Riconoscimento e Recupero dell’informazione per Bioinformatica

Manuele Bicego

Corso di Laurea in Bioinformatica

Dipartimento di Informatica - Università di Verona

Il docente

Manuele Bicego

Dipartimento di informatica

Ufficio: Ca' Vignal 2 – Primo Piano – Stanza 1.55

Telefono: 045 8027072

e-mail: manuele.bicego@univr.it

Ricevimento:

- ⇨ mercoledì ore 11.00 - 13.00
- ⇨ su appuntamento concordato via e-mail

Il corso

Corso da 12 CFU

- ⇨ 9 CFU teoria: 8 CFU Manuele Bicego, 1 CFU Rosalba Giugno
- ⇨ 3 CFU laboratorio: Pietro Lovato

NOTA: lezioni di teoria in aula (alcune in lab, verso la fine), esercitazioni pratiche in laboratorio Alfa

Orario:

Lunedì	10.30 – 13.30	Lab. Alfa (Laboratorio)
Martedì	11.30 – 13.30	Aula C
Mercoledì	8.30 – 10.30	Aula C
Giovedì	15.30 – 18.30	Aula D

Pre-requisiti

Pre-requisiti per le attività in aula:

⇨ Conoscenze di base di Probabilità, Statistica, Analisi

Pre-requisiti per le attività in laboratorio:

⇨ Minima capacità di programmare

⇨ Verranno forniti i fondamenti di Matlab

Punto di vista

- Titolo del corso: Riconoscimento e Recupero dell'informazione per bioinformatica
  - è un titolo molto generico!

- Diversi possibili punti di vista:
  - Ad esempio: teorie e tecniche delle basi di dati

- In questo corso: studio delle tecniche di "Pattern Recognition" per estrarre informazioni (da dati biologici)
  - spesso alla base di programmi largamente utilizzati (ad esempio BLAST, Phylip, HMMER)

Obiettivi formativi

- Fornire le basi delle metodologie di Pattern Recognition
  - Capire cos’è la pattern recognition
  - Capire la differenza tra le diverse tipologie di problemi risolvibili con tecniche di pattern recognition
  - Capire come creare un sistema automatico di pattern recognition
  - Capire come validare i risultati ottenuti
  - Vedere esempi di applicazione di tecniche di Pattern Recognition a problemi di bioinformatica

- L’attenzione è rivolta principalmente alla descrizione delle metodologie piuttosto che ai dettagli dei programmi applicativi (già visti in altri corsi).

Programma (in generale)

Il corso si compone di due parti

- Teoria:
  - in questa parte verranno presentate le diverse metodologie di Pattern Recognition, le motivazioni che portano al loro studio, e i problemi connessi al loro utilizzo.
  - Verranno inoltre analizzati alcuni problemi bioinformatici che sono classicamente risolti con metodologie di pattern recognition

- Laboratorio:
  - verranno implementati in matlab semplici algoritmi di pattern recognition

Materiale didattico

- Materiale didattico: lucidi del corso, appunti presi a lezione (per lezioni alla lavagna), libri suggeriti, articoli, internet in generale.
  - I lucidi del corso saranno messi in linea prima delle lezioni

- Laboratorio di riferimento: VIPS (Vision, Image Processing & Sound), CV2, piano -2

  Info: http://vips.scienze.univr.it

Testi

- R. Duda, P. Hart, D. Stork *Pattern Classification.* Wiley, 2001 (2nd edition).
- P. Baldi, S. Brunak, B*ioinformatics, The Machine Learning Approach.* MIT Press, 2001
- G. Gan, C. Ma, J, Wu: *Data Clustering: Theory, Algorithms and Applications*, ASA-SIAM Series on Statistics and Applied Probability, 2007
- A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, 1988.
  - Disponibile on line http://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf
- S. Theodoridis, K. Koutroumbas: *Pattern Recognition*, Second edition, Academic press, 2003

Altri testi consigliati

- N. Cristianini, M.W. Hahn: *Introduction to Computational Genomics*, Cambridge University Press, 2007
- W.J. Ewens, G.R. Grant: *Statistical Methods in Bioinformatics*, Springer 2001
- C.M. Bishop, P*attern Recognition and Machine Learning*, Springer, 2006.
- W.J. Ewens, G.R. Grant, *Statistical Methods in Bioinformatics.* Springer, 2001
- E. Keedwell, A. Narayanan, *Intelligent Bioinformatics.* Wiley, 2005
- M. Berthold, D.J. Hand, *Intelligent Data Analysis.* Springer, 2003 (2nd edition).

Modalità d’esame

Due parti:

⇨ PARTE 1 (15 punti): breve scritto sugli argomenti del corso

⇨ PARTE 2 (15 punti): seminario di approfondimento, argomento da scegliere autonomamente

- OPZIONE 1: Seminario da fare a fine corso: due persone, 30/35 minuti

- OPZIONE 2: Seminario da fare “quando si vuole”: una persona sola, 30/35 minuti

LABORATORIO: Esame finale (non obbligatorio) con possibilità di prendere un 1 punto di bonus (seguite il lab!)

Per gli studenti particolarmente motivati: possibilità di fare stage e tesi laurea (anche in gruppi di due persone)

NOTA: Tutte le informazioni, il materiale didattico, gli aggiornamenti e gli avvisi inerenti al corso sono pubblicati alla pagina web del corso

Introduzione

Sommario

- Introduzione alla Pattern Recognition

- Pattern Recognition e Bioinformatica: perché?

Pattern Recognition

⇨ Punto di partenza: l'uomo e la Pattern Recognition

Che cos'è questa?

In che città mi trovo?

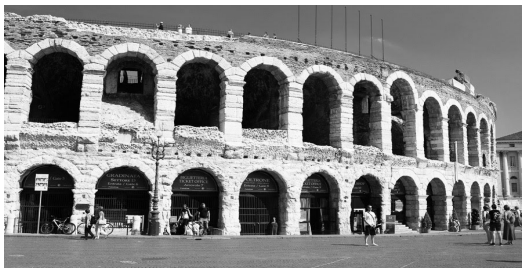
C'è una vespa blu?

Quanti tipi di fiori ci sono?

Pattern Recognition

⇨ Il processo che ci porta a rispondere a queste domande si chiama Pattern Recognition

Riconoscere che si tratta di una mela



Identificare l'oggetto più importante nella foto (l'Arena) ed associarlo alla città di Verona

Trovare nell'immagine tutti gli oggetti di tipo “vespa”, ed identificare se ce n'è una di colore blu

Riconoscere i fiori e distinguerli in due diverse tipologie (anche non sapendo che fiori sono)

Pattern Recognition

Più in generale:

Prendere in ingresso un insieme di dati (un'immagine, un suono, un odore)

Pattern: il dato che viene analizzato, l'entità di interesse

Effettuare un'analisi di tali dati per rispondere ad una domanda tipicamente legata al concetto di categoria o classe (che tipo di oggetto è? Quante categorie di oggetti ci sono? E' presente un dato di una certa categoria?)

Pattern Recognition

- Una definizione storica: “il processo che prende in input dati grezzi (raw) ed effettua un’azione sulla base della categoria dei dati” [Duda et al., 2001]

- E' un problema che l'uomo risolve facilmente (tramite processi complicati non ancora completamente chiari)

Pattern Recognition

La prospettiva informatica: realizzare sistemi AUTOMATICI di Pattern Recognition

- Sistemi che siano in grado di risolvere problemi di Pattern Recognition senza l'intervento dell'uomo

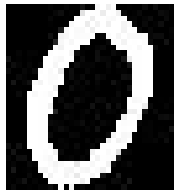
Il problema viene studiato da molti anni, anche se è tipicamente molto difficile!

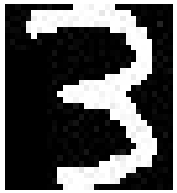
- Esempio: riconoscere caratteri scritti a mano

- Facile per l'uomo, difficile per il calcolatore

Perché è difficile per un calcolatore?

Quello che vede l'uomo:





Quello che vede il calcolatore:

```
0.0 0.0 0.0 0.0 0.0 0.2 0.6 0.3 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.4 0.6 0.6 0.3 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.3 0.6 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.1 0.6 0.6 0.6 0.6 0.5 0.1 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.4 0.6 0.6 0.6 0.5 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.1 0.5 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.2 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.1 0.6 0.6 0.6 0.1 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.3 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.2 0.6 0.6 0.6 0.1 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.5 0.6 0.6 0.2 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.5 0.6 0.6 0.3 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.1 0.2 0.4 0.6 0.6 0.4 0.2 0.1 0.1 0.0 0.0 0.0
0.0 0.0 0.0 0.3 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.1 0.0 0.0
0.0 0.0 0.0 0.2 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.1 0.0 0.0
0.0 0.0 0.0 0.0 0.1 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.2 0.0 0.0 0.0
```

```
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.3 0.4 0.5 0.4 0.1 0.0
0.0 0.4 0.4 0.4 0.4 0.4 0.5 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.1
0.1 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.1
0.0 0.3 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.4 0.4 0.6 0.6 0.6 0.1
0.0 0.0 0.0 0.2 0.2 0.1 0.1 0.1 0.0 0.0 0.0 0.2 0.6 0.6 0.6 0.1
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.4 0.6 0.6 0.6 0.1
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.4 0.6 0.6 0.4 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.6 0.6 0.6 0.3 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.3 0.6 0.6 0.5 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.3 0.6 0.6 0.4 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.2 0.6 0.6 0.5 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.3 0.6 0.6 0.6 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.3 0.6 0.6 0.5 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.5 0.6 0.6 0.4 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.4 0.6 0.6 0.4 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.5 0.6 0.4 0.0 0.0
```

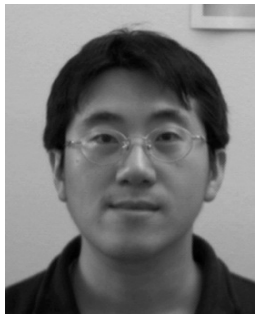
Perché è difficile per un calcolatore?

C'è molta variabilità:

- oggetti della stessa classe possono essere diversi

- oggetti di classi diverse possono essere molto simili

Altri esempi classici



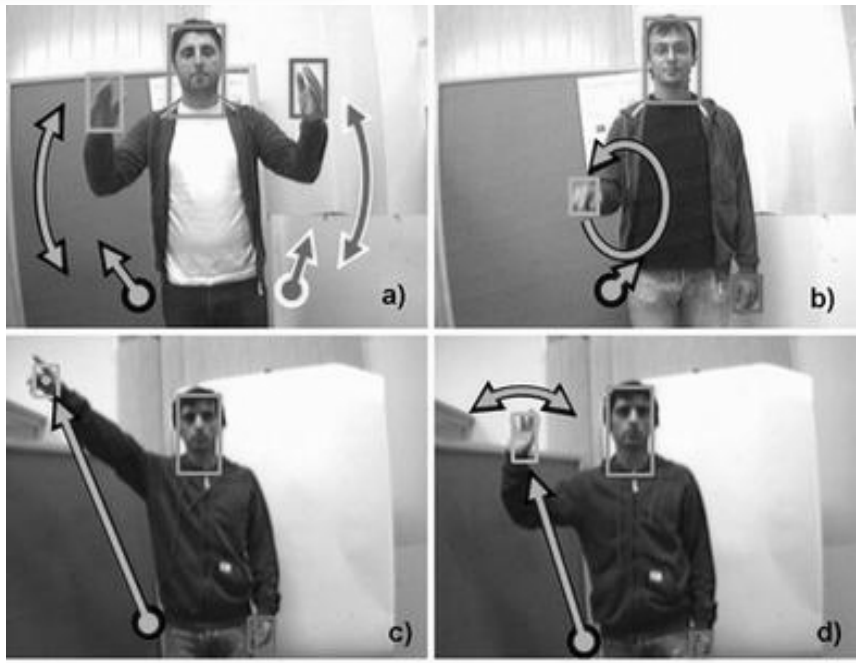
distinguere diverse persone
sulla base del volto

pattern: la parte dell'immagine che contiene la faccia

Altri esempi classici

Riconoscimento del parlato

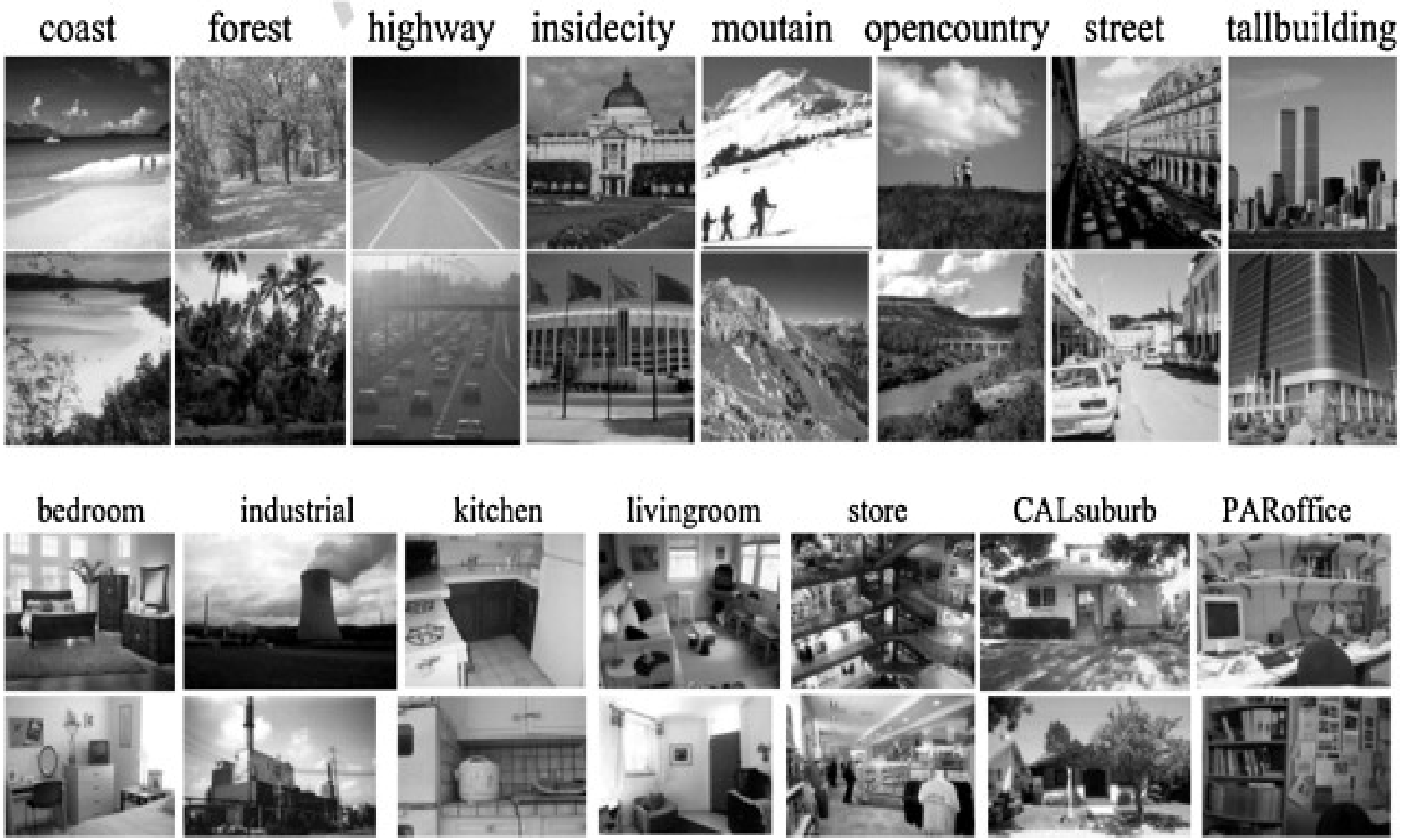
Riconoscimento di impronte digitali



Riconoscimento di gesti

Altri esempi classici

Riconoscimento di Scene a partire da immagini



Altri esempi classici

Classificazione di video: capire in ogni immagine se ci sono oggetti che si muovono (classificare i pixel)

Originale

Classificazione:
bianco = movimento

Altri esempi classici

Videosorveglianza: classificazione di oggetti in movimento

Nel laboratorio VIPS:

- classificazione e clustering di:
  - Oggetti
  - Immagini
  - Audio e video
  - Segnali sismici e naturali
  - ...
- classificazione di situazioni (videosorveglianza)
- Bioinformatica: genetica computazionale, analisi di spettri NMR, analisi di dati da esperimenti microarray, protein remote homology detection
- Analisi di immagini biomedicali: (es. MRI)

Il problema principale

Capire e modellare i diversi pattern di un problema

(tipicamente in termini di classi / gruppi / categorie)

Il paradigma principale

Il problema è risolto usando il cosiddetto paradigma

“apprendimento da esempi”

La conoscenza si deriva da un insieme di esempi campionati dal problema

(il training set – insieme di addestramento)

L’obiettivo principale

GENERALIZZAZIONE: capacità di generalizzare anche a oggetti sconosciuti (non presenti nel training set)

Il vero problema

- Derivare un modello per il problema a partire da esempi
- Tipicamente il problema è risolto con una procedura di ottimizzazione

$$Model \leftarrow \max E\left(T, P, \Theta\right)$$

T = training set
P = informazioni a priori
Θ= parametri

Il vero problema

Problemi da risolvere

- definire la funzione E
  - compromesso tra la capacità di spiegare il training set e la complessità
- ottimizzare E (tipicamente una funzione difficile da ottimizzare)
  - discesa lungo il gradiente
  - Expectation – Maximization
  - Simulated annealing
  - Tabu Search (Reactive Tabu search)
  - Algoritmi genetici

Altri problemi

- Aspetti teorici
  - convergenze del learning
  - comportamenti asintotici
  - ottimalità delle soluzioni
  - ...
- Aspetti pratici
  - accuratezza
  - requisiti computazionali (tempo e spazio)
  - flessibilità
  - usabilità
  - ...

Tipologie di problemi in PR

Ci sono diversi problemi che possono essere risolti con metodologie di pattern recognition

I tre principali sono:

- ⇨ Classificazione
- ⇨ Detection
- ⇨ Clustering

Che cos'è?

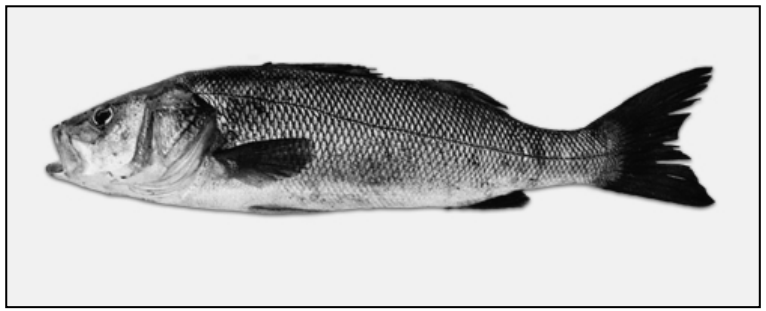
C'è una vespa blu?

Come si raggruppano i fiori di questo campo?

