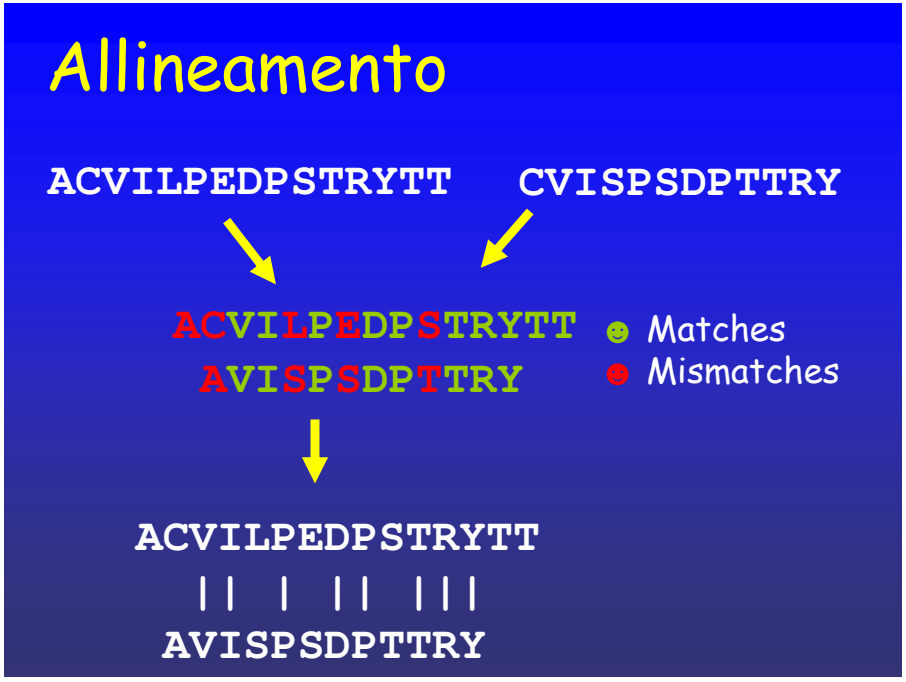
Evoluzione molecolare

Un organismo può possedere nel suo genoma più copie dello stesso gene a causa di eventi di duplicazione. I geni duplicati sono liberi di mutare, poichè la loro funzione continua a essere mantenuta dal gene originale.

Le sequenze evolvono attraverso mutazioni puntiformi, delezioni e inserzioni. La mutazione puntiforme è la sostituzione di una lettera (aminoacidica o nucleotidica) con una lettera diversa. La delezione è l'eliminazione di una o più lettere consecutive da una sequenza. L'inserzione è l'aggiunta di una serie di lettere consecutive, in una posizione di una sequenza, generalmente derivanti da un altro gene. Nella figura si vedono una serie di eventi di mutazione (in rosso) che si accumulano su una copia duplicata (in verde) della sequenza di una proteina.

Confronti di sequenze

In questo capitolo sono trattate solamente sequenze di proteine. Tutte le considerazioni fatte restano comunque valide anche per sequenze di DNA o RNA. Due sequenze proteiche che derivano da un unica sequenza mantengono, nonostante il processo di evoluzione molecolare, per un lungo periodo una certa similarità fra di loro. Confrontare sequenze diverse ha lo scopo di trovare similarità conservate fra le lettere che le compongono e ipotizzare quindi una loro eventuale relazione di parentela o funzionale.



Allineamenti

Per confrontare sequenze proteiche è necessario costruirne un allineamento. Una coppia di sequenze è allineata quando si è stabilita una corrispondenza tra alcuni dei residui dell'una e alcuni residui dell'altra. L'allineamento di due sequenze viene raffigurato affiancandole una sopra l'altra in modo da far corrispondere le coppie di lettere appaiate.

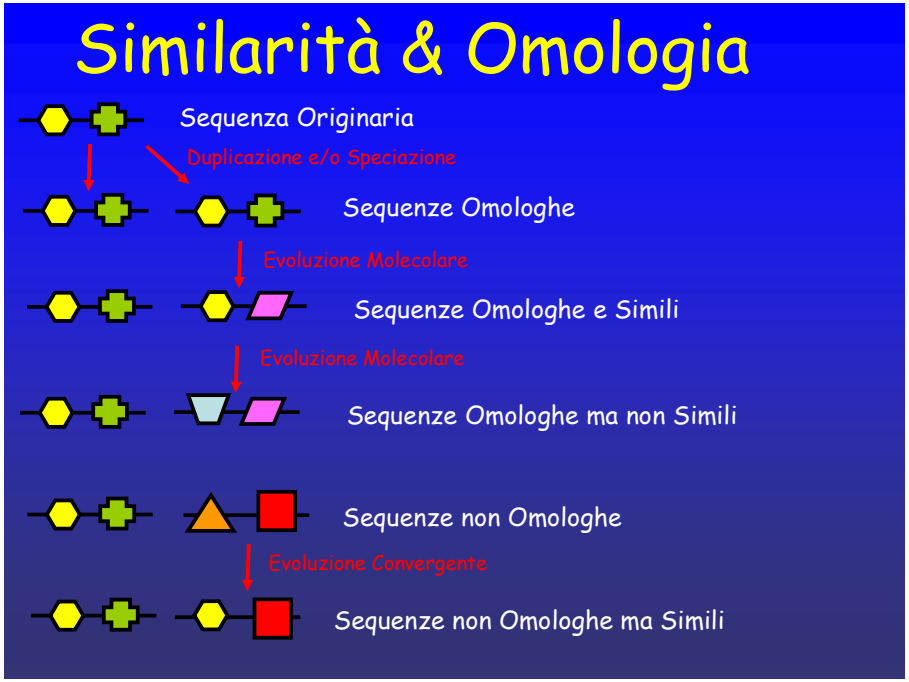
Scopo di un allineamento è l'appaiamento del maggior numero di posizioni degli aminoacidi che probabilmente derivano da un'unica posizione nella sequenza originaria.

Definiamo "matches" le corrispondenze stabilite tra coppie di aminoacidi dello stesso tipo (in verde) e mismatches (in rosso) le coppie di aminoacidi di tipo diverso. Nel rappresentare un allineamento, per dare modo di visualizzare rapidamente il numero di matches si possono usare delle stanghette verticali in corrispondenza delle coppie di aminoacidi identici.

Il modo più semplice per valutare la qualità di un allineamento è il conteggio degli aminoacidi del medesimo tipo che vengono appaiati.

Punteggio di Identità

Il punteggio di Identità serve a dare un valore esatto della qualità di un allineamento. Il punteggio corrisponde al numero di matches fra le due sequenze. Nella figura sono rappresentate due possibili allineamenti delle stesse sequenze. Nel secondo allineamento la sequenza inferiore è stata fatta scorrere di una posizione verso destra, facendo quindi corrispondere coppie di aminoacidi diverse. Nel primo caso il punteggio di Identità è 8 nel secondo caso solamente 2. Il primo allineamento ha quindi maggiore probabilità di aver evidenziato le reali coppie di aminoacidi corrispondenti nell'eventuale sequenza progenitrice.



Similarità & Omologia

Due sequenze si definiscono omologhe se derivano da un'unica sequenza progenitrice. Uno degli scopi di allineare due sequenze è quello di valutare se possano o meno essere omologhe.
Un allineamento però può solo misurare la similarità di due sequenze, ovvero la quantità di residui conservati che possiamo osservare fra di loro. Ad una maggiore similarità tra due sequenze corrisponde una maggiore probabilità che siano omologhe. Quando è trascorso molto tempo dalla loro separazione, anche sequenze omologhe possono non apparire più simili. D'altra parte proteine non omologhe a causa di fenomeni di evoluzione convergente possono acquisire una certa similarità nelle loro sequenze. In figura A è rappresentato il caso di due sequenze omologhe che con il passare del tempo divergono sino a non essere più simili. In figura B è rappresentato il caso di due sequenze non omologhe che per un evento di evoluzione convergete acquistano un tratto di sequenza simile.

