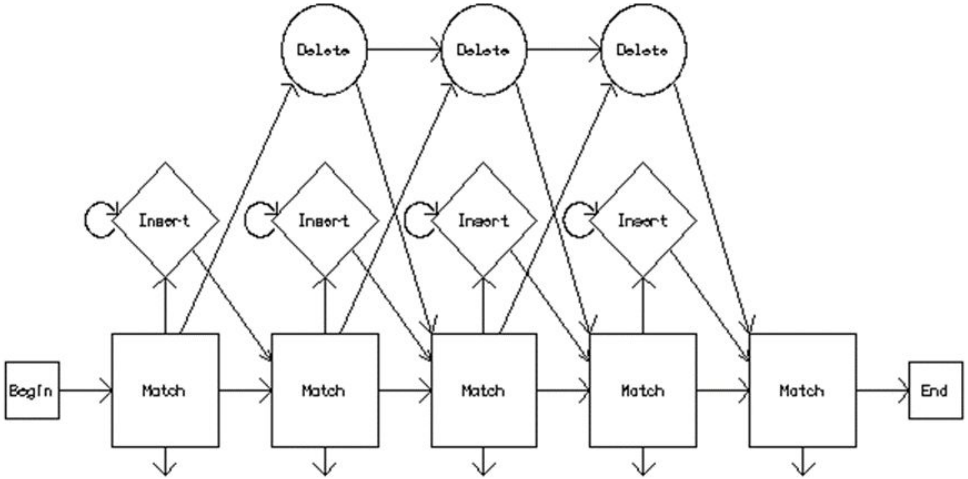
Prendiamo come esempio un allineamento tra sequenze
Partiamo dall'inizio e proviamo a seguire una serie di frecce fino ad arrivare alla fine.
Ogni freccia ci porta ad uno **stato** del sistema. Ad ogni stato viene effettuata una determinata **azione**, cioè viene scelta una freccia che ci porta allo stadio successivo.



Metodi per l'analisi di sequenze proteiche

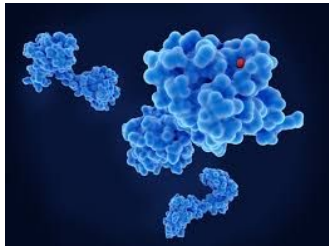


L'azione intrapresa e la scelta dello stato successivo sono governate da un set di **probabilità**. Ad esempio, ogni stato è associato ad una distribuzione della probabilità dei 20 aminoacidi ed una seconda distribuzione di probabilità per la scelta degli stati successivi.
Queste probabilità sono calibrate per codificare informazioni specifiche e caratteristiche di una particolare famiglia di sequenze: il modello generale può essere dunque adattato a molte famiglie proteiche



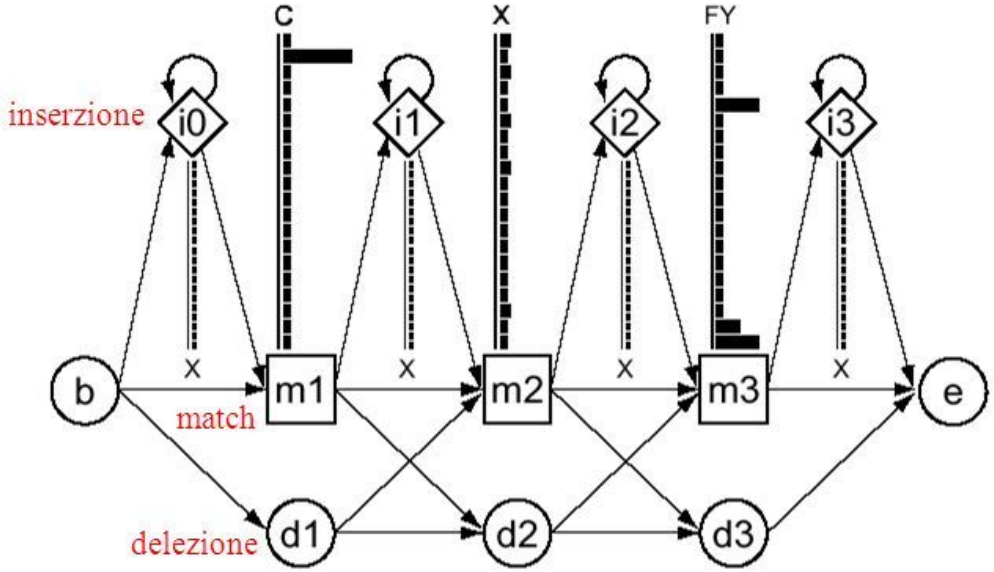
In questo HMM relativo ad un allineamento di sequenze troviamo uno stato match, uno stato insert ed uno stato delete relativi a ciascuna posizione dell'allineamento.

