

# Laboratorio di Bioinformatica

Dispense del corso

**Chiara Solito**

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

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

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

## Contents

<b>1</b>	<b>Il corso</b>	<b>3</b>
<b>2</b>	<b>Cos'è la bioinformatica?</b>	<b>1</b>
2.1	Il flusso dell'informazione biologica . . . . .	1
2.2	Struttura degli acidi nucleici . . . . .	1
2.3	Le proteine . . . . .	2
<b>3</b>	<b>Il cosmo "omico"</b>	<b>2</b>
3.1	La genomica . . . . .	2
3.2	Trascrittomica . . . . .	3
3.3	Proteomica . . . . .	3
3.4	Genomica Strutturale . . . . .	3
3.5	Farmaco-genomica . . . . .	3
<b>4</b>	<b>L'evoluzione ed il confronto tra sequenze</b>	<b>3</b>
<b>5</b>	<b>Le Basi di Dati Biologiche</b>	<b>1</b>
5.1	Introduzione . . . . .	1
5.2	Dati di Sequenza . . . . .	2
<b>6</b>	<b>NCBI</b>	<b>3</b>
6.1	Com'è strutturato il database . . . . .	3
6.2	Operatori Booleani . . . . .	5
6.2.1	Operatore AND (&) . . . . .	5
6.2.2	Operatore OR ( ) . . . . .	6
6.2.3	Operatore NOT (!) . . . . .	6
6.2.4	Combinazione di Operatori Booleani . . . . .	6
6.3	Nel dettaglio . . . . .	6
<b>7</b>	<b>Proteine - Le banche dati proteiche più usate</b>	<b>7</b>
7.1	NCBI Protein - non molto ricco . . . . .	7
7.2	Uniprot . . . . .	7
7.2.1	Struttura del database . . . . .	8
7.3	ExPASy . . . . .	8
<b>8</b>	<b>Allineamenti di Sequenze</b>	<b>1</b>
8.1	Definizione - <u>Allineamento a coppie</u> . . . . .	1
8.2	Altre definizioni . . . . .	1
<b>9</b>	<b>Confrontare due sequenze</b>	<b>3</b>
9.1	Come identificare le zone di somiglianza locale tra due sequenze? . . . . .	4
9.1.1	Matrice a punti - dot plot . . . . .	4
9.1.2	In breve . . . . .	6

<b>10 Algoritmi dinamici di allineamento</b>	<b>7</b>
10.1 Il concetto . . . . .	7
10.2 Regole pratiche . . . . .	7
10.3 Step . . . . .	8
10.4 Needleman-Wunsch: programmazione dinamica . . . . .	10
<b>11 Waterman-Smith</b>	<b>1</b>
11.1 L'algoritmo . . . . .	1
<b>12 Allineamento: globale vs locale</b>	<b>1</b>
12.0.1 In conclusione . . . . .	3
12.1 Significatività statistica di un allineamento . . . . .	3
12.1.1 Significatività allineamento globale: lo Z score . . . . .	3
12.1.2 Significatività allineamento locale . . . . .	4
<b>13 Matrici di punteggio</b>	<b>5</b>
13.1 PAM: Point Accepted Mutation . . . . .	6
13.2 La mutabilità relativa degli amminoacidi . . . . .	8
13.2.1 Matrice PAM1 (probabilità) di Dayhoff . . . . .	9
<b>14 Matrici di sostituzione</b>	<b>9</b>
14.1 Moltiplicare le matrici . . . . .	9
14.2 Approccio Dayhoff . . . . .	11
<b>15 Matrici BLOSUM</b>	<b>13</b>
<b>16 BLOSUM vs. PAM</b>	<b>13</b>
16.1 La "Twilight zone" nell'allineamento di proteine . . . . .	14
<b>17 Allineamento multiplo di sequenze</b>	<b>1</b>
17.1 Visione Generale . . . . .	1
17.1.1 Una definizione . . . . .	1
17.1.2 Alcuni fatti . . . . .	1
17.1.3 Caratteristiche utili per realizzarlo . . . . .	1
17.1.4 Utilizzi e Vantaggi . . . . .	1
17.2 Metodi . . . . .	2
17.2.1 Metodi Euristici . . . . .	2
<b>18 Domande sull'introduzione</b>	<b>1</b>
<b>19 Domande sulle banche dati</b>	<b>1</b>

# 1 Il corso

Il corso si propone di presentare allo studente le basi teoriche e applicative di algoritmi e programmi utilizzati nella ricerca e nell'analisi dei dati contenuti nelle principali banche dati biologiche di uso corrente. Il corso si compone di due moduli di seguito specificati.

Modulo 1: In questo modulo verranno appresi gli strumenti volti all'utilizzo dell'informazione in proteomica, genomica, biochimica, biologia molecolare e strutturale. Si fornisce inoltre un'introduzione all'analisi e la visualizzazione di dati strutturali relativi a macromolecole biologiche e loro complessi e la creazione di semplici modelli dinamici e statici di reti biomolecolari, che avvicinerà lo studente all'emergente disciplina della systems biology.

Modulo 2: In questo modulo lo studente acquisirà conoscenza pratica degli strumenti bioinformatici per l'analisi, l'interpretazione e la predizione di dati biologici in proteomica, genomica, biochimica, biologia molecolare e strutturale. In particolare, gli studenti avranno la possibilità di applicare strumenti della bioinformatica allo stato dell'arte a specifici problemi biologici.

# **Lezione 1: Introduzione**

Ripasso delle basi e introduzione dei concetti fondamentali

## 2 Cos'è la bioinformatica?

La bioinformatica è (oggi) una disciplina scientifica dedicata alla risoluzione di problemi biologici a livello molecolare con metodi informatici. Descrive fenomeni biologici in modo numerico/statistico.

La bioinformatica principalmente:

- Fornisce modelli per l'interpretazione di dati provenienti da esperimenti di biologia molecolare e biochimica al fine di identificare tendenze e leggi numeriche
- genera nuovi strumenti matematici per l'analisi di sequenze di DNA, RNA e proteine (frequenza di sequenze rilevanti, loro evoluzione e funzione).
- organizza le conoscenze acquisite in basi di dati al fine di rendere tali dati accessibili a tutti, ottimizzando gli algoritmi di ricerca dei dati

Condivide alcuni argomenti con:

- **Systems biology**
  - Rappresenta i processi biologici come sistemi per comprenderne le funzioni e i principi in modo olistico per mezzo di modelli matematici
- **Computational biology**
  - Integra i risultati sperimentali con quelli derivanti da esperimenti in silico, ottenuti quindi per mezzo di metodi informatici a partire da dati biologici.

### 2.1 Il flusso dell'informazione biologica

Ad ogni livello di organizzazione (da interazioni fra biomolecole fino a cellule, organismi, popolazioni) l'elemento unificante è l'EVOLUZIONE, unico vero fondamento teorico della disciplina.

- EVOLUZIONE: adattamento progressivo attraverso variabilità genetica casuale e selezione naturale (Darwin, 1859)
- Ad ogni livello biologico, il fenotipo (insieme di tratti e caratteri somatici) è codificato dal genotipo (il patrimonio genetico)
- Genotipo: sorgente primaria di variazione genetica; fenotipo: bersaglio della selezione naturale
- Il genotipo è conservato nel genoma (fatto di DNA, eccezion fatta per virus a RNA)

### 2.2 Struttura degli acidi nucleici

Sono poliesteri composti da nucleotidi (composti da una base azotata, uno zucchero 2'-deossi-ribosio (o ribosio in RNA) e un gruppo fosforico).

2 tipi di basi azotate: purine (adenina, guanina) e pirimidine (timina, citosina, uracile).

L'RNA è meno stabile ma più versatile del DNA; è scarsamente reattivo (meglio per conservare l'informazione) e assume strutture 3D anche molto complesse, ne esistono diverse forme: mRNA, tRNA, rRNA e piccoli RNA; ciò è fondamentale per la trasmissione dell'informazione genetica.

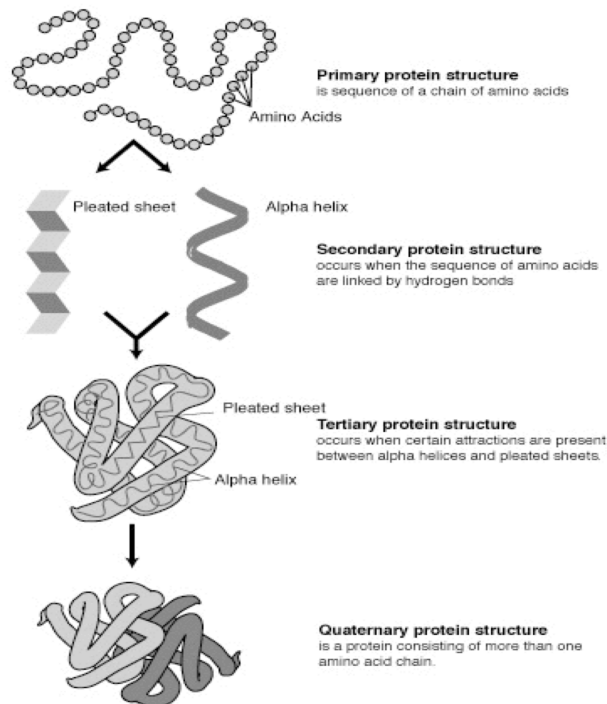
Un gene si trova in una precisa porzione fisica del genoma (**locus genico**). Es. Location: 6p21.1 significa cromosoma 6, braccio corto (p), regione 2 banda 1, sotto-banda 1.

In un gene le **Open Reading Frames** (parti di DNA/RNA codificanti) si trovano comprese fra la sequenza di inizio (codone d'inizio) e la sequenza di stop (codone di stop). Il genoma eucariotico contiene porzioni non codificanti importanti per la regolazione (**promotori**: vi si lega RNA polimerasi; **enhancers**: aumentano x200 la frequenza di trascrizione del gene) e per la costituzione (**introni**, sequenze ripetute). Lo splicing ("saldatura") prepara il pre-mRNA per la traduzione. Nel genoma umano le porzioni non-codificanti sono in netta maggioranza. Diversa è la situazione nei genomi procariotici.

## 2.3 Le proteine

Sono il risultato del flusso dell'informazione genetica. La presenza di 20 amminoacidi naturali con proprietà chimico-fisiche diverse conferisce una variabilità enorme. Il legame peptidico crea il backbone di qualunque proteina.

La struttura di una proteina si organizza in 4 livelli, visibili "srotolando" la matassa della luce di Natale:



**La struttura 3D di una proteina è molto complessa** La determinazione della struttura 3D di proteine è un settore di ricerca molto attivo, come mostra la crescita esponenziale di strutture depositate nel Protein Data Bank.

## 3 Il cosmo "omico"

### 3.1 La genomica

- Genoma: Insieme dei geni di un organismo.
- Genomica: scienza che se ne occupa.
- Genoma Umano: Sequenziato completamente nel 2003.
- Occorre localizzare: Elementi Funzionali:
  - Regioni 'utili' → geni;
  - Sequenze codificanti, comprendere i meccanismi che regolano l'espressione, scoprire la funzione, e cercare d'intervenire specificamente su quest'ultima.

Il costo del sequenziamento del genoma oggi è alla portata di ciascun individuo.

### 3.2 Trascrittomica

- Trascrittoma: l'insieme di tutti i trascritti (RNA messaggeri, mRNA)
- Trascrittomica: scienza che se ne occupa.
- Occorre localizzare: Profili di espressione:
  - più dinamico del genoma
  - microarrays monitorano i livelli di espressione di migliaia di geni allo stesso tempo. Mirano ad individuare correlazioni e legami tra espressione genica, attivazione e inibizione. Esempi: studio nella differenziazione di cellule staminali o evoluzione di tumori.

