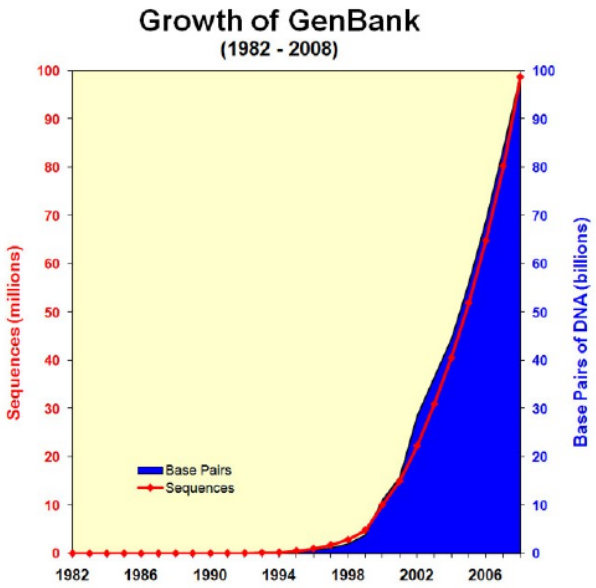
# Informatica e Bioinformatica: Basi di Dati

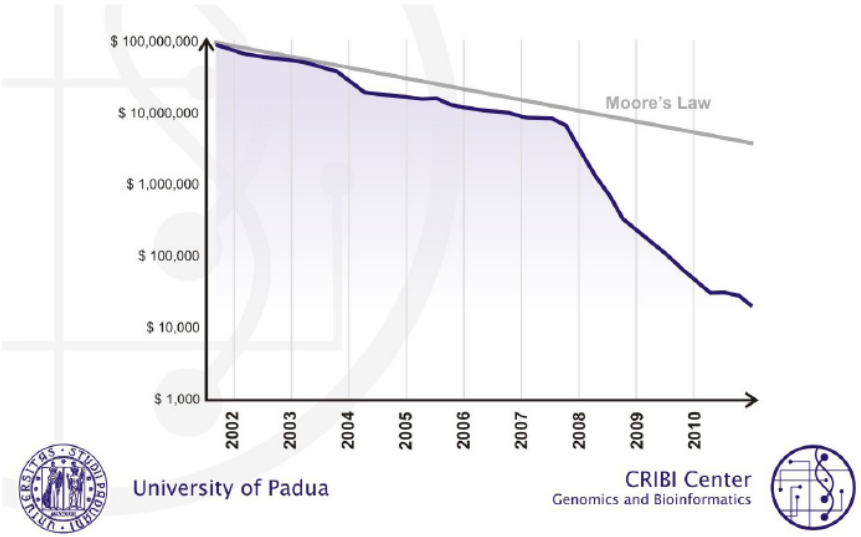
Mauro Conti

Date TBD

# Bioinformatica

- I costi di sequenziamento e di hardware descrescono vertiginosamente




- si hanno a disposizione sempre più dati e hardware sempre più potente e meno costoso. . .
- Bioinformatica: applicazione dell'informatica alla **gestione** e all'analisi dei dati biologici