Percentuale di identità

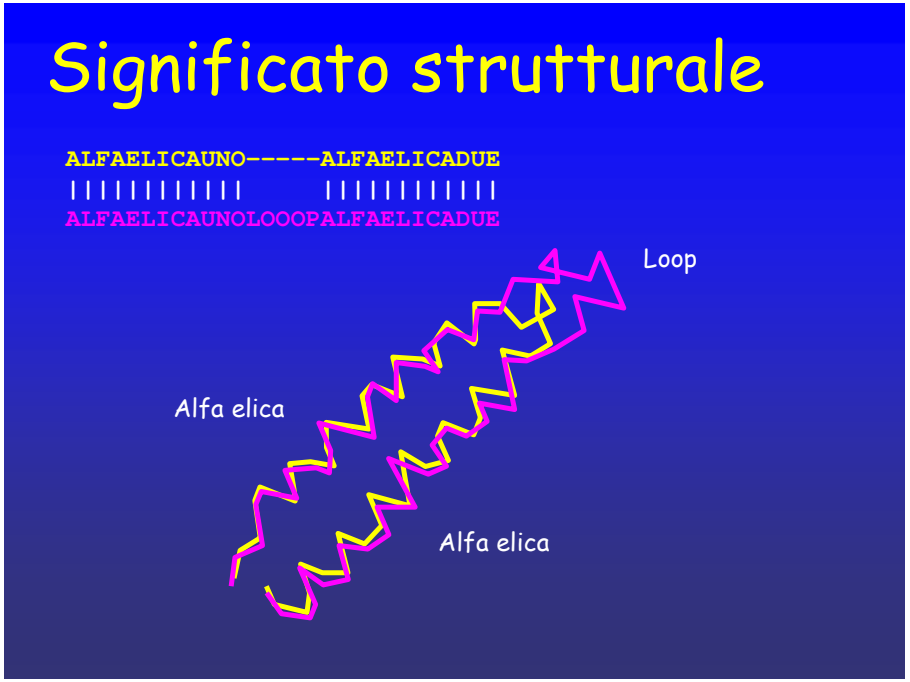
Il punteggio di identità misura il numero dei matches senza considerare la lunghezza delle sequenze. Sequenze più lunghe tendono ad avere un numero maggiore di matches senza che questo indichi necessariamente una loro maggiore similarità. Per normalizzare il punteggio di Identità lo si moltiplica per due e lo si divide per il numero totale degli aminoacidi delle due sequenze. Il numero che si ottiene è la percentuale di coppie identiche all'interno dell'allineamento e si chiama percentuale di identità. Una maggiore percentuale di identità indica una qualità maggiore dell'allineamento indipendentemente dalla lunghezza delle sequenze.

Ricerca del migliore allineamento fra due sequenze

Date due sequenze esistono diversi modi di allinearle. Un metodo semplice per trovare il migliore allineamento tra due sequenze date consiste nel calcolare i punteggi di identità di tutti gli allineamenti possibili e scegliere quello corrispondente al punteggio più alto. Per generare tutti i possibili allineamenti, si fanno scorrere le due sequenze l'una sull'altra spostandosi di una lettera alla volta. Ogni spostamento, facendo corrispondere coppie diverse di aminoacidi, corrisponde ad un diverso allineamento. Il migliore dei possibili allineamenti sarà quello con il punteggio di Identità migliore. Il migliore allineamento delle due sequenze dell'esempio è quello con un punteggio di Identità pari a 4 (i matches sono indicati in verde i mis-matches in rosso).
Il numero totale di possibili allineamenti è pari alla somma delle lettere delle due sequenze meno 1. Nell'esempio, due sequenze lunghe 6 aminoacidi hanno 11 (6+6-1) possibili modi di essere allineate.
Il numero totale di aminoacidi che abbiamo dovuto confrontare (numero delle stanghette) è pari al prodotto delle lunghezza delle due sequenze. In questo caso sono stati fatti 6*6 (36) confronti di aminoacidi.
Il tempo impiegato per trovare il migliore allineamento possibile usando questo metodo è quindi proporzionale a L^2 dove L è la lunghezza media delle sequenze.

I Gaps

Se una proteina ha subito un evento di inserzione o delezione, un suo appaiamento con una proteina omologa non potrà più riassociare tutte le coppie di aminoacidi correttamente, perchè le nuove lettere inserite interrompono l'appaiamento delle coppie simili. Per risovere il problema, quando si allineano due sequenze è possibile inserire dei "gaps" o interruzioni, rappresentati con dei trattini, fra le lettere di una o di entrambe le sequenze. L'apertura di gaps consente di compensare in una delle due sequenze l'inserimento di residui avvenuto nell'altra. L'effetto è quello di ripristinare la distanza corretta fra gli aminoacidi permettendo di appaiare correttamente il maggior numero di coppie. Nella figura, l'inserimento nella prima sequenza di un gap lungo 5 posizioni corrispondente ad un inserimento di 5 residui nella seconda sequenza permette di ottenere un allineamento con identità 18 rispetto ai due possibili allineamenti con punteggio 7 o 11 ottenibili senza l'uso delle inserzioni.



Significato strutturale dei gaps

A sequenze omologhe corrispondono generalmente proteine di struttura simile. E' spesso difficile inserire o rimuovere aminoacidi dalle proteine senza alterarne la struttura globale e mantenendo la corrispondenza tra aminoacidi omologhi anche in struttura. In questo esempio un'inserzione di 5 aminoacidi non cambia la struttura tridimensionale ma semplicemente allunga il "loop" di giunzione fra due alfa eliche.

Allineamenti con gaps

Numero allineamenti $(2n)! / (n!)^2$

ACD EFG	ACD E–FG	ACD E–FG	A–CD EF–G	A–C–D EF–G	AC–D E–FG	A–C–D EFG	ACD EF–G
ACD EFG	AC–D EFG	AC––D EFG	A–CD EF–G	A–CD E–F–G	ACD E–F–G	AC–D EF–G	ACD EF–G
ACD EFG	AC–D EFG	AC––D EFG	A–CD E–FG	AC–D E–F–G	A–C–D EFG	AC–D EF–G	
ACD EFG	AC–D EFG	A––CD EFG	A–CD E–FG	A––CD E–FG	A–C–D E–FG	A–CD EF–G	
	A–CD EFG	A––CD EFG	AC–D E–FG	A–CD EFG	ACD EF––G	ACD E–FG	
ACD EF–G	A–CD EFG	ACD E–F–G	AC–D E–FG	ACD EF–G	ACD E–FG	ACD E–FG	