### 3.3 Proteomica

- Proteoma: l'insieme di tutte le proteine in un sistema biologico o nel suo genoma
- Proteomica: scienza che se ne occupa.
- Occorre localizzare: sia le proteine codificate dai geni che le possibili modificazioni post- traduzionali (gruppi prostetici, multidomini, fosforilazione, ecc).
- Alcune tecniche
  - Gel:
    - 1° dimensione punto isoelettrico
    - 2° massa molecolare
  - Spettrometria di massa: identifica una proteina in base al suo rapporto massa/carica in seguito a ionizzazione

### 3.4 Genomica Strutturale

- Genomica strutturale: determinazione della struttura terziaria e quaternaria (3D e domini) delle proteine.
- Tecniche: cristallografia, NMR, homology modeling, cryoEM (microscopia crioelettronica) + AlphaFold (basato su AI)
- La struttura terziaria di una proteina è essenziale per determinarne la funzione

### 3.5 Farmaco-genomica

- Farmacogenomica: mira a prevedere la reazione di ciascun individuo verso un principio attivo in base al suo genotipo.
- Obiettivo: creare terapie farmacologiche personalizzate per ottimizzare il risultato minimizzando gli effetti collaterali.
- Esempio: previsione di gravi reazione avverse a Abacavir nella terapia dell'HIV

## 4 L'evoluzione ed il confronto tra sequenze

Un allele (variante di un gene presente contemporaneamente nella popolazione) può essere generato, fissato o mutare nel tempo.

**Uno degli obiettivi in senso lato della bioinformatica è stabilire se l'analisi dell'informazione riguardo a due oggetti biologici (e.g. geni o proteine) permette di stabilire una relazione di OMOLOGIA, cioè di discendenza da un antenato comune** Due sequenze che vengono separate fisicamente (per speciazione, duplicazione ecc.) non si scambiano più "informazione" ed evolvono indipendentemente, accumulando mutazioni. Spetta a noi trovare i tratti conservati dal comune antenato.

Un modo per muoversi in tal direzione è allineare le sequenze e determinare la percentuale di identità o sequence



identity (s.i.) (rapporto, in % tra il numero dei amminoacidi/basi identici rispetto al totale) o comunque il grado di similitudine. Di norma, sequenze nucleotidiche non correlate hanno una s.i. 50%; sequenze amminoacidiche non correlate hanno una s.i. 20%. Se tali valori aumentano, aumenta la probabilità che le sequenze siano omologhe. Ma tale indice dovrebbe tener conto anche della lunghezza delle sequenze. Una s.i. del 90% fra due sequenze di 100 a.a. ha un significato diverso rispetto alla stessa s.i. su sequenze di 30 a.a. **Allineare due sequenze significa stabilire se tra esse sussiste una relazione di omologia**

## Lezione 2: Basi di Dati Biologiche

## 5 Le Basi di Dati Biologiche

Il concetto di informazione è strettamente connesso a quello di dato e di struttura. Il dato è un osservabile (insieme di numeri, caratteri, simboli...) La struttura è l'organizzazione ordinata di dati che ne consente l'apprendimento.

Una banca dati è l'insieme di dati elementari, omogenei, ordinati e fruibili. In altre parole: è una collezione organizzata di dati. Esempio: elenco telefonico. L'informazione è strutturata in campi (nome, cognome ecc.). Ogni persona con i propri dati è un record. I dati biologici necessitano di un'organizzazione. Primo tentativo: Margaret Dayhoff (1925-1983): raccolse, nel 1965, le sequenze di 65 proteine (lavoro pionieristico per il tempo!) Le tecniche di sequenziamento rapido ed i progetti *-omici* hanno prodotto una quantità esplosiva di dati, anche di sequenze. L'avvento di Internet ha facilitato di gran lunga l'acquisizione e la distribuzione dell'informazione biologica in banche dati.

### 5.1 Introduzione

- Sono collezioni di dati:
  - strutturati
  - indicizzati
  - aggiornati
  - interconnessi
- I database biologici sono legati a strumenti per:
  - recuperare records al loro interno
  - aggiornare il database
  - combinare le informazioni
- Ci sono 6 principali categorie di basi di dati biologiche:
  - basi di dati di sequenze
    - DNA
    - RNA
    - Proteine
  - basi di dati per il mapping
    - geni
    - cromosomi
    - ...
  - Strutture3d (PDB)
  - Trascrittomica
  - Funzionali (KEGG)
  - Per la letteratura (PubMed), ontologies (GO), ...

A gennaio di ogni anno il Nucleic Acids Research pubblica un Database Issue, a gennaio:

- nel 2020 contiene 89 nuovi database e l'aggiornamento di 90 database
- classificati nelle seguenti categorie
  - Nucleotide Sequence Databases
  - RNA sequence databases
  - Protein sequence databases
  - Structure Databases

- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
- Cell biology
- COVID-19 databases

Le banche dati si strutturano e si integrano per favorire lo studio del dogma centrale della biologia. Tre enti al mondo sono i principali.

- EMBL
- NCBI
- DDBJ

Integrando collegamenti esterni (Swiss-prot, ExPASy, UCSC, ecc, ecc...) sono un punto ideale di partenza.

## 5.2 Dati di Sequenza

Che dati si possono trovare?

- Principalmente sono presenti
  - sequenze di caratteri (nucleotidi, amminoacidi)
  - o strutture
- L'uso della rappresentazione dei dati biologici di varia natura come sequenze è la forma di gran lunga più diffusa.
- Sequenze di DNA: formate da 4 tipi di lettere (a,c,g,t), convenzionalmente minuscole
- Sequenze di RNA: formate da 4 tipi di lettere (A,C,G,U), convenzionalmente maiuscole
- Sequenze proteiche: formate da 20 lettere (A, C, D, E, F, G, H, I,K, L, M, N, P, Q, R, S, T, V, W, Y), convenzionalmente maiuscole

Il formato FASTA-Pearrson:

- Rappresentazione mediante testo di sequenze nucleotidiche o peptidiche (lettere MAIUSCOLE).
- La prima riga (di lunghezza arbitraria) è preceduta da ">" e rappresenta la descrizione della sequenza.
- Le linee precedute da ">" o ";" sono considerate di commento e non vengono interpretate come dato di sequenza
- Le linee successive (ciascuna di 80 caratteri) rappresentano la sequenza.
- Un file fasta può avere estensione (non c'è uno standard)

Il formato XML (eXtensible Markup Language).

- Replica la struttura logica del record nella banca dati
- I tag permettono di delimitare e definire campi e sottocampi

## 6 NCBI

NCBI (National Center for Biotechnology Information) presso il National Institute of Health. Offre accesso a tante risorse di vario tipo:

- Sequenze geniche e proteiche
- Strutture terziarie
- Genomi completi
- Pathways
- EST (expressed sequence tags)
- Profili trascrittomici
- Cataloghi tassonomici

Fornisce accesso a numerosi database attraverso il sistema Entrez:

- GenBank
- Swissprot
- PubMed
- GEO
- ...

Fornisce accesso anche a diversi software bioinformatici.

### 6.1 Com'è strutturato il database

Una ricerca qualunque dall'home page apre ENTREZ, interfaccia per l'accesso ai database presenti in NCBI.

- PubMed è l'interfaccia di accesso a MEDLINE. Con i suoi
  - 20 milioni di record fino agli anni '50
  - 4600 riviste da più di 70 paesi
 È la banca dati per la letteratura biomedica più completa. (Accessibile anche tramite EBI tramite 17 CiteXplore)
- Nucleotide è un database che raccoglie sequenze da diversi altri database di NCBI. Per sequenze nucleotidiche
  - EST ( expressed sequence tag )
  - GSS ( genome sequence surveys Gene è orientato ai geni, ai loci altre sequenze, B act A rtif C hromosome , Y east A rtif C hromosome ,... )
 Inoltre:
  - RefSeq ( sistema di identificazione )
  - Unigene ( sequenze raggruppate )
  - UniProt ( proteine )
- Gene è orientato ai geni, ai loci
- Proteins è la sezione focalizzata sulle proteine, alle quali possono corrispondere strutture
- PubChem dedicato ai composti chimici
- In Genome genomi completi con riferimenti alla ricerca effettuata, varianti genomiche, ecc

- Informazioni su profili di espressione genica in diverse condizioni, modifiche post-traduzionali GEO (Gene Expression Omnibus) repository

GenBank è la banca dati di tutte le sequenze in NCBI (sincronizzata con EMBL e DDBJ). Le sequenze derivano da diverse fonti e tipi:

- Geni (regioni di regolazione, esoni, introni: unità ereditarie)
- EST (Expressed Sequence Tags) brevi segmenti di DNA trascritti e sequenz. da cDNA (ottenuto da mRNA retrotrascritto)
- STS (sequence tagged site, dove l'informazione genetica è mappata fisicamente)
- GSS (Genome Survey Sequence, vettori sequenze solo parzialmente sequenziate)
- HTGS (High Throughput Genomic Sequence, sequenze prodotte da tecniche di seconda generazione per il sequenziamento veloce, messe qui in "preview")
- Sequenze di proteine (sezione nr, non redundant)

Così tanto materiale ha provocato l'esigenza di ordine: **RefSeq**.

RefSeq è stato ideato per far corrispondere a ciascun trascritto normalmente prodotto da un gene e a ciascuna proteina una sequenza di riferimento, un identificatore (accession number).

Altri esempi di identificatori NON RefSeq sono:

- X02775 GenBank/EMBL/DDBJ nucleotidic sequence
- Rs7079946 dbSNP (single nucleotide polymorphism)
- N91759.1 An expressed sequence tag
- AAC02945 GenBank protein
- Q28369 SwissProt protein
- 1KT7 Protein Data Bank structure record

Refseq fornisce un identificatore per la sequenza di riferimento, curato dal personale dell'NCBI. formati principali degli id RefSeq sono:

- Complete genome/chromosome/plasmid **NC\_#####**
- Genomic contig (segmenti sovrapposti di DNA segments che rappresentano una sequenza consenso) **NT\_#####**
- mRNA (DNA format) **NM\_#####**
- Protein **NP\_#####**

**Un primo esempio di ricerca - L'Emoglobina** Una delle prime proteine ad essere studiata (anni '30 e '40, da Mulder, Liebing et al.).

È stata la prima proteina ad essere usata negli allineamenti multipli di sequenza: voglio fare dei confronti di sequenze (ad esempio per confrontare la stessa proteina prodotta da diverse specie). Con le prime tecniche di sequenziamento abbiamo scoperto che è stata localizzata in due loci, uno sul cromosoma 16 (subunità alfa) e 11 (subunità beta). I due geni sono regolati sia in base all'età che in base ai diversi tessuti.

È quindi un problema complesso che ha poi originato una serie di considerazioni. La mioglobina, una globina (struttura globulare a 8 eliche) che lega l'ossigeno nei tessuti muscolari, è stata la prima proteina la cui struttura tridimensionale è stata risolta tramite cristallografia.

L'emoglobina è un tetramero (due catene alfa e due beta negli adulti) è il principale trasportatore di ossigeno nei vertebrati. Assieme alla mioglobina è stata usata nei primi studi sugli allineamenti multipli.

Negli anni '80 con le prime tecniche di sequenziamento è stata localizzata in due loci, uno sul cromosoma 16 (subunità alfa) e 11 (subunità beta). I due geni sono regolati sia in base all'età che in base ai diversi tessuti.

## Ricerca dell'emoglobina

1. Inseriamo "beta globin" nella barra di ricerca
2. Seguiamo poi il link a "Gene"
3. Entrez Gene (ex LocusLink) è un portale curato che descrive loci genetici
  - nomenclatura
  - alias
  - accession numbers
  - fenotipi
  - OMIM (ereditarietà dei caratteri)
  - HomoloGene
  - mappatura sul genoma
  - collegamenti esterni
4. In generale ad oggi questa ricerca trova 126 entries
5. Intestazione: Entrez Gene, Noa: "Official Symbol", HBB per la beta globina
6. Limitiamoci alla ricerca per Homo Sapiens (selezionando sulla destra da Results by taxon)
7. Cliccando la specie si aggiorna automaticamente la stringa di ricerca: (beta globin) AND "Homo sapiens" [ porgn:\_txid9606 ]
8. Con il limite Homo Sapiens le entries sono solo 41
9. Apriamo la prima entry
10. Sulla dx in basso troviamo numerosi link a database esterni
11. Abbiamo una sezione sulle regioni genomiche, una sulla bibliografia
12. Sezione interessante: GeneRif (intended to facilitate access to publications documenting experiments that add to our understanding of a gene and its function)
13. E ancora Fenotipi, Variazione Genica, Pathways per Biosistemi e Interazioni note con altri geni.
14. Ontologia: (fondamentale per sistemi automatici di apprendimento). Classificazione e organizzazione dei dati in categorie predefinite così da agevolare l'individuazione di analogie e caratteristiche primarie. Può essere di diversi tipi, ma la principale distingue:
  - Funzione molecolare
  - Localizzazione cellulare
  - Processo biologico
15. Catalogazione RefSeq (a fine pagina)

## 6.2 Operatori Booleani

### 6.2.1 Operatore AND (&)

Restringe il campo di ricerca, inserendo ad es. la stringa: `equus caballus AND hemoglobin alpha`

La banca dati ci mostrerà una lista di sequenze proteiche i cui campi di descrizione contengono entrambe le parole. Quindi le sequenze proteiche del cavallo che non contengono nella descrizione la parola hemoglobin non vengono selezionate.