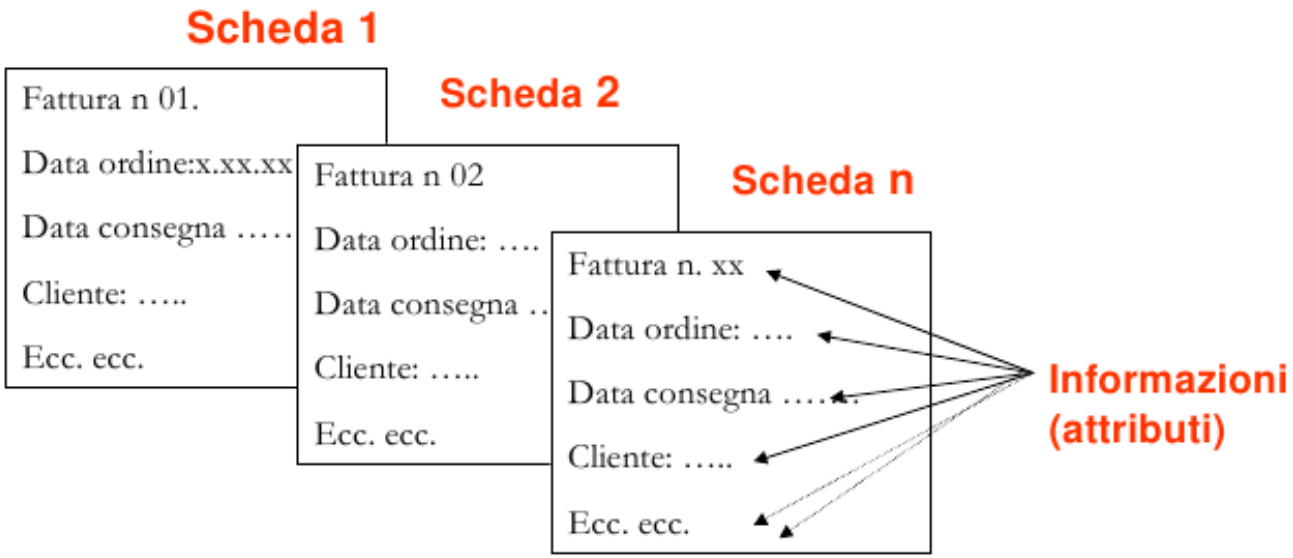
# Archivi di Dati: Introduzione

- Gestione: conoscere **come** e dove sono archiviati i dati biologici e saper effettuare velocemente ricerche esaustive per trovare ciò che interessa.
- Oggi ci occuperemo del "come" sono archiviati i dati biologici.
- Prima dell'avvento dei computer, i dati erano salvati su supporti cartacei:
  - quaderni, memorizzazione sequenziale dei dati. È difficile riordinarli secondo criteri diversi da quello in cui sono stati scritti
    - Ad esempio un quaderno di fatture, scritte secondo l'ordine di emissione, riordinarle alfabeticamente.
  - schedari, dove i dati sono organizzati in schede (ad esempio una scheda per ogni fattura)

# Archivio

- Un archivio deve essere specifico per l'ambito a cui fa riferimento, non può contenere tutto
  - se si vuole costruire uno schedario clienti ad ogni cliente dovrà essere associata una scheda
- Ogni scheda dovrà contenere le informazioni, chiamate attributi, che costituiscono l'elemento
  - la scheda di una fattura potrà avere i seguenti attributi: data di emissione, ammontare, nome cliente, ecc. . .

# Archivio: Caratteristiche




- Una scheda rappresenta una singola fattura e c'è una sola scheda per ogni fattura
  - Ogni scheda rappresenta un elemento unico (ogni fattura è univoca e rappresenta una singola transazione di merce)
- Però informazioni presenti su una scheda (attributi) possono essere ripetute
  - lo stesso cliente può comparire in più schede

# Database

- Un database (base di dati) è l'equivalente software di uno schedario
- Più formalmente, il termine database (base di dati) indica un archivio di dati riguardanti uno o più argomenti correlati tra loro e strutturato in modo da consentire la gestione dei dati stessi da parte di applicazioni software gestite da un elaboratore.
- Per gestione dei dati si intende linsieme di operazioni che permette di effettuare interrogazioni (queries): l'inserimento, la cancellazione, l'aggiornamento, la ricerca, la stampa di report personalizzati.

# Database: Definizioni

Differenze tra i database e gli archivi cartacei:

- Nel gergo dei database si utilizzano i seguenti termini corrispondenti a quelli che abbiamo visto:
  - Attributo = Campi (Fields)
  - Scheda = Record
  - Database = Archivio di Schede
- I database sono salvati su memorie di massa, mentre gli archivi su supporti cartacei
- La gestione dei dati è gestita da programmi applicativi per i database, è manuale per gli archivi cartacei
  - nei database è più veloce la ricerca di dati, la selezione di una lista di record appartenenti a schede diverse, l'ordinamento di record o campi secondo criteri arbitrari.

