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 LaTeXcon cui è stato generato.

Contents

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 cor-rente. 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 prote-omica, 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 stru-menti della boinformatica 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.

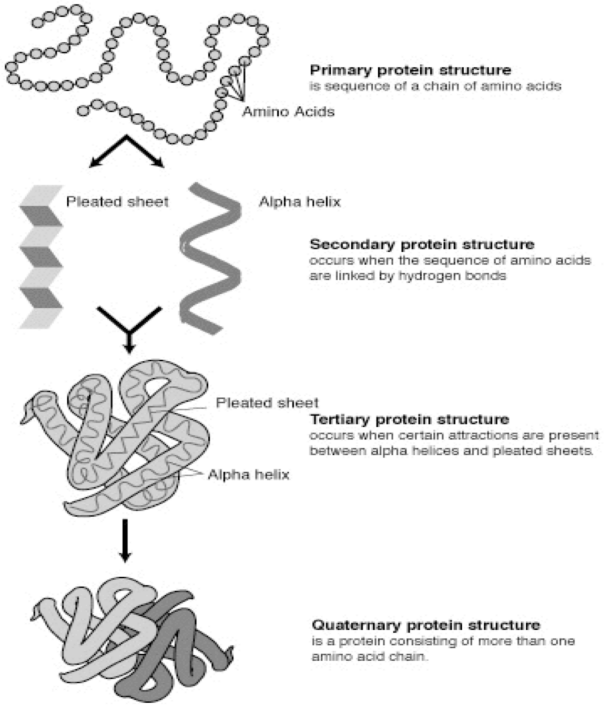
- EVOLUZIONE: adattamento progressivo attraverso variabilita' genetica casuale e selezione naturale (Darwin, 1859)
- Ad ogni livello biologico, il fenotipo (insieme di tratti e caratteri somatici) è codificato dal genotipo (il patrimonio genetico)
- Genotipo: sorgente primaria di variazione genetica; 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 piu' 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 proprieta' 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 piu' "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 identita' 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 pioneristico per il tempo!) Le tecniche di sequenziamento rapido ed i progetti -*omici* hanno prodotto una quantita' 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 Accids 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 N**C**_######
- Genomic contig (segmenti sovrapposti di DNA segments che rappresentano una sequenza consenso) N**T**_######
- mRNA (DNA format) N**M**_######
- Protein N**P**_######

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 quesa 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 automaticmente la stringa di ricerca: `(beta globin) AND "Homo sapiens" [ porgn:_txid9606 ]`
8. Con il limite Homo Sapiend 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 detagliato, 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. dentificazione degli a.a. che svolgono un ruolo strutturale o funzionale analogo (allineamento).

Esempio 1: la catena β dell'emoglobina e miogobina «si somigliano»

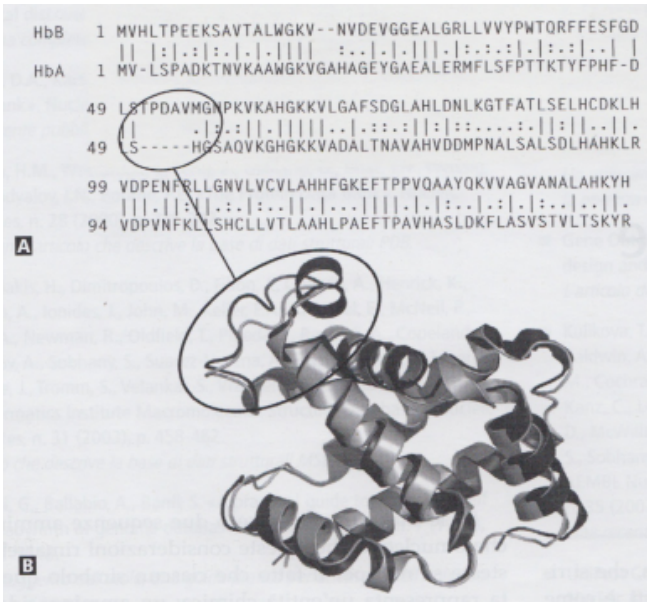


Figure 1: Le zone con indel nelle sequenze sono strutturalmente dissimili

- **Ortologhi**

  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.