### 6.2.2 Operatore OR (|)

Estende il campo di ricerca, digitando ad esempio: Restringe il campo di ricerca, inserendo: `homo sapiens OR mus musculus`

Otterremo una lista di sequenze i cui campi contengono la parola `homo sapiens` o la parola `mus musculus`. L'operatore allarga l'insieme delle sequenze che incontrano le nostre esigenze.

### 6.2.3 Operatore NOT (!)

Restringe il campo di ricerca, inserendo: `homo sapiens NOT hemoglobin` Richiederemo sequenze i cui campi contengono la parola `homo sapiens` ma non la parola `hemoglobin`.

### 6.2.4 Combinazione di Operatori Booleani

Gli operatori booleani si possono combinare, vengono letti da sinistra a destra. Per questo sono utili le parentesi. Ad esempio: `globin AND promoter OR enhancer` produce quasi 5000 hits. Ma se si scrive `globin AND (promoter OR enhancer)` se ne ottengono circa 70.

Altre possibilità sono:

- Specificare un organismo (human, nella query: `human[ORGN]`)
- Usare l'asterisco: `glob *` restituisce tutte le entry che contengono una stringa che inizia per "glob"
- Usare le virgolette " ". La ricerca di "toxin B1" restituirà le entries che contengono esattamente la stringa intera.
- ...

## 6.3 Nel dettaglio

**Homologene** la risorsa ideale per individuare gruppi di geni omologhi negli eucarioti presenti in NCBI

**OMIM** Catalogo di geni umani e disordini genetici

**SNP** Single Nucleotide Polimorfism



## 7 Proteine - Le banche dati proteiche più usate

**Uniprot** (Universal Protein Resource) raccoglie le informazioni dei database:

1. Swiss-prot (SIB)
2. TrEMBL (EBI)
3. PIR

Offre la possibilità di effettuare Text Search o Blast Search. Viene curato anche un database NON RIDONDANTE (UniRef).

**Swissprot** Molto curato e dettagliato, con annotazioni circa funzione, struttura, modificazione e altre informazioni utili.

**TrEMBL** È la traduzione in silico di ogni entry codificante del database primario dell'EMBL, non è accurato, ma è ricchissimo.

**PIR** È il discendente diretto del database della Dayhoff, è curato a mano e le annotazioni sono molto ricche e precise.

### 7.1 NCBI Protein - non molto ricco

Entrez Protein: Contiene diverse informazioni su proteine

- 147 amminoacidi
- PRI: primates
- *NP\_000509* (protein accession number)
- *NM\_000518.4* (mRNA, RefSeq)
- Riferimenti bibliografici
- Sequenze FASTA (Opzione Display)
- Siti di modificazione post-traduzionale (AA94, AA121)
- Riferimenti ad altri database
- Sequenza amminoacidica (1 lettera)

È un record non molto ricco dal punto di vista dei dati delle proteine.

### 7.2 Uniprot

Uniprot è il più completo database centralizzato per le sequenze proteiche.

È organizzato su 3 livelli:

1. Uniprot Knowledge Base
  - Swiss-Prot (curato)
  - TrEMBL (automatico)
2. UniProt Reference clusters (UniRef)
  - Cluster di proteine che condividono il 50%, 90%, 100% di identità di sequenza
3. UniProt Archive (UniParc)
  - Archivio di sequenze proteiche stabile, non ridondante, da diverse 58 fonti

### 7.2.1 Struttura del database

Nella homepage abbiamo la classica barra di ricerca e subito sotto i link di accesso alle diverse informazioni contenute in Uniprot.

#### Un esempio di ricerca

1. Inseriamo "hbb" nella barra di ricerca.
2. Sulla sinistra possiamo selezionare gli organismi a cui restringere la ricerca. Selezioniamo Humans.
3. Questo aggiornerà automaticamente la stringa di ricerca: `hbb AND organism: "Homo sapiens (Human) [9605]"`
4. Selezioniamo la prima entry.
5. Sulla sinistra troviamo la tavola con tutti i contenuti disponibili.
6. Tra i più importanti abbiamo: "Function" (che specifica la funzione della proteina), "Pathology & Biotech", "Expression", "Interaction", "Family & Domains", ...
7. In "Structure" e altre sezioni troviamo i link a PDB (Protein Data Bank), database di strutture proteiche.
8. In "Sequence" troviamo tutta la sequenza proteica, scaricabile in formato FASTA.
9. Abbiamo inoltre vari link di collegamento ad altri database di sequenze (EMBL, GeneBank, DDBJ), varianti, ...

### 7.3 ExPASy

(Expert Protein Analysis System)

È una risorsa curata, espressione del SIB (Swiss Institute of Bioinformatics). Principalmente dedicata alle proteine.

La risorsa principale che ha prodotto è SwissProt (confluita in Uniprot). Rimane un punto di riferimento per molti tools.

## Lezione 3: Allineamenti di Sequenze - concetti e algoritmi

## 8 Allineamenti di Sequenze

Un primo e precoce allineamento di sequenze si ha nel 1961: H.C. Watson and J.C. Kendrew, "Comparison Between the Amino-Acid Sequences of Sperm Whale Myoglobin and of Human Hæmoglobin." Nature 190:670-672, 1961.

**L'allineamento di sequenze a coppie è un'operazione fondamentale in bioinformatica** È utilizzato per decidere se due proteine (o geni) sono correlate strutturalmente e funzionalmente. Viene utilizzato per identificare i domini o motivi che sono condivisi tra le proteine. È alla base della ricerca con BLAST (prossime lezioni) e viene utilizzato anche per l'analisi dei genomi.

**Allineamento a coppie: sequenze di proteine possono essere più informative del DNA** Le proteine sono più informative del DNA (20 vs 4 caratteri); molti aminoacidi condividono proprietà biofisiche. Ricordiamo che i codoni sono degenerati: i cambiamenti in terza posizione spesso non alterano l'amminoacido che ne è specificato (mutazioni sinonime). Le sequenze di proteine offrono un più lungo tempo di "look-back" e le sequenze di DNA possono essere tradotte in proteine, e poi utilizzate negli allineamenti a coppie.

### 8.1 Definizione - Allineamento a coppie

Il processo che allinea due sequenze per raggiungere livelli massimi di identità (e conservazione, nel caso di sequenze di amminoacidi) al fine di valutare il grado di similitudine e la possibilità di omologia.

### 8.2 Altre definizioni

- Identità

La misura in cui due sequenze (di nucleotidi o aminoacidi) sono invarianti. (es. identità del 32% => 32 a.a. su 100 sono ordinatamente identici)

- Conservazione

In una sequenza, modifiche in una specifica posizione di un amminoacido (o meno comunemente, di un nucleotide) che preservano le proprietà fisico-chimiche del residuo originale.

- Similitudine

La misura in cui due sequenze (di nucleotidi o aminoacidi) sono correlate. Si basa su identità + conservazione.

- Omologia

Similitudine attribuita a discendenti da un antenato comune.

**! Nota bene:**

- OMOLOGIA indica che due entità (es. 2 sequenze) hanno una stessa origine filogenetica, cioè derivano da un antenato comune. È un carattere QUALITATIVO.

- SIMILITUDINE indica che due entità (es. 2 sequenze), in relazione ad un certo criterio comparativo, hanno un certo grado di similitudine. È un carattere QUANTITATIVO (vedremo tra breve come definirla).

**! Osservazioni:**

- La struttura di una proteina dipende della sua sequenza di a.a. (concetto alla base del Protein Folding).

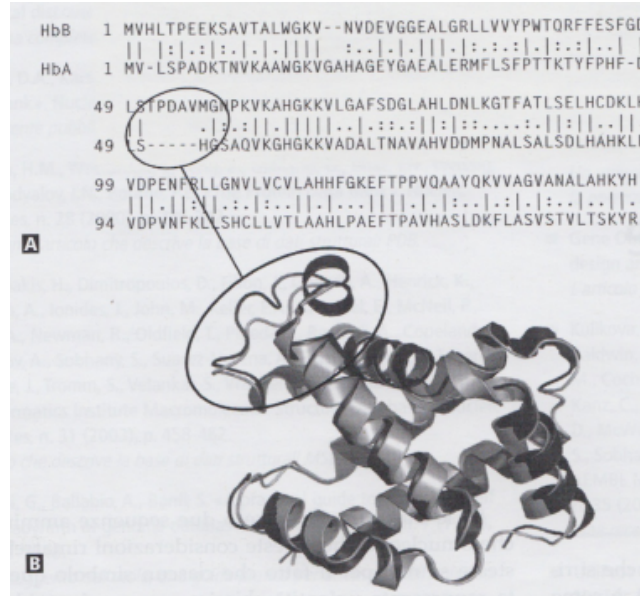
- La struttura determina la funzione molecolare della proteina.

- Se una sequenza proteica è conservata durante l'evoluzione ed è quindi presente in organismi diversi (famiglia di proteine) è ragionevole assumere che le funzioni che svolge siano simili o per lo meno correlate.

**! Passi per predizione di funzione:**

1. Identificazione delle proteine di una famiglia (evolute da un progenitore comune → sequenza di a.a. abbastanza simile.)
2. Identificazione degli a.a. che svolgono un ruolo strutturale o funzionale analogo (allineamento).

**Esempio 1:** la catena  $\beta$  dell'emoglobina e mioglobina «si somigliano»



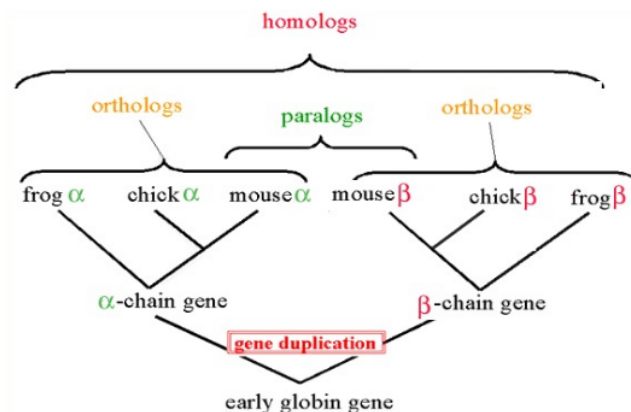
**Figure 1:** Le zone con indel nelle sequenze sono strutturalmente dissimili

- **Ortologi**

Sequenze omologhe in diverse specie che derivano, tramite la speciazione, da un gene ancestrale comune. La funzione può essere o non essere simile.

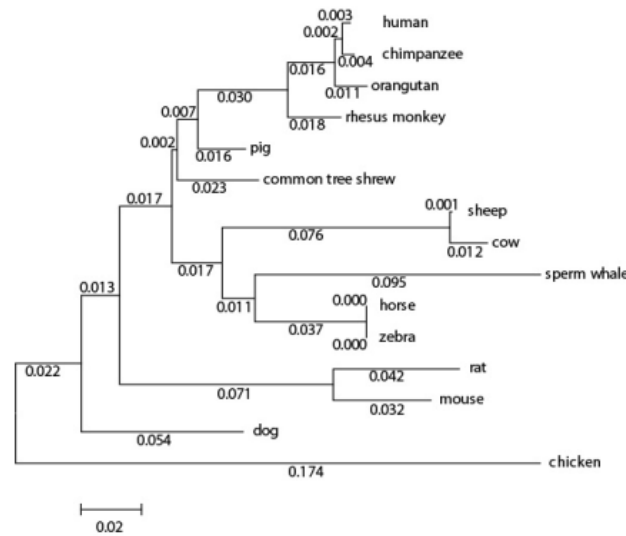
- **Paraloghi**

Sequenze omologhe all'interno di una singola specie sorte dalla duplicazione genica.