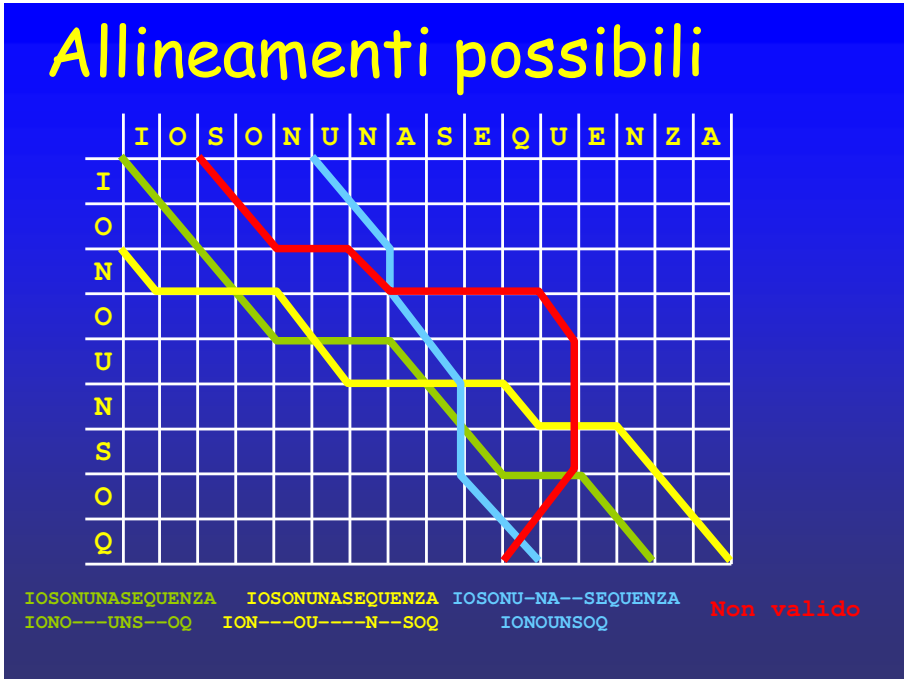
Numero di allineamenti con inserzioni

Il migliore allineamento tra due sequenze, se si prende in considerazione l'uso dei gaps, non può più essere trovato calcolando il punteggio di tutti gli allineamenti possibili. Infatti due sequenze, considerando tutte le possibili combinazioni di gaps inseriti, possono essere allineate in un numero altissimo di modi, pari a $(2n)! / (n!)^2$. Non è più possibile quindi provare tutte le combinazioni se non per sequenze molto corte, ma è necessario trovare un altro metodo.

Matrici di allineamento

Matrici di allineamento

Spesso cambiando il modo di vedere le cose possiamo trovare soluzioni inattese a problemi che sembravano irrisolvibili. Nel caso della ricerca dei migliori allineamenti con gaps fra due sequenze la soluzione viene dalla rappresentazione diversa dell'allineamento. Esiste un modo alternativo rispetto a quello utilizzato sino a questo punto di scrivere le due sequenze una sull'altra con gli aminoacidi affiancati a coppie. Il metodo consiste nel rappresentare un allineamento usando una matrice. Per generare una matrice di allineamento ad ogni colonna si fa corrispondere un aminoacido della prima sequenza e ad ogni riga un aminoacido della seconda. A questo punto un allineamento è rappresentato da una linea continua che attraversa la matrice partendo dalla prima riga o colonna e terminando sul'ultima riga o colonna. La linea può correre lungo le linee orizzontali e verticali che disegnano la griglia della matrice o attraversare le celle in diagonale. Dove la linea di allineamento attraversa una cella della matrice in diagonale (in verde) gli aminoacidi che stanno sulla riga e sulla colonna della cella sono appaiati. Dove la linea segue i bordi di una cella in direzione verticale o orizzontale (in rosso) l'amminoacido che si trova sulla riga o sulla colonna corrispondente è allineato con un gap.
La rappresentazione di un allineamento usando una matrice è alternativa in tutto e per tutto alla rappresentazione classica. A partire infatti da una qualsiasi delle due è facilmente ricavabile l'altra.



Allineamenti possibli con gaps

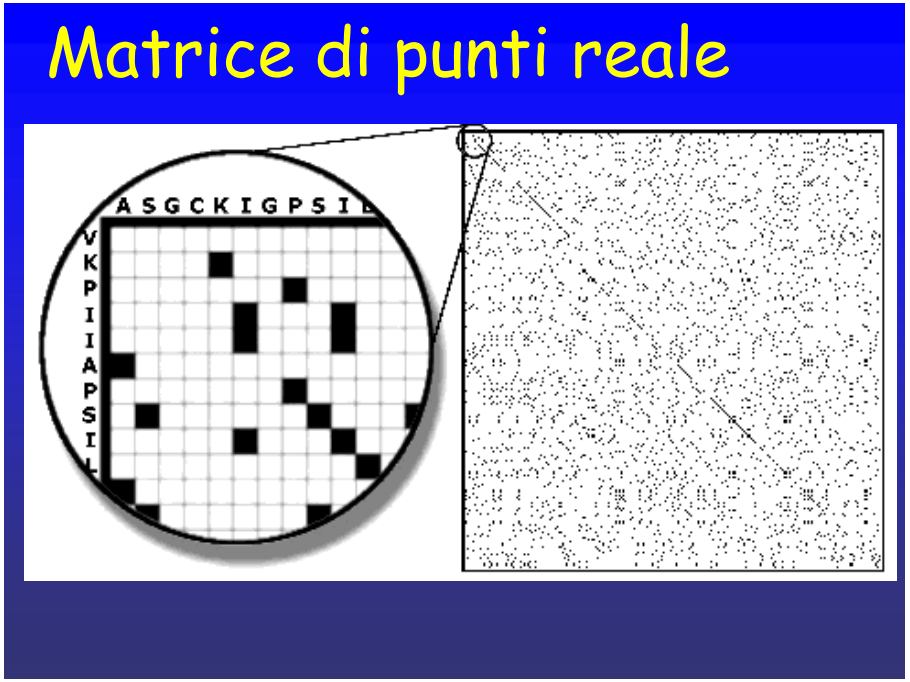
Una matrice di allineamento permette di rappresentare tutti i possibili allineamenti fra due coppie di sequenze. Ogni allineamento è rappresentato da una diversa linea continua che parte dalla prima riga o colonna e termina nell'ultima riga o colonna. Le direzioni permesse per un allineamento sono verso destra, verso il basso o in diagonale verso destra in basso. Qualunque direzione diversa da queste tre genera un allineamento non valido in cui ad esempio lo stesso aminoacido di una sequenza potrebbe essere allineato contemporaneamente a più di un residuo sulla seconda. Nell'esempio, l'allineamento errato (in rosso) appaia la lettera U contemporaneamente sia alla seconda che alla terza O dell'altra sequenza.

Migliore allineamento su matrice

L'allineamento fra due sequenze con il migliore punteggio di Identità è quello che appaia il maggior numero possibile di coppie di aminoacidi identici. Su una matrice di allineamento questo significa che la linea che rappresenta il migliore allineamento è quella che attraversa il maggior numero possibile di celle che si trovino all'incrocio di righe e colonne che contenenti lo stesso tipo di aminoacido.

Matrice di punti (Dot matrix)

La costruzione di una matrice di punti è il primo dei passaggi necessari per trovare il miglior allineamento possibile con gaps fra due sequenze. In una matrice di punti sono rappresentate tutte le coppie possibili di aminoacidi identici fra due sequenze. Per costruire una matrice di punti si procede come per la costruzione di una matrice di allineamento, scrivendo sulla prima riga gli aminoacidi della prima sequenza e sulla seconda riga gli aminoacidi della seconda. A differenza del caso della matrice di allineamento, quello che non possiamo fare è tracciare la linea corrispondente all'allineamento perchè ancora non conosciamo il percorso migliore. Possiamo però facilmente identificare le celle dove è conveniente che il migliore allineamento passi. Queste celle sono quelle che corrispondono ad aminoacidi uguali nelle due sequenze. Indicheremo quindi con un punto (dot) tutte le celle che hanno una lettera uguale sulla riga e sula colonna (vedi figura).

