I dati ed i problemi della bioinformatica

Giorgio Valentini
DSI – Università degli Studi di Milano

1

Caratteristiche dei dati della biologia molecolare

- Diverse tipologie di dati bio-molecolari
- Per ogni tipo di dato diverse tipologie di misurazioni ottenute tramite bio-tecnologie
- In generale diverse bio-tecnologie evidenziano diverse caratteristiche dei dati

⇩

Necessità di metodi per
l'integrazione di dati eterogenei

2

Dati di sequenza

- Il processo di sequenziamento del DNA:
  - Necessario gran numero di molecole identiche (PCR)
  - Sono sequenziabili frammenti con poche centinaia di nucleotidi → assembly computazionale successivo
- Sequenziamento proteine:
  - diretto
  - indiretto tramite mRNA→cDNA e calcolo prodotti traduzione
  - Indiretto da genoma completo e calcolo prodotti traduzione
- Allineamenti:
  - Globali (Needleman e Wunsch, 1970)
  - Locali (Smith e Waterman, 1980)
  - Euristiche "veloci" (FASTA, BLAST, BLAT, …)
  - Allineamenti multipli (ClustalW, …)

3

Database di sequenze

- Tabella 1.5
- Esistono molti altri DB specializzati
- Ogni anno la rivista *Nucleic Acid Research* fornisce report su molti DB
- Molti DB accessibili da portali: SRS (EBI, europeo) ed Entrez (NBI, USA)

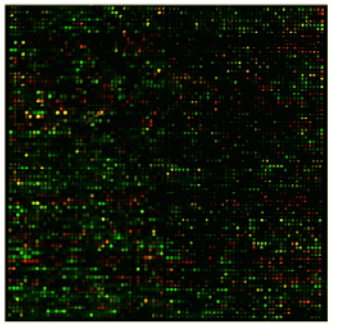
4

Dati di espressione genica

- Microarray (DNA chip):
  - Misurano il *livello di espressione* di migliaia di geni simultaneamente
  - Forniscono un'"istantanea" dello *stato funzionale* di una cellula
- Spot (cDNA e oligonucleotidi)
- Probe e target
- Ibridizzazione
- Misurazione del segnale emesso da ciascuno probe ibridizzato con il target

5

RGB overlay of Cy3 and Cy5 images



6

Caratteristiche dati DNA microarray

- Rappresentati in matrici a valori reali
- Rumore
- Bassa cardinalità ed elevata dimensionalità
- Necessità di pre-processing e "normalizzazione"
- Applicati alla ricerca bio-molecolare di base, alla ricerca bio-medica ed alla pratica clinica (ma anche altre applicazioni …)

7

Altre tecnologie per la misurazione dell'espressione genica

- SAGE (Serial Analysis of Gene Expression):
  - mRNA → cDNA → amplificazione →sequenziamento
  - Clustering delle sequenze
  - Stima livello espressione dalla dimensione del clustering
- qPCR (quantitative Polymerase Chain Reaction)

8

Data base espressione genica

Database	URL	Note
Data base generali		
ArrayExpress	www.ebi.ac.uk/arrayexpress	EBI
GEO	www.ncbi.nlm.nih.gov/geo	NCBI
Database organismo-specifici		
MGI GXD	www.inforamtics.jacs.org	*M. musculus*
TAIR	www.arabidopsis.org	*A. thaliana*
WormBase	www.wormbase,org	*C. elegans*
Database laboratorio-specifici		
SMD	Genome-www.stanford.edu/microarray	Stanford
YMD	Info.med.yale.edu/microarray	Yale

9

Dati relativi alle proteine

- Dati di sequenza (str. primaria)
- Dati relativi a struttura secondaria
- Dati struttura terziaria
- Dati di interazione :
  - Interazioni molecolari di legame
  - Interazioni di regolazione
  - Link a pathway metabolici
  - In generale network bio-molecolari (grafi)
- Dati di espressione:
  - Elettroforesi 2D
  - Spettrometria di massa