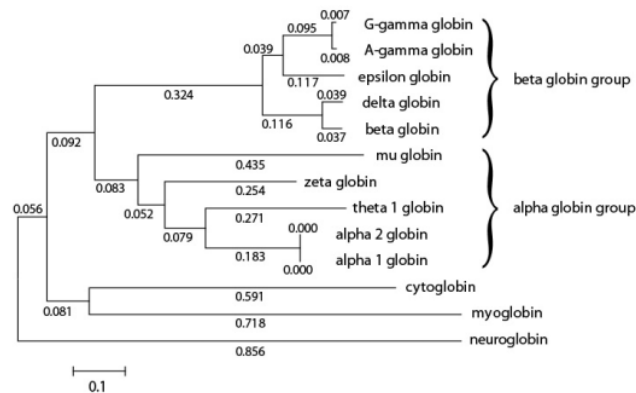
**Figure 2:** Ortologi e paraloghi sono spesso rappresentati in un albero singolo.

Ortologhi: membri di una famiglia di geni (proteine) in vari organismi.



**Figure 3:** Questo albero mostra gli ortologhi della globina.

Paraloghi: i membri di una famiglia di geni (proteine) all'interno di una specie.



**Figure 4:** Questo albero mostra i paraloghi della globina umana.

## 9 Confrontare due sequenze

**Come posso trasformare una stringa in un'altra?** Un modo semplice per capirlo è allineare le due stringhe:

ESEMPIO: 1 - LA CASA NUOVA; 2 - LA CASSA VUOTA

```

1 - L A C A - S A N U O V A
2 - L A C A S S A V U O T A
    oppure
1 - L A C A - S A - N U O V A
2 - L A C A S S A V - U O T A
  
```

Nel secondo caso c'è un'operazione in più.

- Il numero minimo di operazioni necessarie per allineare due sequenze ne misura la distanza.

- La Natura dispone di varie operazioni per trasformare un oggetto nell'altro (mutazioni, indel...)
- L'evoluzione sceglie la via piu' breve (principio di massima parsimonia); cio' si manifesta tramite l'analisi dell'allineamento.

Dobbiamo avere chiari i concetti di match (residui appaiati), mismatch (sostituzioni) e gap (indel).

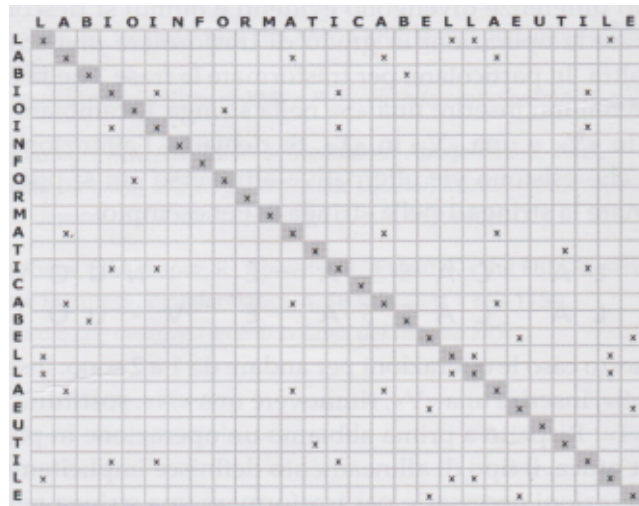
## 9.1 Come identificare le zone di somiglianza locale tra due sequenze?

### 9.1.1 Matrice a punti - dot plot

È un modo relativamente semplice.

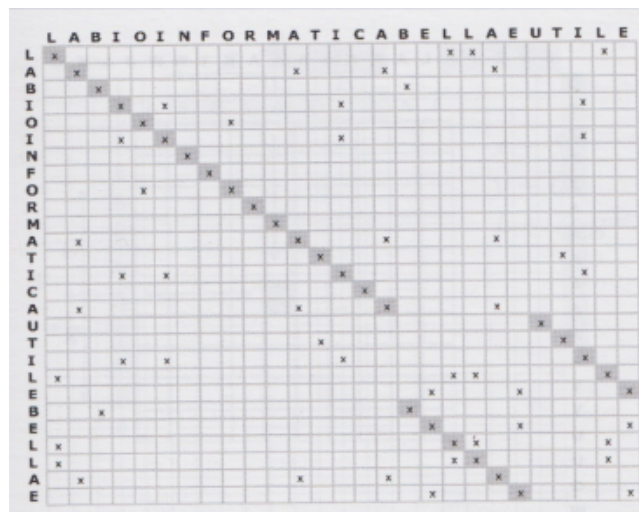
Confrontiamo la stringa con se stessa (autoconfronto):

1. Mettiamo una x per ogni identità



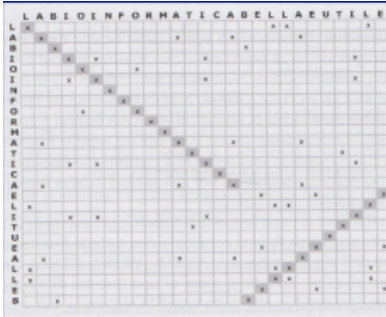
**Figure 5:** ... piuttosto banale. Cambiamo la seconda stringa.

Effettuiamo un'inversione. Il pattern delle diagonali lo mostra chiaramente: la diagonale principale si spezza ma porzioni delle stringhe sono identiche.

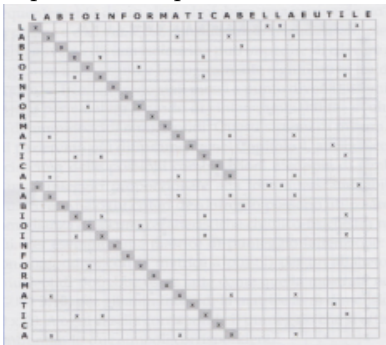


Possiamo individuare facilmente alcuni patterns mediante le matrici a punti (la prima stringa in alto resta immutata):

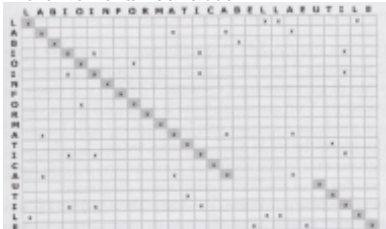
- Inversione di parole



- Ripetizione di parole



- Delezione di caratteri

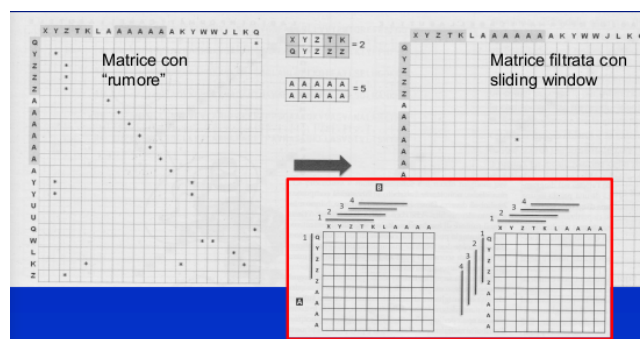


Se però allineamo sequenze di acidi nucleici (solo 4 lettere) il “segnale” di similitudine è mascherato dal grande rumore di fondo. Servono FILTRI per ridurre il rumore.

**Una semplice osservazione:** le zone delle sequenze più simili localmente si distribuiscono su diagonal; le altre somiglianze puntiformi si distribuiscono casualmente. Allora, è meglio confrontare le sequenze non per singole posizioni, ma per interi segmenti (FINESTRE).

Potremmo usare delle finestre scorrevoli.

ESEMPIO: confrontiamo una finestra di 5 residui in una seq con una finestra di 5 residui nell'altra. Confronto tutte le finestre (facendole scorrere), e metto una \* al centro della finestra solo se ho un match totale.



In generale: si fa scorrere una finestra alla volta.



## La procedura

1. Definiamo la posizione di una casella  $(x, y)$ .
2. Fissiamo il centro della finestra, di raggio  $g$ .
3. La lunghezza della finestra è dunque:  $L = 2g + 1$
4. Il numero  $N$  di residui identici in quella finestra è allora:  $N(x, y) = \sum_{h=-g}^{+g} S(x+h, y+h)$   
 $S = 1$  se il carattere in  $x+h$  è identico a quello in  $y+h$ ; altrimenti  $S = 0$

Questa regola è però molto restrittiva. A noi interessa anche la similitudine, non solo l'identità. Potremmo definire una soglia  $s$ , per cui:

- Se  $N(x, y) > s$  mette un simbolo nella casella in posizione  $x, y$ .

Dobbiamo quindi misurare la similitudine, e.g. tra aa o basi. Un esempio di matrice di punteggio (non di punti!!) per seq di nucleotidi può essere:

**Esempio: secondo la matrice di punteggio a sx, l'allineamento ha punteggio 8**

	A	T	C	G		A	A	A	T	C	C	G	A	A
A	2	1	0	0		A	T	A	C	A	G	A	T	T
T	1	2	0	0		2	+1	+2	+0	+0	+1	+0	+1	+1
C	0	0	2	1										
G	0	0	1	2										

Misurata la similitudine, possiamo allora attribuire un nuovo punteggio al confronto fra due finestre:

- Se  $N(x, y)$  è il numero dei residui simili, ed è la media dei punteggi delle singole coppie prelevati dalla matrice di punteggio scelta:
- $N(x, y) = \sum_{h=-g}^{+g} \frac{S(x+h, y+h)}{L}$

$S$  ora dipende da quale matrice di punteggio scegliamo e dalla lunghezza della finestra. Possiamo quindi essere un po' più "elastici" pur mantenendo la regola:

- Se  $N(x, y) > s$  mette un simbolo nella casella in posizione  $x, y$ .

$s$  lo decide la matrice di punteggio (cioè la similitudine).

### 9.1.2 In breve

In conclusione, la visualizzazione (ed il calcolo) di una matrice a punti dipende da:

1. La lunghezza  $L$  della finestra scorrevole scelta
2. Il metodo per misurare la similitudine  $S(x, y)$
3. La soglia  $s$  per "marcare" la casella rispettiva

In pratica conviene fissare 2 parametri e variare il terzo per rendere le zone di similitudine più evidenti. Molti programmi fanno questo.

Es. DOTTER/Dotlet - assegna un colore dipendentemente da  $S$ .

**Nota:** Analogamente all'identità di sequenza, che si può misurare in percentuale (%), anche la similitudine, una volta quantificata, si può misurare in %.

$S_1$  e  $S_2$  sono due sequenze lunghe, rispettivamente,  $L_1$  ed  $L_2$ .

Scelta  $L_1$  come riferimento si ha:

$$SequenceIdentity(s.i.) = \left( \frac{\#matches}{L_1} \right) * 100$$

$$SequenceSimilarity(s.s.) = \left( \frac{S_1 vs. S_2 s.c.}{S_1 vs. S_2 i.c.} \right) * 100$$

*s.c.* = *similarity score*, ed è ottenuto allineando  $S_1$  con  $S_2$ , e attribuendo il punteggio ottenuto dalla matrice di score.

*i.c.* = *identity score* ed è ottenuto allineando  $S_1$  con sè stessa ed attribuendo il punteggio ottenuto dalla matrice di score

*Nota:* se sono presenti indel, a denominatore metto la lunghezza dell'allineamento (compresi gli indel), e non la lunghezza originale della sequenza.

## 10 Algoritmi dinamici di allineamento

I dot plots non tengono in considerazione gli indel. Occorrono altri algoritmi che, passo a passo e seguendo una certa direzione, trovino l'allineamento con:

- Maggior numero di simboli identici
- Minor numero di indel (sfavorite evolutivamente)

Esempio: proviamo tutte le combinazioni da sx a dx, riempiendo una colonna alla volta (qui mostriamo solo i primi 3 residui!!)