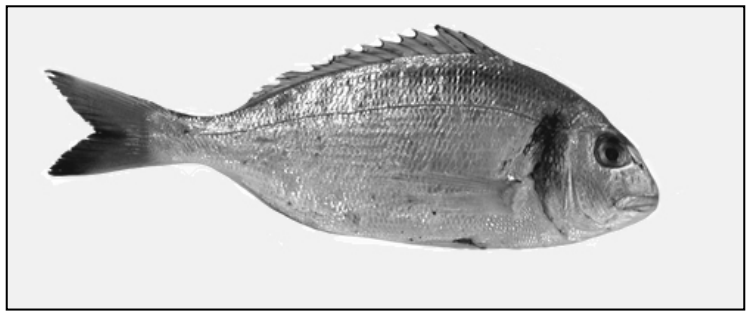
Più nel dettaglio...

[Duda Hart Stork, Pattern Classification, Second Edition, Wiley 2001]

Problema: modellare pesci



spigola

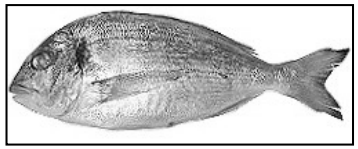


orata

Classificazione

Trovare due modelli M1 e M2, uno per l’orata e uno per la spigola

Che cos'è?

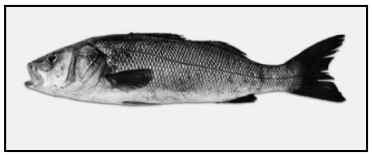


spigola o orata?

M1, M2

Detection

Trovare un modello M per tutte le spigole

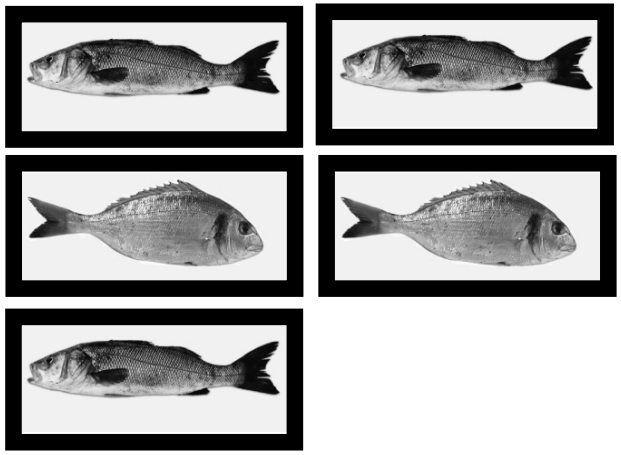


Questo pesce è una spigola?

M

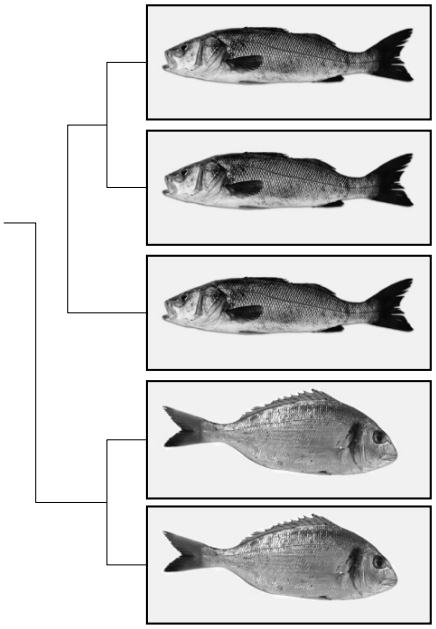
C'è una vespa blu?

Clustering



1. identificare pesci simili (identificare tutti i gruppi “naturali” e creare i modelli)

2. identificare le relazione tra i pesci (clustering gerarchico)



Come si raggruppano i fiori di questo campo?

In ogni caso.....

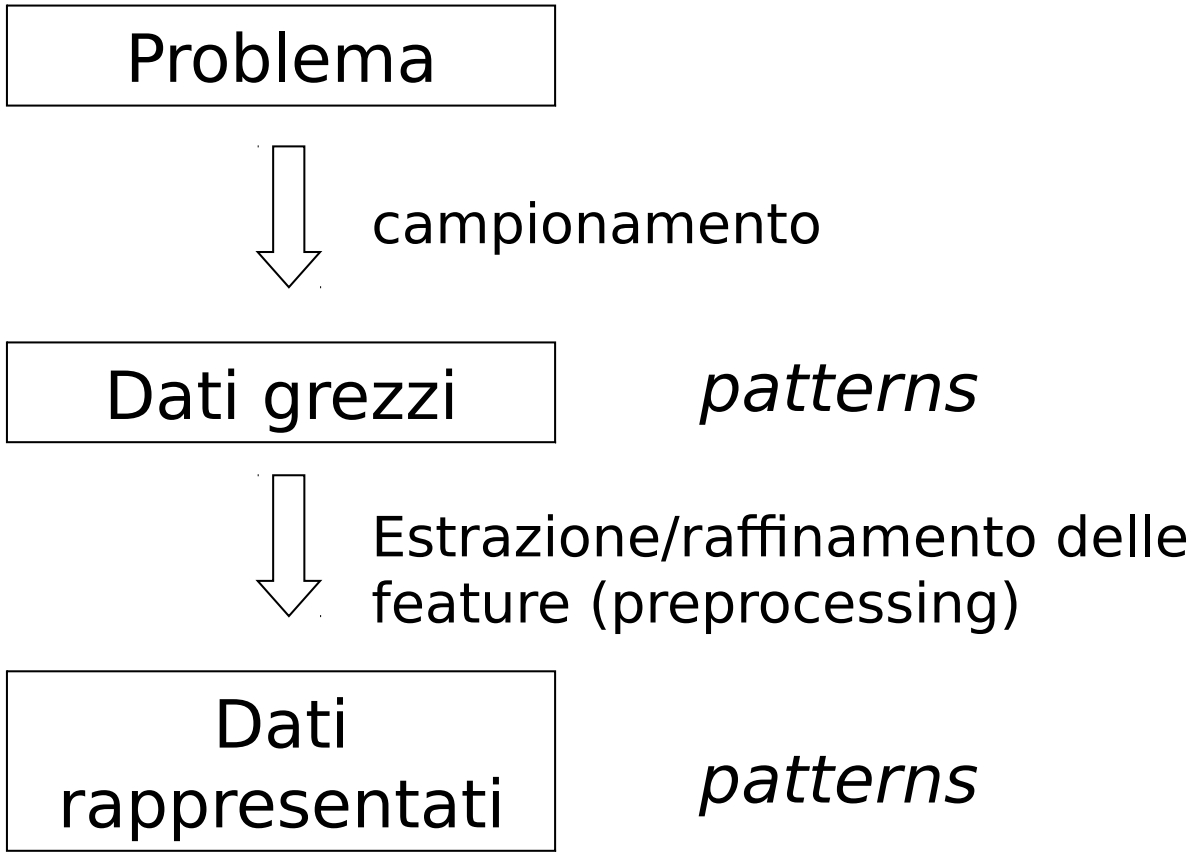
occorre costruire un modello a partire dai dati!

Aspetti principali

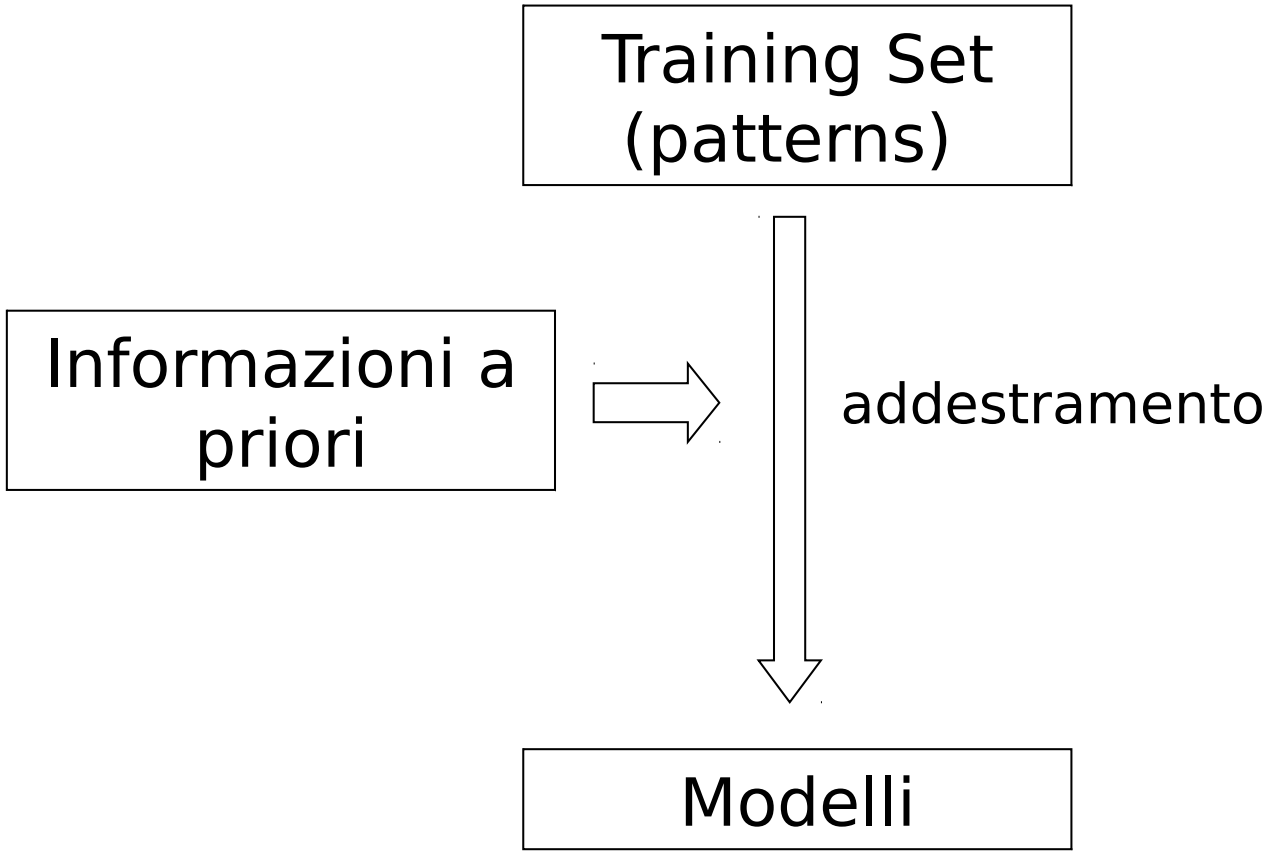
La realizzazione di un sistema di Pattern Recognition implica la soluzione dei seguenti problemi:

- Rappresentazione: come rappresentare in modo digitale gli oggetti del problema
- Costruzione del modello: come costruire un modello a partire da un insieme di dati (training set)
- Testing: come utilizzare il modello per “spiegare qualcosa” dei dati
  - tipicamente per fare classificazione, clustering o detection

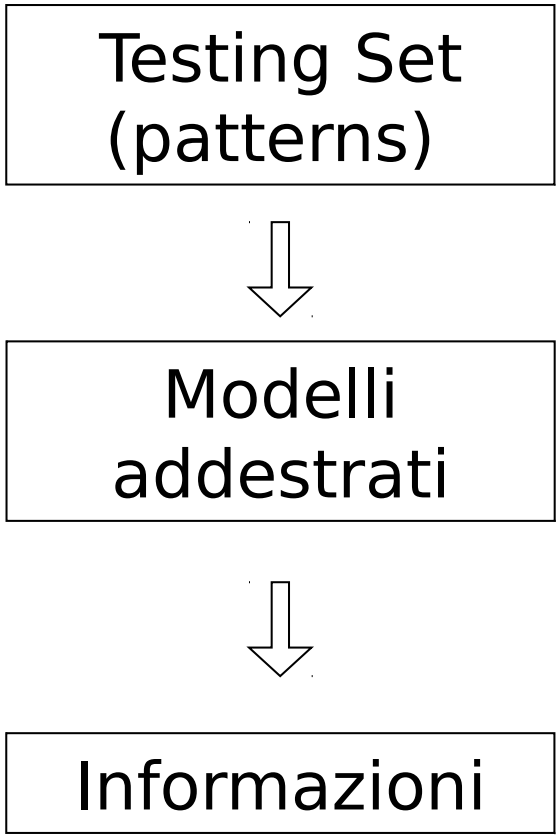
Rappresentazione



Costruzione del modello



Testing



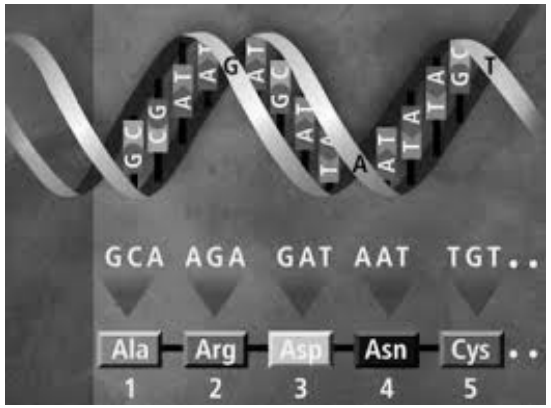
Più nel dettaglio....

Rappresentazione

- Obiettivo: trovare una rappresentazione digitale per gli oggetti del problema in esame
- Tipicamente si effettuano una serie di misure sull'oggetto, utilizzando dei sensori
- L'insieme di queste misure è detto PATTERN, ogni singola misura è detta FEATURE

- Esempio

L'immagine è il pattern, ogni pixel è una feature (viene misurato il colore)



La sequenza di DNA è il pattern, ogni nucleotide è una feature (viene misurato il tipo – A,T,C,G)

Rappresentazione

- Le misure sono spesso “grezze”
  - Immagine: migliaia di pixels!
  - Sequenze di DNA: migliaia di basi!

- Pre-processing dei dati: “migliorare” la rappresentazione:
    - ridurre la dimensione del pattern (per visualizzare, per ridurre il carico computazionale, ...)
    - mettere in evidenza particolari strutture o migliorare le capacità discriminative dello spazio
  - Estrazione di feature: trasformazione dello spazio originale
  - Selezione di features: selezionare le feature migliori

Rappresentazione: esempio

Il problema

campionamento

dati grezzi (un'immagine)

estrazione di due features

(altezza,lunghezza)

$\mathbf{x}_1 = [5, 10]$

dati pre-processati

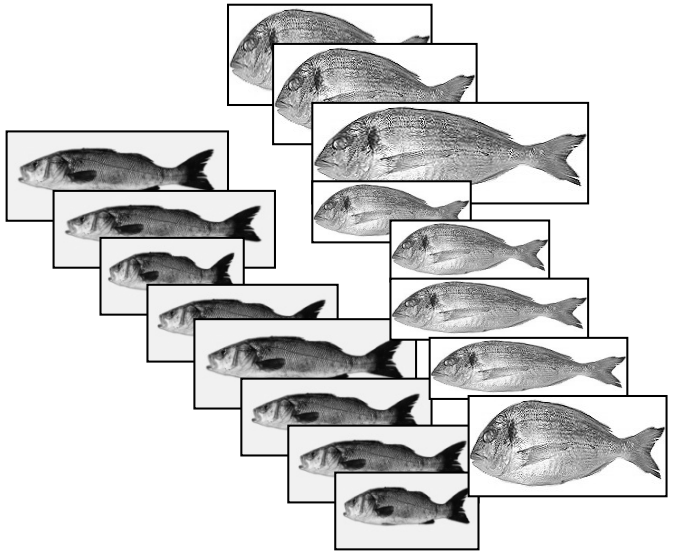
Costruzione del modello

- Problema da risolvere: costruire un modello in grado di spiegare i dati del training set
  - training/learning/addestramento
- “Il modello deve spiegare il problema”: capacità di generalizzare anche a pattern mai visti (capacità di generalizzazione)

- La costruzione del modello (paradigma di apprendimento da esempi) si basa su:
  - Le misure (il training set)
  - La conoscenza a priori (le etichette del training set, o altro)

Costruzione del modello

- Il training set deve essere adeguatamente:
  - largo (molti pattern)
  - completo (tutte le categorie devono essere ragionevolmente rappresentate)
  - variabile (deve tenere in considerazione la variabilità dei pattern nelle categorie)



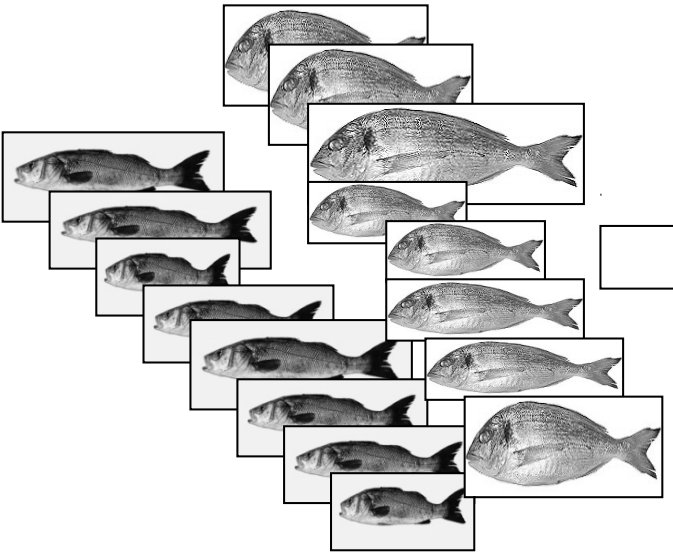
Costruzione del modello

- Scelte da effettuare
  - tipo di modello
  - parametri del modello
  - dimensione del modello
  - metodo di addestramento (funzione da ottimizzare, metodo di ottimizzazione)
  - metodo di validazione (come capire se il modello scelto effettivamente rappresenta il fenomeno in questione)

- procedura diversa a seconda che si parli di classificazione, clustering o detection
  - Diversa tipologia di informazione a priori disponibile (cosa conosco degli esempi del training set)

Esempio: classificazione

Info a priori: di tutti gli esempi del training set conosco la classe (patter recognition supervisionata)



Insieme di addestramento

$\mathbf{x}_1, y_1$

$\mathbf{x}_2, y_2$

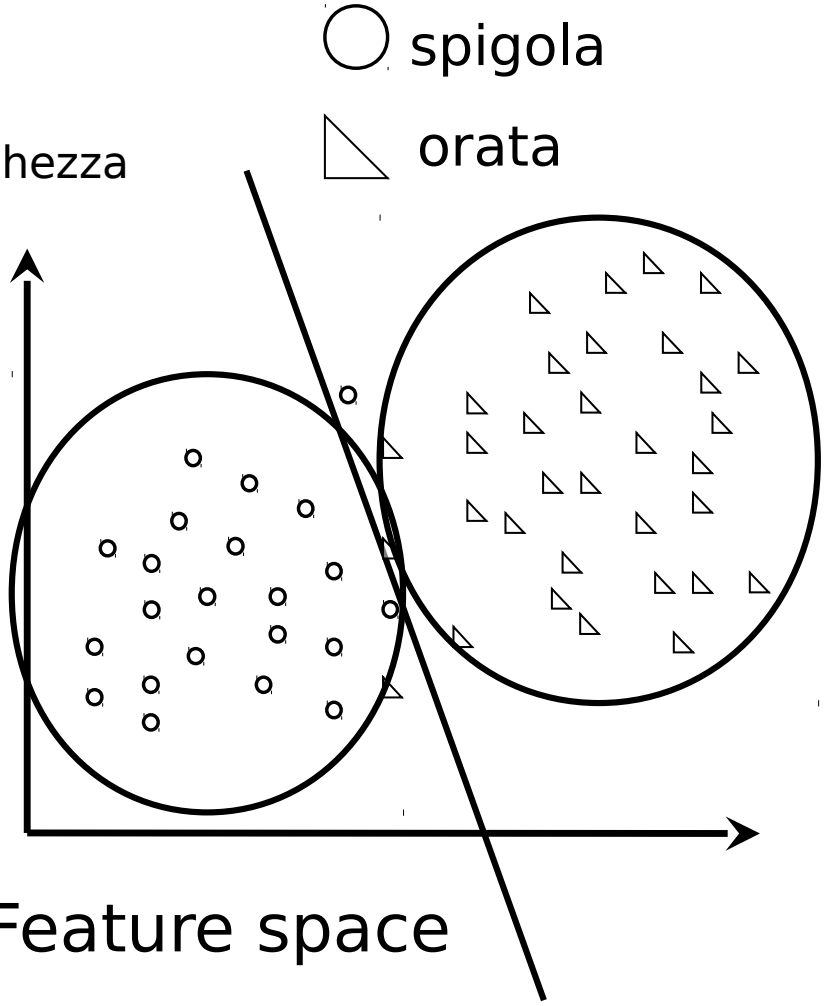
...