Metodi per l'analisi di sequenze proteiche

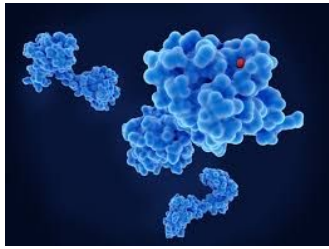


Negli HMM utilizzati per l'analisi di sequenze, gli **stati principali** (match) sono gli stati corrispondenti ai residui della sequenza. Gli **stati di inserzione** e **delezione**, invece, sono stati "fittizi" introdotti per modellare l'evoluzione delle sequenze in modo più accurato.

1	2	3
C	A	F
C	G	W
C	D	Y
C	V	F
C	K	Y

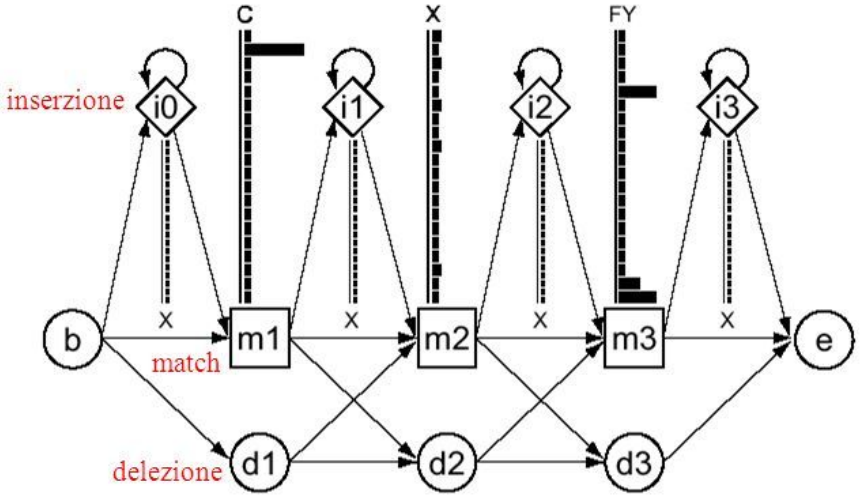


Metodi per l'analisi di sequenze proteiche

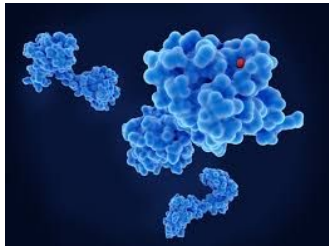


Gli stati di inserzione modellano l'idea che in una sequenza il residuo successivo è uguale a quello precedente, mentre gli stati di delezione modellano la possibilità che un residuo possa essere stato rimosso o perso durante l'evoluzione della sequenza.

1	2	3
C	A	F
C	G	W
C	D	Y
C	V	F
C	K	**Y**

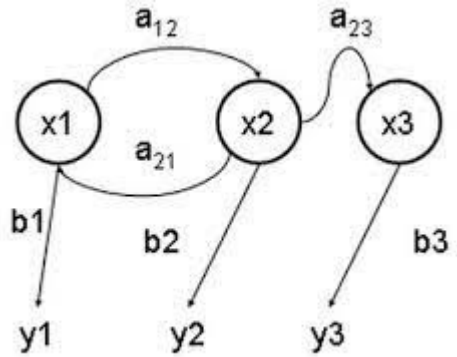


Metodi per l'analisi di sequenze proteiche

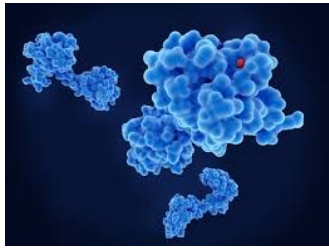


Gli HMM sono composti da un certo numero di stati che possono, per esempio, corrispondere a residui di una sequenza, a colonne di un allineamento multiplo oppure a posizioni in una struttura proteica tridimensionale.

Data la loro versatilità hanno trovato un amplissimo utilizzo in bioinformatica, soprattutto per quanto riguarda la ricerca di profili e predizioni strutturali.



Metodi per l'analisi di sequenze proteiche

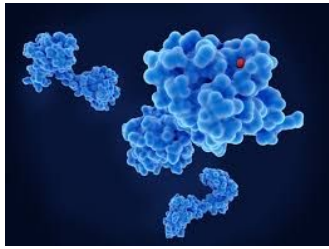


L'algoritmo forward-backward
è un algoritmo utilizzato nell'analisi di sequenze biologiche per stimare le probabilità di stato nascosto all'interno di un modello nascosto di Markov (HMM).

L'algoritmo forward-backward si basa sulla decomposizione di un HMM in due processi: un processo di **forward** e un processo di backward. Il processo forward calcola la probabilità di essere in un particolare stato nascosto al tempo t, date tutte le osservazioni dalla posizione 1 alla posizione t.