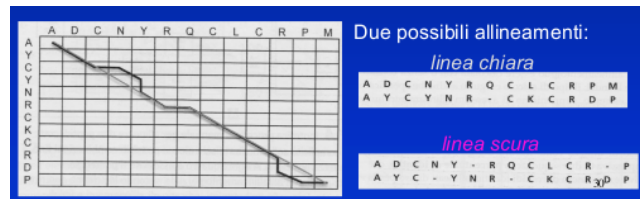
Stabiliamo inoltre i punteggi:

- -1 per indel
- 0 per residui diversi
- +1 per residui identici

Il numero delle combinazioni possibili è molto grande, specie per sequenze lunghe. Nel 1970 NEEDLEMAN e WUNSCH creano un algoritmo, poi migliorato ed esteso. Noi analizziamo l'originale.

### 10.1 Il concetto

distribuiamo le due sequenze in una matrice. Il possibile allineamento tra le due identifica un percorso che unisce le caselle dei residui appaiati.



### 10.2 Regole pratiche

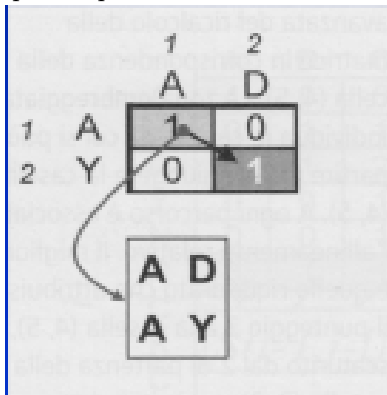
1. Il percorso ha una direzione e procede solo in avanti (NON si torna indietro!)
2. Occorre trovare il percorso con il **maggior** numero di aa identici e il **minor** numero di indel
3. Occorre anche tener conto della similitudine fra amminoacidi (significato evolutivo)
4. IMPORTANTE: un allineamento ottimale è sempre composto da suballineamenti ottimali (cioè: togliendo uno ad uno i residui dal fondo, l'allineamento deve restare ottimale, per poter ricostruire il percorso a ritroso)

### 10.3 Step

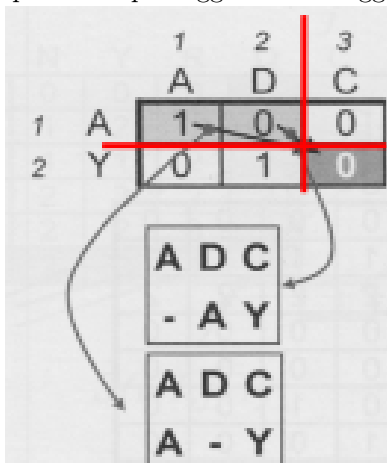
1. INIZIALIZZAZIONE della matrice: regola molto semplice; 1 se identici, 0 altrimenti. N.B. ora  $(x, y)$  identifica: (residuo in colonna, residuo in riga) (N.B. dopo aver trattato le matrici di score potremo normalizzare diversamente)

		1	2	3	4	5	6	7	8	9	10	11	12	13
		A	D	C	N	Y	R	Q	C	L	C	R	P	M
1	A	1	0	0	0	0	0	0	0	0	0	0	0	0
2	Y	0	0	0	0	1	0	0	0	0	0	0	0	0
3	C	0	0	1	0	0	0	0	1	0	1	0	0	0
4	Y	0	0	0	0	1	0	0	0	0	0	0	0	0
5	N	0	0	0	1	0	0	0	0	0	0	0	0	0
6	R	0	0	0	0	0	1	0	0	0	0	1	0	0
7	C	0	0	1	0	0	0	0	1	0	1	0	0	0
8	K	0	0	0	0	0	0	0	0	0	0	0	0	0
9	C	0	0	1	0	0	0	0	1	0	1	0	0	0
10	R	0	0	0	0	0	1	0	0	0	0	1	0	0
11	D	0	1	0	0	0	0	0	0	0	0	0	0	0
12	P	0	0	0	0	0	0	0	0	0	0	0	1	0

2. Partiamo da (1,1) (in alto a sx: la direzione è importante!!):  $(1,1) \rightarrow (2,2)$  L'unico percorso possibile è questo: per "ricordarmelo" sommo il valore della cella (1,1) a quello della cella (2,2) da matrice inizializzata



3. Prossimo step: verso (2,3). Ci sono due possibili percorsi, corrispondenti a due diversi allineamenti: scelgo quello con punteggio finale maggiore, sempre sommando casella precedente a (2,3)



4. Procediamo analogamente: ad esempio, verso (4,4)

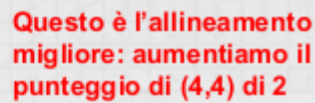


Figure 1 shows a 12x12 matrix of binary values (0 or 1) representing the relationship between 12 amino acids (A, D, C, N, Y, R, Q, C, L, C, R, P, M) and 12 nucleotides (A, D, C, N, Y, R, Q, C, L, C, R, P, M). The matrix is labeled with 'N - Y' and 'AYCY' above the columns and 'N - Y' and 'YCY' above the rows. A 4x4 sub-matrix is highlighted with a dashed box, and arrows point from the labels 'ADCNY', 'DCNY', 'CNY', and 'NY' to the corresponding columns of this sub-matrix. The sub-matrix is labeled 'C - - Y' below the columns and 'CY' below the rows.

	1	2	3	4	5	6	7	8	9	10	11	12	13
	A	D	C	N	Y	R	Q	C	L	C	R	P	M
1	A	1	0	0	0	0	0	0	0	0	0	0	0
2	Y	0	1	1	0	1	1	1	1	1	1	1	1
3	C	0	1	2	1	1	2	2	3	2	2	2	2
4	Y	0	0	1	2	1	0	0	0	0	0	0	0
5	N	0	0	0	1	0	0	0	0	0	0	0	0
6	R	0	0	0	0	0	1	0	0	0	0	1	0
7	C	0	0	1	0	0	0	0	1	0	1	0	0
8	K	0	0	0	0	0	0	0	0	0	0	0	0
9	C	0	0	1	0	0	0	0	1	0	1	0	0
10	R	0	0	0	0	0	1	0	0	0	0	1	0
11	D	0	1	0	0	0	0	0	0	0	0	0	0
12	P	0	0	0	0	0	0	0	0	0	0	1	0

ADCNY      DCNY      CNY      NY  
C - - Y      C - - Y      C - Y      CY

		1	2	3	4	5	6	7	8	9	10	11	12	13
		A	D	C	N	Y	R	Q	C	L	C	R	P	M
1	A	1	0	0	0	0	0	0	0	0	0	0	0	0
2	Y	0	1	1	1	2	1	1	1	1	1	1	1	1
3	C	0	1	2	1	1	2	2	3	2	3	2	2	2
4	Y	0	1	1	2	3	2	2	3	3	3	3	3	3
5	N	0	1	1	3	2	3	3	3	3	3	3	3	3
6	R	0	1	1	2	3	4	3	3	3	3	4	3	3
7	C	0	1	2	2	3	3	4	5	4	5	4	4	4
8	K	0	1	1	2	3	3	4	4	5	5	5	5	5
9	C	0	1	2	2	3	3	4	5	5	6	5	5	5
10	R	0	1	1	2	3	4	4	5	5	7	6	6	6
11	D	0	2	1	2	3	3	4	4	5	5	6	7	7
12	P	0	1	2	2	3	3	4	4	5	5	6	8	7

Pagina 9

A	D	C	N	Y		R	Q	C	L	C	R		P	M	
A	Y	C		Y	N	R		C	K	C	R	D	P		
1	0	1	0	1	0	1	0	1	0	1	1	0	1	0	= 8

Qui il punteggio totale coincide con il numero di residui identici.

$$\frac{8}{15} = 0.53 = 53\% \text{ identità di sequenza}$$

**In termini formali . . .** La casella (i,j) avrà lo score  $S(i,j)$  ricalcolato a partire dalla matrice di inizializzazione in questo modo:

		A	D	C	N	Y	R	Q	C	L	C	R	P	M	
A		1	0	0	0	0	0	0							
Y		0	1	1	1	2	1	1							
C		0	1	2	1	1	2	2							
Y		0	0	1	2	3	2	2							
N		0	0	1	2	2	3	3							
R		0	1	1	2	2	4	3							
C									1						
K															
R															
D															
P															

$$S(i, j) = s(a, b) + \max[S(i-1, j-1), S(i-k, j-1), S(i-1, j-l)]$$

Dovremmo però trovare un modo più efficace di inizializzare la matrice tenendo conto della similarità fra aa.

#### 10.4 Needleman-Wunsch: programmazione dinamica

NW garantisce l'ottimalità dell'allineamento, anche se l'algoritmo non calcola tutti i possibili allineamenti. È un esempio di un algoritmo di programmazione dinamica: un percorso ottimale (allineamento) è identificato dall'estensione graduale di sottopercorsi localmente ottimali. Dunque, una serie di decisioni è effettuata ad ogni passo dell'allineamento per trovare la coppia di residui con il miglior punteggio per quel passo.

L'algoritmo di Needleman-Wunsch è disponibile presso EBI, che ospita molti tools per allineamenti locali e globali di sequenze (Pairwise Sequence Alignment).

## Lezione 4: Allineamenti di Sequenze 2

### Matrici di Sostituzione

## 11 Waterman-Smith

Un problema dell'algoritmo di Needleman-Wunsch: non si tiene conto della penalizzazione delle indel.

L'algoritmo di WATERMAN-SMITH (1976) introduce una funzione di penalizzazione delle indel, per migliorare l'algoritmo NW, serve un sistema di pesatura delle indel, ad esempio:

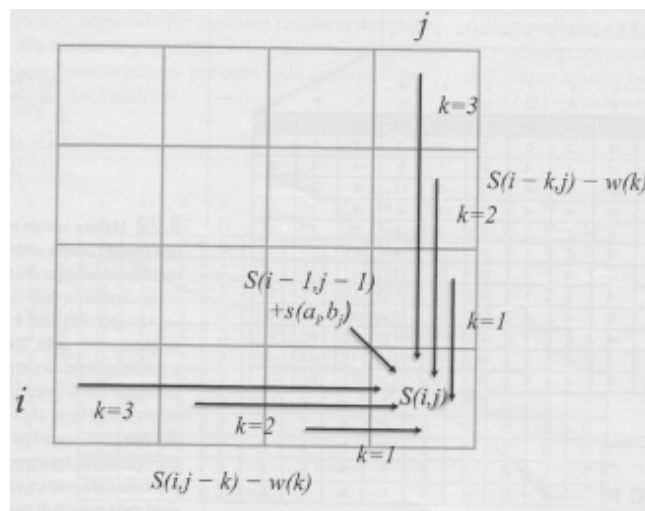
$$w(k) = g + e(k - 1)$$

Il peso  $w$  di una indel di lunghezza  $k$  dipende dalla penalizzazione per l'apertura di una singola indel ( $g$ ) e dalla penalizzazione per l'allungamento ( $e$ ).