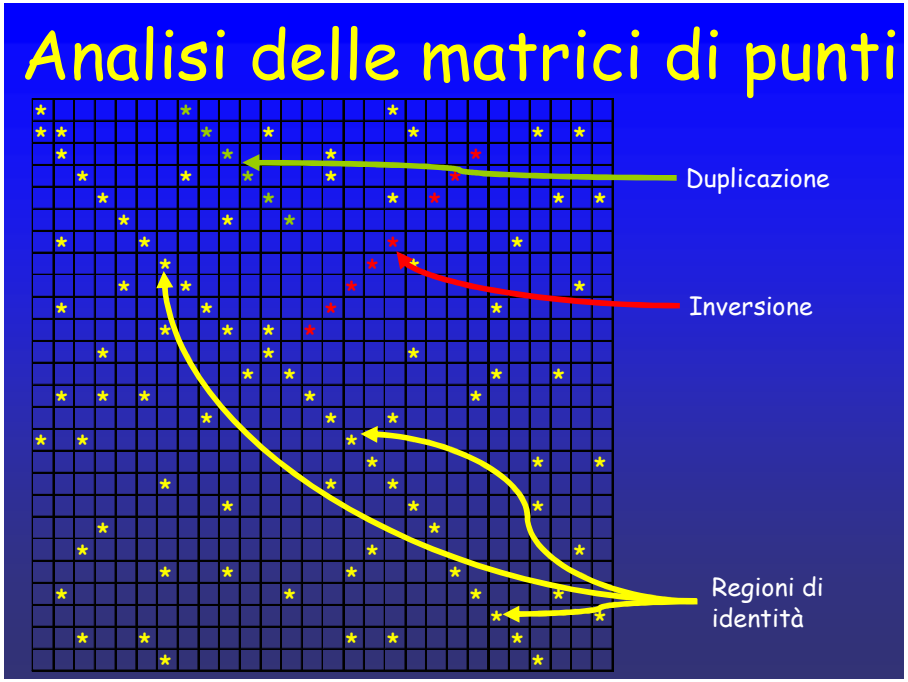


Ricerca del migliore allineamento

Su di una matrice di punti il miglior allineamento fra le due sequenze è spesso visibile a occhio. Il migliore allineamento è rappresentato infatti dalla linea che attraversa in diagonale il maggior numero possibile di punti, appaiando quindi il maggior numero di coppie di aminoacidi identici.
Se ogni punto di una "dot matrix" fosse una moneta d'oro, trovare il migliore allineamento corrisponderebbe a trovare la strada attraverso la matrice che ci permette di raccoglierne il maggior numero possibile.
Nell'esempio l'allineamento migliore fra le due sequenze è rappresentato dalla linea rossa che attraversa sette punti. Una volta tracciato il percorso, possiamo scrivere l'allineamento in forma estesa e verificare che effettivamente ha un punteggio di identità pari a 7.

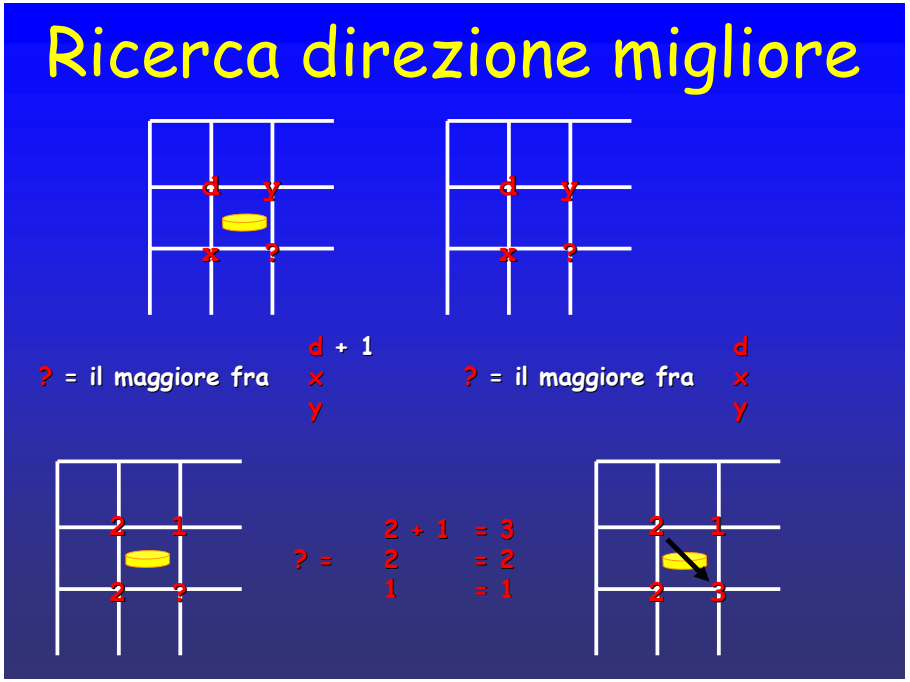
Matrici di punti di sequenze reali

Lunghe sequenze reali di proteine generano matrici di punti molto complesse. Se le due sequenze utilizzate sono molto simili fra di loro, si genera una lunga linea di punti più o meno continua parallela alla diagonale della matrice(vedi figura). Se le due sequenze hanno invece un basso livello di similarità, spesso non è possibile individuare velocemente regioni si similarità che indichino il percorso del migliore allineamento.



Analisi delle matrici di punti

Le matrici di punti, oltre a fornire indicazione sulle lunghe regioni di identità, ovvero i tratti composti da aminoacidi identici (in giallo) nelle due sequenze, possono fornire altre informazioni sui loro tratti comuni. Inversioni presenti fra le due sequenze, ovvero regioni identiche in sequenza ma invertite nella direzione, appariranno come segmenti diagonali (in rosso) che correranno in direzione perpendicolare rispetto alle regioni di identità. Duplicazioni, ovvero tratti di una sequenza presenti in più copie su di un altra, appariranno come tratti diagonali (in verde) che correranno parallelli rispetto ad un'altra regione di identità.



Massimo punteggio in un punto della matrice

Dato un qualunque punto sulla griglia della matrice di punti (indicato da un punto interrogativo nella figura), percorrendo un allineamento valido, possiamo giungerci da solo tre possibli direzioni, dall'alto, da sinistra o dalla diagonale (punti x, y e d della figura). Il numero di monete d'oro che possiamo sperare di aver raccolto arrivando in quel punto dipenderà quindi solamente dal numero di monete già raccolte in uno degli altri tre punti. In particolare se vi arriveremo provenendo dall'alto o da sinistra (inserzione di un gap), il numero di monete raccolte resterà invariato, se vi arriveremo invece in diagonale potremmo raccogliere l'eventuale nuova moneta contenuta nella cella (appaiamento di una coppia di aminoacidi identici), incrementando il totale di monete raccolte di +1. Il massimo numero di monete raccolte nel punto 'd' sarà quindi il massimo fra i valori x, y e d+1 o d+0 a seconda che la cella contenga o meno una moneta.

Algoritmo di programmazione dinamica

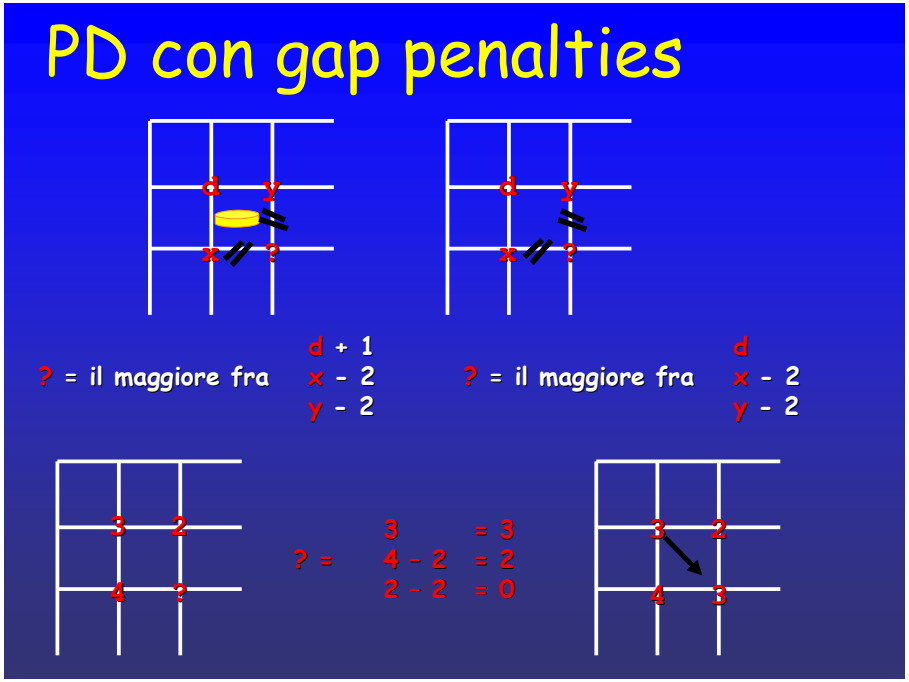
Partendo da questa considerazione è possibile sviluppare un metodo automatico che garantisca di trovare il migliore allineamento possibile in modo esatto e veloce.
Il metodo calcola il maggior numero di monete recuperabili per ogni punto sulla matrice dove può passare un allineamento e quindi cerca il punto con il totale più alto, questo sarà il punto di arrivo della linea che rappresenta il migliore allineamento.
Prima di iniziare a calcolare i valori massimi di monete in ogni punto il metodo assegna un punteggio iniziale pari a 0 a tutti i possibili punti di partenza per un allineamento, ovvero a tutti gli incroci sopra la prima riga e a sinistra della prima colonna.
Dal momento che per conoscere il punteggio in ogni punto è necessario prima conoscere il punteggio nei tre punti precedenti, il metodo di programmazione dinamica comincia a calcolare i punteggi dal punto in alto a sinistra sulla griglia. Il metodo si muove poi, mano a mano che vengono calcolati nuovi valori, verso l'ultimo punto della griglia in basso a destra, che è l'ultimo a poter essere determinato.
Quando sono stati calcolati i punteggi massimi per tuti i punti di incrocio sulla griglia, viene cercato il punto di arrivo dell'allineamento, ovvero il punto sulla griglia dove è possibile passare dopo aver raccolto il maggior numero di monete. Questo punto sarà posto obbligatoriamente o sull'ultima riga o sull'ultima colonna.