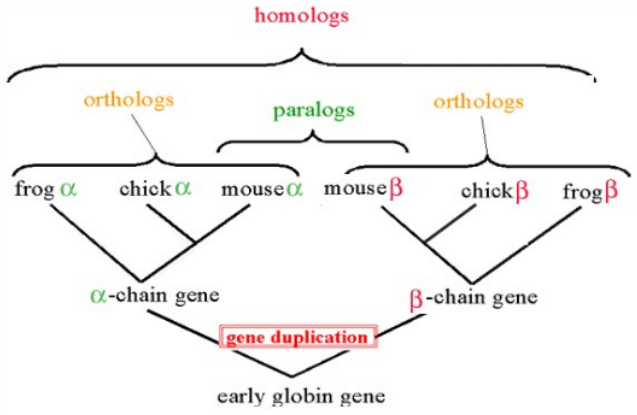


Figure 2: Ortologhi e paraloghi sono spesso rappresentati in un albero singolo.

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

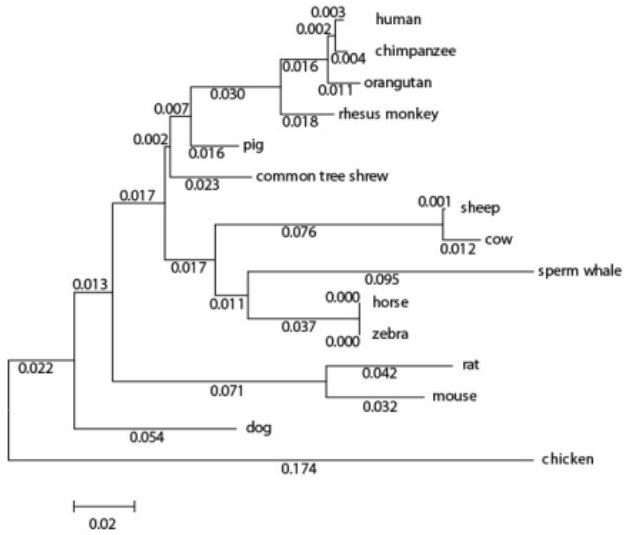


Figure 3: Questo albero mostra gli ortologhi della globina.

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

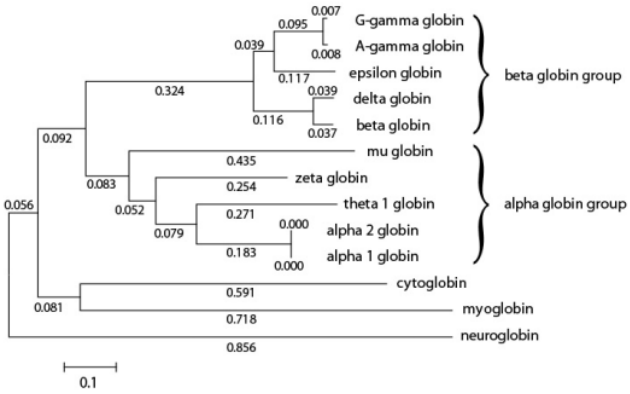


Figure 4: Questo albero mostra i paraloghi della globina umana.