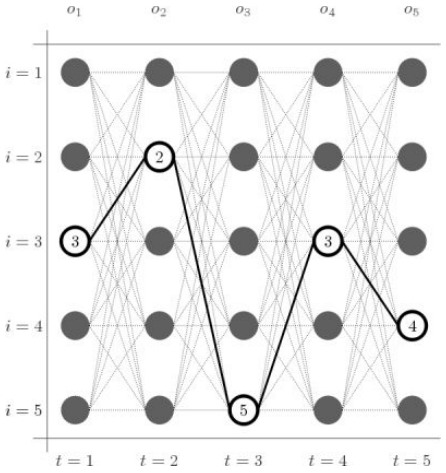
Il processo **backward**, al contrario, calcola la probabilità di osservare tutte le osservazioni dalla posizione t+1 alla fine della sequenza, dato che si parte dallo stato nascosto t.

Metodi per l'analisi di sequenze proteiche

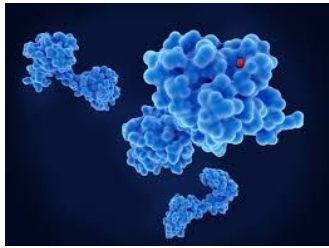


L'algoritmo di Viterbi
è un algoritmo utilizzato per trovare la sequenza di stati nascosti più probabile in un modello nascosto di Markov (HMM). In sostanza, dato un HMM e una sequenza di osservazioni, l'algoritmo di Viterbi calcola la sequenza di stati nascosti che massimizza la probabilità a posteriori condizionata delle sequenze di stati, data la sequenza di osservazioni.

L'algoritmo di Viterbi utilizza la programmazione dinamica per trovare la sequenza di stati nascosti più probabile in modo efficiente, evitando di dover esaminare tutte le possibili sequenze di stati.



Metodi per l'analisi di sequenze proteiche

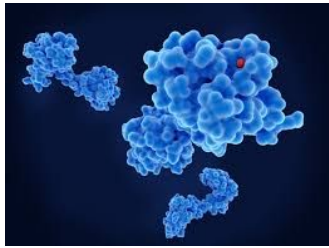


Il modello OOPS (One Observation Per Sequence)
è un tipo di modello nascosto di Markov (HMM) utilizzato per l'allineamento di sequenze. In questo modello, ogni sequenza viene vista come una singola osservazione e ogni posizione all'interno delle sequenze viene vista come uno stato dell'HMM.

Il modello OOPS si differenzia dagli altri modelli HMM in quanto **non tiene conto delle correlazioni** tra le sequenze durante il processo di allineamento. Questo significa che il modello considera ogni sequenza come indipendente dalle altre e non tiene conto delle informazioni sulle relazioni evolutive tra le sequenze.

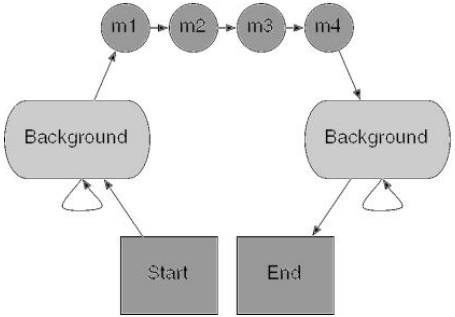
Pertanto, il modello OOPS è utile quando si devono allineare sequenze non correlate tra loro, come ad esempio sequenze di proteine con funzioni completamente diverse.

Metodi per l'analisi di sequenze proteiche

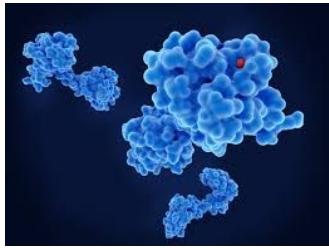


Lo ZOOPS (Zero or One Occurrence Per Sequence)
è un tipo di modello nascosto di Markov (HMM) utilizzato nell'analisi di sequenze di DNA o proteine.

In pratica, lo ZOOPS è una variante del modello OOPS, che tiene conto della possibilità che alcune sequenze possano avere una zona particolarmente conservata, ovvero una regione che è presente con la stessa lunghezza e la stessa posizione in tutte le sequenze.
Questa zona conservata viene definita "zero" e, a differenza del modello OOPS, in cui tutte le posizioni delle sequenze contribuiscono allo stesso modo alla costruzione del profilo, nello ZOOPS le posizioni della zona conservata **non** contribuiscono affatto alla costruzione del profilo.