Percorso migliore

Dopo aver identificato il punto di arrivo del migliore allineamento possibile per le due sequenze (cerchio nero nella figura), il metodo deve ricostruire la strada (allineamento) attraverso la quale si è arrivati in quel punto.
Ci sono due modi per determinare da quale delle tre posizioni si era arrivati a un determinato punto del percorso: ricordarselo o calcolarlo nuovamente. Infatti il punto da cui si è arrivati è per forza l'unico dei tre da cui si poteva ottenere il punteggio attuale.
Se arrivando da due punti diversi si può ottenere lo stesso punteggio, vuol dire che la linea si biforca, dando luogo a due diversi allineamenti possibili (entrambi caratterizzati dal punteggio migliore).
Quando si arriva indietro fino alla prima riga o colonna l'allineamento è determinato completamente.
Se si trova più di un allineamento alternativo, tutti gli allineamenti trovati possono essere scritti per esteso .

Penalità per inserimento di gaps

Gli eventi di inserzione e delezione rispetto a quelli di sostituzione aminoacidica, sono più infrequenti dal punto di vista evolutivo, perchè più difficili da accettare all'interno della struttura tridimensionale di una proteina. Nel calcolo del punteggio di identità di un allineamento che è stato adottato sino a questo punto, un residuo appaiato con un gap vale 0 punti come una coppia di residui differenti. Per migliorare il modello di punteggio è necessario che l'inserzione di un gaps sia più costosa del semplice appaiamento di un aminoacido con uno diverso. Assegnamo quindi ad ogni trattino (gap) inserito in un allineamento un punteggio negativo che abbassi il punteggio di identità dell'allineamento stesso. Questa penalità si chiama penalità per l'inserimento di gaps, a cui assegnamo per adesso il valore di -2 punti.



Programmazione dinamica con penalità per i gaps.

Per inserire il calcolo delle penalità per i gaps nel punteggio di un allineamento trovato con l'algoritmo di programmazione dinamica, bisogna modificare il metodo usato per determinare il miglior punteggio raggiungibile in ogni punto. La modifica consiste nel diminuire per il valore della penalità (-2) (i trattini neri nella figura) il numero delle monete raccolte, ogni volta che si viene dalla direzione verticale o da quella orizzontale (dal punto y o dal punto x in figura). Il punteggio del punto d sarà quindi calcolato come il valore massimo tra x -2, y - 2 o d (+1 o +0 a seconda della presenza della moneta). Questa è l'unica modifica. Per il resto l'algoritmo di programmazione dinamica rimane uguale al caso degli allineamenti senza gaps.

Penalità per estensione gaps

```
a) IOSONOUNASEQUENZA
1) ISONOUNSEQENZO
2) IOSONOSEQUENZO

a) IOSONOUNASEQUENZA
   | |||||| ||| |||        Identità = 13 -2 -2 -2 = 7
1) I-SONOUN-SEQ-ENZO

a) IOSONOUNASEQUENZA
   ||||||    |||||||      Identità = 13 -2 -1 -1 = 9
2) IOSONO---SEQUENZO
```

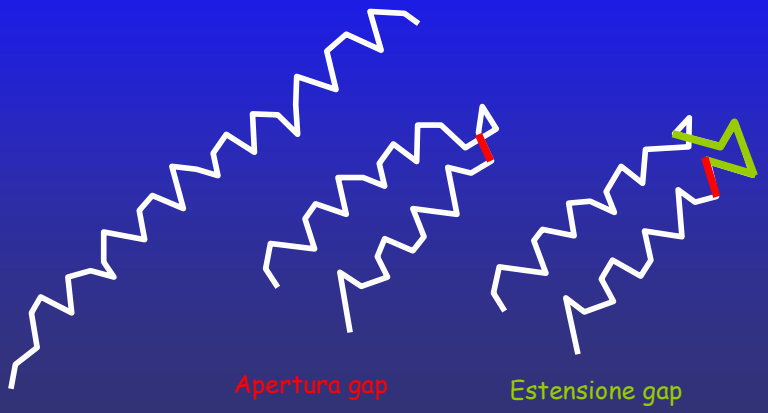
GAP extension penalty = -1 per ogni estensione di un gap già presente

Penalità per estensione dei gaps

La quantità di residui inseriti o rimossi in un singolo evento di inserzione o delezione è meno determinante per la funzionalità di una proteina rispetto all'evento in se stesso. Questo significa che tre inserzioni diverse di un aminoacido in tre punti diversi di una sequenza sono molto più difficili da accettare rispetto ad un'unica inserzione di tre residui consecutivi.
Per considerare questo fattore la penalità per il primo gap inserito in un allineamento deve essere maggiore della penalità per i gaps successivi adiacenti al primo. La penalità per l'inserzione di gaps adiacenti ad un gap già esistente si chiama "penalità per l'estensione di un gap" ed èsempre inferiore alla "penalità per l'inserzione di un gap". Essendo di -2 la penalità per l'inserzione, decidiamo di assegnare un valore di -1 alla penalità per l'estensione.