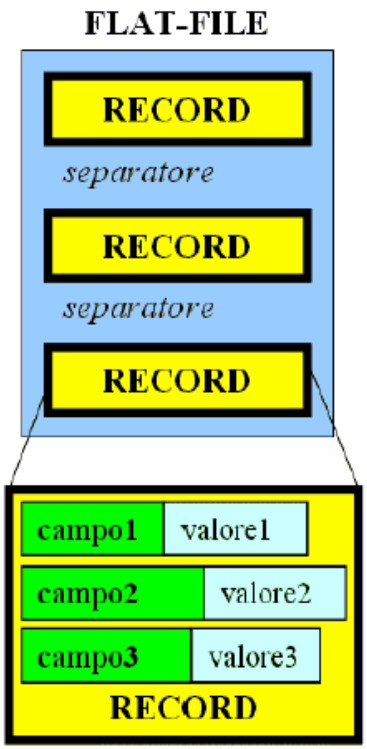
# Classi di Database

Esistono principalmente 2 tipi di database (2 modi diversi di realizzare un database)

- Flat file
- Database relazionali o a oggetti

# Database Flat File

- Un database può essere costituito da un semplice file di testo ordinato secondo un preciso schema.
- All'interno del file esistono delle sequenze di caratteri, gli spaziatori (separatori nella figura a fianco), che permettono di individuare i singoli record.
- All'interno di un record esistono delle parole chiave che permettono di individuare i campi di quel record. Questi identificatori di campo possono corrispondere semplicemente alla posizione di una casella di testo oppure possono essere delle "etichette" che indicano il tipo di campo.
- In genere esiste un campo (ID, identificatore, chiave...) che identifica univocamente i record del DB (non possono esistere 2 record con lo stesso valore nel campo chiave!)



# Flat File Database

Esempio di Flat File

```
ID                  : 28877
PARENT ID           : 28876
RANK                : no rank
GC ID               : 1
SCIENTIFIC NAME     : IDIR agent
SYNONYM             : Infectious disease of infant rats
SYNONYM             : Rotavirus (GROUP B / STRAIN IDIR)
SYNONYM             : infectious diarrhea of infant rats agent IDIR
//
ID                  : 55279
PARENT ID           : 6607
RANK                : family
GC ID               : 1
MGC ID              : 5
SCIENTIFIC NAME     : Idiosepiidae
//
ID                  : 82764
PARENT ID           : 82761
RANK                : family
GC ID               : 1
MGC ID              : 5
SCIENTIFIC NAME     : Idoteidae
//
```

Nell'esempio ci sono 3 record separati da //, i record hanno 8,6,6 campi rispettivamente (ciascun campo occupa una riga).
L'identificatore del campo è separato dal valore del campo da ":".
Il campo chiave ha come identificatore ID.

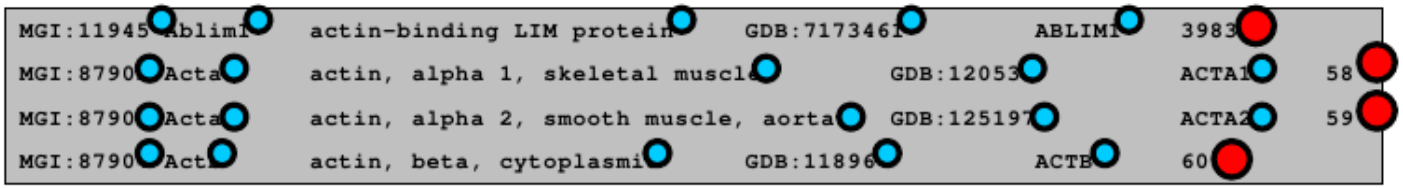
# Flat File: un altro esempio

```
MGI:11945 Ablim1    actin-binding LIM protein     GDB:7173461         ABLIM1    3983
MGI:87902 Acta1     actin, alpha 1, skeletal muscle         GDB:120535          ACTA1     58
MGI:87909 Acta2     actin, alpha 2, smooth muscle, aorta    GDB:125197          ACTA2     59
MGI:87904 Actb      actin, beta, cytoplasmic      GDB:118964          ACTB      60
```

Qual è il separatore di record?
Qual è il separatore di campo?