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

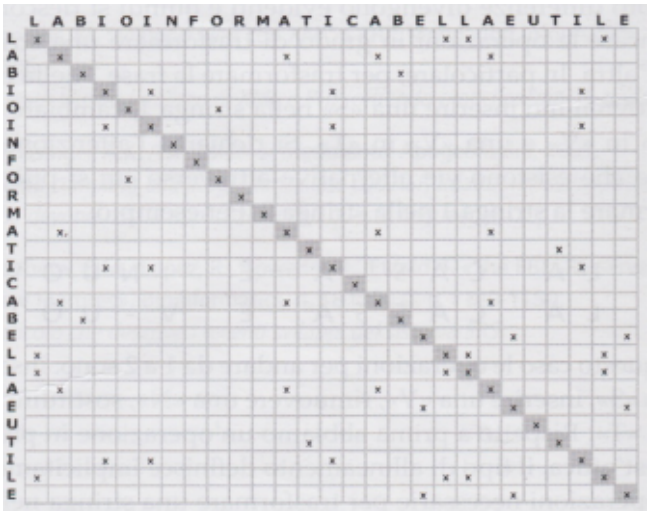
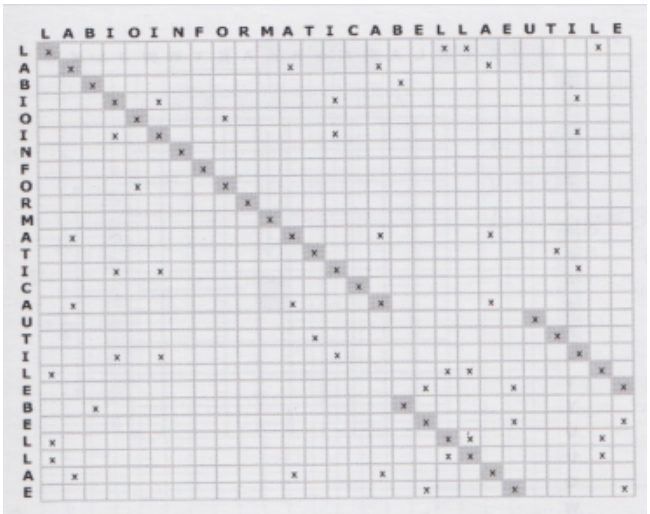


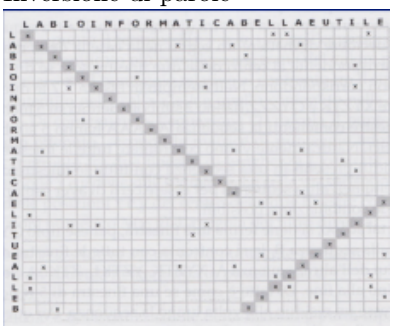
Figure 5: ... piuttosto banale. Cambiamo la seconda stringa.

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

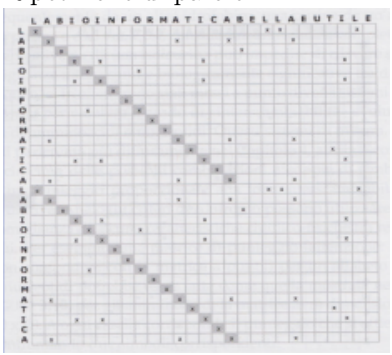


Possiamo individuare facilmente alcuni patterns mediante le matrici a punti (la prima stringa in alto resta immutata):

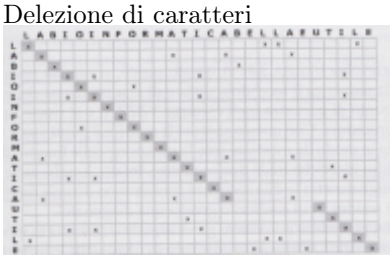
- Inversione di parole



- Ripetizione di parole

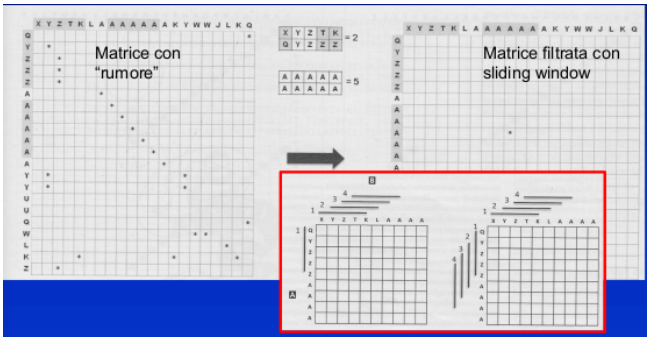


- Delezione di caratteri



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 diagonali; 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.