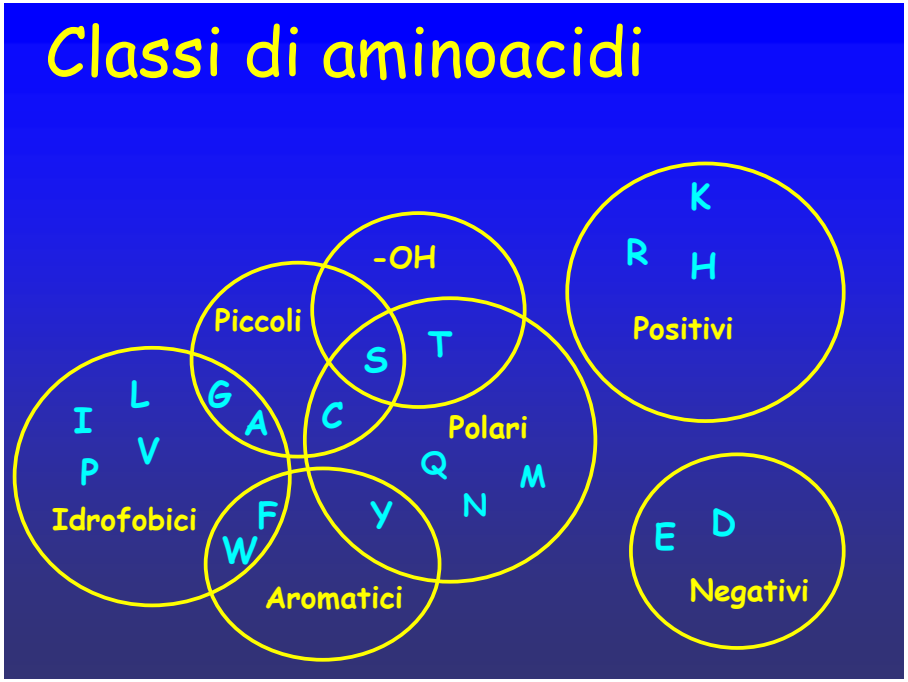
Significato strutturale

```
ALFAELICAUNOALFAELICADUE
ALFAELICAUNOLALFAELICADUE        apertura gap
ALFAELICAUNOLOOOPALFAELICADUE    estensione gap
```



Apertura ed estensione di gaps

La ragione per cui è più facile inserire più residui in un singolo evento di inserzione, rispetto all' inserzione degli stessi residui in più eventi diversi e in in punti diversi, ha anche delle motivazioni di tipo strutturale. Ad ogni inserzione aminoacidica la struttura di una proteina può subire drastici cambiamenti per adattare il residuo al suo interno. Nell'esempio della figura, una singola alfa elica in seguito all'inserzione di un singolo residuo (in rosso) può spezzarsi in due alfa eliche interrotte da una regione di collegamento "loop". Dopo questo primo evento una seconda inserzione di altri 4 residui nello stesso punto (in verde) non altera più la struttura, limitandosi ad allungare la regione di collegamento.



Classi di aminoacidi simili

Gli aminoacidi in base alle loro proprietà biochimiche possono essere classificati in classi di aminoacidi simili. Eventi di mutazione puntiforme che comportano la sostituzione di un singolo aminoacido con un altro possono essere più o meno sfavorevoli. Sostituzioni di aminoacidi con altri dalle proprietà biochimiche simili, hanno maggiori probabilità di mantenere la proteina funzionante e di essere quindi accettate.
Possiamo definire simili due aminoacidi se sono classificati insieme in almeno una delle classi di similarità. Ad esempio (G)licina e (S)erina sono simili perchè appartengono entrambi alla classe degli aminoacidi "Piccoli".

Punteggio di similarità di un allineamento

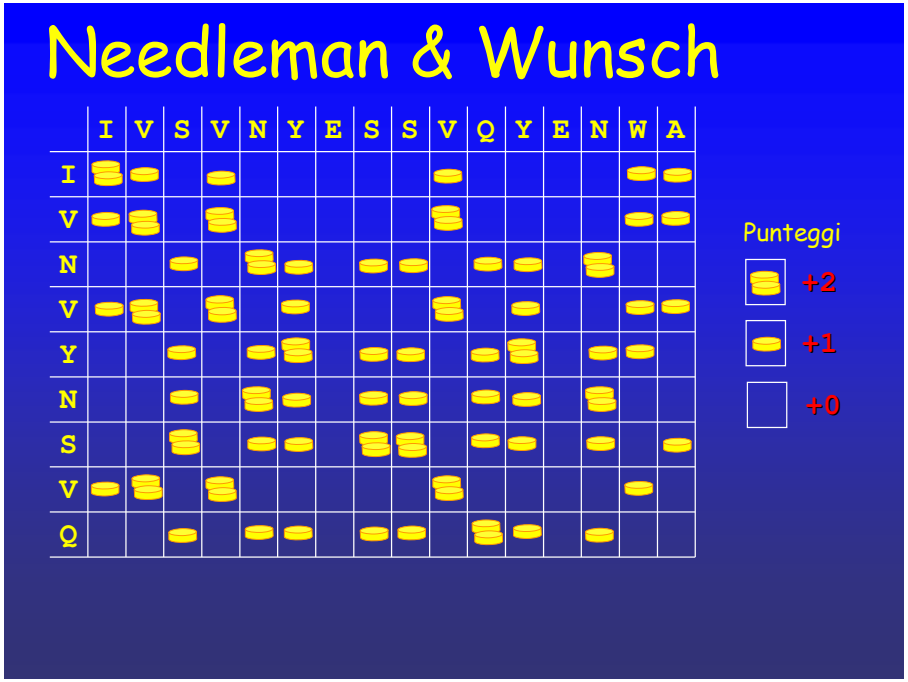
Il punteggio di identità non premia sostituzioni di aminoacidi simili fra di loro rispetto a sostituzioni fra aminoacidi molto differenti. Il punteggio di similarità, calcolato sommando un valore di similarità per ogni coppia di aminoacidi, è quindi una migliore misura della qualità di un allineamento rispetto al punteggio di Identità.
Il più semplice tipo di valori che possiamo utilizzare per calcolare un punteggio di similarità è di 2 punti per ogni coppia di aminoacidi uguali, 1 punto per ogni coppia di aminoacidi simili e 0 punti per coppie di aminoacidi non simili.

Matrici di sostituzione

Ogni coppia di aminoacidi è associata ad un suo valore di similarità, che dipende dal grado di similarità biochimica dei due aminoacidi. Per comodità di utilizzo i valori di sostituzione per tutte le possibili (20x20) 400 coppie di aminoacidi sono conservati in una sorta di tabelle, chiamate "Matrici di Sostituzione"
Una Matrice di Sostituzione è una matrice 20 x 20 che contiene in ogni riga ed in ogni colonna uno dei 20 tipi di aminoacidi. All'interno di ogni cella della matrice è contenuto il punteggio di similarità che si deve usare quando si fa un allineamento per la sostituzione della coppia di aminoacidi. Sulla diagonale della matrice ci sono i valori di similarità per le coppie di aminoacidi identici. Poichè il valore di similarità per la sostituzione dell'aminoacido A1 con l'aminoacido A2 sarà lo stesso della sostituzione dell'aminoacido A2 con quello A1, i possibili valori della matrice sono solamente 210. Una metà della matrice, contenendo solo informazioni ridondanti, può quindi essere lasciata vuota. Nella figura è rappresentata la matrice di sostituzione che contiene i semplici valori di similiarità tra gli aminoacidi, definiti consideranto le classi di appartenenza biochimica. Ovvero 2 punti per gli aminoacidi identici, 1 punto per quelli che hanno una classe in comune, 0 punti per tutti gli altri.

Punteggio di similarià di un allineamento

Per calcolare il punteggio di Similarità di un allineamento usando una matrice di sostituzione, si procede nel seguente modo: per ogni coppia di aminoacidi che sono stati appaiati, si consulta la matrice di sostituzione nella cella che ha come riga uno dei due aminoacidi e come colonna l'altro. Il valore trovato nella matrice viene sommato al punteggio totale dell'allineamento. Nell'esempio della figura, la coppia AA dell'allineamento ha un valore di sosttuzione (pari a 2) trovato nella matrice di sostituzione all'incrocio fra la riga A e la colonna A. La coppia AD il vaolre 0 trovato all'incrocio fra la riga A e la colonna D, e così via.