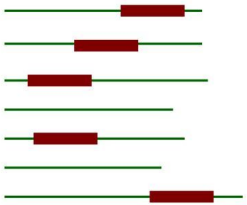
Metodi per l'analisi di sequenze proteiche



Ad esempio, se consideriamo un insieme di sequenze proteiche che contengono un motivo funzionale (ad esempio una sequenza di riconoscimento del DNA), possiamo ipotizzare che ci sia una zona di lunghezza costante e posizione conservata in tutte le sequenze che partecipa alla funzione di riconoscimento. In questo caso, utilizzando il modello ZOOPS, potremmo costruire un profilo che tenga conto di questa zona zero, in modo da migliorare la capacità di predizione del modello.

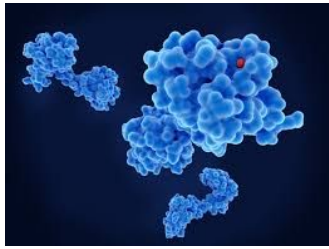
The ZOOPS Model

- The approach as we've outlined it, assumes that each sequence has exactly one motif occurrence per sequence; this is the OOPS model
- The ZOOPS model assumes *zero or one* occurrences per sequence



27

Metodi per l'analisi di sequenze proteiche

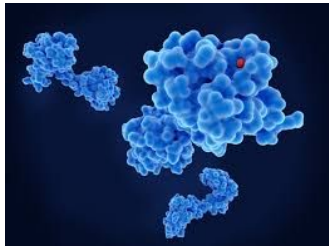


L'algoritmo MEME (Multiple EM for Motif Elicitation)
è un algoritmo utilizzato per identificare i motivi funzionali nelle sequenze di DNA, RNA e proteine. MEME utilizza un approccio di apprendimento automatico basato su algoritmi EM (Expectation-Maximization), che cerca di trovare i motivi funzionali più conservati nelle sequenze analizzate.

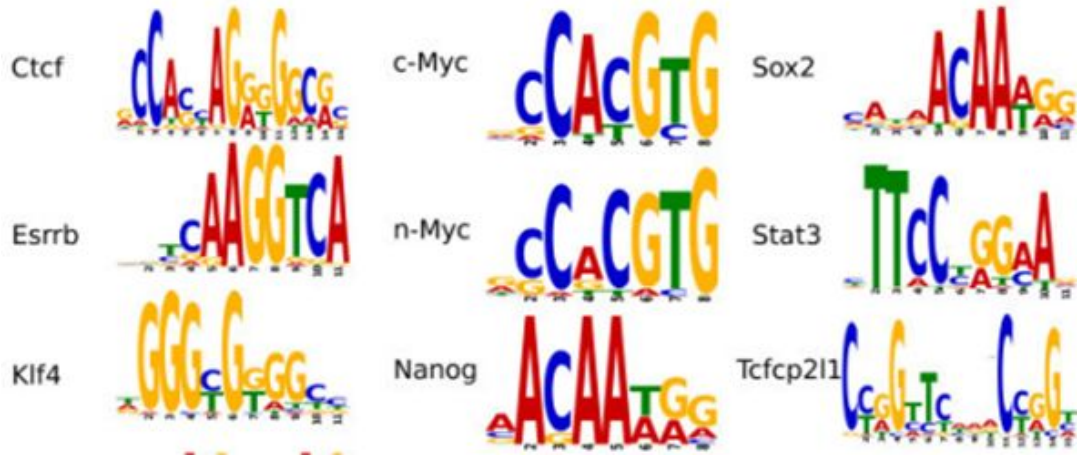
MEME
Multiple Em for Motif Elicitation
Version 5.5.2

https://meme-suite.org/meme/tools/meme

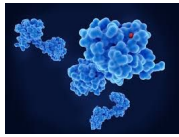
Metodi per l'analisi di sequenze proteiche



L'algoritmo MEME lavora identificando le posizioni più conservate all'interno di un insieme di sequenze e definendo un modello di motivo che spiega la variabilità delle sequenze in questi punti. Il modello di motivo è rappresentato come un'istogramma di probabilità per ogni posizione del motivo, dove ogni bin rappresenta un diverso nucleotide o amminoacido.