### 11.1 L'algoritmo

Nella pratica l'algoritmo procede in questo modo:

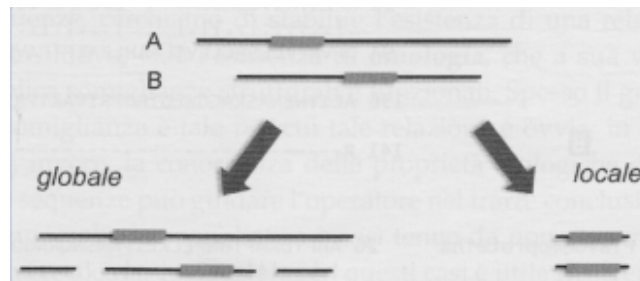
1. Inserisce una riga e una colonna 0-ime alla matrice di inizializzazione (calcolata ad esempio partendo da BLOSUM o PAM che vedremo, ecco perché non ha solo 0 o 1).  
Nella riga e colonna ombreggiate è sviluppata la funzione di penalizzazione:  $w(k) = -12 - 4(k - 1)$   
La riga e la colonna 0-sime contengono il punteggio che la sequenza avrebbe se allineata a una delezione lunga fino alla cella corrispondente.
2. Tiene conto dei possibili modi per arrivare alla casella  $(i, j)$ . Il suo punteggio  $S(i, j)$  dipende da essi:
  - (a) Mi muovo in diagonale: no indel e punteggio dato da: punteggio della casella di partenza + punteggio della casella  $(i, j)$  secondo la matrice di inizializzazione (come in NW)
  - (b) Mi muovo in verticale o orizzontale: inserisco indel nella sequenza  $i$  e  $j$ . Il punteggio sarà dato da: punteggio della casella di partenza - funzione di penalizzazione  $w(k)$  ( $k$  è la lunghezza della indel).
  - (c) Scelgo alla fine il percorso che dà il punteggio migliore



## 12 Allineamento: globale vs locale

**Allineamento globale** (NW e SW visto finora) si estende da un capo all'altro di ogni sequenza.

**Allineamento locale** trova le regioni (sottosequenze) di due sequenze che si allineano in modo ottimale.

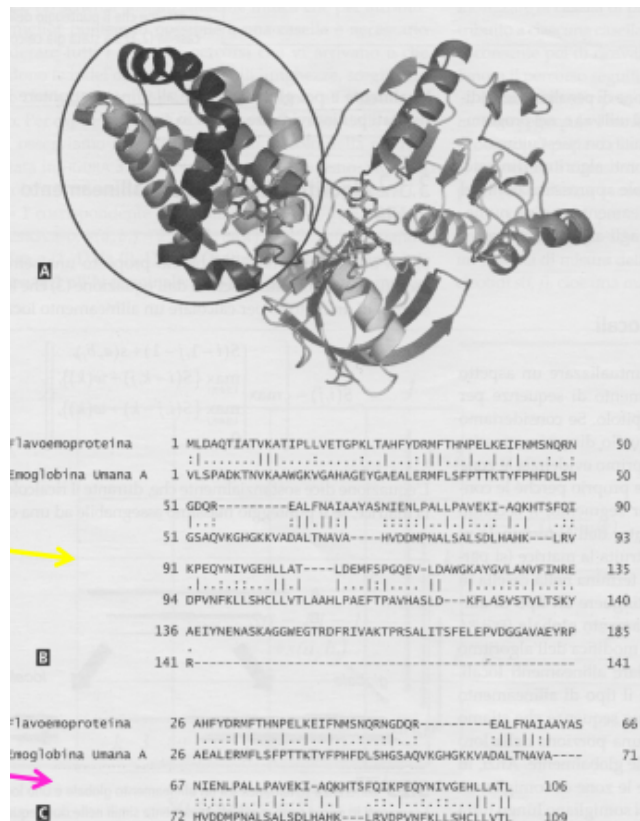


*Qui l'allineamento globale maschera la corrispondenza tra zone somiglianti*

SW si può modificare per renderlo capace di calcolare allineamenti locali: basta introdurre fra i casi possibili  $S(i, j) = 0$  nel caso in cui nell'allineamento globale fosse negativo

**Esempio:** L'allineamento fra una flavoemoproteina (con un dominio di tipo emoglobinico) e la catena A dell'emoglobina umana

- **globale:** più difficile notare quantitativamente la similitudine
- **locale:** più apparente



**Esempio:** SW locale: dopo aver ricalcolato la matrice cerco la cella con il valore massimo assoluto e parto da lì.

Gli stessi due peptidi di prima, allineati con Waterman-Smith globale e locale danno luogo a matrici ed allineamenti diversi. Partendo dalle caselle con score maggiore il percorso a ritroso individua allineamenti differenti (non sempre AL è sottoinsieme di AG)



	A	D	C	N	Y	R	Q	C	L	C	R	P	M		A	D	C	N	Y	R	Q	C	L	C	R	P	M				
A	1	-12	-18	-20	-24	-28	-32	-36	-40	-44	-48	-52	-56	A	1	-12	-18	-20	-24	-28	-32	-36	-40	-44	-48	-52	-56	A			
D	-12	1	-10	-14	-18	-22	-26	-30	-34	-38	-42	-46	-50	D	-12	0	0	0	0	0	0	0	0	0	0	0	0	D			
C	-18	-10	1	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44	C	-18	0	0	0	0	0	0	0	0	0	0	0	0	C			
N	-20	-14	-8	1	-4	-8	-12	-16	-20	-24	-28	-32	-36	N	-20	0	12	0	0	0	0	0	0	0	0	0	0	N			
Y	-24	-18	-12	-4	1	-8	-12	-16	-20	-24	-28	-32	-36	Y	-24	0	0	0	10	0	0	0	0	0	0	0	0	Y			
R	-28	-22	-16	-8	-12	1	-4	-8	-12	-16	-20	-24	-28	R	-28	0	2	0	2	8	18	0	0	0	7	0	7	R			
Q	-30	-24	-18	-10	-12	-8	1	-4	-8	-12	-16	-20	-24	Q	-30	0	0	0	0	0	14	11	0	0	0	0	0	Q			
C	-34	-28	-22	-14	-16	-10	-4	1	-4	-8	-12	-16	-20	C	-34	0	0	0	0	0	0	2	9	23	11	12	3	C			
L	-40	-34	-28	-18	-16	-10	-4	-3	1	9	18	8	0	-3	L	-40	0	0	0	0	0	13	1	3	3	11	28	8	L		
C	-44	-38	-32	-20	-18	-10	-4	-8	13	6	38	18	14	10	C	-44	0	0	0	0	0	13	1	0	15	6	34	20	C		
R	-48	-42	-36	-24	-22	-12	-4	-7	1	19	18	38	24	20	R	-48	0	0	0	0	0	12	1	19	7	3	12	29	25	R	
D	-52	-46	-40	-28	-24	-16	-2	-3	-2	14	24	38	23	23	D	-52	0	4	0	2	8	7	21	9	5	18	28	37	35	D	
P	-56	-50	-44	-34	-31	-20	-14	-5	-6	10	20	30	20	33	P	-56	0	0	0	0	0	0	0	18	6	12	22	32	35	35	P

ALLINEAMENTO WS-GLOBALE

ALLINEAMENTO WS-LOCALE

ADCN YRQC CLRCPM

AYC - YNRKC CLRCP

YRQC CLR

YNRKC

8

### 12.0.1 In conclusione

- L'allineamento locale è quasi sempre utilizzato per la ricerca su database (tramite BLAST). È utile per trovare domini (o regioni limitate di omologia) all'interno di sequenze.
- Smith e Waterman (1981) hanno risolto il problema dell'allineamento locale ottimale di sequenze.
- Altri metodi (BLAST, FASTA) sono più veloci ma meno accurati. Li vedremo in seguito.
- In ogni caso, qualunque metodo di allineamento si scelga esso fornirà un punteggio  $S$  all'allineamento. Ricordiamo sempre che lo score  $S$  dipende dal metodo di allineamento e non è assoluto!

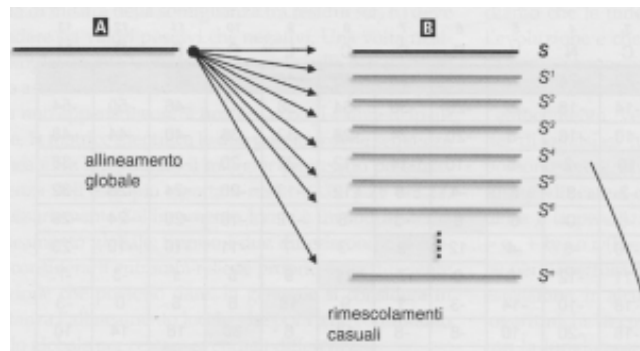
## 12.1 Significatività statistica di un allineamento

**Domanda:** Ho allineato due sequenze  $A$  e  $B$ , ottenuto il punteggio  $S$ . Come posso capire se sono omologhe? Che probabilità ho di trovare il punteggio  $S$  "per caso"?

Il problema è più facilmente risolvibile per gli allineamenti locali, e meno per quelli globali.

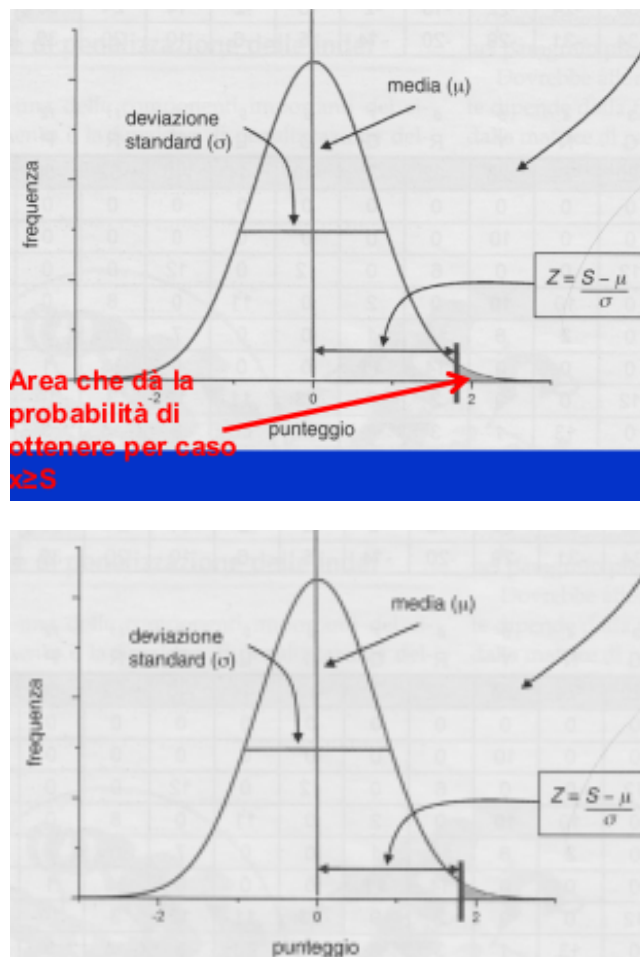
### 12.1.1 Significatività allineamento globale: lo Z score

La seq  $A$  è mantenuta fissa; la  $B$  è "anagrammata"  $n$  volte ed ogni volta globalmente allineata ad  $A$ , calcolando lo score  $S_i$  per l'allineamento  $i$ .



$S_i$  si distribuisce su una curva di cui si calcola la media  $\mu$  e la deviazione standard  $\sigma$ . Si definisce allora la distanza  $Z$  del punteggio  $S$  dell'allineamento dalla media in termini di dev. standard:

$$Z = \frac{S - \mu}{\sigma}$$



- Uno Z-score 0 = significa che la somiglianza osservata non è migliore rispetto alla media di permutazioni casuali della sequenza, e può anche essere casuale.
- Problema con Z-score: si assume una distribuzione normale, ma ciò può non esser corretto. Perciò Z deve essere considerato come una soglia di significatività.

### 12.1.2 Significatività allineamento locale

Teoria abbastanza complessa, sviluppata da Karlin e Altschul partendo da questa osservazione:

Date due sequenze casuali, di lunghezza  $m$  ed  $n$ , il numero atteso  $E$  di sottosequenze allineate localmente senza indel che ottengono un punteggio  $S \geq x$  è:

$$E(S \geq x) = K m n e^{-\lambda x}$$

$m, n$  : lunghezze delle due sequenze.

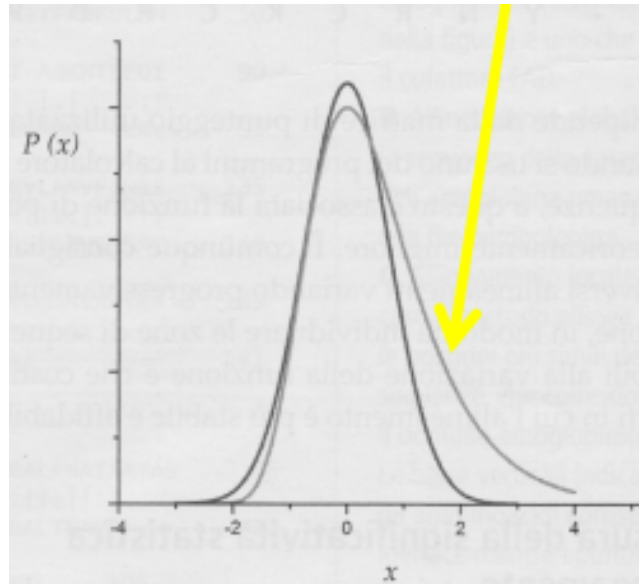
$K$ : dipende dalla matrice di punteggio.

$\lambda$ : dipende dalla composizione amminoacidica.