# Flat File: un altro esempio

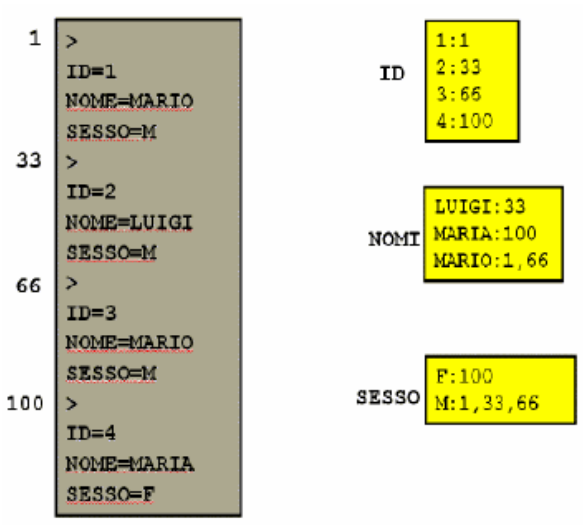
```
MGI:11945 Ablim1    actin-binding LIM protein     GDB:7173461         ABLIM1    3983
MGI:87902 Acta1     actin, alpha 1, skeletal muscle          GDB:120535          ACTA1     58
MGI:87909 Acta2     actin, alpha 2, smooth muscle, aorta     GDB:125197          ACTA2     59
MGI:87904 Actb      actin, beta, cytoplasmic      GDB:118964          ACTB      60
```



Il separatore di record è il carattere di "a capo" (pallino rosso).
Il separatore di campo è il carattere "TAB" (pallino blu)

# Flat File: Ricerca di Record

- Un flat file biologico può contenere migliaia o milioni di record.
- È inefficiente percorrere tutto il file dall'inizio alla fine per cercare un record (ad esempio cercando nel campo ID il valore 82764).




- Si prepara, per ogni campo, una lista di tutti i valori presenti nell'archivio con la loro posizione nel file (indice).
- Gli indici sono più piccoli dei file originali, quindi fare le ricerche sugli indici è più veloce!

# Flat File: Considerazioni

Vantaggi:

- Implementato in banche dati progettate ed ottimizzate per la consultazione rapida nel caso di formulazioni di query complesse.
- Risposta rapida alle domande che si pongono alla banca dati che può contenere Gigabyte di informazioni (come le banche dati biologiche primarie).
- Si può scegliere di indicizzare solo parte dei dati per rendere il file di indici più compatto e veloce da consultare.
- La scelta è stata fatta ed utilizzata dalle banche dati biologiche basate su sistemi di interrogazioni come ENTREZ e SRS.

Svantaggi:

- È complicato aggiornare il database, se si aggiungono nuovi record o si aggiornano i dati pre-esistenti si deve calcolare nuovamente l'indice dei dati.

# Database Flat File

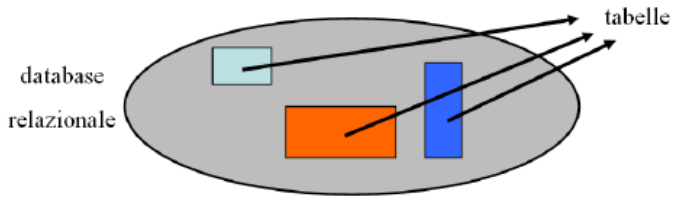
- Un database Flat File è composto da un'unica tabella (qua sotto ancora l'esempio delle fatture)

| Numero | data | nome cliente | indirizzo cliente | prezzo |
|---|---|---|---|---|
| 1 | 1/1/2003 | Mario Rossi | via Trieste 63 | 185 |
| 2 | NULL | Gino Verdi | vicolo stretto 1 | 2100 |
| 3 | 19/7/2011 | Mario Rossi | via Trieste 95 | 100 |

- Ogni riga della tabella è un record (una scheda), ogni cella della tabella un campo
- In ogni record devono essere presenti tutti i campi, nel caso di dati mancanti si deve utilizzare un valore speciale NULL.

