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.

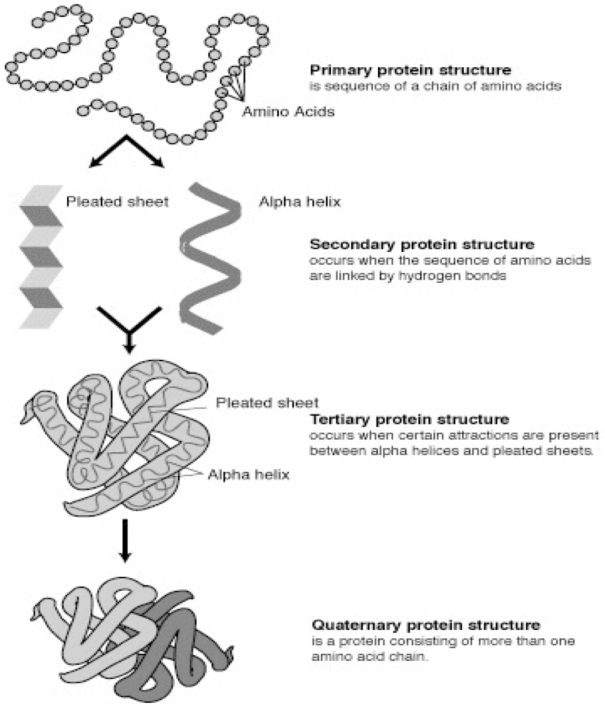
- EVOLUZIONE: adattamento progressivo attraverso variabilita' genetica casuale e selezione naturale (Darwin, 1859)
- Ad ogni livello biologico, il fenotipo (insieme di tratti e caratteri somatici) è codificato dal genotipo (il patrimonio genetico)
- Genotipo: sorgente primaria di variazione genetica; 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.
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**_######

Lezione 6: Allineamenti Multipli di Sequenze

7 Allineamento multiplo di sequenze

7.1 Visione Generale

7.1.1 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.

7.1.2 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

7.1.3 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.

7.1.4 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

7.2 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)

7.2.1 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)
-