Dalla definizione di  $E$  si può calcolare la probabilità di osservare un allineamento locale con punteggio  $S \geq x$ :

$$p(S \geq x) = -\exp(K m n e^{-\lambda x})$$

Distribuzione del valore estremo o di Gumbel: è diversa dalla gaussiana



In pratica:

- allineamo localmente due seq
- otteniamo il punteggio  $x$
- calcoliamo  $p(S \geq x)$ , la probabilità di ottenere un punteggio maggiore di  $x$  nell'ipotesi: le due seq NON sono omologhe
- se  $p < \text{soglia}$  (es.  $0.01 = 1\%$ ) siamo confidenti che siano omologhe.
- SEMPRE: serve significatività BIOLOGICA oltre che statistica

## 13 Matrici di punteggio

Abbiamo già visto che per dare un punteggio a un allineamento dobbiamo misurare la similitudine fra aa. Usiamo perciò matrici di punteggio o di sostituzione: saranno matrici  $20 \times 20$ . Sono matrici simmetriche:  $A \rightarrow B = B \rightarrow A$  (non sappiamo evolutivamente chi si è trasformato dei due).

**Come quantificare la somiglianza degli amminoacidi** Difficile stabilire criteri oggettivi per le somiglianze fisico-chimiche degli amino acidi. Non è possibile sapere a priori quali delle varie caratteristiche fisico-chimiche sono più importanti per le proteine.

Emile Zuckerkandl e Linus Pauling (1965) considerarono frequenze di sostituzione in 18 globine (mioglobine e emoglobine da uomo a lampreda).

- Nero: identità
- Grigio: sostituzione molto conservativa (occorrenza  $> 40\%$ )
- Bianco: sostituzione abbastanza conservativa (occorrenza  $> 21\%$ )
- Rosso: non è possibile osservare sostituzioni

Substituent residue  
(Percentage of total residue sites at which the substituent occurs)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A				28			31	33								31				
R								50			58					25				
N	33			47				33		33						33	33			
D	44		22			47	34	22			28					25				
C	(66)																			
Q				56			30	40			70									
E	50			44			38				41			24						
G	51			33			30				27					36				
H				26							26	30				22	22			
I	39										58									46
L	21									23	23	28								30
K	23	21		28			31	23			21					21				
M	22									22	89			22						45
F									22		61									
P	50			43			57	43				21								
S	49			24			24	36			24						40			
T	32						28	24			24					52				
W	(40)										(40)			(60)						
Y								(33)						(50)						
V	36									21	43	21								

Sequence (original amino acid)

**Cosa vogliamo ottenere?** una matrice (PAM250) di score. Matrice di calcolo che assegna i punteggi e tollera le discordanze. Inoltre ...tutta una serie di matrici di punteggio, (fino a PAM10) che è via via meno tollerante con i disallineamenti.

### 13.1 PAM: Point Accepted Mutation

Mutazione puntuale accettata.

- È l'evento in cui il DNA subisce una mutazione che produce il cambiamento di un aminoacido
- Tale mutazione diviene prevalente in una specie

Dayhoff ha osservato famiglie di sequenze identiche all'85% (omologhe e molto simili). Le ha allineate e ha creato alberi di sequenze in cui ha dedotto le sequenze dei progenitori. Piccoli passi evolutivi, per osservare l'evoluzione e dedurne le caratteristiche.

Le matrici PAM sono basate su allineamenti globali di proteine strettamente correlate. Il PAM1 è la matrice calcolata dal confronto di sequenze con non più di 1% di divergenza. Ad un intervallo evolutivo di PAM1, un cambiamento si è verificato su una lunghezza di 100 aminoacidi.

Altre matrici PAM sono estrapolate da PAM1 ( PAM1 non ha utilità pratica ). Per PAM250, 250 sostituzioni si sono verificate tra due proteine su una lunghezza di 100 aminoacidi, nel passo evolutivo che essa rappresenta.

*Nota bene:* Tutti i dati PAM provengono da proteine strettamente correlate (> 85% di identità degli aminoacidi).

**Dayhoff: 34 superfamiglie di proteine**

<b>Proteina</b>	<b>PAMs per 100 milioni di anni</b>
Ig kappa chain	37
Kappa casein	33
luteinizing hormone b	30
lactalbumin	27
complement component 3	27
epidermal growth factor	26
proopiomelanocortin	21
pancreatic ribonuclease	21
haptoglobin alpha	20
serum albumin	19
phospholipase A2, group IB	19
prolactin	17
carbonic anhydrase C	16
Hemoglobin $\alpha$	12
Hemoglobin $\beta$	12
apolipoprotein A-II	10
lysozyme	9.8
gastrin	9.8
myoglobin	8.9
nerve growth factor	8.5
myelin basic protein	7.4
thyroid stimulating hormone b	7.4
parathyroid hormone	7.3
parvalbumin	7.0
trypsin	5.9
insulin	4.4
calcitonin	4.3
arginine vasopressin	3.6
adenylate kinase	3.2
triosephosphate isomerase 1	2.8
vasoactive intestinal peptide	2.6
glyceraldehyde phosph. dehydrogease	2.2
cytochrome c	2.2
collagen	1.7
troponin C, skeletal muscle	1.5
alpha crystallin B chain	1.5
glucagon	1.2
glutamate dehydrogenase	0.9
histone H2B, member Q	0.9
ubiquitin	0

**Dayhoff e i numeri di “point accepted mutations”:  
Quali sostituzioni si verificano nelle proteine?**

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A								
R	30							
N	109	17						
D	154	0	532					
C	33	10	0	0				
Q	93	120	50	76	0			
E	266	0	94	831	0	422		
G	579	10	156	162	10	30	112	
H	21	103	226	43	10	243	23	10

Conteggio delle mutazioni osservate (PAM1) Dayhoff (1978) p.346.

### 13.2 La mutabilità relativa degli amminoacidi

Quanto spesso mutano nelle proteine? Definiamo la Frequenza relativa di mutazione.

AA	Freq	AA	Freq
Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Phe	41
Ile	96	Phe	41
Gln	93	Cys	20
Val	74	Trp	18

Frequenze normalizzate

#### Ogni quanto occorrono nelle proteine?

<b>Gly</b>	<b>8.9%</b>	<b>Arg</b>	<b>4.1%</b>
<b>Ala</b>	<b>8.7%</b>	<b>Asn</b>	<b>4.0%</b>
<b>Leu</b>	<b>8.5%</b>	<b>Phe</b>	<b>4.0%</b>
<b>Lys</b>	<b>8.1%</b>	<b>Gln</b>	<b>3.8%</b>
<b>Ser</b>	<b>7.0%</b>	<b>Ile</b>	<b>3.7%</b>
<b>Val</b>	<b>6.5%</b>	<b>His</b>	<b>3.4%</b>
<b>Thr</b>	<b>5.8%</b>	<b>Cys</b>	<b>3.3%</b>
<b>Pro</b>	<b>5.1%</b>	<b>Tyr</b>	<b>3.0%</b>
<b>Glu</b>	<b>5.0%</b>	<b>Met</b>	<b>1.5%</b>
<b>Asp</b>	<b>4.7%</b>	<b>Trp</b>	<b>1.0%</b>

- **blu**=6 codoni; **rosso**=1 codone
- Le frequenze  $f_i$  si sommano a 100

**Esempio** Prendiamo un allineamento multiplo di sequenze, ad esempio della deidrogenasi gliceraleide 3-fosfato.

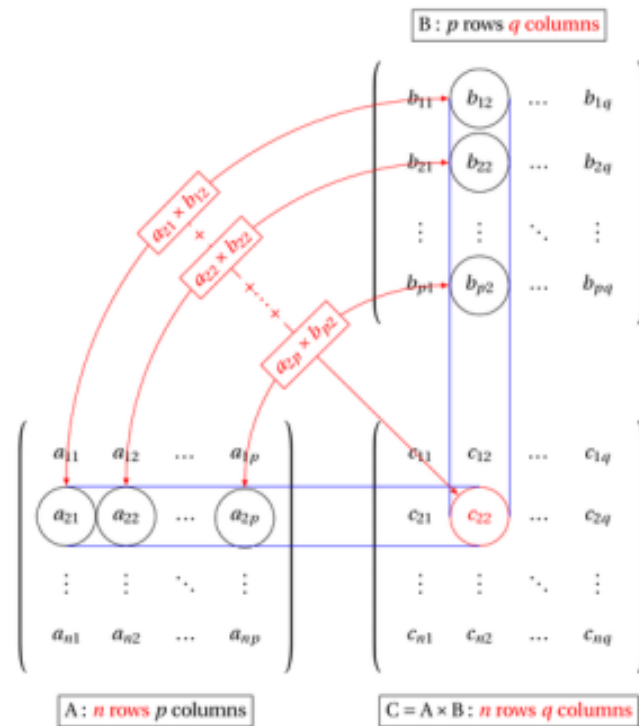
**OSSERVAZIONE:** le colonne di residui possono avere conservazione alta o bassa.

### 13.2.1 Matrice PAM1 (probabilità) di Dayhoff

Ogni elemento della matrice mostra la probabilità che un amminoacido (in alto) venga sostituito da un altro amminoacido (a lato).

Una matrice di sostituzione contiene valori proporzionali alla probabilità che l'amminoacido  $i$  muti nell' amminoacido  $j$  per tutte le coppie possibili di amminoacidi. Le matrici di sostituzione sono costruite assemblando un campione ampio e diversificato di allineamenti a coppie (o allineamenti multipli di sequenza) di amminoacidi. Le matrici di sostituzione dovrebbero riflettere la probabilità reale di mutazione in un periodo di evoluzione. I due principali tipi di matrici di sostituzione: PAM e BLOSUM.

Con (PAM1)<sup>n</sup> si può simulare il passaggio di n passi di evoluzione.



**Matrice di sostituzione PAM0 (probabilità)** Ovvero: nulla cambia.

		Aminoacido originale							
Aminoacido mutato	PAM0	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
	A	100%	0%	0%	0%	0%	0%	0%	0%
	R	0%	100%	0%	0%	0%	0%	0%	0%
	N	0%	0%	100%	0%	0%	0%	0%	0%
	D	0%	0%	0%	100%	0%	0%	0%	0%
	C	0%	0%	0%	0%	100%	0%	0%	0%
	Q	0%	0%	0%	0%	0%	100%	0%	0%
	E	0%	0%	0%	0%	0%	0%	100%	0%
	G	0%	0%	0%	0%	0%	0%	0%	100%

Si sono verificati 0 passi di evoluzione: non è cambiato nulla!

