Docente: **Matteo Re**
re@di.unimi.it
matteo.re@unimi.it

Università degli studi di milano

Banche dati biologiche

A.A. 2014-2015 semestre I

4 INTRODUZIONE, STRUMENTI WEB

bioinformatica:

Come mai esistono numerosi strumenti web dedicati in modo specifico alla manipolazione ed elaborazione di dati di tipo biomolecolare e biotecnologico?

Lo sviluppo di questi strumenti è iniziato a metà degli anni ottanta del secolo scorso per rispondere all'esigenza di rendere possibile l'estrazione di informazioni da collezioni di schede testuali che descrivevano molecole presenti nelle cellule di diversi organismi.

Lo sviluppo di questi strumenti di estrazione e di metodi per confrontare biomolecole ha portato alla nascita di una disciplina che prende il nome di bioinformatica.

Bioinformatica è l'**applicazione** di strumenti propri delle scienze dell'informazione (es. algoritmi, intelligenza artificiale, databases) a problemi di interesse biologico, biotecnologico e biomedico.

Banche dati biologiche e bioinformatica:

Attualmente la bioinformatica suscita grande interesse perché:

•La creazione di nuove biotecnologie ha permesso di ridurre il costo ed il tempo necessario per l'acquisizione di informazioni sulle biomolecole. → **esistono banche dati contenenti informazioni riguardanti milioni di biomolecole,.**

•La dimensione dei problemi biologici è sufficiente a motivare lo sviluppo di algoritmi efficienti

•I problemi sono accessibili (elevata quantità di dati pubblici e letteratura inerente) ed interessanti

•Le scienze biologiche si avvalgono sempre più spesso di strumenti computazionali

Bioinformatica:
Scienze dell'Informazione o Biologia? (I)

Gli sviluppi delle scienze biomediche, (in particolare per quanto riguarda la biologia molecolare) si verificano ad un ritmo tale da porre seri problemi:

La nostra capacità tecnologica di acquisire nuovi dati (spesso in quantità elevate) rende impossibile la loro analisi **in assenza di strumenti efficienti**.

Bioinformatica: Scienze dell'Informazione o Biologia? (II)

Cosa hanno **in comune** scienze biologiche e scienze dell'informazione?

La **biologia**, ed in particolare la biologia molecolare, si occupa dei fenomeni che avvengono nei viventi a livello di atomi e molecole. L'unità di base dei viventi è la **cellula**. La costruzione di una cellula richiede la **lettura e la manipolazione di informazioni** …

Scienze dell'informazione è la disciplina che si occupa di calcolo (in senso generico) e delle sue applicazioni. Essa si basa sullo studio sistematico di fattibilità, struttura e automatizzazione di metodi che permettono l'acquisizione, accesso e manipolazione dell'informazione (intesa in senso generico).

Bioinformatica:
Scienze dell'Informazione o Biologia? (III)

Cosa hanno **in comune** scienze biologiche e scienze dell'informazione?

Scienze dell'informazione	Biologia
Dati di input	DNA
Lettura	copiatura (trascrizione)
Elaborazione	traduzione
Dati di output	PROTEINA

Bioinformatica:
Scienze dell'Informazione o Biologia? (IV)

Cosa hanno **in comune** scienze biologiche e scienze

Scienze dell'informazione

Biologia

STUDIANO ENTRAMBE:

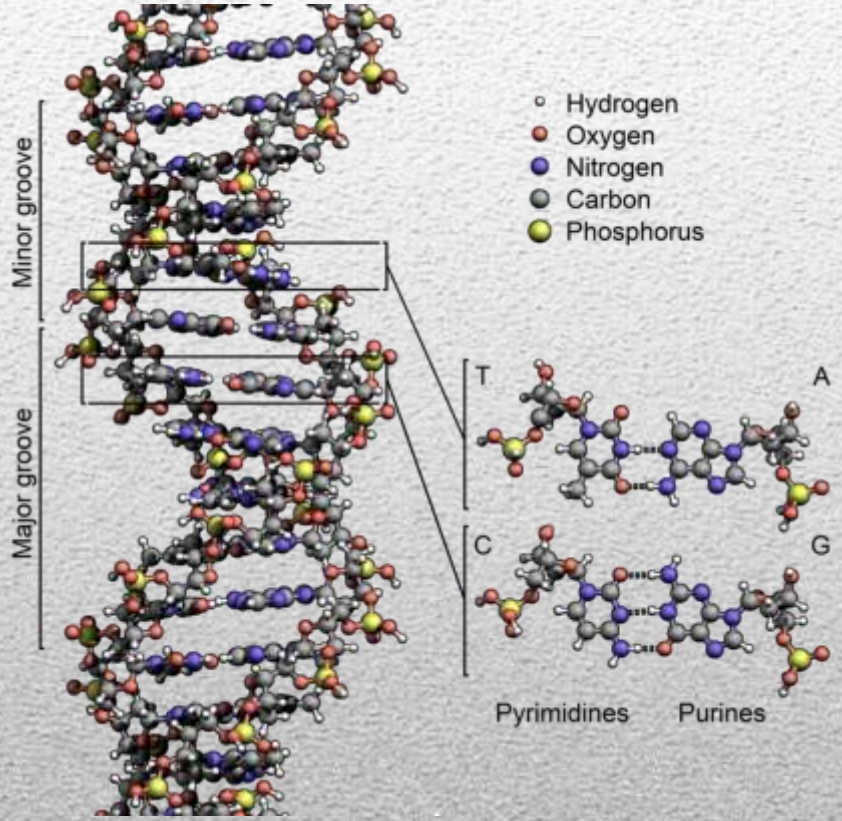
- **Flussi di informazione** (biologia nei viventi, scienze dell'informazione all'interno di processi di calcolo)
- **Manipolaziolazione dell'informazione** (biologia nei viventi, scienze dell'informazione all'interno di processi di calcolo)
- **Organizzazione e rappresentazione dell'informazione** (biologia nei viventi, scienze dell'informazione all'interno di processi di calcolo)