# Database Relazionali

- Un Database Relazionale è un insieme di TABELLE (table), in origine chiamate relazioni, collegate logicamente l'una con l'altra.
- Una tabella di un DB relazionale è l'equivalente di un flat-file.




DB clienti

| ID cliente | Nome | Indirizzo | Ecc.ecc. |
|---|---|---|---|
| 01 | Rossi Guido | Padova, via …. | .. |
| 02 | Barca Mario | Milano, …. | .. |
| Ecc | Ecc | Ecc | |

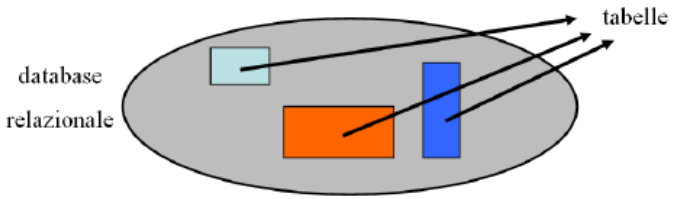
DB prodotti

| ID prodotto | Descrizione | Costo unitario | Ecc ecc |
|---|---|---|---|
| 01 | Armadio ….. | 550 | --- |
| 02 | Tavolo ….. | 100 | --- |
| Ecc ecc | --- | --- | --- |

DB ordini

| ID ordine | ID cliente | Data ordine | ID Prodotto | Ecc ecc |
|---|---|---|---|---|
| 01 | 02 | 24/12/2005 | 02 | --- |
| 02 | --- | --- | --- | --- |
| --- | --- | --- | --- | --- |

# Database Relazionali




DB clienti

| ID cliente | Nome | Indirizzo | Ecc.ecc. |
|---|---|---|---|
| 01 | Rossi Guido | Padova, via …. | .. |
| 02 | Barca Mario | Milano, …. | .. |
| Ecc | Ecc | Ecc | |

DB prodotti

| ID prodotto | Descrizione | Costo unitario | Ecc ecc |
|---|---|---|---|
| 01 | Armadio ….. | 550 | --- |
| 02 | Tavolo ….. | 100 | --- |
| Ecc ecc | --- | --- | --- |

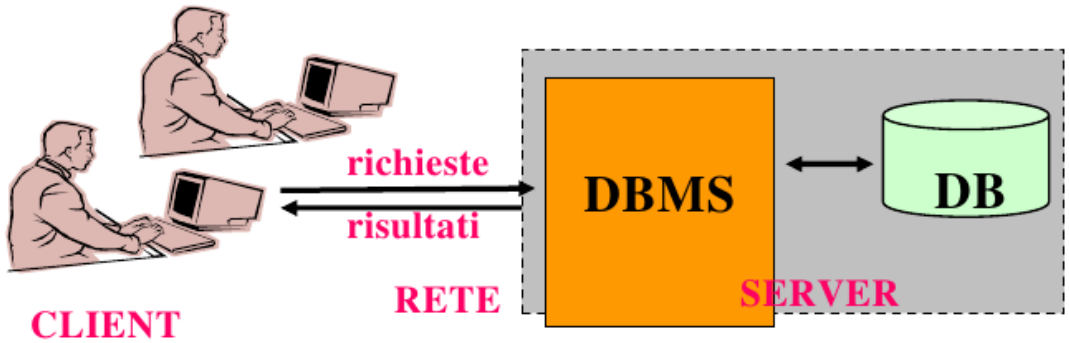
DB ordini

| ID ordine | ID cliente | Data ordine | ID Prodotto | Ecc ecc |
|---|---|---|---|---|
| 01 | 02 | 24/12/2005 | 02 | --- |
| 02 | --- | --- | --- | --- |
| --- | --- | --- | --- | --- |

- Si utilizzano più tabelle per evitare di duplicare i dati.
- Ad esempio, se dobbiamo modificare il nome del cliente basta modificare un solo record!