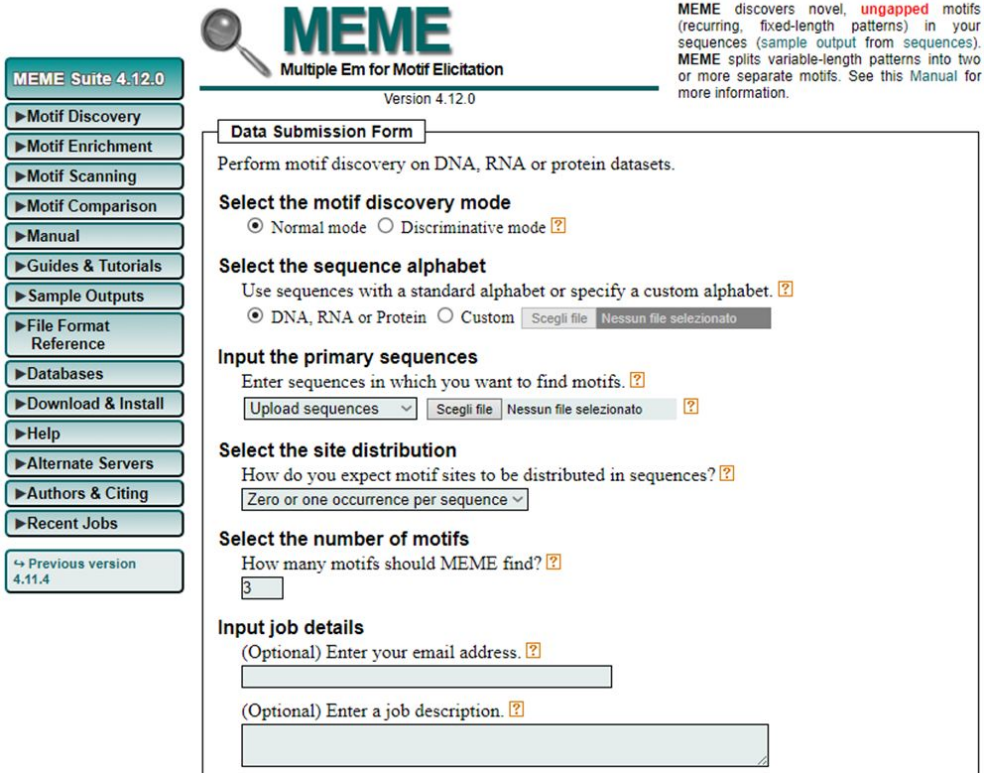


Metodi per l'analisi di sequenze proteiche

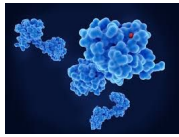


Quindi lo posso utilizzare quando ho un gruppo di sequenze che presuppongo possano essere ad esempio coregolate (sotto il controllo dello stesso promotore o regolate dallo stesso fattore di trascrizione), ma non conosco la sequenza consensus dello stesso

- Possono essere identificati più motivi, anche di lunghezza diversa e sono tollerate alcune variazioni
- I risultati vengono riportati come sequenze consensus (loghi) associati a p-value

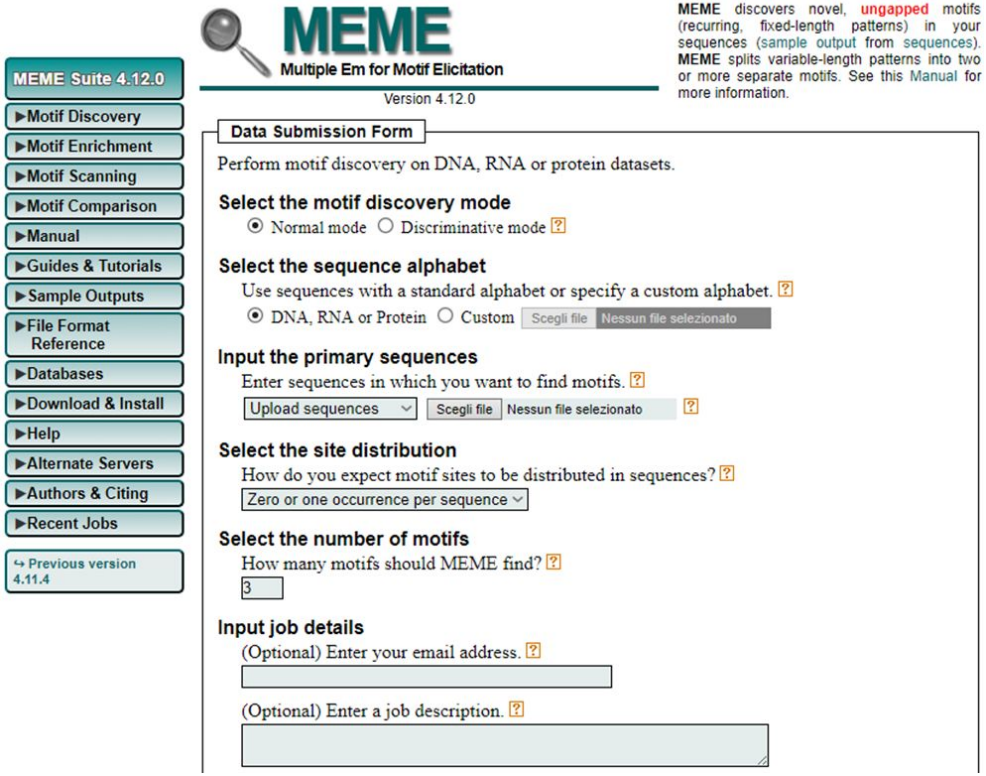


Metodi per l'analisi di sequenze proteiche

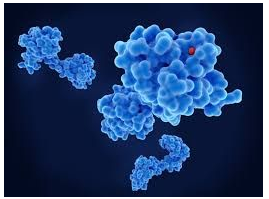


Quindi lo posso utilizzare quando ho un gruppo di sequenze che presuppongo possano essere ad esempio coregolate (sotto il controllo dello stesso promotore o regolate dallo stesso fattore di trascrizione), ma non conosco la sequenza consensus dello stesso