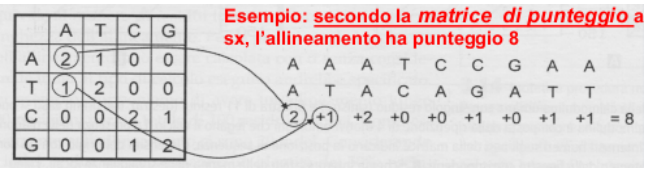
La procedura

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:



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:

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

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

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

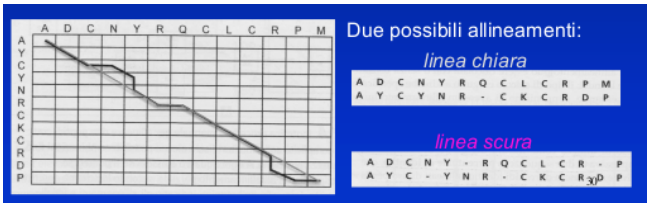
Esempio: proviamo tutte le combinazioni da sx a dx, riempiendo una colonna alla volta (qui mostriamo solo i primi 3 residui!!)
Stabiliamo inoltre i punteggi:

- -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.

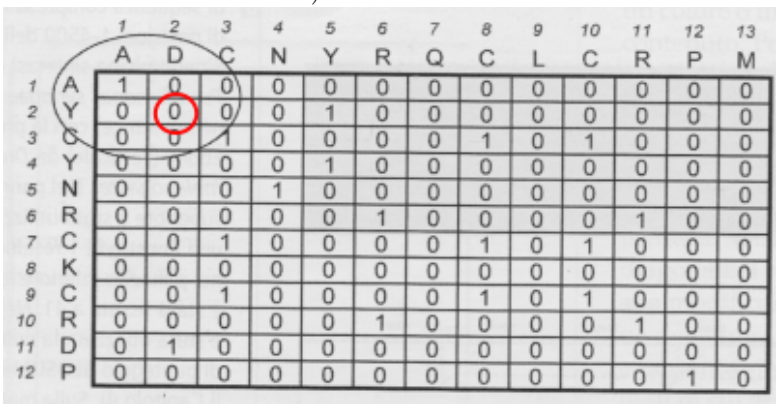


10.2 Regole pratiche

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)

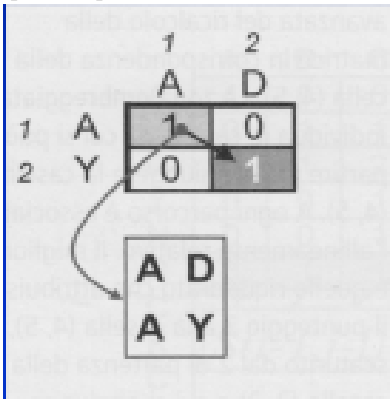
10.3 Step

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)

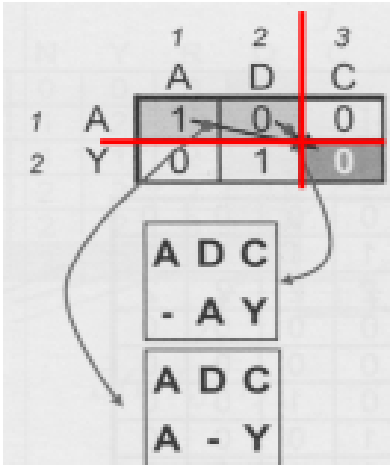


		1	2	3	4	5	6	7	8	9	10	11	12	13
		A	D	C	N	Y	R	Q	C	L	C	R	P	M
1	A	1	0	0	0	0	0	0	0	0	0	0	0	0
2	Y	0	0	0	0	1	0	0	0	0	0	0	0	0
3	C	0	0	1	0	0	0	0	1	0	1	0	0	0
4	Y	0	0	0	0	1	0	0	0	0	0	0	0	0
5	N	0	0	0	1	0	0	0	0	0	0	0	0	0
6	R	0	0	0	0	0	1	0	0	0	0	1	0	0
7	C	0	0	1	0	0	0	0	1	0	1	0	0	0
8	K	0	0	0	0	0	0	0	0	0	0	0	0	0
9	C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