$\mathbf{x}_N, y_N$

Rappresentazione

x_i patterns

y_i etichette



Feature space

Addestramento: modellare (separare) le due classi

Esempio: classificazione/testing

oggetto

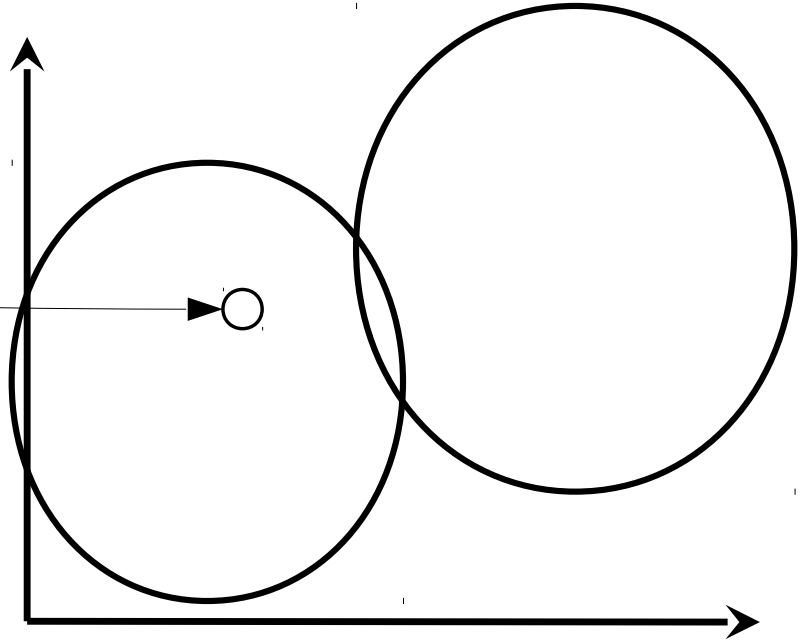
categoria: spigola

rappresentazione

$\mathbf{x}_1 = [3, 12]$

dati pre-processati

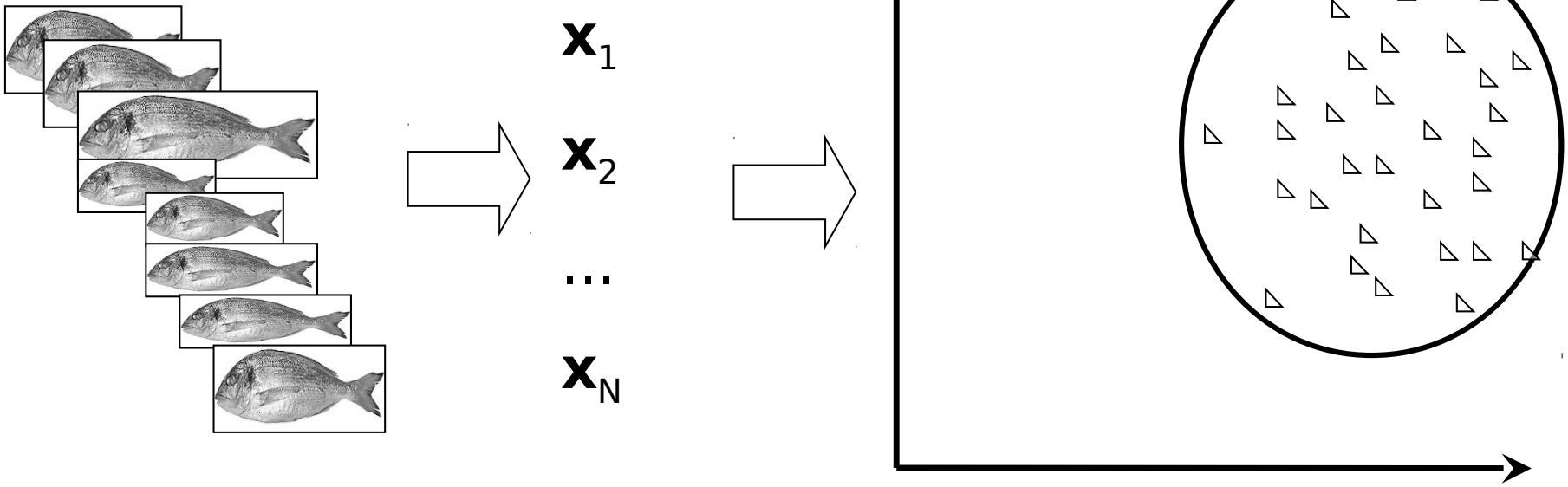
testing



Modelli

Esempio: detection

Info a priori: tutti gli esempi del training set sono nella stessa classe (patter recognition supervisionata)



Insieme di addestramento

Addestramento: modellare la classe

Esempio: detection/testing

oggetto

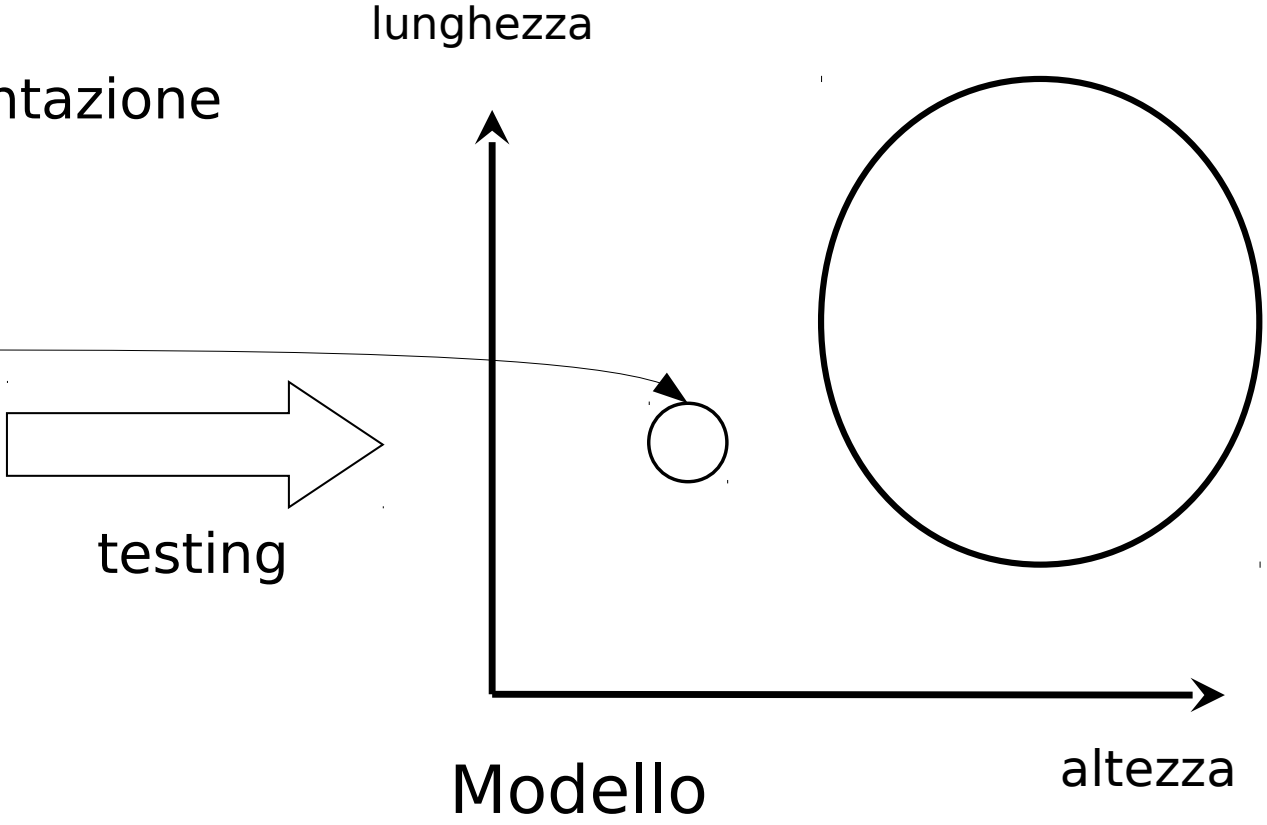
rappresentazione

$\mathbf{x}_1 = [3,12]$

dati pre-processati

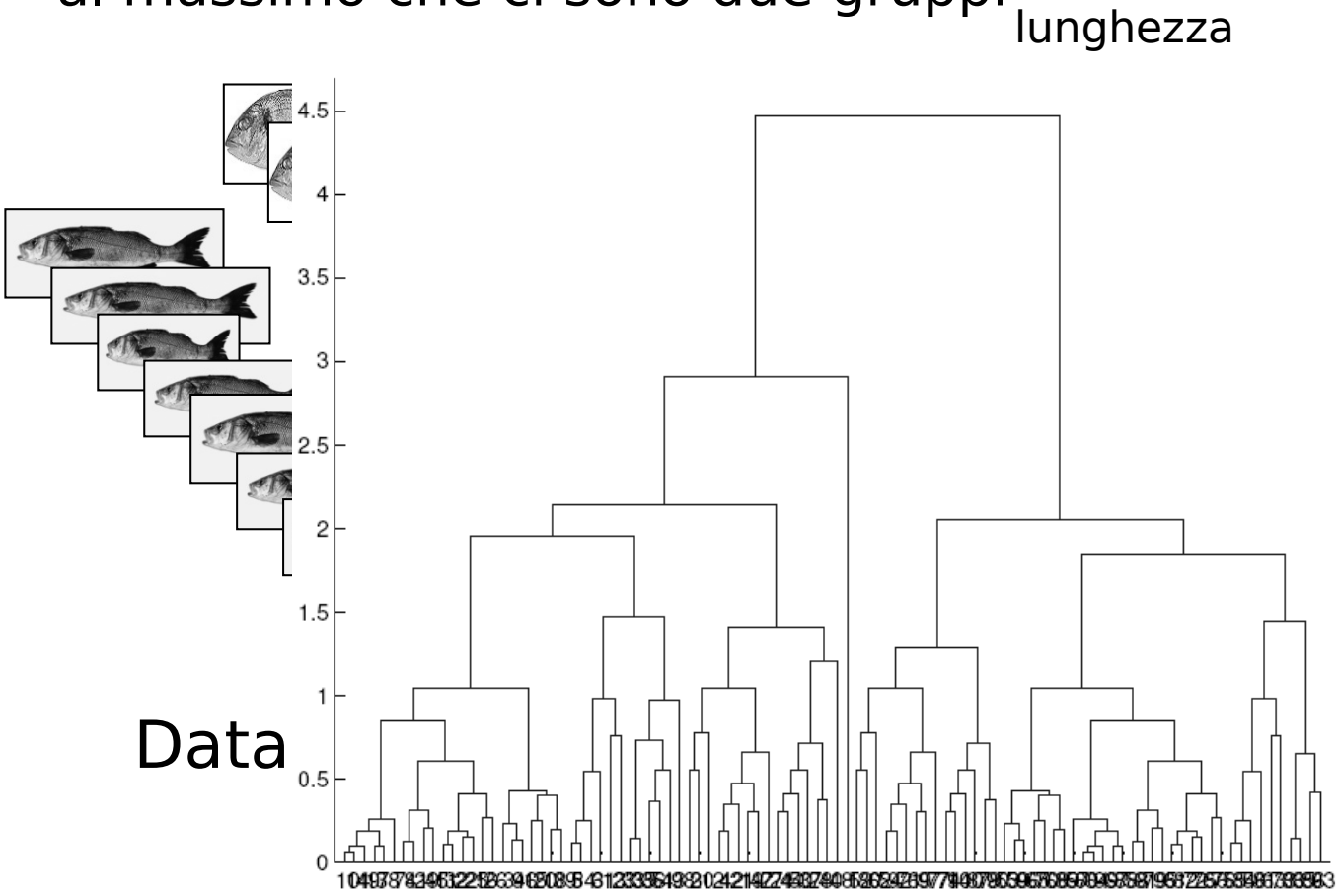
testing

questo pesce non appartiene al gruppo

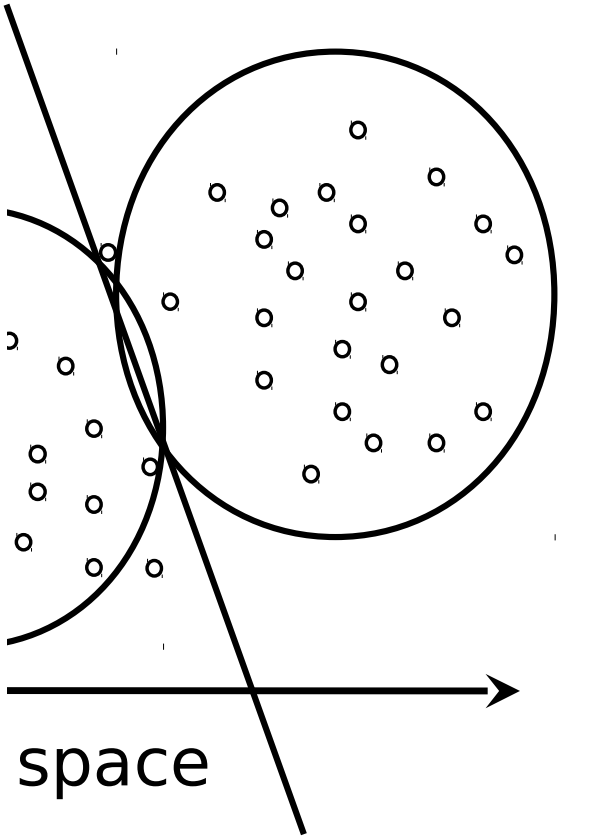


Esempio: clustering

Info a priori: nessuna
(pattern recognition non supervisionata)
al massimo che ci sono due gruppi



Data



Goal1: scoprire i gruppi naturali

Goal2: descrivere le relazioni tra i patterns

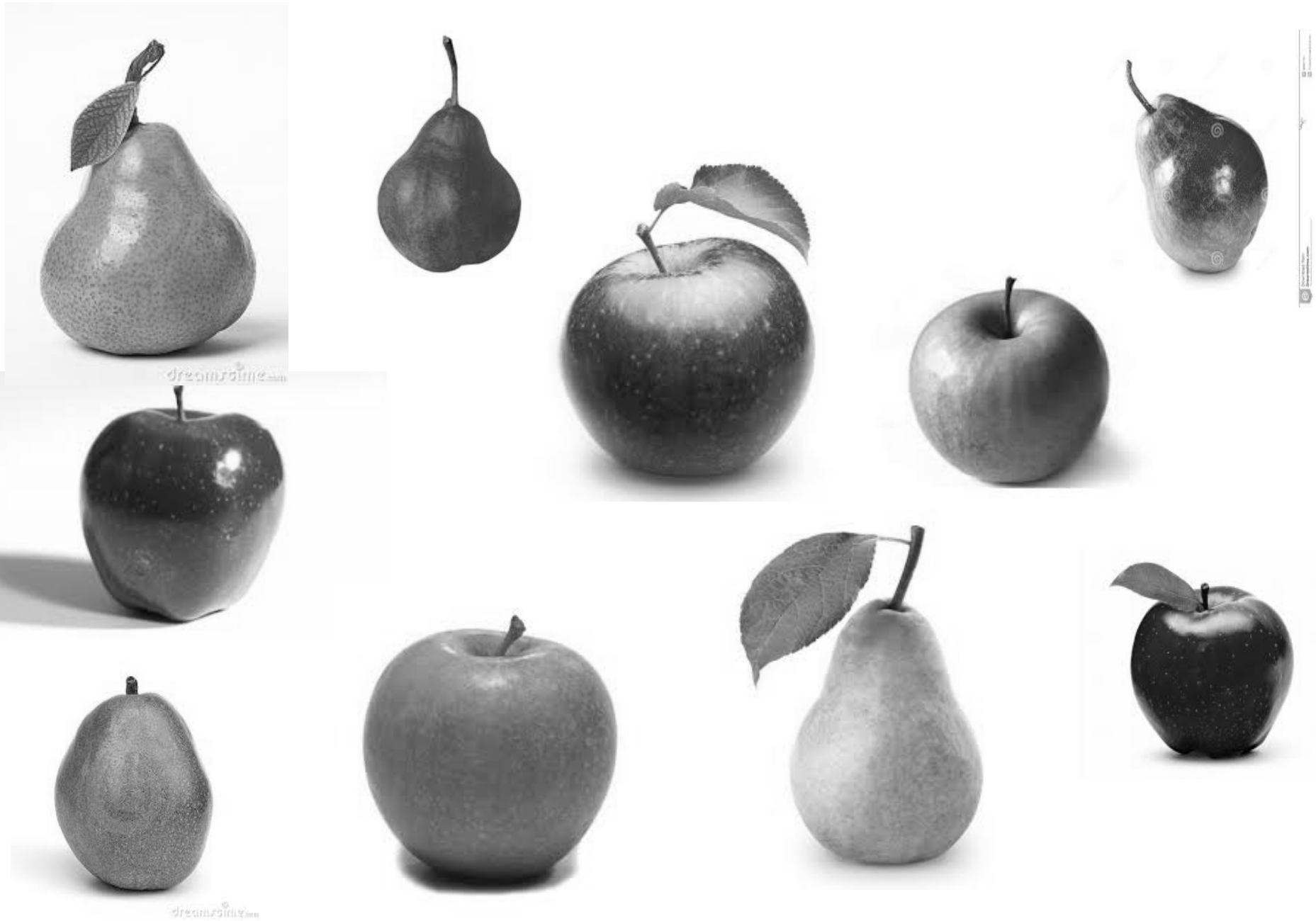
Un commento sul clustering

- Il clustering è un problema più difficile della classificazione
  - Il processo è non supervisionato: non è possibile misurare la correttezza del risultato! (differentemente dalla classificazione)

