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.

Lezione 6: Allineamenti Multipli di Sequenze

3 Allineamento multiplo di sequenze

3.1 Visione Generale

3.1.1 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.

3.1.2 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

3.1.3 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.

3.1.4 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

3.2 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)

3.2.1 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.