- Possono essere identificati più motivi, anche di lunghezza diversa e sono tollerate alcune variazioni
- I risultati vengono riportati come sequenze consensus (loghi) associati a p-value



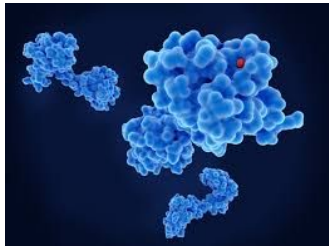
Metodi per l'analisi di sequenze proteiche



Come possono essere applicati gli HMM in biologia?

- Training: dato un set di sequenze omologhe non allineate è possibile allinearle ed aggiustare i parametri di transizione e output di stato per definire un HMM che ben rappresenti i pattern evolutivi inferiti dalle sequenze prese in esame.
- Individuazione di omologhi distanti: dato un HMM ed una sequenza da testare, è possibile calcolare la probabilità che l'HMM abbia generato la sequenza stessa. Se un HMM «allenato» su una familglia di sequenze note può generare la sequenza testata con una probabilità piuttosto alta, c'è una buona probabilità che questa sequenza appartenga effettivamente alla famiglia.
- Allineamento di ulteriori sequenze: la probablità che una sequenza di stati si verifichi può essere computata a partire delle probabilità di transizione da stato a stato. Trovare la successone di stati che l'HMM userebbe con maggiore probabilità per generare un allineamento tra due o più sequenze prese in esame può ottimizzare l'allineamento stesso secondo criteri probabilistici

Metodi per l'analisi di sequenze proteiche

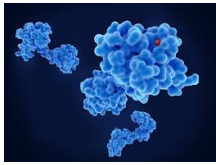


Gli HMM presentano diversi vantaggi rispetto ad altre tecniche di analisi di sequenze. Innanzitutto, permettono di modellare la variabilità delle sequenze che rappresentano una famiglia di proteine o di nucleotidi, e quindi di catturare la diversità di sequenze all'interno di una stessa famiglia.

Inoltre, gli HMM sono in grado di modellare sequenze di lunghezza variabile, rendendoli adatti per l'analisi di famiglie di sequenze eterogenee.

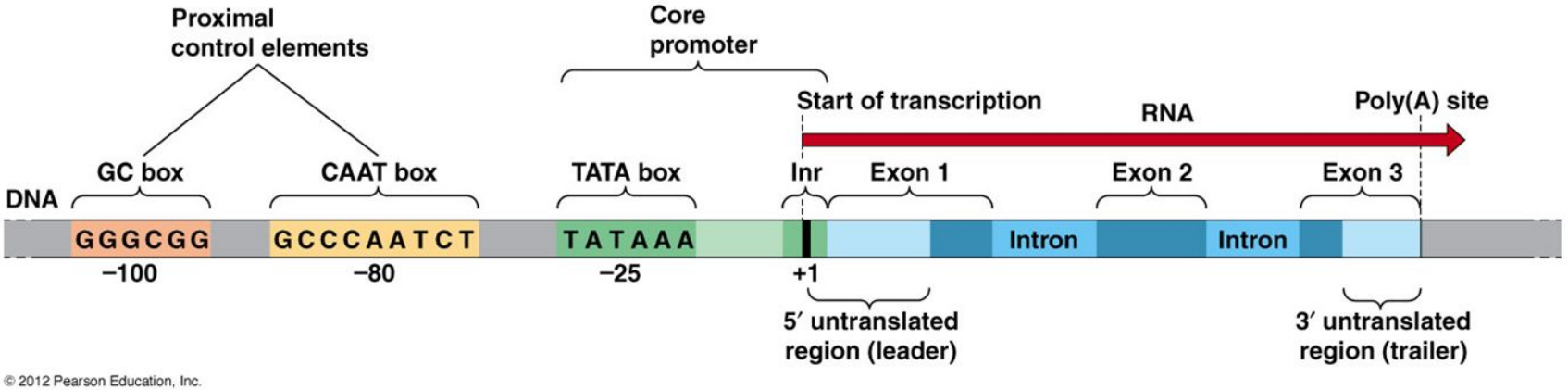
Infine, gli HMM sono in grado di fare previsioni sulle sequenze sconosciute basandosi sulla somiglianza con le sequenze già note, consentendo quindi di identificare nuove sequenze all'interno di una stessa famiglia di proteine o nucleotidi.