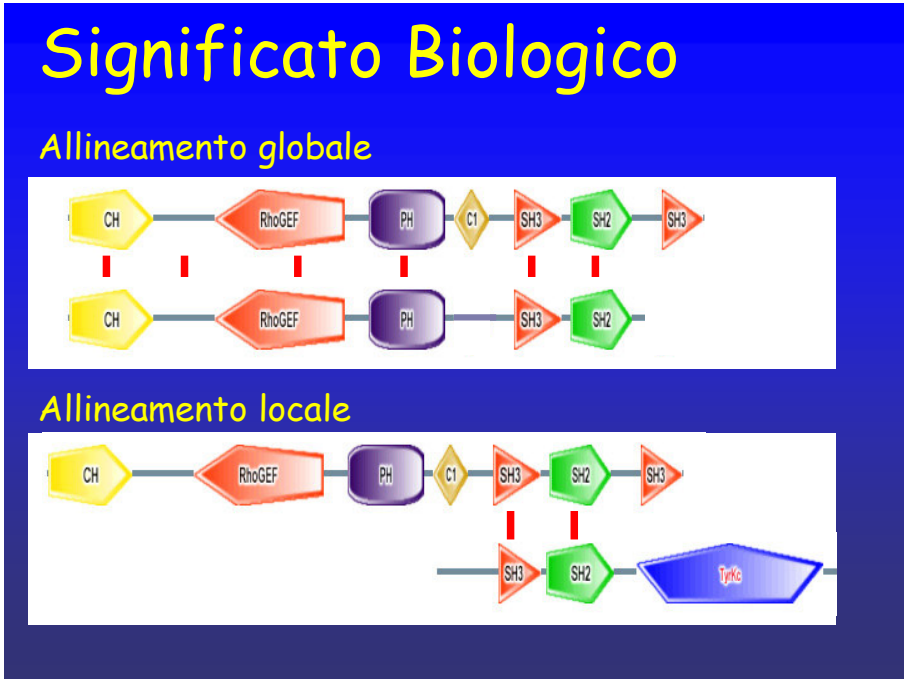


Algoritmo di Needleman e Wunsch

Per tenere conto dei diversi valori di sostituzione fra gli aminoacidi quando si cerca il miglior allineamento usando l'algoritmo di programmazione dinamica, bisogna utilizzare invece di 1 punto (1 moneta d'oro) per le coppie di aminoacidi identici e 0 punti (nessuna moneta d'oro) per le coppie di aminoacidi diversi, punteggi di sostituzione diversi per differenti tipi di aminoacidi.
Ad esempio utilizzando i valori di sostituzione semplificati, metteremo 2 monete d'oro in ogni cella in corrispondenza delle coppie di aminoacidi identici, 1 moneta in corrispondenza delle coppie di aminoacidi simili e nessuna moneta in tutte le altre. La procedura di allineamento è identica a quella vista precedentemente con l'ovvia differenza che un allineamento che attraversa una cella con 2 monete contribuirà con 2 punti al punteggio dell'allineamento invece che con 1. L'algoritmo di programmazione dinamica che utilizza i punteggi di sostituzione per aminoacidi simili e le penalità per i gaps si chiama di Needlemann e Wunsh dai nomi dei suoi ideatori.

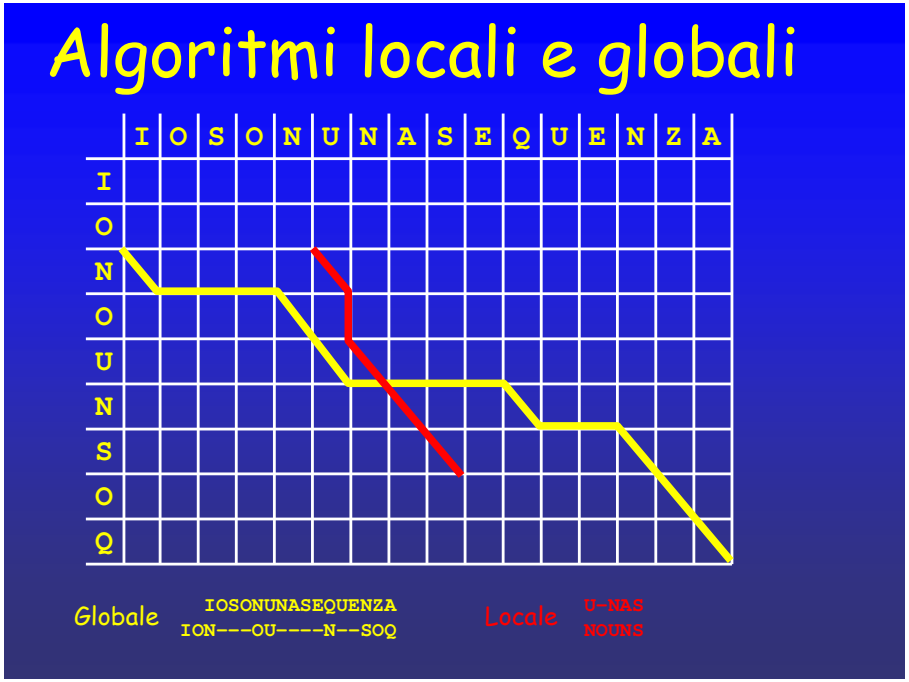
Allineamenti locali e globali

Due sequenze possono essere allineate in due modi diversi, ma altrettando validi: cercando di allinearle nella loro interezza accoppiando il maggior numero possibile di residui, oppure cercando corte regioni all'interno delle due sequenze con livello di identità più alto rispetto ai restanti tratti. Il primo dei due allineamenti si chiama "Globale" il secondo "Locale".
La figura mostra come le stesse due sequenze possono essere allineate, validamente, in entrambi i modi.
L'allineamento di tipo globale si usa per cercare similarità fra proteine che si pensa siano omologhi vicini, e che quindi conservino una similarità che si estende sull'intera loro sequenza.
Per confrontare proteine poco simili o addirittua non omologhe ma che si suppone possano contenere dei corti tratti di sequenza comune, si usa invece l'allineamento di tipo locale.



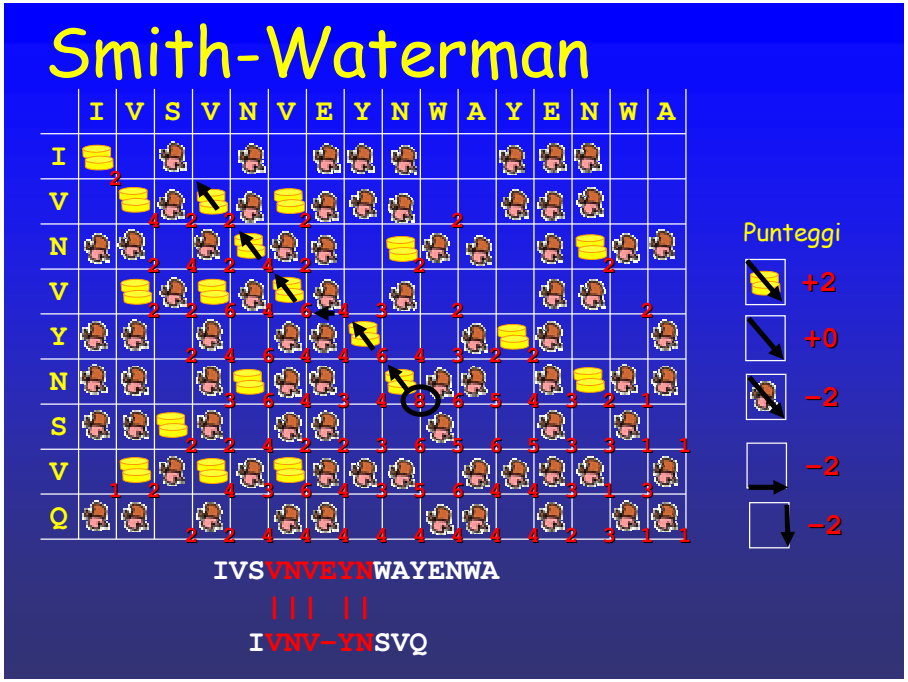
Significato biologico di similarità globale e locale

Sia le proteine che i geni sono composti da moduli che evolovono e possono essere scambiati in modo indipendente: gli esoni dei geni e i corrispondenti domini sulle proteine. Proteine che hanno la stessa funzione in organismi diversi, spesso sono simili in tutta la loro lunghezza, essendo la loro sequenza composta dalla stessa serie di domini tutti conservati. Queste proteine hanno una similarità globale che può essere facilmente identificata.
Quando si confrontano proteine con funzioni diverse, può capitare che queste abbiano comunque al loro interno alcuni domini funzionali in comune. Per essere in grado di identificare questi corti moduli è necessario usare metodi di allineamento "Locale" che rinuncino ad allineare le due proteine nella loro interezza (cosa impossibile) per cercare invece di trovare le corte regioni di similarità locale, corrispondenti ai domini comuni.