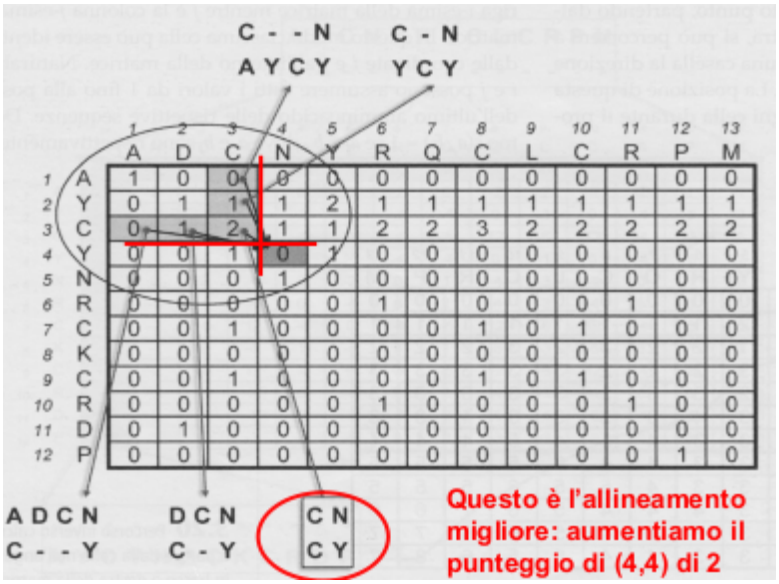
2. Partiamo da (1,1) (in alto a sx: la direzione è importante!!): (1,1) $\rightarrow$ (2,2) L'unico percorso possibile è questo: per "ricordarmelo" sommo il valore della cella (1,1) a quello della cella (2,2) da matrice inizializzata



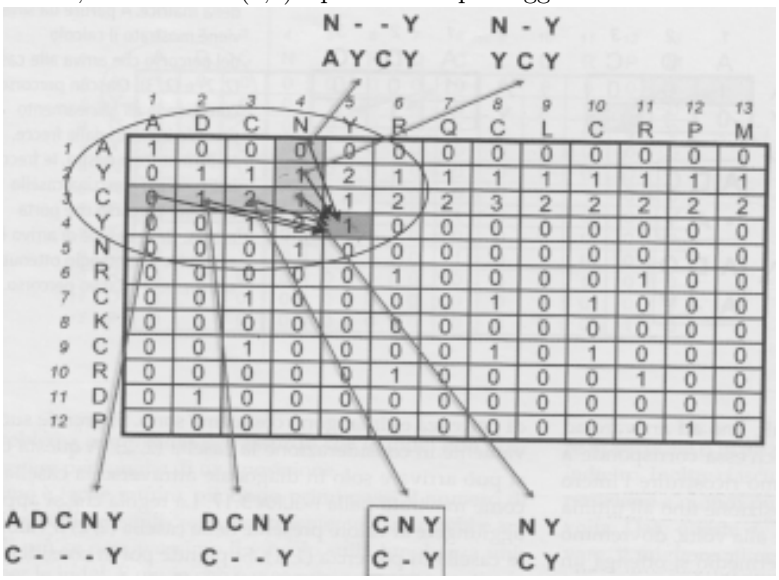
3. Prossimo step: verso (2,3). Ci sono due possibili percorsi, corrispondenti a due diversi allineamenti: scelgo quello con punteggio finale maggiore, sempre sommando casella precedente a (2,3)