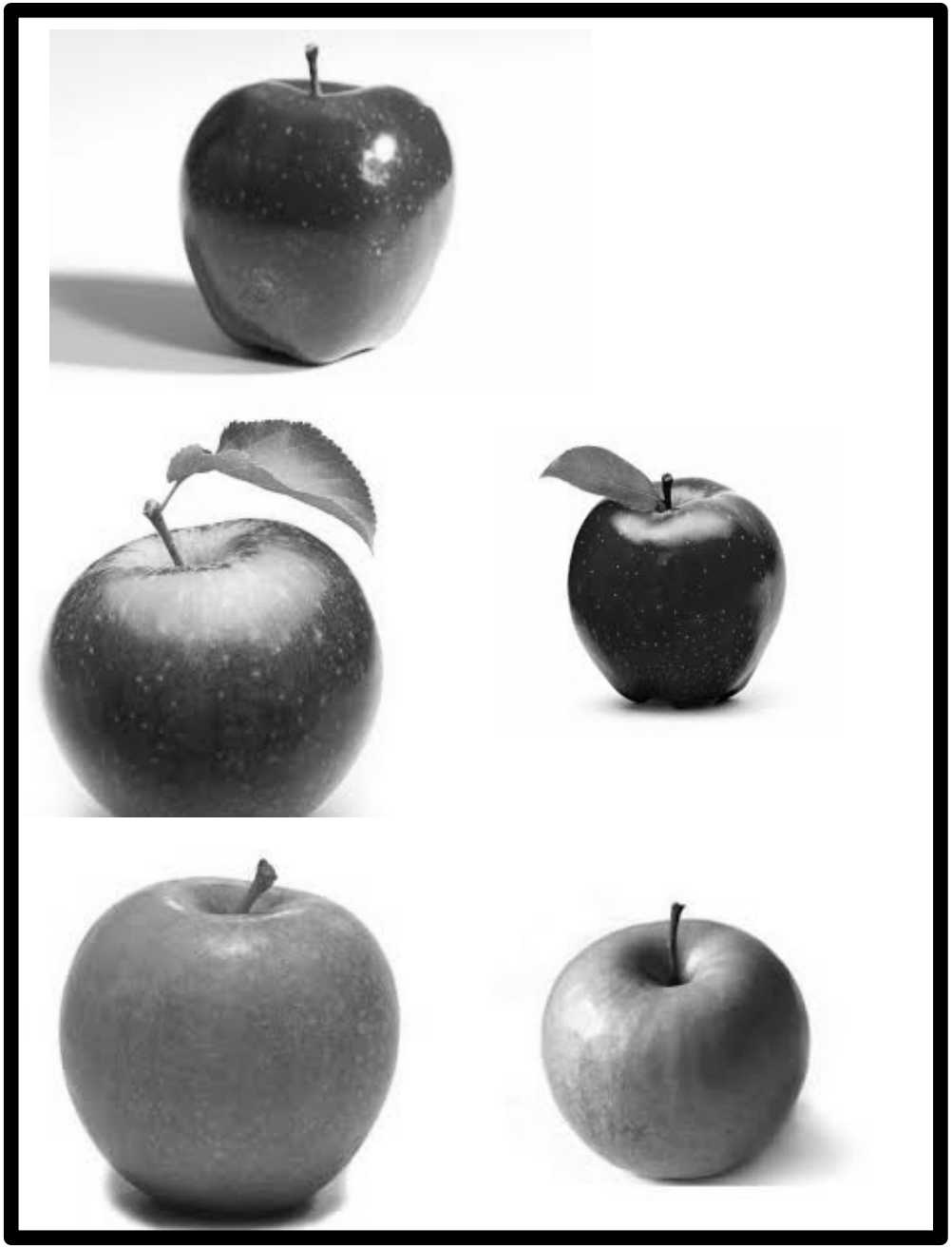
- Il clustering rappresenta l'organizzazione di un insieme di patterns (entità) in gruppi (clusters) sulla base della similarità

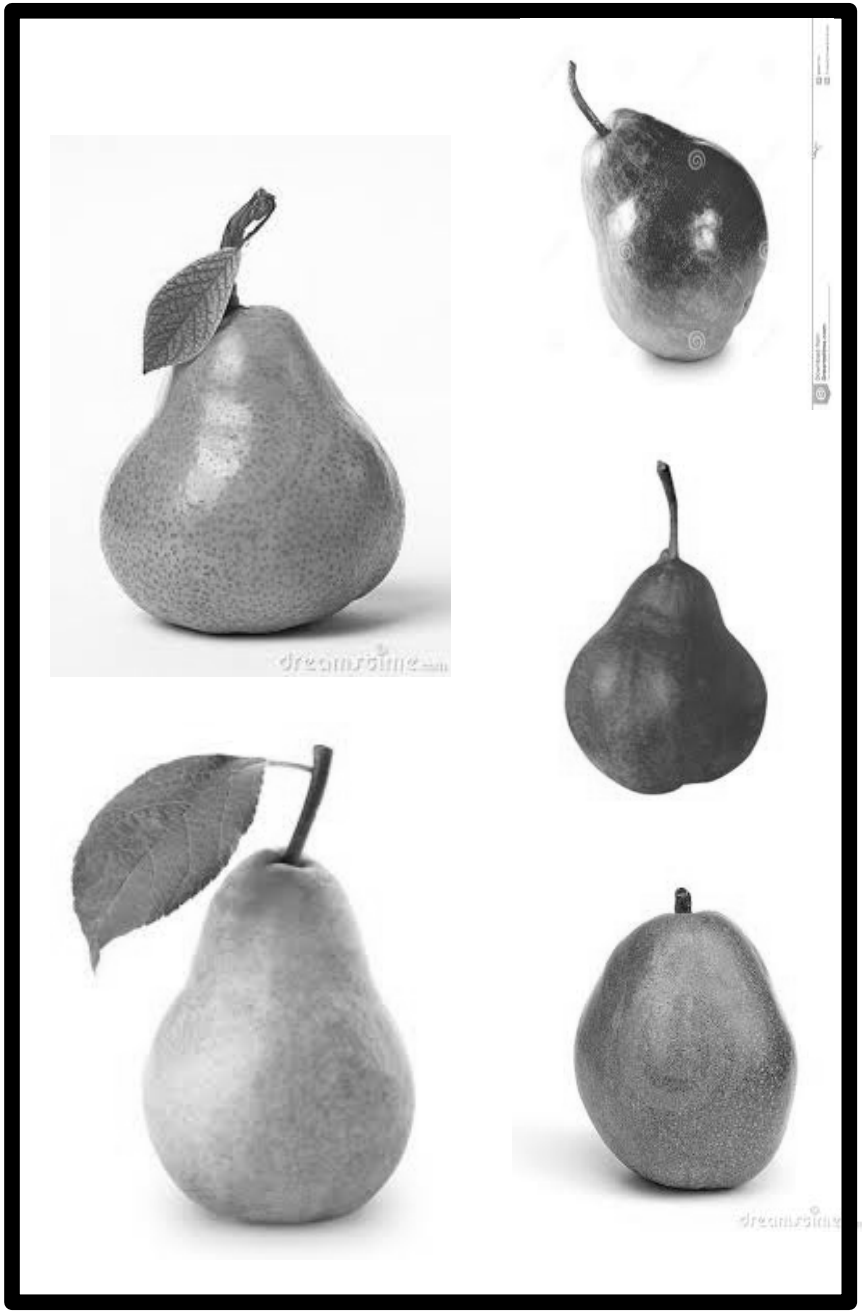
- Qual'è la similarità più appropriata?
  - Cambiare la similarità cambia il risultato
- Cosa deve rappresentare un "buon gruppo"?
  - Il concetto di gruppo è definito in modo vago e assolutamente soggettivo

Esempio: Oggetti da clusterizzare

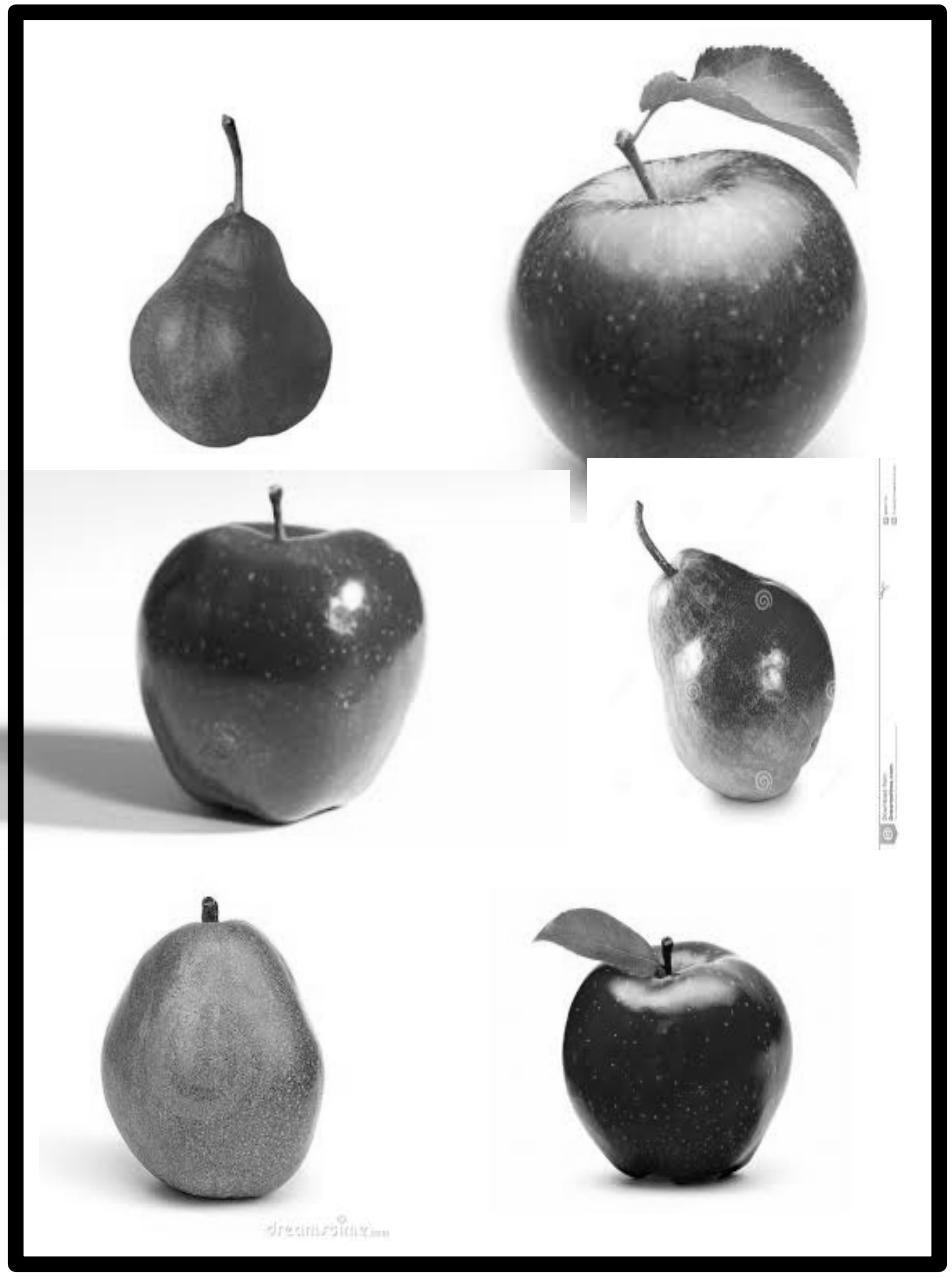


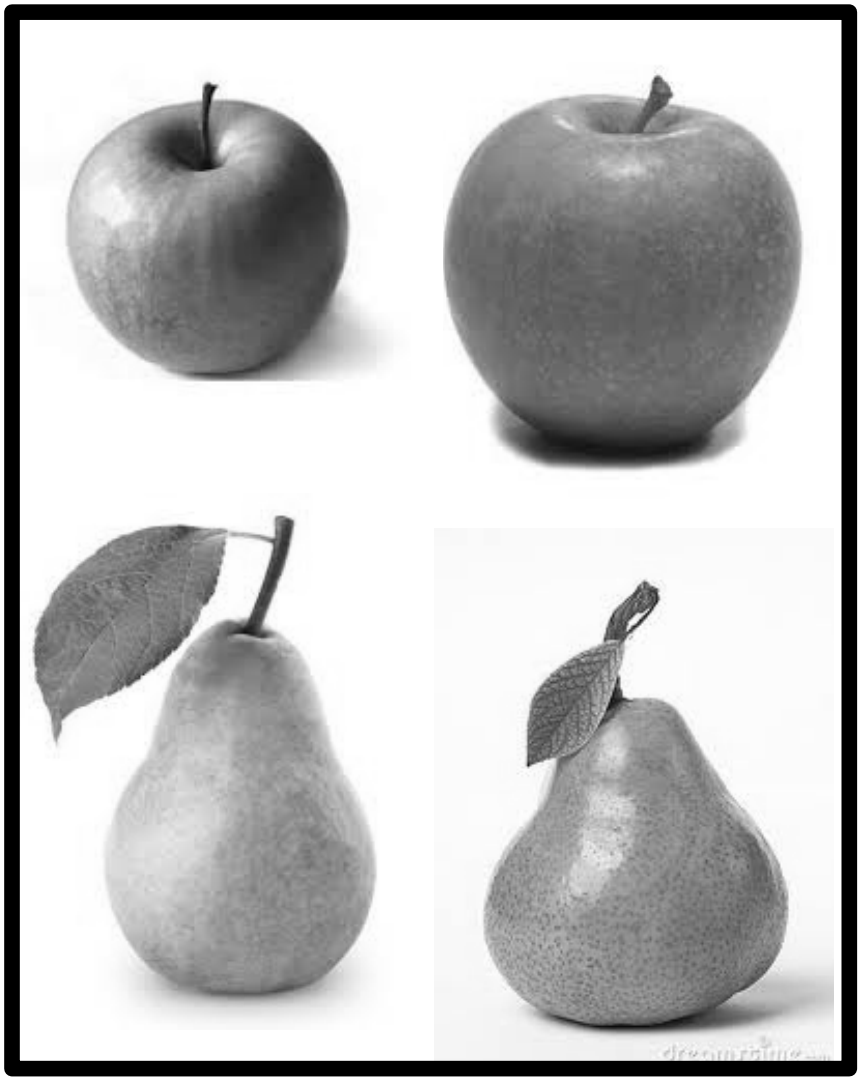
Ci sono 2 gruppi: mele e pere





Altra possibilità: frutta rossa e frutta verde





Quindi

- Il concetto di cluster è vago:
  - Dipendentemente dalle misure di similarità utilizzate cambia il risultato
- La scelta della misura di similarità è cruciale.
  - Dovrebbe essere fatta in modo da inglobare la maggior quantità possibile di informazione a priori.
- Il risultato può cambiare anche a seconda della metodologia utilizzata per fare clustering (il concetto sarà più chiaro in seguito)

Sommario

- La costruzione del modello può avvenire in modo supervisionato (classificazione e detection) o non supervisionato (clustering)
  - Supervisionato (*Supervised learning*): per ogni oggetto del training set si conosce l'esatta categoria
  - Non supervisionato (*Unsupervised learning*): non si conosce nulla

- *Reinforcement learning* (per classificazione)
  - a metá strada tra le due: non viene fornita alcuna informazione sulla categoria esatta, viene dato un giudizio sulla correttezza della classificazione

Interpretazione dei risultati

- L’obiettivo finale è quella di estrarre / recuperare conoscenza
  - ottenere intuizioni dal data set

- Il fuoco deve essere sulla ”interpretabilità” dei prodotti
  - interpretabilità dei metodi
    - mette a proprio agio l’utente
  - interpretabilità delle soluzioni
    - permette di capire gli errori

Pattern Recognition e bioinformatica: perché?

PR e bioinformatica: perché?

Ci sono molti buoni motivi per utilizzare tecniche di Pattern Recognition nella Bioinformatica...

⇨ LA MOTIVAZIONE PRINCIPALE: la caratterizzazione di una popolazione in termini di gruppi/classi/categorie può essere utilizzata per inferire alcune proprietà di oggetti sconosciuti guardando ad oggetti conosciuti nello stesso gruppo

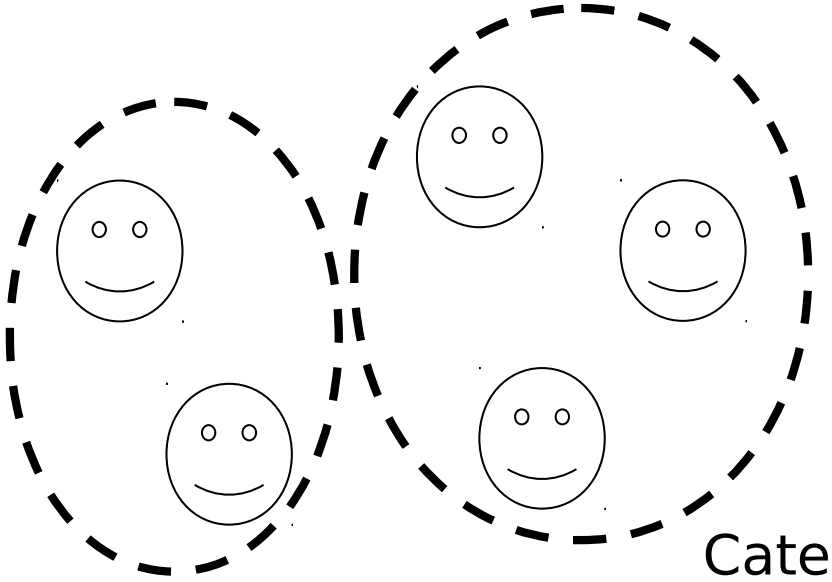
⇨ In altre parole: vengono inferite informazioni su entità sconosciute a partire da informazioni note su entità conosciute che siano “simili”

Esempio 1

Problema:

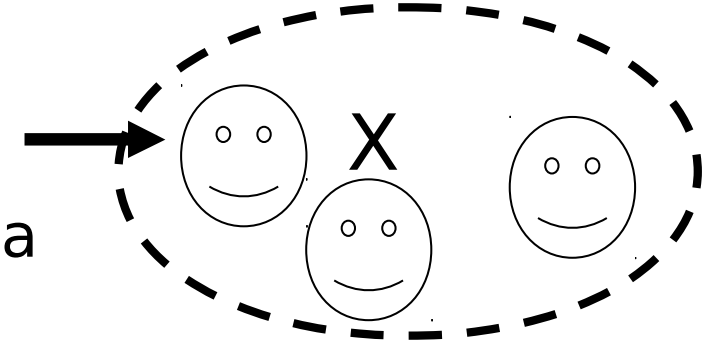
Da che continente proviene X?

Popolazione (dati – patterns)



Categorizzazione sulla base del colore

Questo è conosciuto: si chiama John e viene dall’Europa

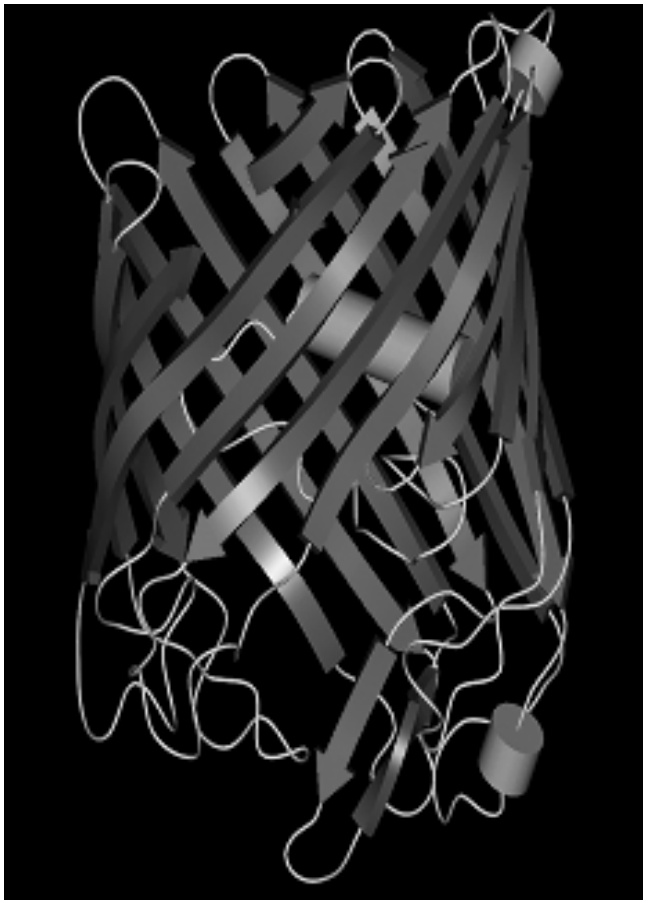


X e John sono simili, sono nello stesso gruppo/classe

Posso ipotizzare: anche X viene dall’Europa

Esempio 2

- Esempio:
  - ho una proteina B la cui funzione è sconosciuta
  - trovo una proteina A che ha una struttura/sequenza molto simile (misura di similarità, clustering)
  - Posso ipotizzare che la proteina B abbia una funzione simile.