**Matrice di sostituzione PAM2000 (probabilità)**  $PAM1^{2000}$ , ovvero: il caso.

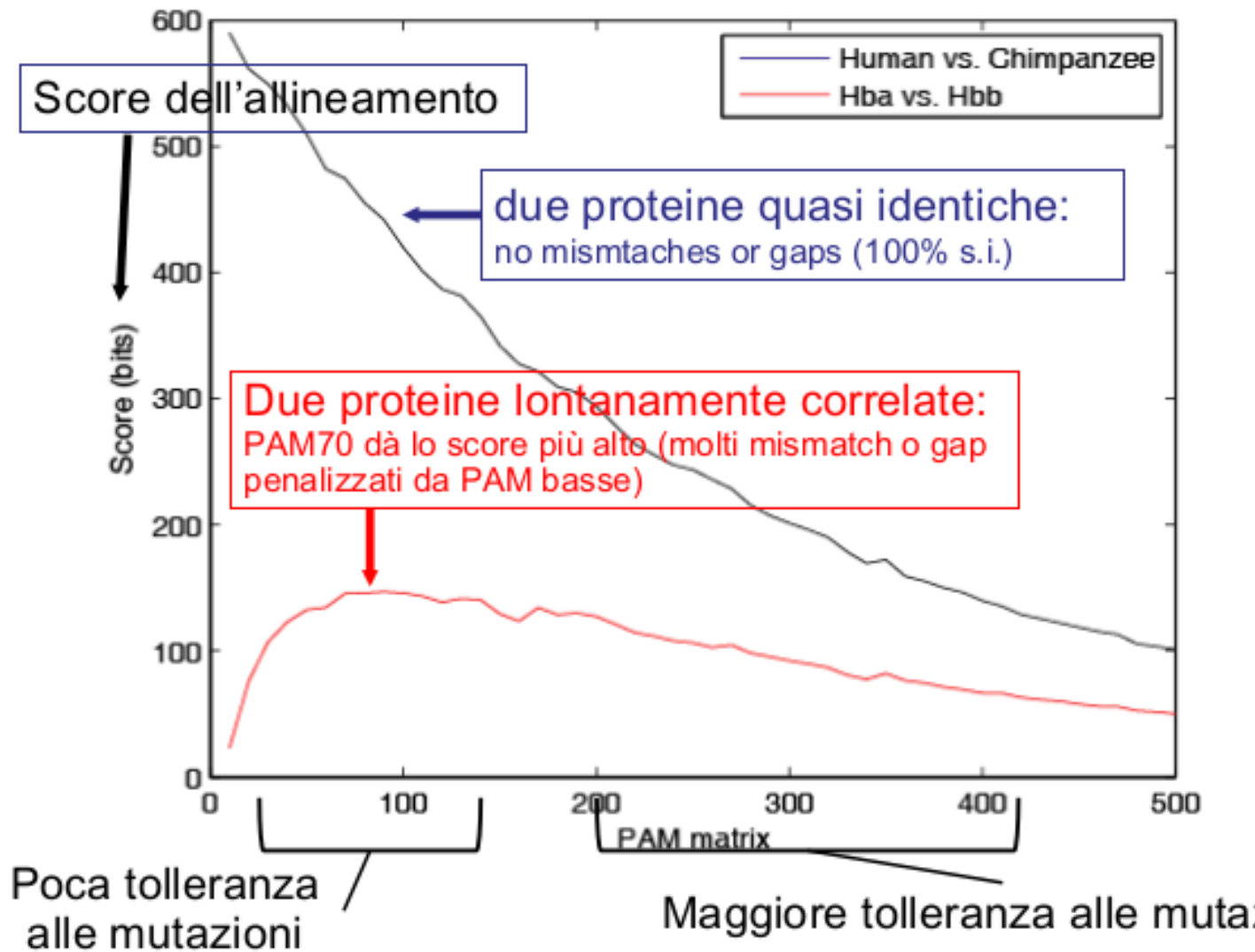
		Aminoacido originale							
Aminoacido mutato	PAM $\infty$	A Ala	R Arg	N ASN	D ASP	C Cys	Q Gln	E Glu	G Gly
	A	8,7%	8,7%	8,7%	8,7%	8,7%	8,7%	8,7%	8,7%
	R	4,1%	4,1%	4,1%	4,1%	4,1%	4,1%	4,1%	4,1%
	N	4,0%	4,0%	4,0%	4,0%	4,0%	4,0%	4,0%	4,0%
	D	4,7%	4,7%	4,7%	4,7%	4,7%	4,7%	4,7%	4,7%
	C	3,3%	3,3%	3,3%	3,3%	3,3%	3,3%	3,3%	3,3%
	Q	3,8%	3,8%	3,8%	3,8%	3,8%	3,8%	3,8%	3,8%
	E	5,0%	5,0%	5,0%	5,0%	5,0%	5,0%	5,0%	5,0%
	G	8,9%	8,9%	8,9%	8,9%	8,9%	8,9%	8,9%	8,9%

Moltiplicando PAM1 per 2000 (passi di evoluzione) si arriva ad una situazione in cui la probabilità converge alla frequenza osservata

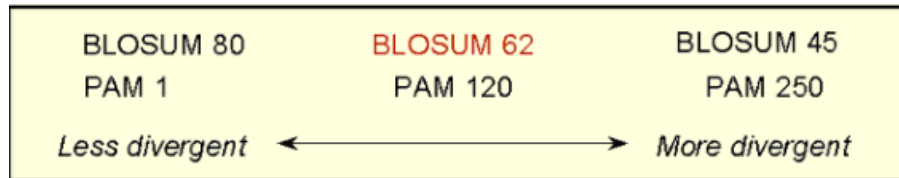
**Matrice di sostituzione PAM250 (probabilità) di mutazione**









**Più conservativo**

es. Globina: topo vs ratto

Nella BLOSUM 80 le sequenze identiche per l'80% finiscono in un unico blocco e gli score sono applicati considerando gli altri allineamenti → score adatti per proteine simili (come per la PAM10)

**Meno conservativo**

es. Globina: topo vs batterio

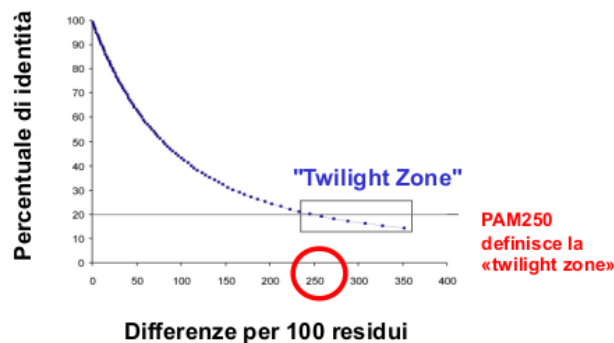
Nella BLOSUM 45 le sequenze identiche per il 45% finiscono in un unico blocco e gli score sono applicati considerando gli altri allineamenti → score adatti per proteine diverse (come per la PAM250)

## 16.1 La "Twilight zone" nell'allineamento di proteine

Quando compariamo due sequenze proteiche simili, quante mutazioni possono accumularsi prima che le differenze le rendano irriconoscibili? L'identità tra due sequenze ciascuna di 100 aminoacidi cala come un esponenziale negativo all'accumularsi delle mutazioni.

- A PAM1, due proteine sono al 99% identiche
- A PAM10.7, ci sono 10 differenze ogni 100 residui
- A PAM80, ci sono 50 differenze ogni 100 residui
- A PAM250, ci sono 80 differenze ogni 100 residui

Oltre (20-25% identità) non è più distinguibile una similitudine (Allineamenti multipli, modeling)



## Lezione 6: Allineamenti Multipli di Sequenze

## 17 Allineamento multiplo di sequenze

### 17.1 Visione Generale

#### 17.1.1 Una definizione

Un allineamento multiplo è una collezione di tre o più sequenze proteiche (o nucleotidiche) parzialmente o completamente allineate

- I residui e le zone omologhe sono allineate in colonne per tutta la lunghezza delle sequenze
- Il senso dell'omologia dei residui è evolutivo
- Il senso dell'omologia dei residui è strutturale

Si tratta di un argomento di ricerca attivo dagli anni '90.

#### 17.1.2 Alcuni fatti

Non c'è necessariamente un allineamento "corretto" per una famiglia di proteine.

**Perché?**

- Le sequenze di proteine evolvono
- Le corrispondenti strutture tridimensionali evolvono, anche se più lentamente
- Può essere particolarmente difficile identificare i residui che si sovrappongono nello spazio (strutturalmente) in un allineamento multiplo di sequenze.

Due proteine che condividono il 30% di identità di sequenza avranno circa il 50% dei residui sovrapponibili nelle due strutture

#### 17.1.3 Caratteristiche utili per realizzarlo

Alcuni residui allineati, come cisteine che formano ponti disolfuro, o i triptofani, possono essere altamente conservati

- Ci possono essere motivi conservati come un dominio transmembrana
- Alcune caratteristiche come le strutture secondarie, siti attivi e di legame per ligandi o complessi sono spesso conservate
- Ci possono essere regioni con inserimenti o delezioni propagati in parte della famiglia.
- I principi che vedremo sono focalizzati sulle proteine ma sono validi in generale anche per sequenze nucleotidiche.

#### 17.1.4 Utilizzi e Vantaggi

- Il MSA è più sensibile di quello a coppie nel rilevamento di omologie, per questo è uno strumento essenziale nella costruzione di modelli strutturali per omologia
- L'output di BLAST può assumere la forma di un MSA, e possono essere individuati residui conservati o motivi
- In un MSA si possono analizzare i dati di una popolazione
- Una singola query può essere cercata contro un database di MSA (ad esempio Pfam)
- Le regioni regolatorie dei geni sono spesso identificabili da MSA

## 17.2 Metodi

I metodi esatti non vengono trattati in questa sede: non ci sono soluzioni efficienti e già con 5 sequenze il tempo di computazione è eccessivo (esponenziale)

### 17.2.1 Metodi Euristici

**Metodi progressivi:** usano un albero guida (analogo ad un albero filogenetico) per determinare come combinare uno per uno allineamenti a coppie (progressivamente) per creare un allineamento multiplo.

Esempi: CLUSTAL OMEGA (W), MUSCLE (usato da HomoloGene)

**Il MSA progressivo di Feng-Doolittle (1987) alla base di Clustal (W) avviene in 3 fasi**

1. Realizzare una serie di allineamenti a coppie globali (Needleman e Wunsch, algoritmo di programmazione dinamica) di cui si calcola la distanza (matrice delle distanze)
2. Creare un albero guida a partire dalla matrice delle distanze
3. Allineare progressivamente le sequenze

**MSA progressivo, fase 1 di 3:**

generare allineamenti a coppie globali

Esempio: allineare 5 globine (1, 2, 3, 4, 5).

**Primo step:** a due a due e valutare gli score di ogni possibile allineamento a coppie

**Numero di allineamenti a coppie necessari per coprire tutte le possibili combinazioni**

- Per n sequenze,  $(n-1)(n) / 2$
- Per 5 sequenze,  $(4)(5) / 2 = 10$
- Per 200 sequenze,  $(199)(200) / 2 = 19.900$

... Quindi per molte sequenze ClustalW è molto lento ed è preferibile usare metodi più veloci (MUSCLE è molto veloce).

**Secondo step:** albero guida

**Convertire i punteggi di similitudine in punteggi di distanza:** è matematicamente più semplice, oltre che più intuitivo, lavorare con le distanze. Una semplice definizione di distanza è data dalla percentuale di residui diversi (100-SI in %) che viene inserita nella matrice delle distanze.

- Dalla matrice delle distanze si calcola l'albero guida con il metodo di clustering neighbor joining che vedrete nel modulo 2.
- Vediamo un semplice esempio di clustering e costruzione di albero guida

**Il clustering alla base di CLUSTAL(W)** È una matrice di distanze, minore è il numero, maggiore è la similitudine. *Nota: tutte le distanze tra la I e la IV riga sono minori di quelle riportate nella V*

- Tutte le sequenze vengono poi allineate progressivamente, seguendo le indicazioni dell'albero guida: prima si allineano le più simili (vicine) e poi progressivamente le più distanti.
- A ogni passaggio si utilizza un algoritmo dinamico di allineamento molto efficiente che accoppia sequenze o gruppi di sequenze
- Il MA è composto a partire a tanti allineamenti a coppie, anche fra gruppi di sequenze.
- *Nota: Le indel presenti negli allineamenti già effettuati restano fisse.*
- Come allineare progressivamente due gruppi di sequenze? Si usa sempre una matrice. È simile all'allineamento dinamico di due sequenze visto.
- Lo score S in ogni casella è la media degli score ottenuti confrontando tutte le possibili coppie di a.a. nella riga e colonna corrispondenti (secondo ad es. BLOSUM62)
-

## Domande stile esame



## 18 Domande sull'introduzione

1. Differenza tra omologia e similitudine. È possibile che due proteine abbiano un'identità di sequenza del 57% e una similitudine del 21%? Perché?

## 19 Domande sulle banche dati

1. ) Cos'è Uniprot? Oltre ad una descrizione generale si descrivano i suoi livelli e si elenchino almeno 5 sezioni che si possono trovare nelle pagine delle singole proteine (entry). Si descrivano inoltre brevemente i database PDB e ExPASy. Cosa hanno in comune queste tre risorse?
2. Si citino e discutano brevemente 4 banche dati presenti in NCBI. Se possibile escludere NCBI protein.
3. Si descrivano le banche dati NCBI Gene, Unigene e Genbank. Inoltre, in che relazione sono tra di esse?
4. Descrivere Uniprot, Pfam, Prosite, CATH e PDB, con particolare attenzione alla più importante tra di esse (qual è?)
5. Cos'è Pubmed? Descrivere la banca dati, i suoi contenuti, e gli strumenti messi a disposizione degli utenti (illustrati a lezione), sia per la ricerca che per la gestione dei risultati.
6. Descrivere brevemente le caratteristiche del Protein Data Bank ed il contenuto di un file pdb.
7. Si descrivano il formato FASTA, il formato ASN.1 e il formato XML nell'ambito della bioinformatica.
- 8.