Metodi per l'analisi di sequenze proteiche

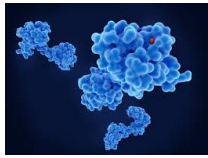


Applicazione – pattern e motivi funzionali

- La RNA polimerasi riconosce l'inizio del gene. Viene diretta sul TSS (Transcription Start Site) sulla base della sua affinita' per la specifica sequenza upstream al gene, ovvero il promotore. La doppia elica viene aperta dove inizia la sintesi del messaggero.
- I TF (fattori di trascrizione) legano la sequenza promotrice (e gli enhancers) formando un complesso multiproteico
- Il complesso recluta la pol II complessata ad alcuni GTF (general transcription factors e questa si lega al promotore core.



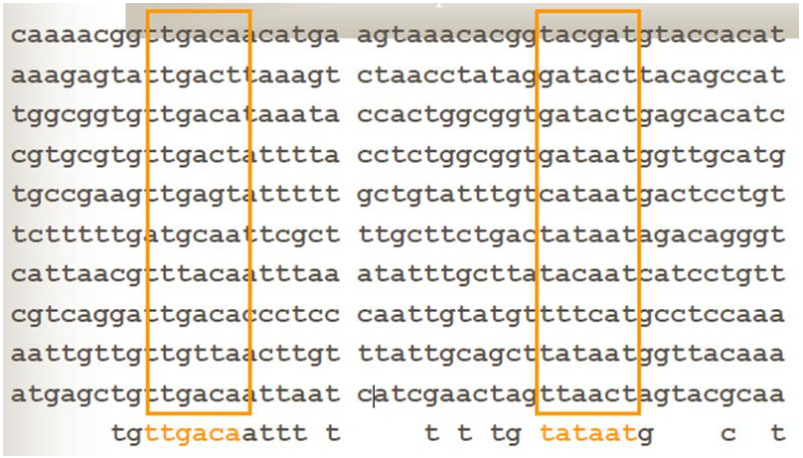
Metodi per l'analisi di sequenze proteiche



Schematicamente un promotore per la Pol II è composto da:

-
- PROMOTORE CORE - regione sufficiente a deteminare il TSS esatto
- PROMOTORE PROSSIMALE - 200-300 bp upstream al TSS, responsabile, almeno in parte, della modulazione dell'espressione
- PROMOTORE DISTALE - 100 bp – 2 Mb

Queste regioni sono, almeno in parte, conservate e riconducibili ad un consensus, in quanto vengono riconosciute da fattori che solitamente regolano la transcrizione di svariati geni

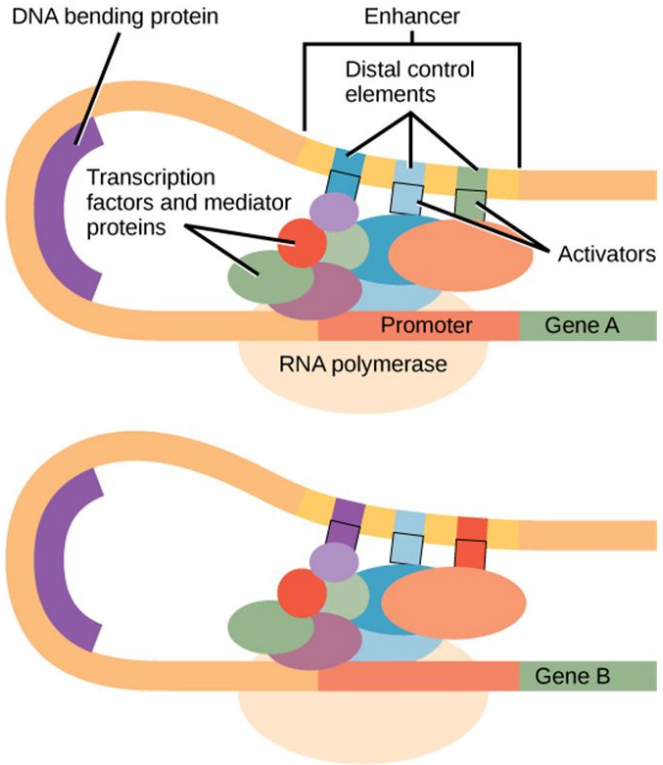


Nell'esempio a fianco vediamo un allineamento tra i promotori di alcuni geni batterici trascritti tramite il fattore housekeeping sigma70

Due regioni in cui la sequenza e' conservata: -10 - 35 dallo start site (motivi TTGACA e TATAAT