PR e bioinformatica: perché?

Altre motivazioni:

1. In bioinformatica ci sono molti problemi di classificazione, clustering e detection
2. Possibilità di derivare modelli per i dati tramite esempi (paradigma di apprendimento da esempi)
3. Ci sono problemi di classificazione (onerosi in termini di tempo) che possono essere automatizzati
   - ⇨ apprendimento da esempi che possono essere giudicati da esperti / validati sperimentalmente

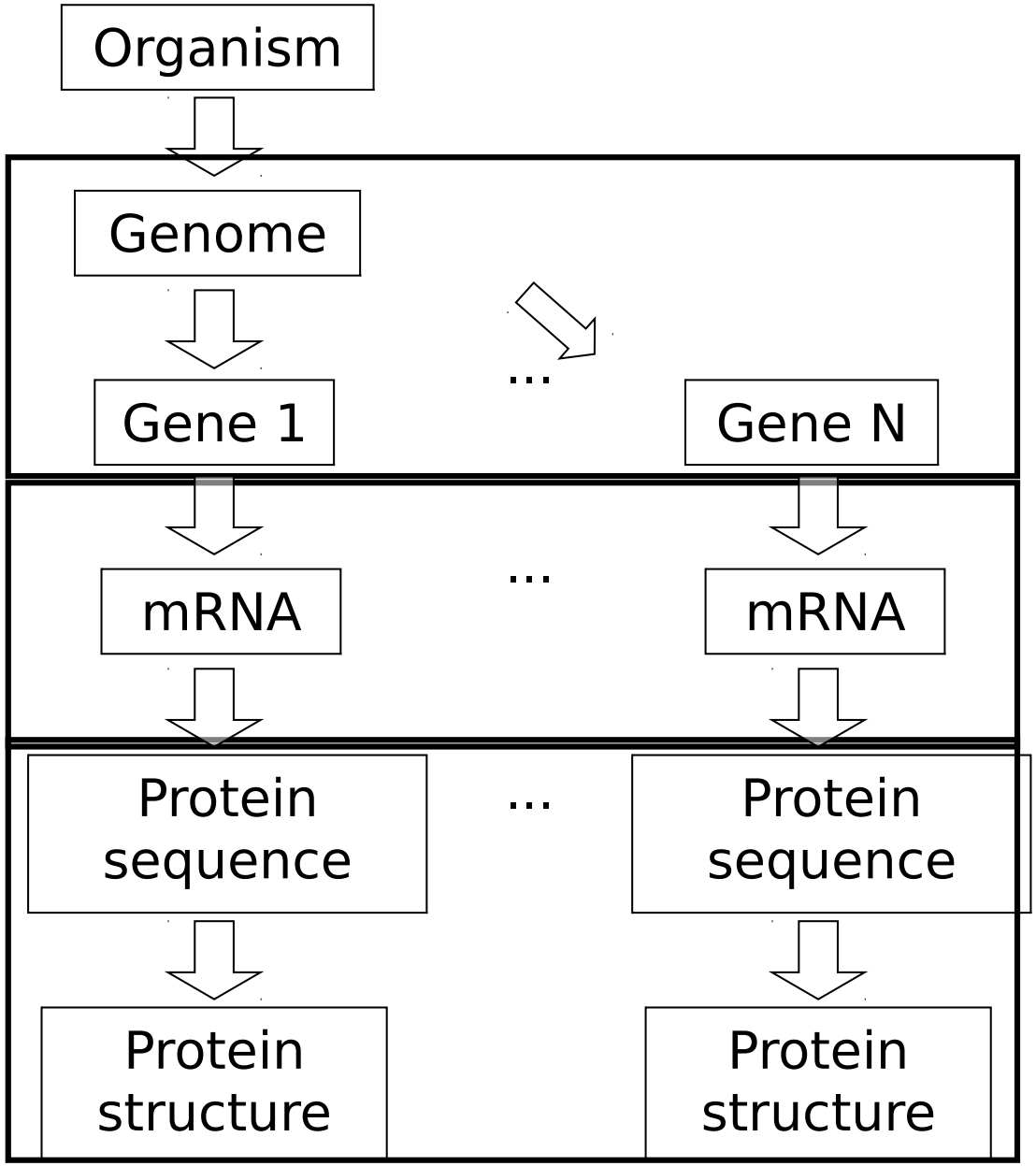
più in dettaglio...

Sommario

Genomica

Trascrittomica

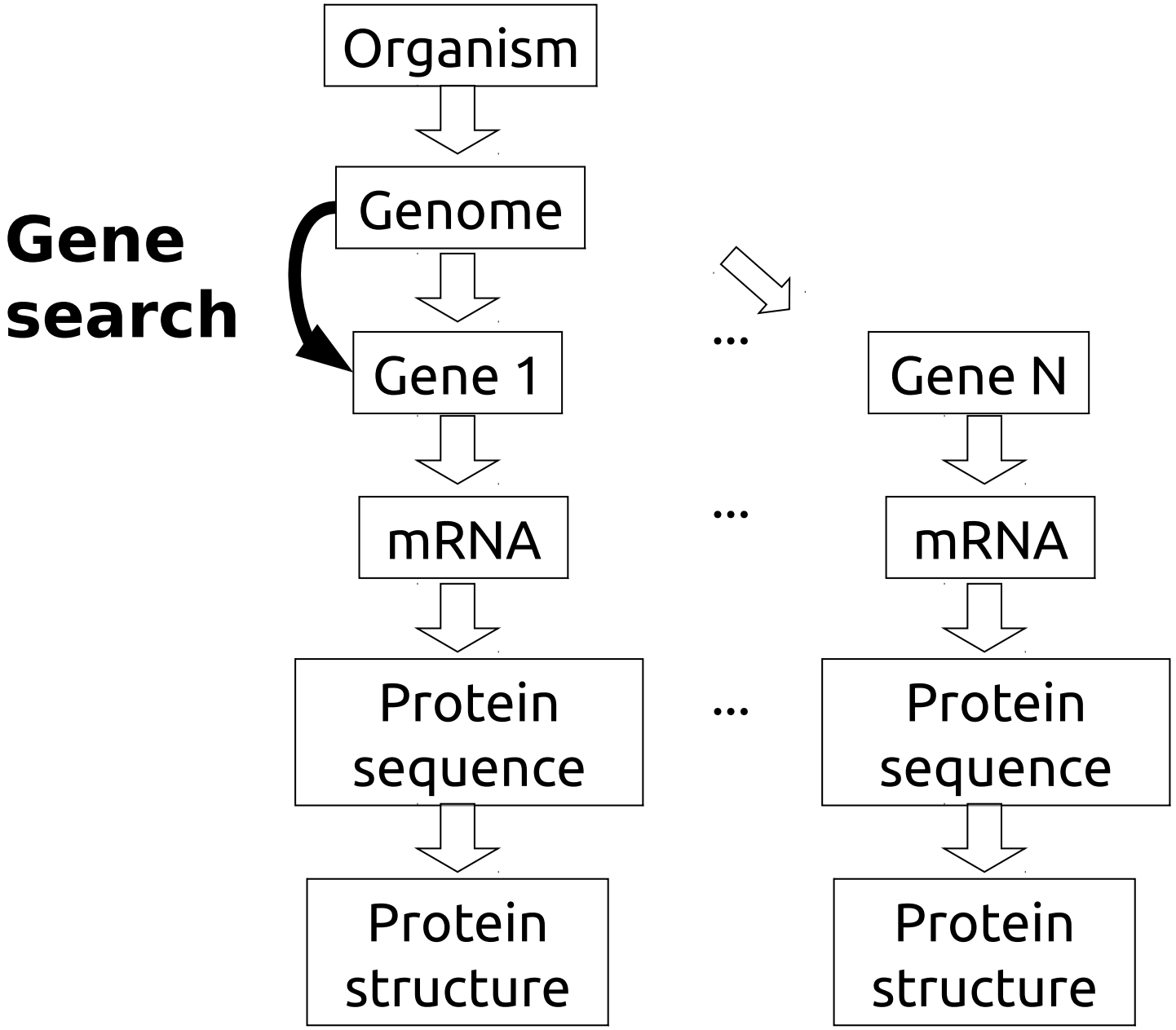
Proteomica



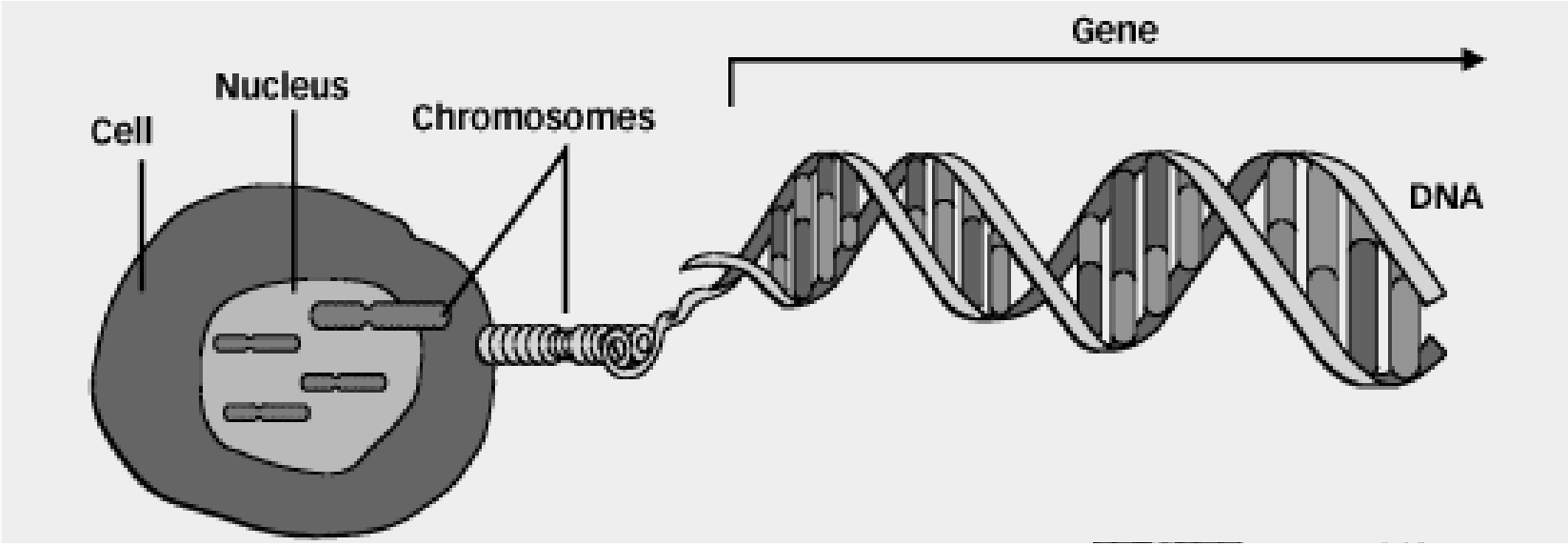
PR e bioinformatica: perché?

1. In bioinformatica ci sono molti problemi di classificazione, clustering e detection
2. Possibilità di derivare modelli per i dati tramite esempi (paradigma di apprendimento da esempi)
3. Ci sono problemi di classificazione (onerosi in termini di tempo) che possono essere automatizzati

PR e Bioinformatica



Gene search

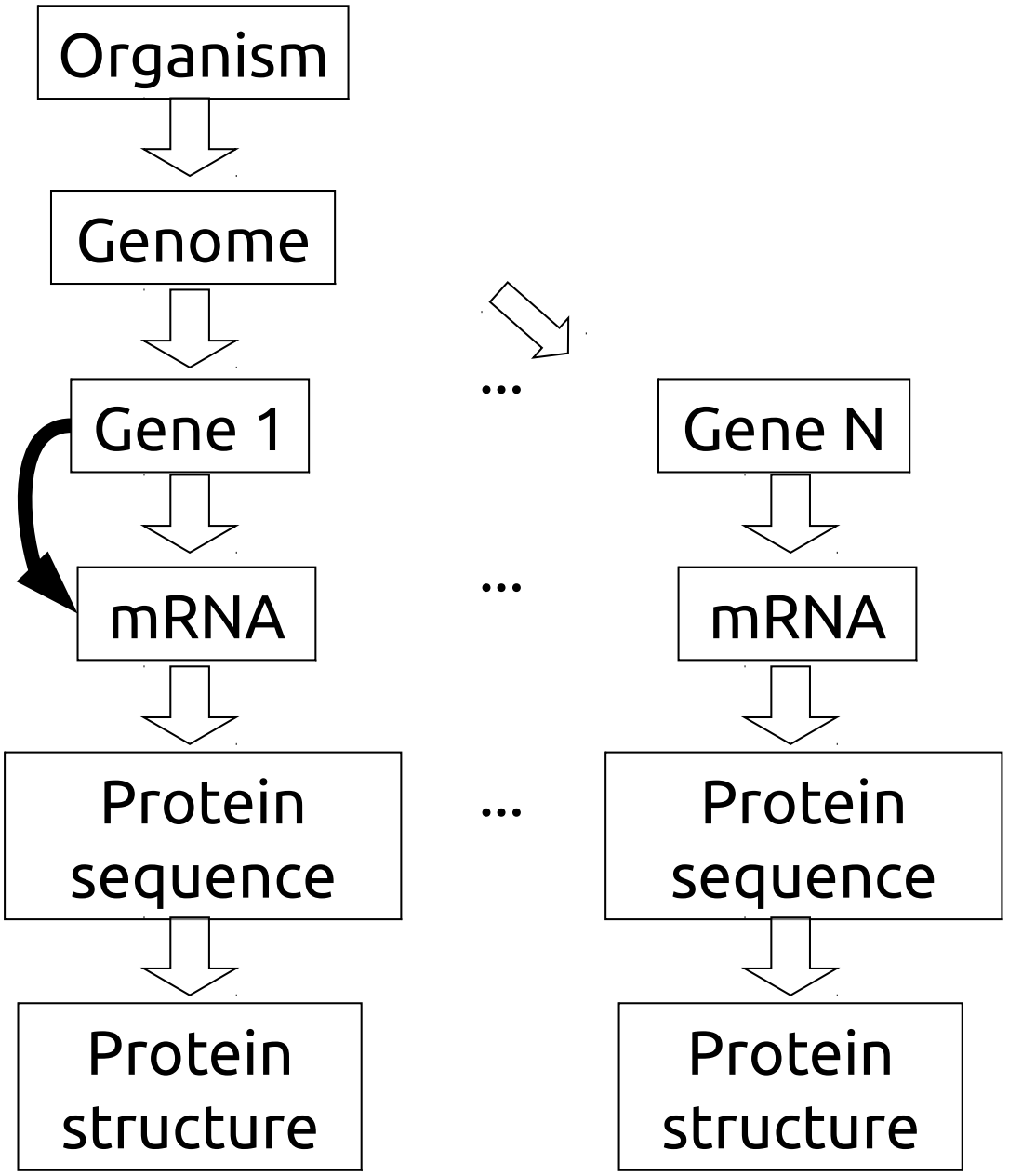


Non tutto il DNA del genoma è “geni”

PR: detection di geni

PR e Bioinformatica

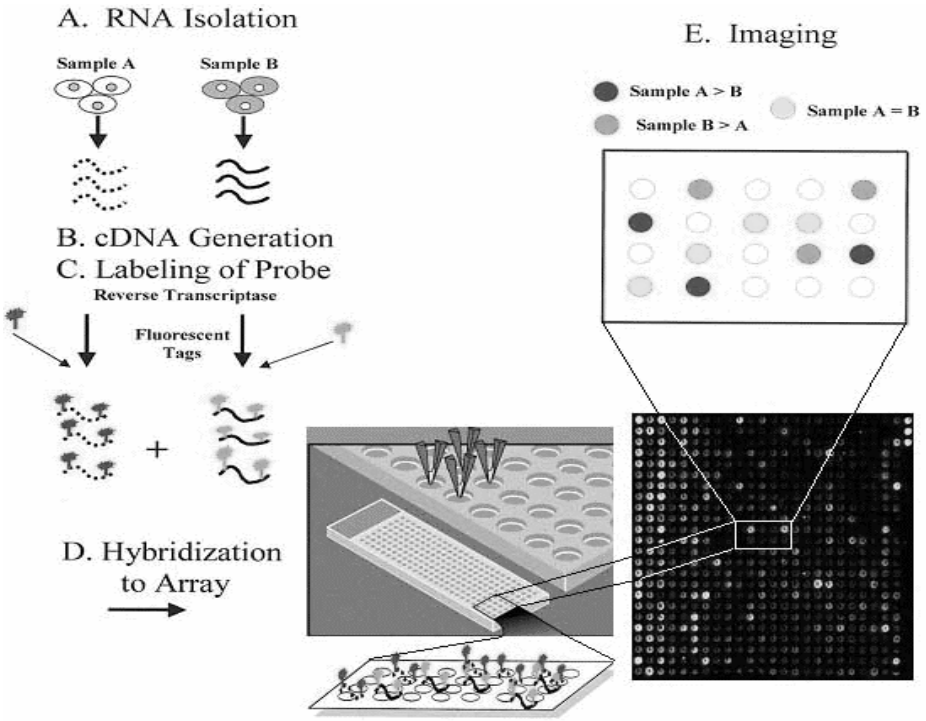
Analisi dell'espressione e della regolazione genica (microarrays)