Matrici di allineamento locali

Su di una matrice di allineamento tutti gli allineamenti che partono dalla prima riga o colonna e continuano sino a terminare nell'ultima riga o colonna sono di tipo globale (linea gialla). Questo vuol dire che ogni possibile appaiamento di residui continua sempre fino ad almeno una delle due estremità di entrambe le sequenze.
Gli allineamenti locali (linea rossa) sono invece rappresentati da tutte le linee che partono e/o terminano in punti della matrice diversi dalla prime o ultime righe o colonne. In questi allineamenti, come si vede in figura, solamente alcuni dei residui che compongono le sequenze sono allineati, e non debbono necessariamente essere posti ad un estremità della sequenza.
Tutti gli allineamenti trovati con l'algoritmo di Needleman e Wunsh sono sempre globali. Il motivo per cui questo avviene e che non esistono movimenti in diagonale che possano dare un punteggio negativo. Sarà quindi sempre possibile estendere la linea di un allineamento muovendosi in diagonale fino a raggiungere l'ultima riga o colonna, dove si troveranno sempre i migliori punteggi possibili.
Per permettere al metodo di Needleman e Wunsh di trovare similarità locali è necessario introdurre nella matrice di sostituzione dei punteggi negativi, in modo da rendere conveniente in alcuni casi fermarsi al centro della matrice piuttosto che estendere il percorso diminuendone il punteggio.



Smith e Waterman

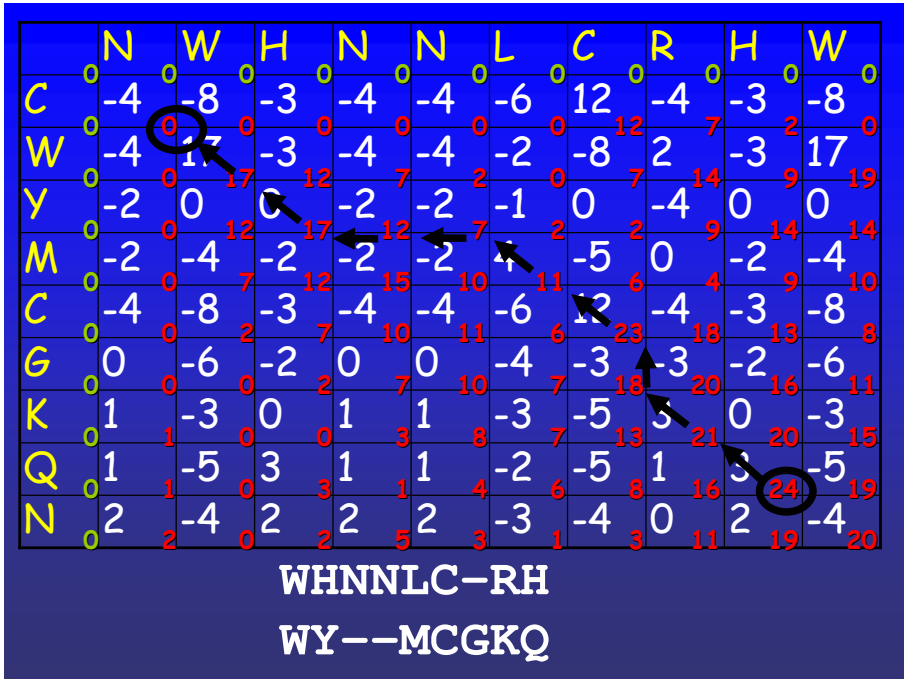
L'algoritmo di Smith e Waterman con lievi differenze rispetto a quello di Needleman e Wunsh permette di trovare allineamenti locali.
Primo, vengono introdotti punteggi negativi per la sostituzione di due aminoacidi. Bisogna quindi usare una matrice di sostituzione che assegni punteggi positivi alle sostituzioni favorite e negativi a quelle sfavorite. Ad esempio nella figura sono stati usati punteggi di +2 per aminoacidi identici (2 monete d'oro), 0 per quelli simili e -2 per le coppie di aminoacidi non simili (faccine di brigante).
Secondo, il valore del migliore punteggio per ogni punto della griglia non può mai scendere sotto lo zero. Per cui se il miglior punteggio calcolato per una cella è un numero negativo, questo viene considerato comunque zero (i punteggi pari a 0 non sono scritti in figura).
Terzo, la cella di arrivo del migliore percorso possibile non dovrà più necessariamente trovarsi sull'ultima riga o colonna, ma ovunque sulla matrice (cerchio nero nella figura).
Quarto, la cella di partenza dell'allineamento non dovrà necessariamente trovarsi nella prima riga o colonna, ma, nel ripercorrere la linea del migliore allineamento a ritroso, bisognerà fermarsi non appena si raggiunge un punto sulla griglia con un valore di monete trovate pari a 0.
Quinto, trascrivendo l'allineamento in forma estesa bisognerà rappresentare solo i residui effettivaemente inclusi nell'allineamento (in rosso) e non le intere sequenze (in bianco).

Una vera matrice di sostituzione

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3
C		12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0
D			4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-6	-4
E				4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-6	-4
F					8	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	6
G						5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-6	-5
H							6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0
I								5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1
K									5	-3	0	1	-1	1	3	0	0	-2	-3	-4
L										6	4	-3	-3	-2	-3	-3	-2	2	-2	-1
M											6	-2	-2	-1	0	-2	-1	2	-4	-2
N												2	-1	1	0	1	0	-2	-4	-2
P													6	0	0	1	0	-1	-6	-5
Q														4	1	-1	-1	-2	-5	-4
R															6	0	-1	-2	2	-4
S																2	1	-1	-2	-3
T																	3	0	-5	-3
V																		4	-6	-2
W																			17	0
Y																				10

Una matrice di sostituzione reale

Le matrici di sostituzioni che vengono utilizzate per allineamenti di sequenze reali contengono un valore diverso per ogni tipo possibile di sostituzione fra due aminoacidi. I punteggi sono calcolati considerando le sostituzioni aminoacidiche che si osservano in famiglie di proteine simili nel corso dell'evoluzione. Il modo esatto in cui sono costruite è descritto nel prossimo capitolo. Le coppie di aminoacidi identici si trovano sulla diagonale ed hanno un valore più alto rispetto alle altre sostituzioni dello stesso residuo. Gli altri punteggi sono di tipo positivo o negativo a seconda della similarità della coppia di aminoacidi.



Un allineamento reale

Nella figura è rappresentata la matrice usata per allineare due corte sequenze usando l'algoritmo di Smith e Waterman. I valori di sostituzione con cui viene riempita inizialmente la matrice (in bianco) sono presi dalla matrice di sostituzione reale della figura precedente, e corrispondono alle monete d'oro o alle penalità per i briganti dell'esempio precedente. I valori 0 inseriti all'inizio nella prima riga e colonna sono indicati in verde. Le somme del maggior numero di punti ottenibile in ogni cella sono scritti in rosso. Per le inserzioni dei gaps (movimenti in orizzontale e verticale) è stata usata una penalità di -5 punti. I cerchi neri rappresentano la cella di partenza e di arrivo del migliore allineamento (dal punteggio più alto, 24, fino al primo 0 incontrato). Le frecce nere indicano il percorso a ritroso che rappresenta il migliore allineamento locale possibile fra le due sequenze. In bianco sotto la matrice è scritto l'allineamento trovato per esteso, deducibile dal percorso delle frecce.