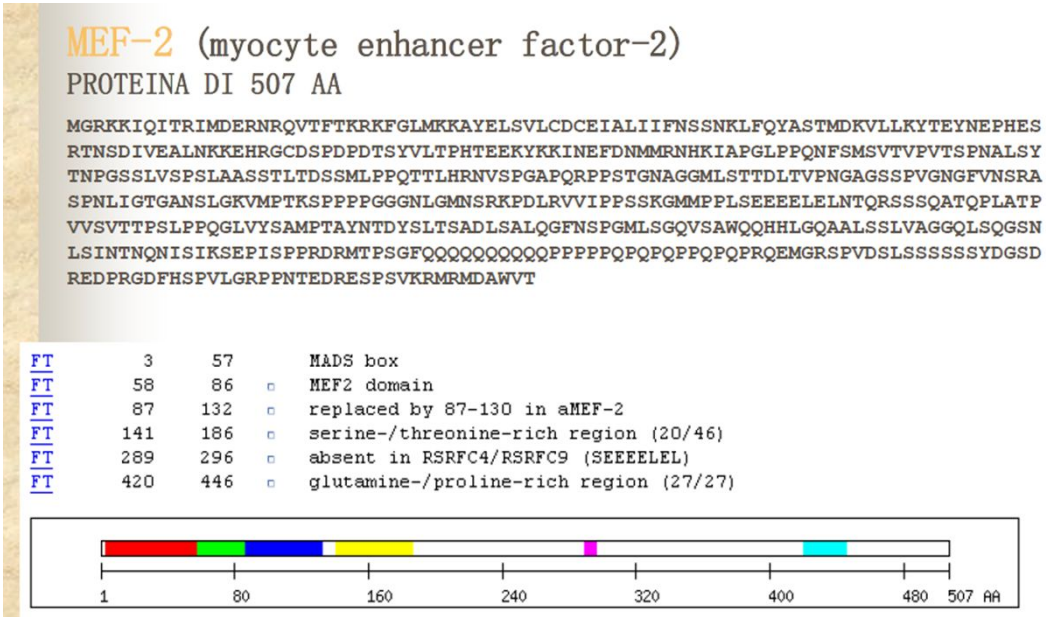
Metodi per l'analisi di sequenze proteiche

- I motivi TTGACA and TATAAT che abbiamo visto sono i segnali che vengono riconosciuti dalla subunita' sigma70 della polimerasi.
- La "forza" relativa di un promotore e' proporzionale alla sua similarita' ad una specifica sequenza consenso.
- Mutazioni nelle regioni -10 and –35 alterano la "forza" del promotore
- Esperimenti tipo footprinting o methylation interference confermano la loro attivita'
- Oltre a questi elementi «generali» ne esistono moltissimi ben più specifici che regolano positivamente o negativamente la trascrizione, andando a fungere da siti di legame per una serie di TF e fattori accessori facenti parte di macrocomplessi molecolari



Metodi per l'analisi di sequenze proteiche

- I TF sono proteine in grado di diffondere nelle cellule e di interagire con sequenze bersaglio che possono essere presenti su numerose molecole di DNA in siti diversi di un genoma, anche su cromosomi diversi
- I geni che presentano questi elementi sono quindi coregolati
- Esempio IRF (interferon responsive facors) legano gli IRE (Interferon Responsive Elements) che sono posizionati a monte di svariati geni coinvolti nei processi infiammatori



Metodi per l'analisi di sequenze proteiche

- MEF-2 riconosce una determinata sequenza consensus, il cui intorno non è ben definito
- Nell'esempio a fianco vediamo la matrice relativa al sito consensus di legame con le frequenze di occorrenza diìei 4 nucleotidi in determinate posizioni
- Sulla base d queste osservazioni, risulta evidente che il consensus di legame, ovvero la stringa di nucleotidi riconosciuta in modo specifico da MEF-2, è:

NNNNNNKCTAWAAATAGMNNNN

- Tante più mutazioni saranno presenti in questa stringa, tanto minore sarà la capacità di riconoscimento ed interazione... in sostanza ogni posizione può essere associata ad una probabilità di osservazione di uno dei 4 nucleotidi

POS.	A	C	G	T	
01	5	28	25	42	N
02	16	32	31	21	N
03	18	36	27	19	N
04	19	25	33	23	N
05	22	12	43	23	N
06	33	9	21	37	N
07	20	4	43	33	K(G o T, Keto)
08	3	85	3	9	C
09	3	8	0	89	T
10	85	0	0	15	A
11	57	0	0	41	W(A O T, Weak)
12	91	0	1	8	A
13	96	0	0	4	A
14	93	0	1	6	A
15	0	0	0	100	T
16	100	0	0	0	A
17	9	0	90	1	G
18	34	46	11	9	M(A o C,Amino)
19	36	28	8	28	N
20	20	37	15	28	N
21	30	34	13	23	N
22	23	23	22	32	N