# DBMS

- I database relazionali necessitano di particolari programmi di gestione (Database Management System o DBMS)
  - che siano in grado di saltare da una tabella all'altra e di capire le relazioni ed i vincoli ad esse associati.
  - Devono inoltre occuparsi di gestire l'aggiunta, la modifica e la gestione degli indici
- Il DBMS funge da interfaccia verso il database, in una tipica configurazione client-server. Il server è residente su un computer remoto, mentre i client sono in generale altri computer.



# SQL Structured Query Language

- SQL È sicuramente il DBMS relazionale più diffuso.
- SQL é uno standard di cui esistono alcune implementazioni:
  - ORACLE (commerciale)
  - Mysql (gratuita)
- Al fine di rendere più semplice la consultazione dei dati, molti database biologici sono estratti dai loro sistemi di gestione relazionale e sono "appiattiti" in file di testo, i "flat files" leggibili come semplici file di testo. Ricordo che i "flat files" possono essere indicizzati e utilizzati per ricerche anche molto complesse (ENTREZ, SRS).

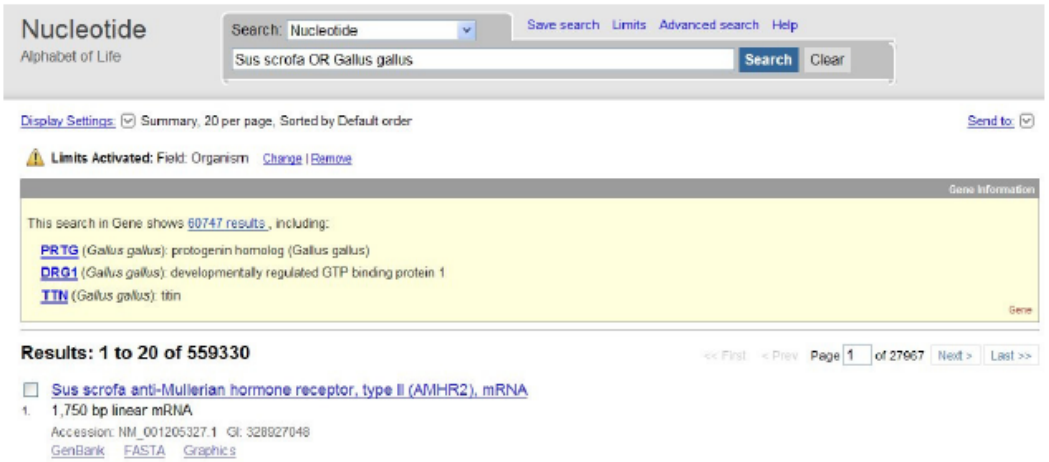
# Interrogazioni delle basi di dati

- RICERCHE TESTUALI (QUERY)
  - Utilizzano dei programmi di RETRIEVAL (cioè di ricerca, reperimento dati) per restituire i record di un database che soddisfano i criteri richiesti.
  - Sfruttano una ricerca di tipo letterale ed individuano termini uguali.
- RICERCHE PER SIMILARITÀ (su sequenze nucleotidiche o proteiche)
  - Restituiscono le sequenze di un database più simili ad una sequenza fornita come query.
  - Non sono delle vere e proprie query in quanto richiedono l'esecuzione di programmi anche piuttosto complessi (ad esempio BLAST o FASTA).

# Interrogazioni delle basi di dati

- È generalmente possibile utilizzare gli operatori booleani AND, OR, NOT per costruire interrogazioni più complesse

- **Esempio di OR**




- **Esempio di AND (non esistono sequenze contemporaneamente di maiale e di pollo)**