10

Database per le proteine

- Tab. 1.7

11

Altri tipi di dati

- Metaboliti, molecole-segnale
- SNP (Small Nucleotide Polymorphism)
- Testi scientifici (target per algoritmi di text mining)

12

Tipi di dati genomici e loro rappresentazione per l'analisi computazionale

- Tab. 1.4

13

Organismi modello

- Tab. 1.10

14

I problemi della bioinformatica

- Analisi della struttura e delle funzioni del genoma
- Analisi della struttura e delle funzioni del proteoma
- Simulazione di sistemi biologici
- Analisi delle relazioni fra dati bio-molecolari e fenotipi
- Chemioinformatica

15

Analisi della struttura e delle funzioni del genoma

- Ricerca ed analisi della struttura dei geni
  - Individuazione delle regioni codificanti
  - Identificazione dei siti di splicing
  - Identificazione dei promoter
- Comprensione della regolazione della trascrizione
  - Predizione dei livelli di espressione dai promoter e dai TF
  - Ricerca degli elementi funzionali e delle interazioni fra elementi funzionali
- Comparazione di interi genomi
  - Comparazione di coppie di genomi
  - Comparazione multipla di genomi
  - Ricostruzione della storia evolutiva delle specie

16

Analisi della struttura e delle funzioni del proteoma (1)

- Problemi distinti per:
  - Il tipo di proprietà da predire
  - Il tipo di dati utilizzato per la predizione

Tipo di dato:	**Tipo di Proprietà da predire**
• sequenza	• Struttura
• struttura	• Funzione
• espressione	• Interazione
• filogenetico	• Localizzazione

17

Analisi della struttura e delle funzioni del proteoma (2)

- Predizione della struttura:
  - Secondaria
  - Terziaria (3D)
- Predizione della funzione
  - Classificazione gerarchica (GO)
  - Utilizzo di diverse sorgenti di dati
- Ricostruzione di reti genetiche
  - Grafi che specificano interazioni fra molecole
  - Reti booleane, lineari e bayesiane
- Docking
  - Predizione di legami proteina-proteina
  - Predizione di legami ligando-proteina
  - Predizione dell'intensità del legame

18

Analisi delle relazioni fra dati bio-molecolari e fenotipi per la medicina bio-molecolare

- Diagnosi
  - Diagnosi basata su dati di espressione genica
  - Diagnosi basata su SNP
  - Diagnosi basate sull'integrazione di dati bio-molecolari e dati clinici tradizionali
  - Ricerca non supervisionata di sottoclassi patologiche a livello bio-molecolare
- Ricerca di target
  - Metodi di feature selection per l'individuazione di geni/molecole target
  - Individuazione di target per lo sviluppo di nuovi farmaci
- Terapia individualizzata
  - Ricerca di farmaci sulla base del profilo bio-molecolare di pazienti e/o agenti patogeni
- Genetica di popolazioni
  - Alberi genealogici annotati con informazioni genotipiche e fenotipiche degli individui
  - Ricerca di regioni cromosomiche e marker genetici associati a fenotipi specifici

19

Chemioinformatica

Analisi computazionale di composti chimici (sottodisciplina della bioinformatica?)

- HTS virtuale
  - **H**igh **T**hroughput **S**creening: ampi db di composti sono testati virtualmente rispetto a proteine recettore per individuare potenziali ligandi
- Inferenza della proprietà chimiche di molecole dalla loro struttura
  - QSAR, Quantitative Structure-Activity Relationship)
- Tossicologia predittiva
  - Predizione in silico della tossicità acuta e cronica delle sostanze chimiche
  - Predizione delle proprietà di molecole candidate per farmci (ADME: **A**bsorption, **D**istribution, **M**etabolism and **E**xcretion)

20

I problemi della bioinformatica

- Si possono “naturalmente” porre come problemi di machine learning
- Richiedono spesso l’elaborazione di dati complessi ed eterogenei

---

- Come modellare appropriatamente i problemi di bioinformatica come problemi di machine learning?
- Come elaborare strutture dati complesse?
- Come combinare differenti tipi di dati?

21