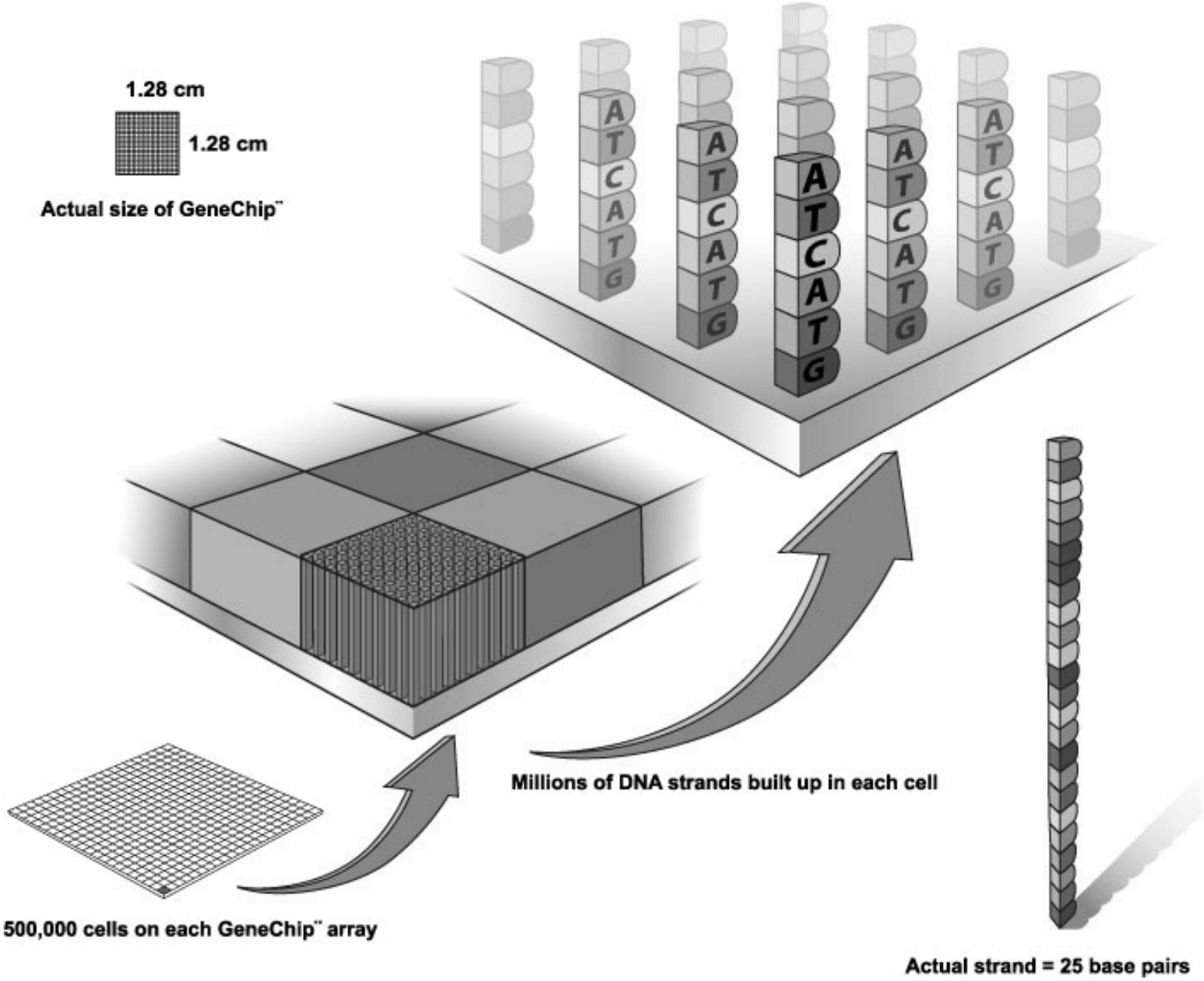
Microarray

Microarray: tecnologia in grado di analizzare simultaneamente migliaia di geni

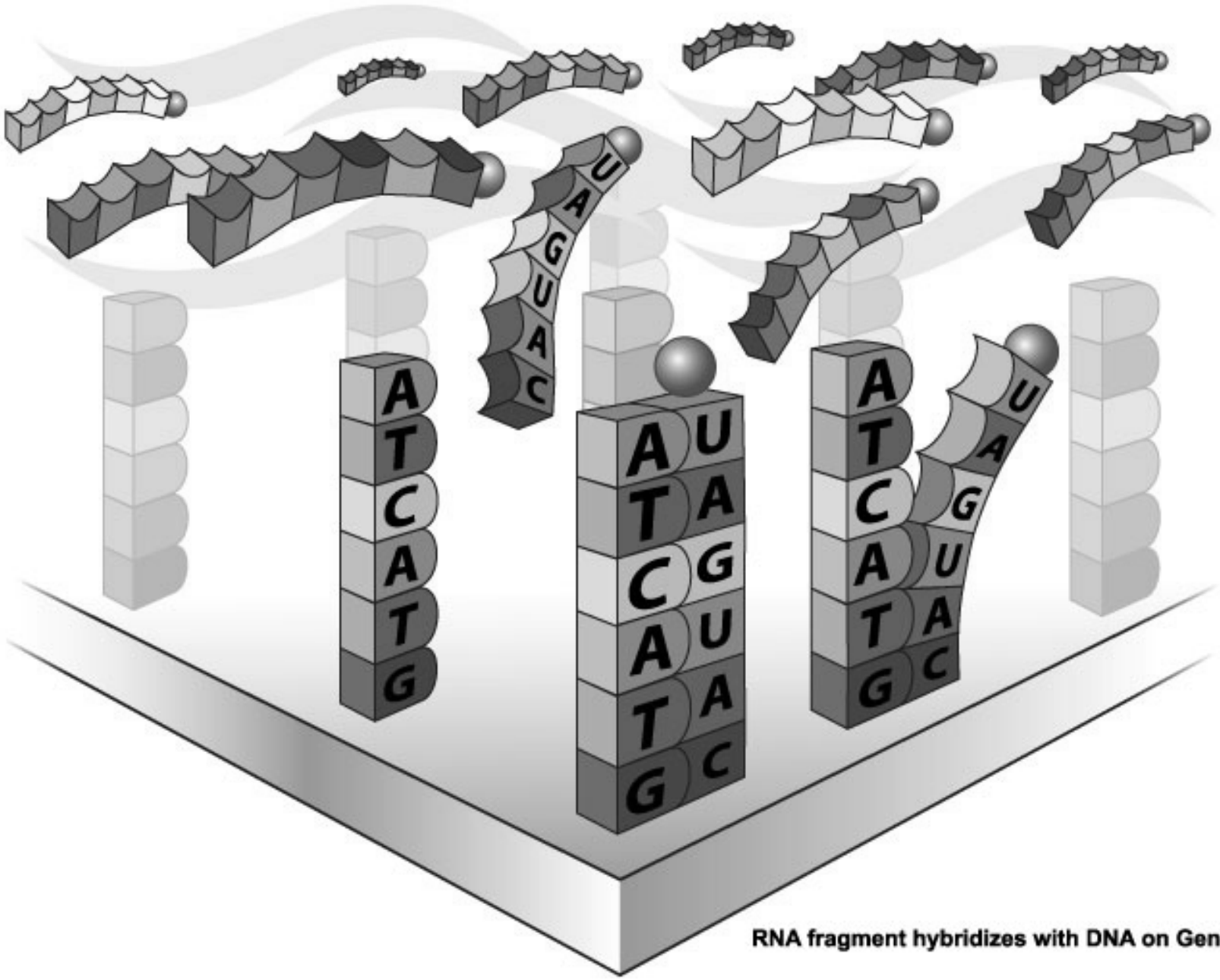
Expression microarrays: misurano il livello di espressione dei geni



Microarray

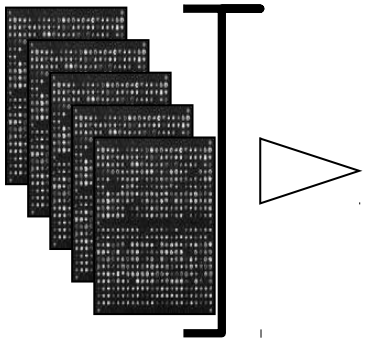

1.28 cm
1.28 cm
Actual size of GeneChip™
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

RNA fragments with fluorescent tags from sample to be tested



RNA fragment hybridizes with DNA on GeneChip

- Di solito vengono effettuati diversi esperimenti:
  - differenti condizioni di crescita
  - soggetti diversi
  - malattie diverse



esperimenti

geni

- Risultato
  - Matrice di espressione

    e(g,s)

    lo spot e(g,s) rappresenta quanto il gene e è espresso nell'esperimento s

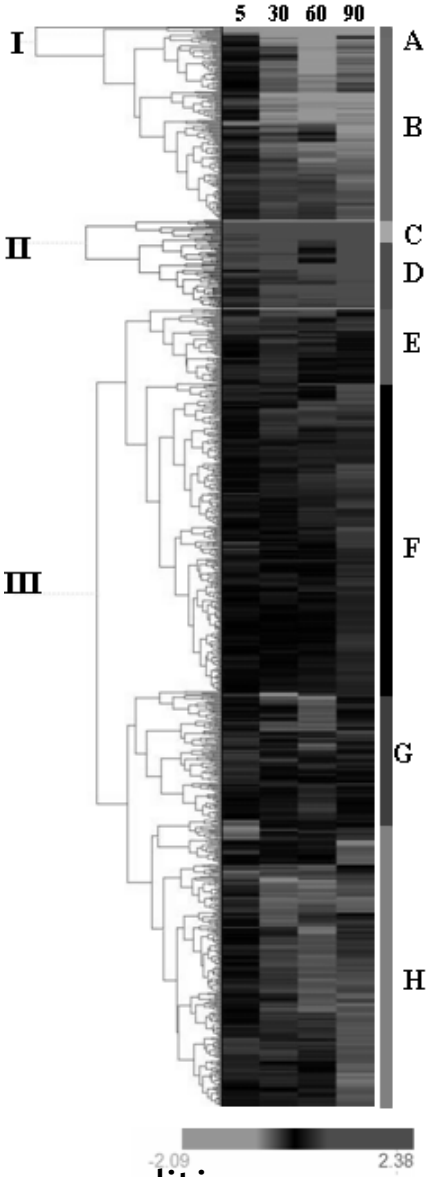
Microarray

PR:

1. classificazione di campioni

2. clustering

Two conditions

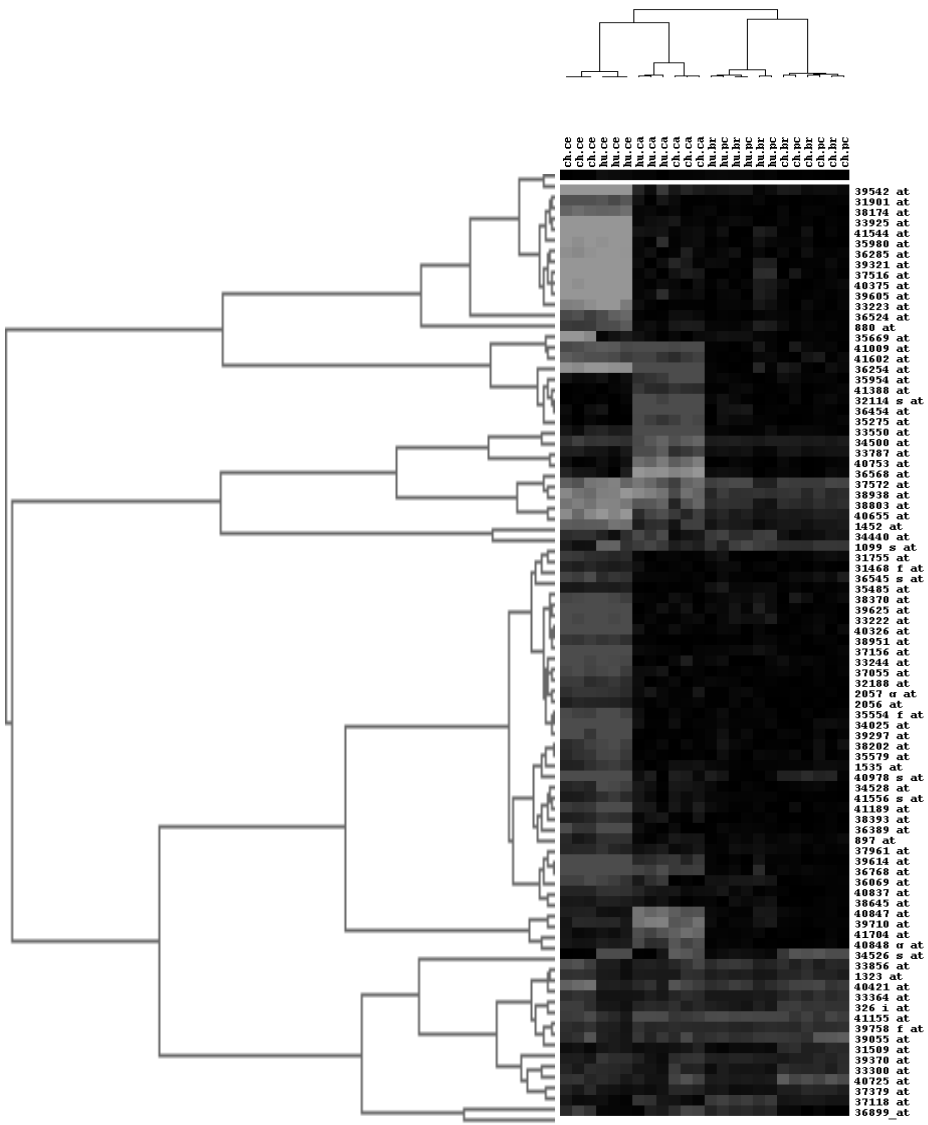


Four conditions

Clustering

clustering di esperimenti

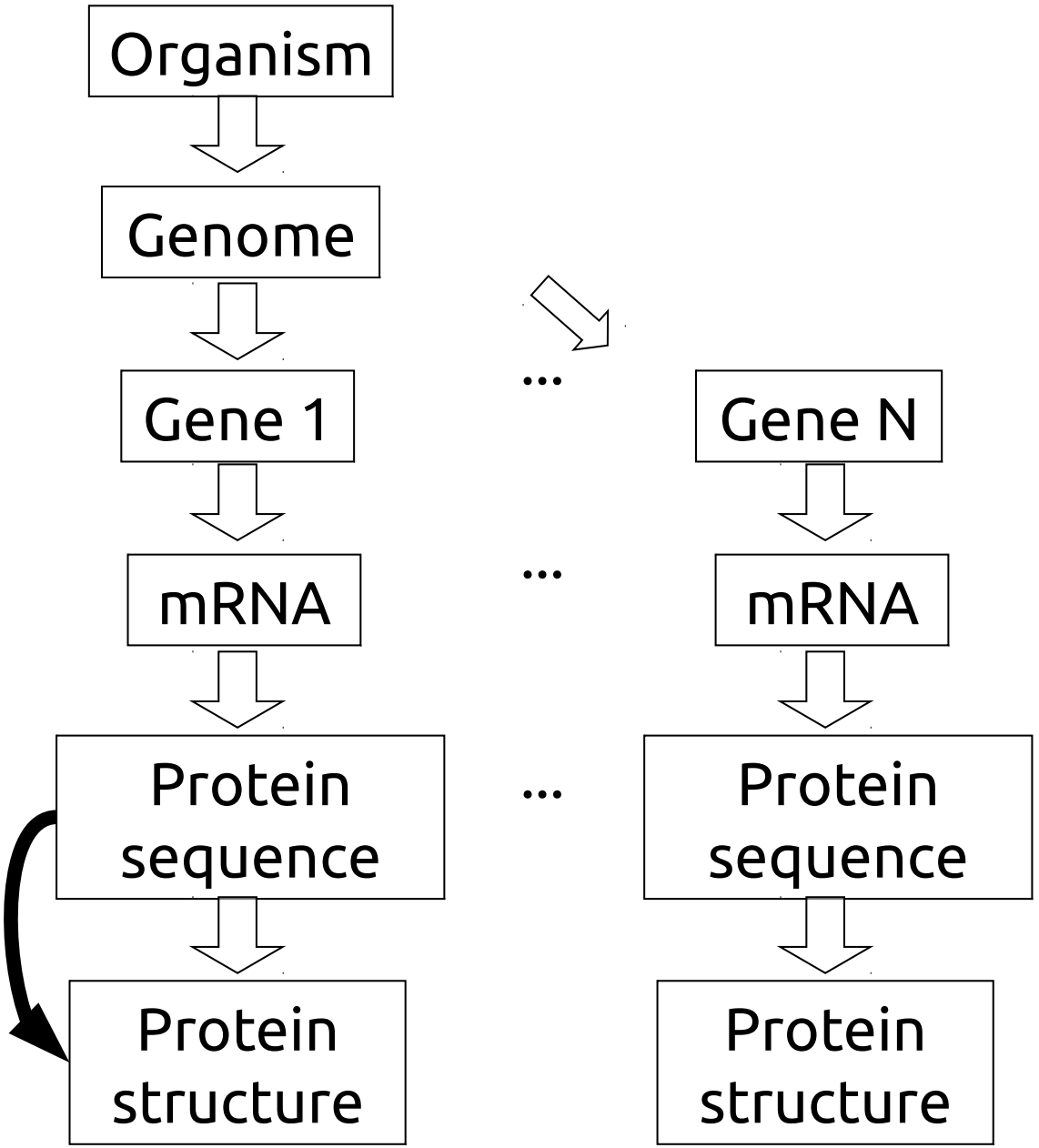
trovare esperimenti con geni espressi in modo simile



clustering di geni

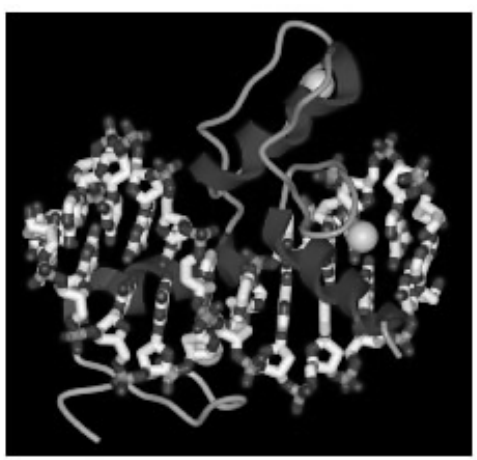
Trovare geni con pattern di espressione simili (quindi con funzioni ipoteticamente correlate)