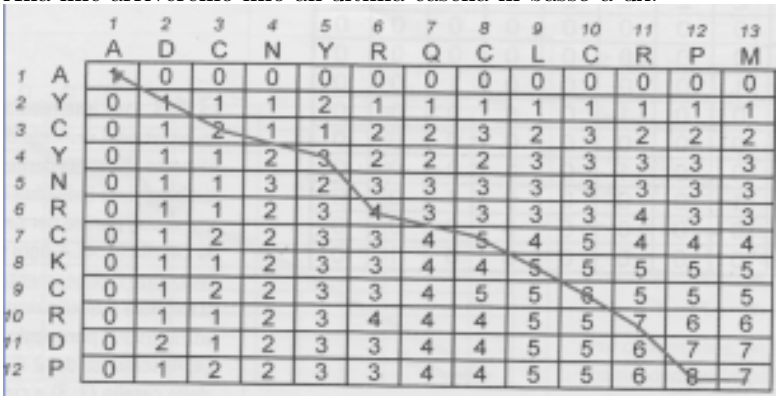
4. Procediamo analogamente: ad esempio, verso (4,4)



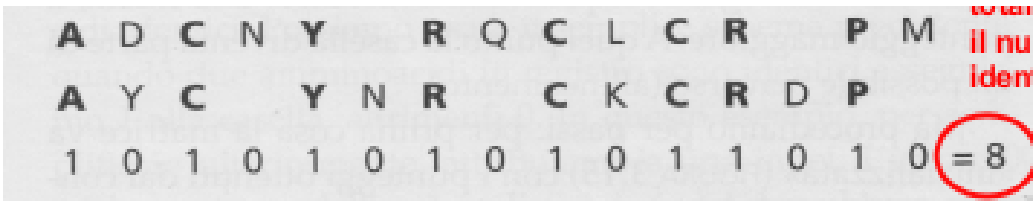
5. e oltre, ad es. verso (4,5): qui il nuovo punteggio sarà 3



6. Alla fine arriveremo fino all'ultima casella in basso a dx:

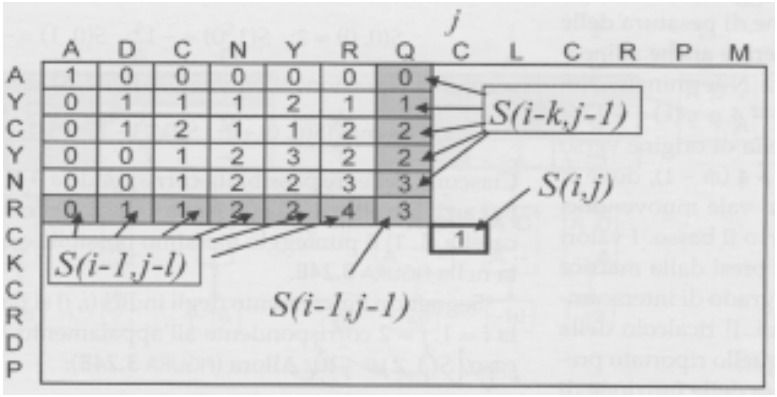


7. Da essa possiamo spostarci a ritroso poiché abbiamo memorizzato i migliori punteggi dalle caselle precedenti. Abbiamo così determinato il miglior allineamento:



Qui il punteggio totale coincide con il numero di residui identici.
$\frac{8}{15} = 0.53 = 53\%$ *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:



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

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 6: Allineamenti Multipli di Sequenze

11 Allineamento multiplo di sequenze

11.1 Visione Generale

11.1.1 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 evoluzionistico
- Il senso dell'omologia dei residui è strutturale

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

11.1.2 Alcuni fatti

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

- 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

11.1.3 Caratteristiche utili per realizzzarlo

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.

11.1.4 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

11.2 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)

11.2.1 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 pogressivamente 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

12 Domande sull'introduzione

1. Differenza tra omologia e similitudine. È possibile che due proteine abbiano unìidentità di sequenza del 57% e una similitudine del 21%? Perchè?

13 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.