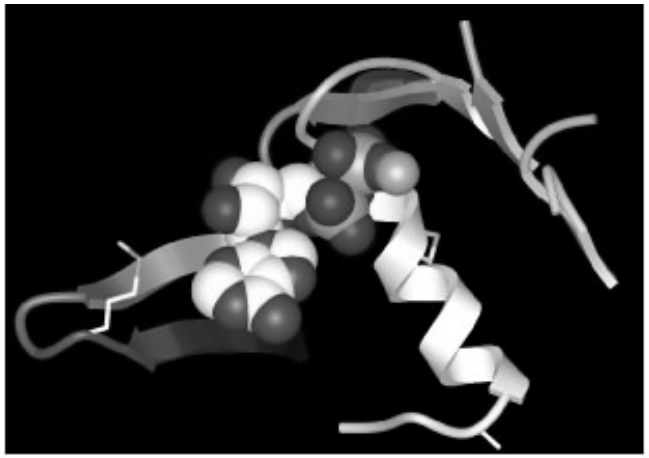
PR e Bioinformatica



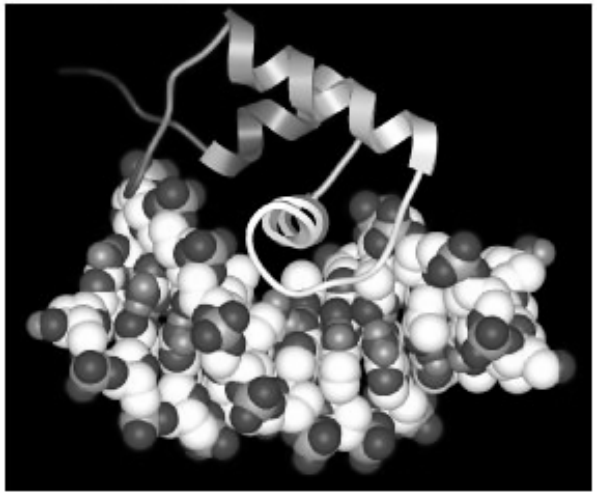
Motif discovery

Motif Discovery

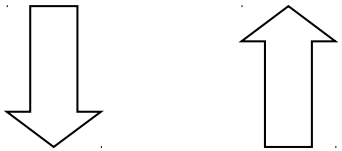




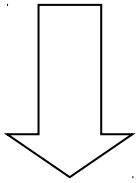
-PHE-ALA-ARG-SER-ASP-GLU-ARG-LYS-ARG-HIS-



parti simili in strutture di diverse proteine

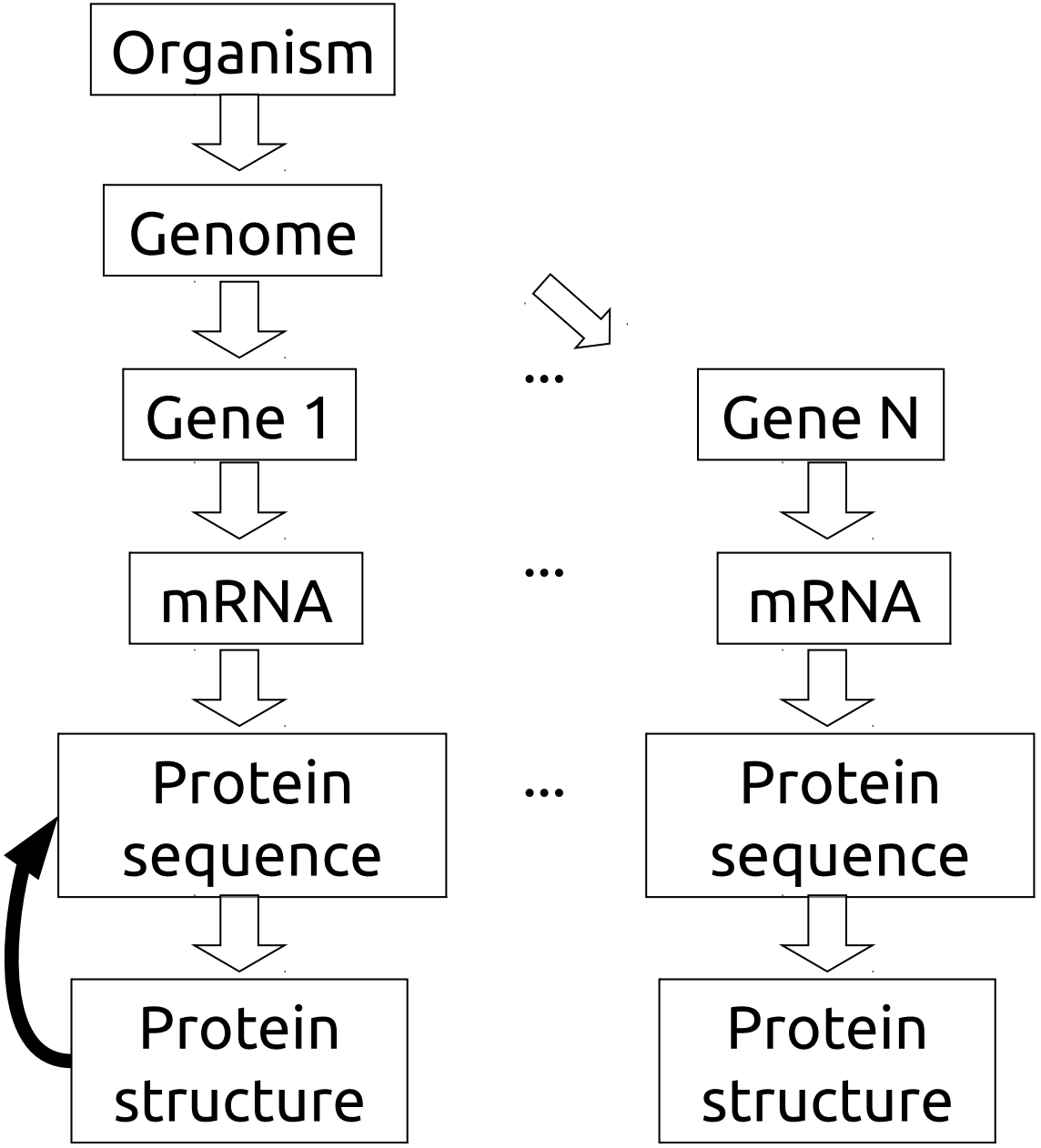


pattern simili nelle sequenze



PR: detection di questi pattern (motif)

PR e Bioinformatica



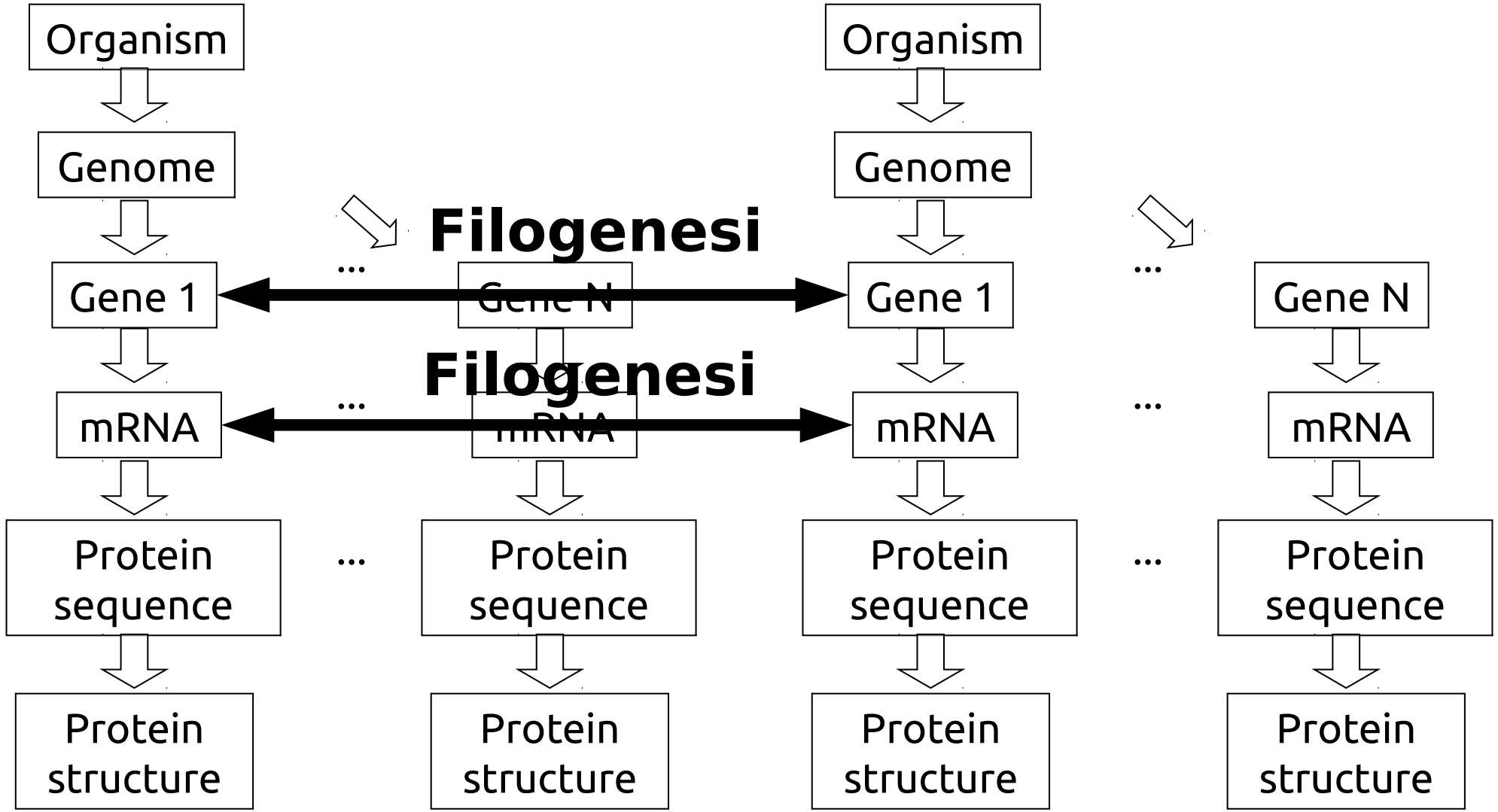
Protein Remote Homology Detection

Protein Remote Homology Detection

Stessa Funzione “A” ⇨ Molte proteine con strutture simili ⇨ Le sequenze corrispondenti condividono una similarità remota

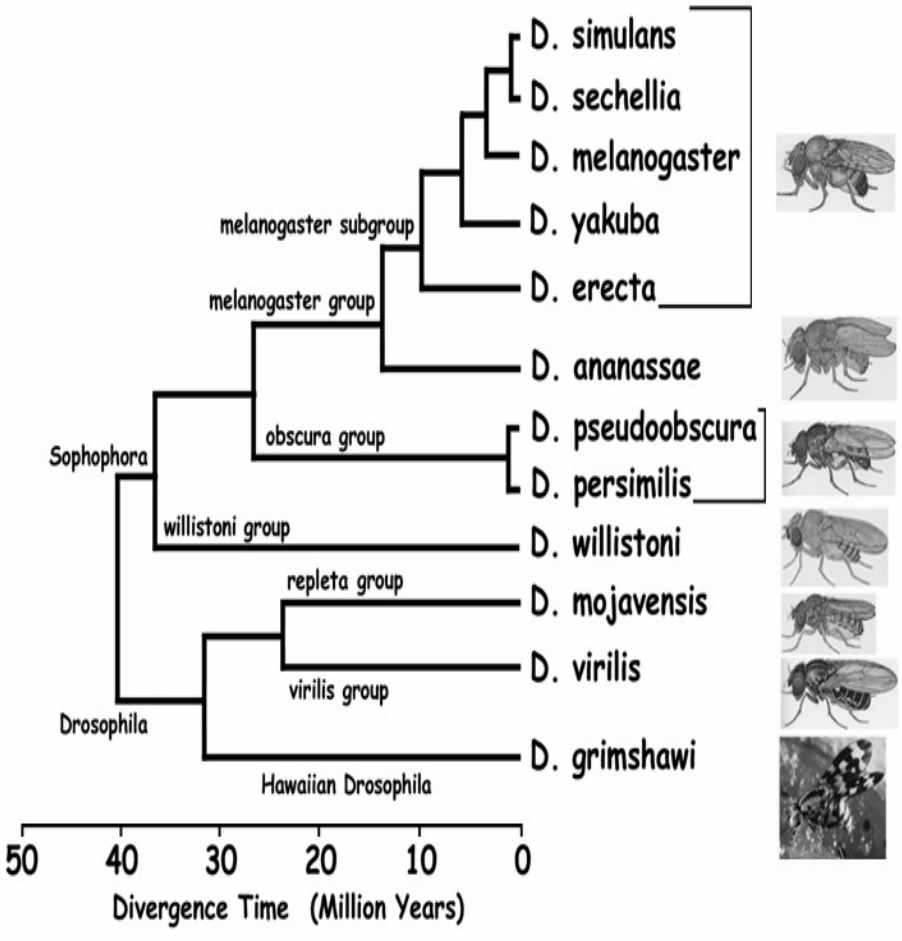
PR: caratterizzare ogni classe di sequenze omologhe (in senso remoto)

PR e Bioinformatica



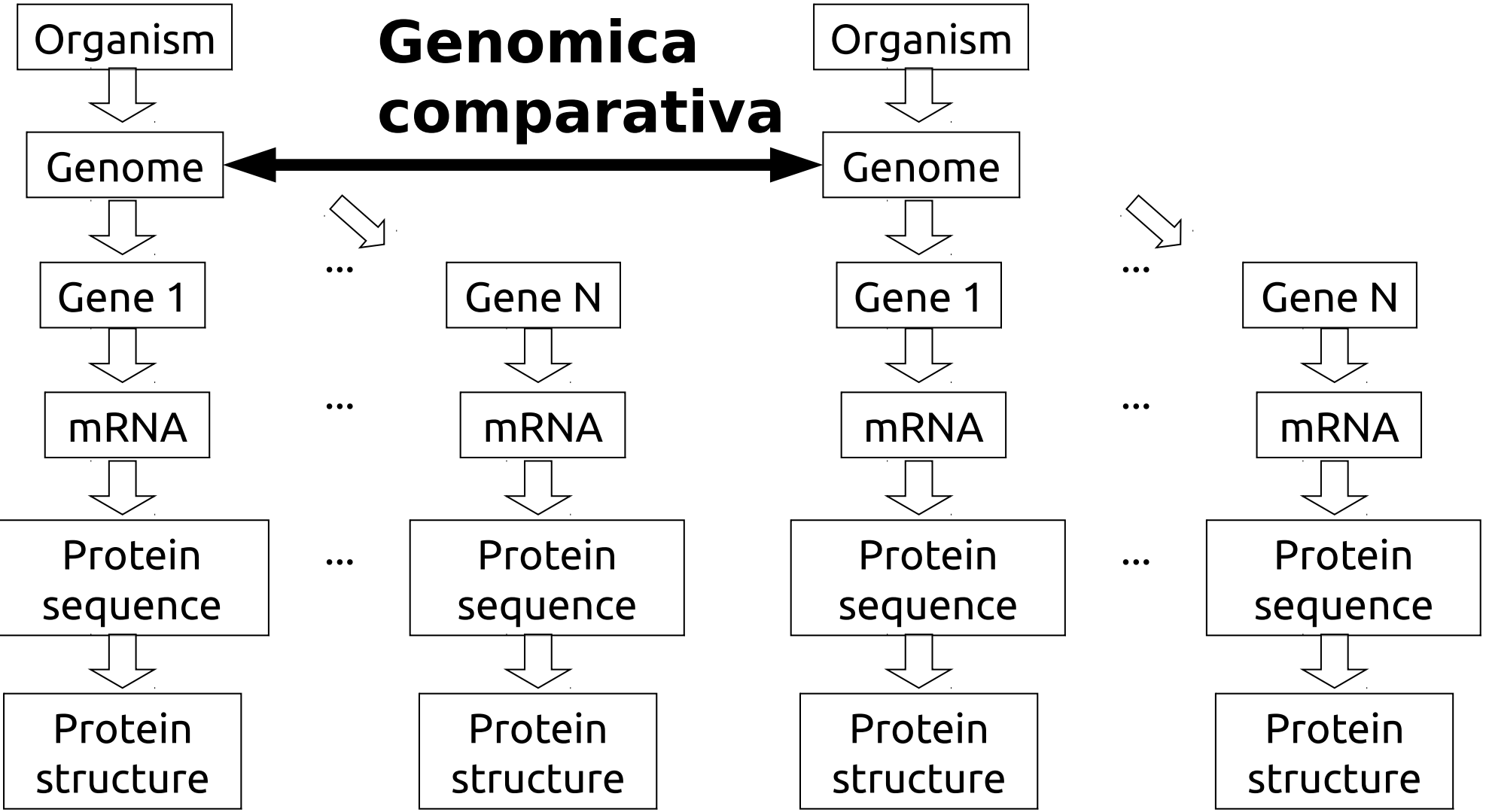
Filogenesi

Filogenesi: inferire le relazioni genealogiche tra gli organismi



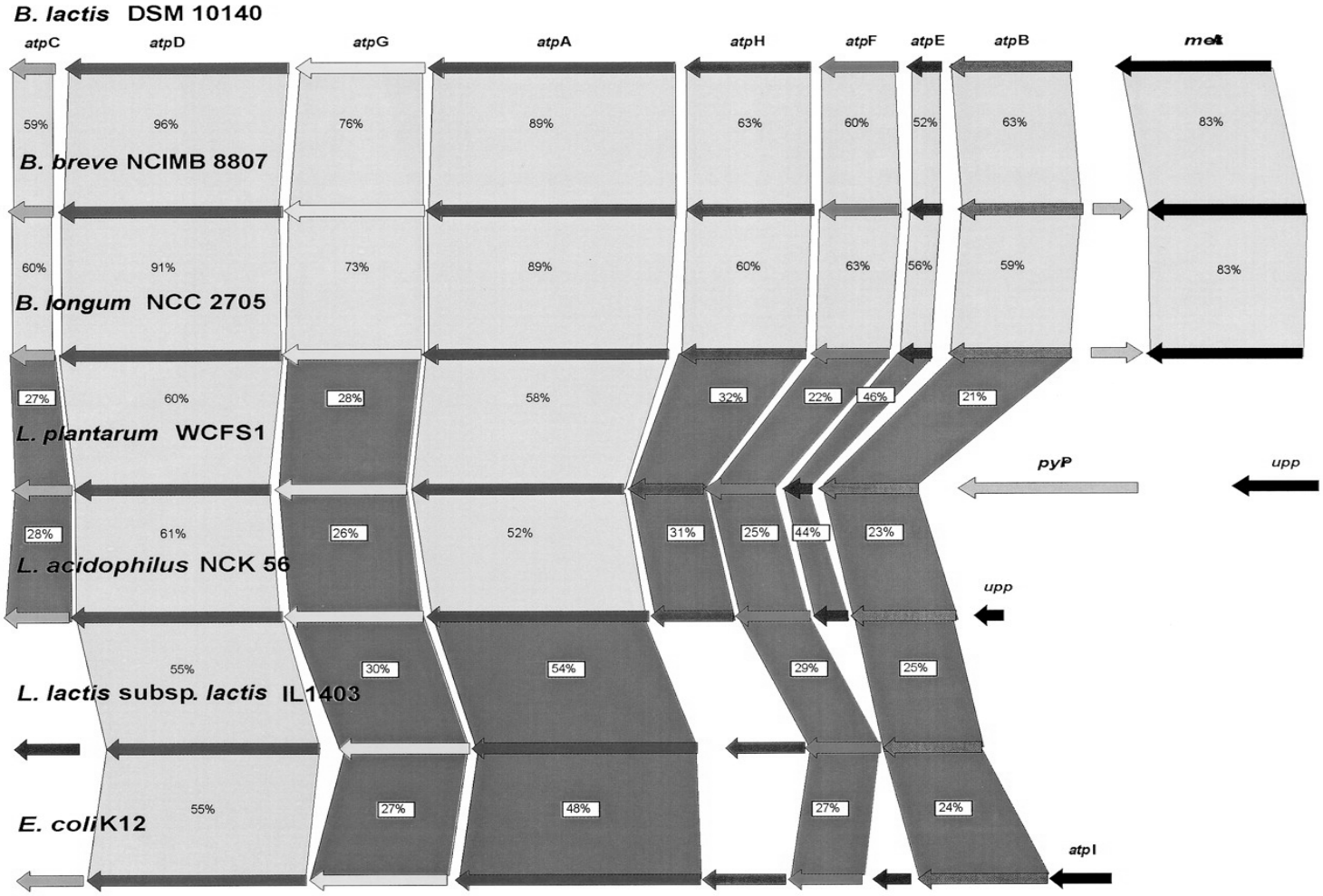
PR: clustering di sequenze geniche o proteiche

PR e Bioinformatica



Genomica comparativa (filogenomica)

Filogenomica: ha lo stesso obiettivo della filogenesi ma viene effettuata a livello di genoma



Problemi:

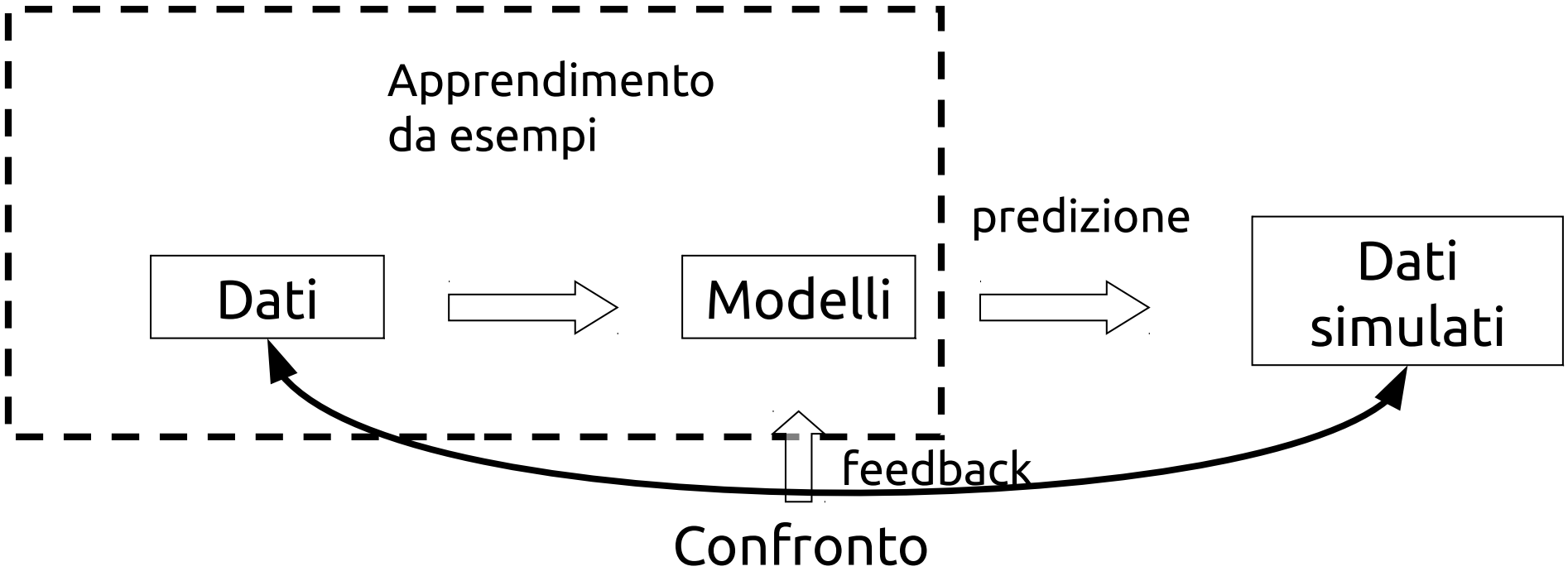
- ⇨ Presenza assenza di geni
- ⇨ traslazioni spaziali di geni
- ⇨ duplicazione di geni

PR e bioinformatica: perché?

1. In bioinformatica ci sono molti problemi di classificazione, clustering e detection
2. Possibilità di derivare modelli per i dati tramite esempi (paradigma di apprendimento da esempi)
3. Ci sono problemi di classificazione (onerosi in termini di tempo) che possono essere automatizzati

Modelli dai dati

- Modelli dai dati con il paradigma di “apprendimento da esempi”
  - Permette l'estrazione di informazioni semplificate o riassuntive
- Loop “simulazione & feedback”



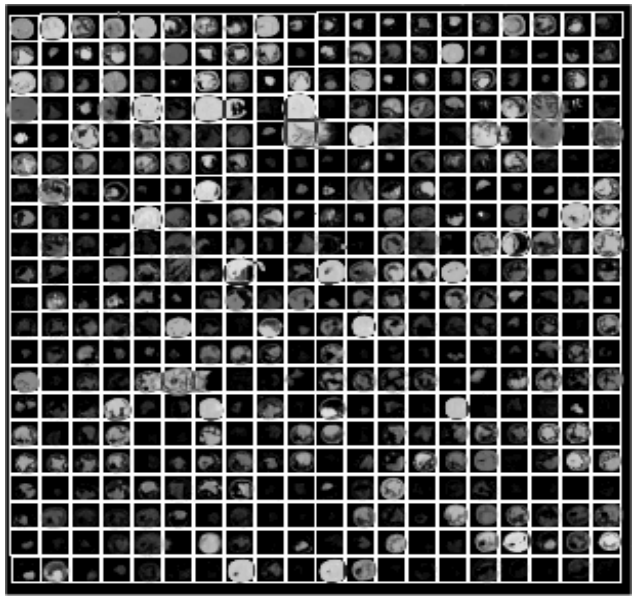
PR e bioinformatica: perché?

1. In bioinformatica ci sono molti problemi di classificazione, clustering e detection
2. Possibilità di derivare modelli per i dati tramite esempi (paradigma di apprendimento da esempi)
3. Ci sono problemi di classificazione (onerosi in termini di tempo) che possono essere automatizzati

Automazione di procedure

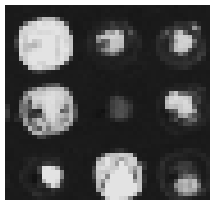
- Possibilità di automatizzare procedure di classificazione onerose dal punto di vista del tempo richiesto
  - Si può “imparare” come uno specialista esegue tali operazioni
  - Addestramento da un training set “annotato” da esperti

Esempio: qualità degli spot dei microarray

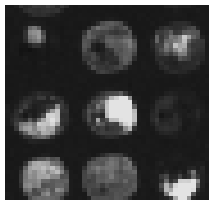


Problema:

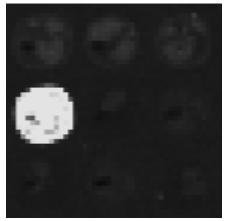
- ⇨ Trovare gli spot dei microarray con bassa qualità
- ⇨ Spot: immagine che contiene l'espressione di un gene



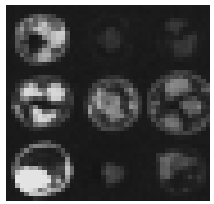
size



roundness



intensity



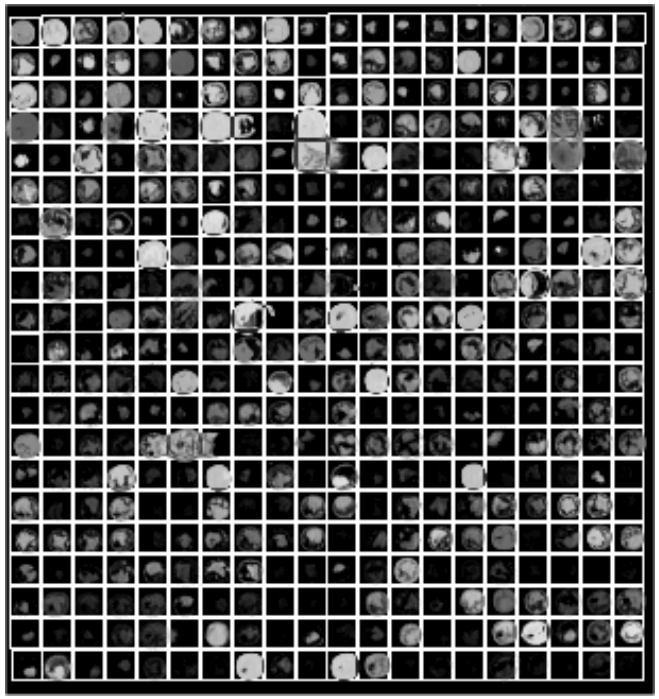
pixel distribution

Approccio tipico:

- ⇨ Annotazione manuale da parte di esperti

L’approccio PR

Imparare un modello, usando i giudizi dell’esperto, in un esperimento



spots (raw data)

$\mathbf{x}_1$ y_1

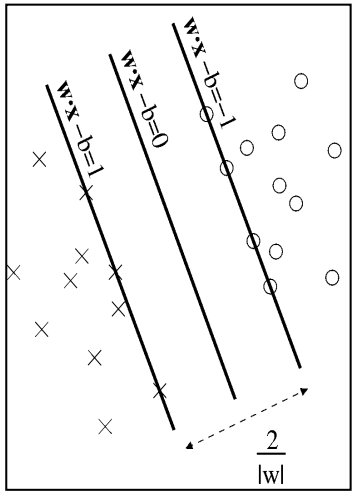
$\mathbf{x}_2$ y_2

... ...

$\mathbf{x}_N$ y_N

features

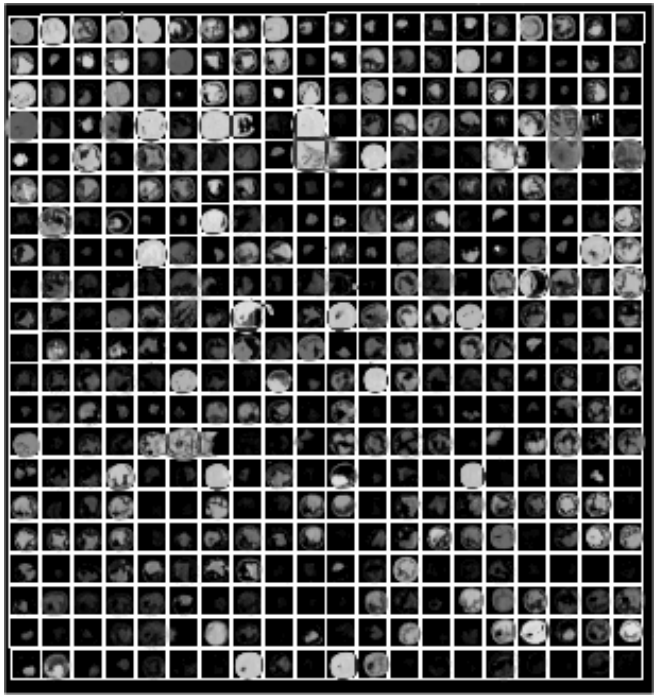
etichette degli esperti



Addestramento del modello

L’approccio PR

⇨ Testing: per un esperimento qualsiasi



spots (raw data)

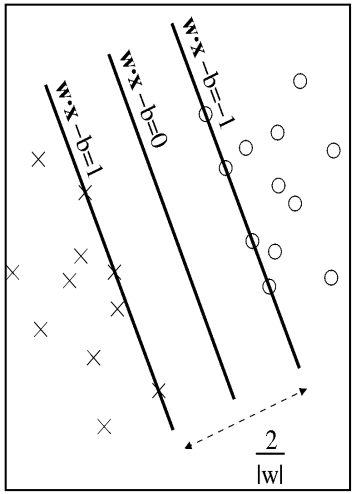
$\mathbf{x}_1$

$\mathbf{x}_2$

...

$\mathbf{x}_N$

features



modello addestrato

per ogni spot: buono o non buono

PR e Bioinformatica (more)

- Applicazioni legate alle immagini:
  - Rilevamento (detection) di parti interessanti in immagini
    - spots in microarray
    - gel
    - immagini mediche
  - misure
    - calcolo di feature (e.g. rotondità degli spot nei microarray)
    - presenza / assenza di elementi

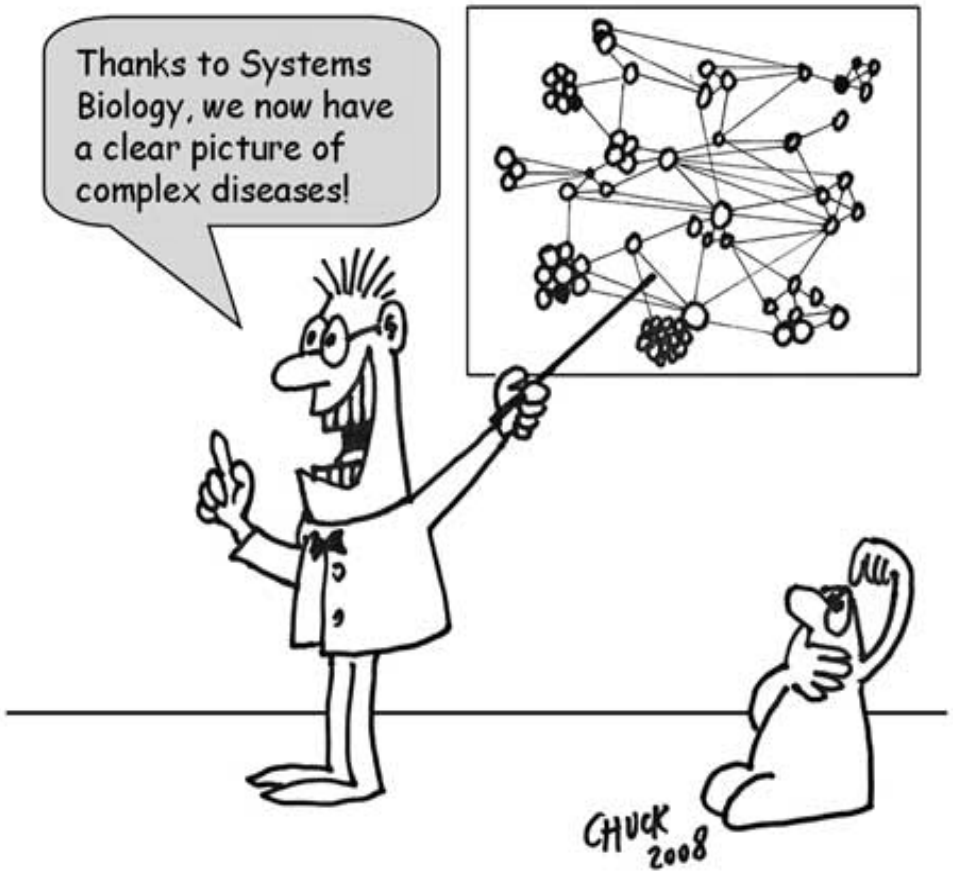
PR e Bioinformatica (more)

Ricerche nei database (GenBank, PDB)

- sequenze:
  - trovare similarità tra sequenze (e.g. BLAST, FASTA)
  - PR: similarità trovate utilizzando modelli addestrati
- documenti:
  - estrazione di informazioni rilevanti
  - PR: retrieval by content (invece della classica ricerca per keywords)
  - PR: classificazione di documenti

Sfide

- Enorme complessità e diversità dei sistemi biologici
- Enorme quantità di dati
  - Esempio: > 13K Completed Genome Projects - www.genomesonline.org
- Potenziale crescita esplosiva (e.g. il 95% della biodiversità microbica è sconosciuta)

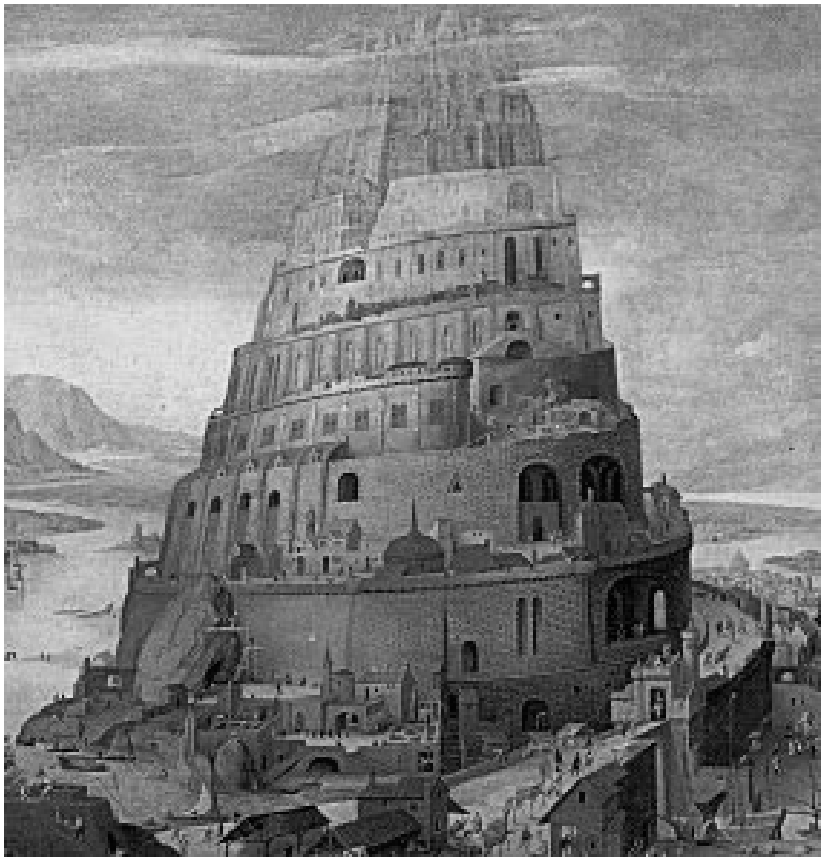


Sfide

Necessità di interagire con medici e biologi

E' difficile comunicare

- ⇨ Aspettative differenti
- ⇨ Background differenti
- ⇨ Linguaggi differenti!!



Sfide

- Necessità di utilizzare il più possibile le informazioni biologiche note a priori

- Ricerca della “interpretabilità biologica” di:
  - metodologie sviluppate
  - soluzioni ottenute

Il programma del corso

Programma dettagliato

INTRODUZIONE:

- ⇨ Introduzione generale alla Pattern Recognition: cos’è, cosa serve, com'è fatto un tipico sistema di PR
- ⇨ Rappresentazione e visualizzazione dei dati

CLASSIFICAZIONE:

- ⇨ Teoria di Bayes per la classificazione
- ⇨ Classificatori generativi
- ⇨ Classificatori discriminativi
- ⇨ Validazione

Programma dettagliato

CLUSTERING:

- ⇨ Misure di similarità tra dati
- ⇨ Tecniche di clustering: tassonomia e dettagli delle tecniche più utilizzate
- ⇨ Validazione

APPLICAZIONI:

- ⇨ classificazione e clustering di dati microarray (con cenni alle tecniche di biclustering)
- ⇨ analisi di immagini biomedicali (cenni)
- ⇨ *(Classificazione di omologia remota tra proteine)*

Riviste e convegni principali
(sul tema generale)

- Convegni
  - NIPS, ICML, ECML, CVPR, AI, ICPR, ICCV, ECCV, ICIP, etc.
- Journals
  - PAMI, IEEE Trans. on Pattern Analysis & Machine Intelligence
  - Artificial Intelligence
  - Machine Learning
  - Journal of Machine Learning Research
  - CVIU, Computer Vision and Image Understanding
  - GMIP, Graphical Models & Image Processing
  - IVC, Image and Vision Computing
  - PR, Pattern Recognition
  - PRL, Pattern Recognition Letters
  - IEEE Trans. on Image Processing
  - IEEE Trans. on Systems, Man, & Cybernetics
  - Int. J. on Pattern Recognition & Artificial Intelligence
  - IEEE Trans. on Neural Networks
  - Neural Computation
  - Proceedings of the IEEE

Riviste e convegni principali
(specifici su Bioinfo)

- Convegni
  - molti, parole chiave bioinformatics, computational biology, medical informatics, pattern matching, systems biology, AI in Medicine, etc.
- Journals
  - Bioinformatics
  - BMC Bioinformatics
  - Journal of Bioinformatics & Computational Biology
  - IEEE/ACM Trans. on Computatinal Biology & Bioinformatics
  - Int'l Journal of Data Mining & Bioinformatics
  - Eurasip Journal of Bioinformatics & Systems Biology
  - Int'l Journal of Bioinformatics Research and Applications
  - Journal of Biomedical Informatics
  - Journal of Computational Biology
  - Journal of Proteomics & Bioinformatics
  - Journal of Integrative Bioinformatics
  - The Open Bioinformatics journal