L’INFORMAZIONE NEI VIVENTI

I viventi sono formati da un numero estremamente ampio di molecole. Ognuna di esse è funzionalmente unica e costruita rispettando criteri ben definiti.

Se ogni molecola è costruita seguendo un «progetto» questo implica che, nei viventi, **deve** esistere un luogo il cui ruolo è quello di **immagazzinare** e **rendere disponibili** al momento del bisogno le informazioni relative ai «progetti» delle molecole.

A livello biomolecolare come sono realizzate l'**organizzazione** e la manipolazione dell'informazione?



Depositario informazione: DNA

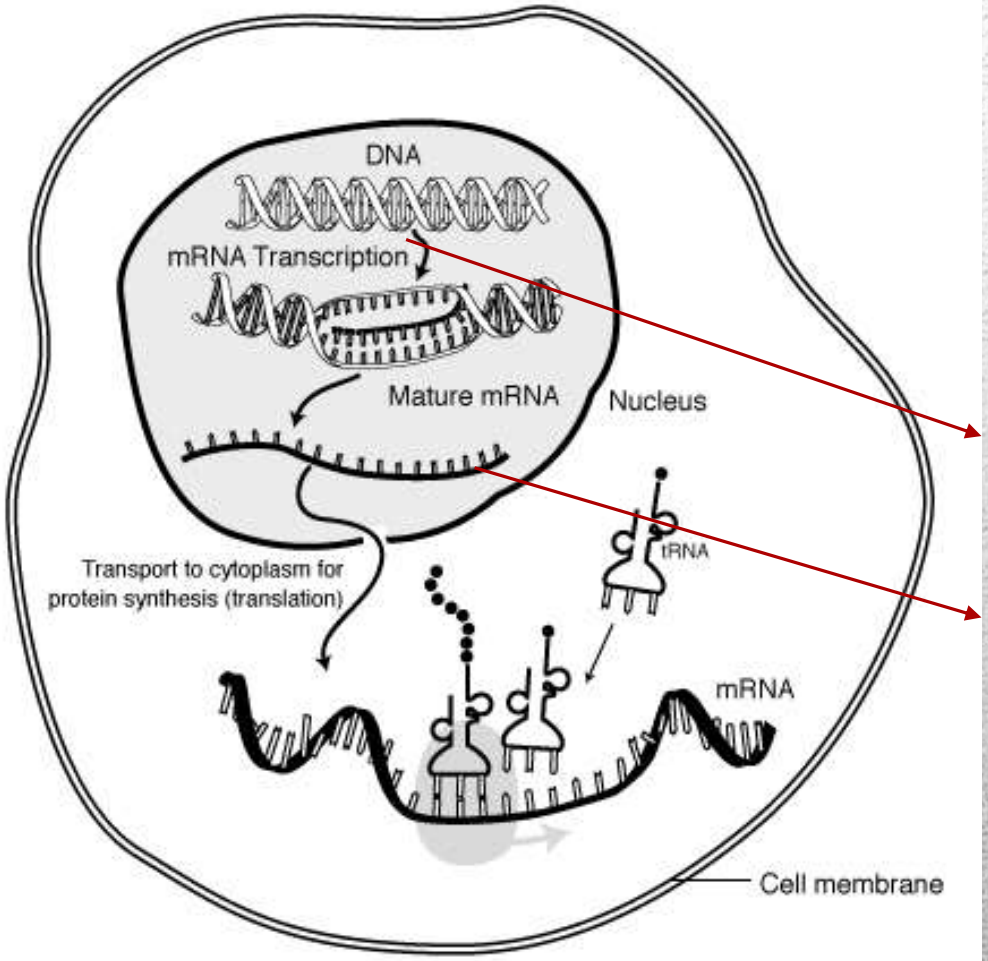
- Doppia elica: ognuna delle due eliche può essere ricostruita a partire dall'informazione presente nell'altra (banca dati con sistema di «backup» incorporato)
- Informazione: ogni elica è una lunga catena formata da una **sequenza** di 4 elementi : adenina (**A**), Timina (**T**), citosina (**C**) e guanina (**G**). Essi vengono detti **nucleotidi**.

…AGCGGAGGAGCATGCGGATTAGGCTTCGGATCGGAT…

…TCGCCTCCTCGTACGCCTAATCCGAAGCCTAGCCTA…

Lunghezza DNA umano? … più di 3 miliardi di caratteri.

A livello biomolecolare come sono realizzate l'**organizzazione** e la **manipolazione** dell'informazione?

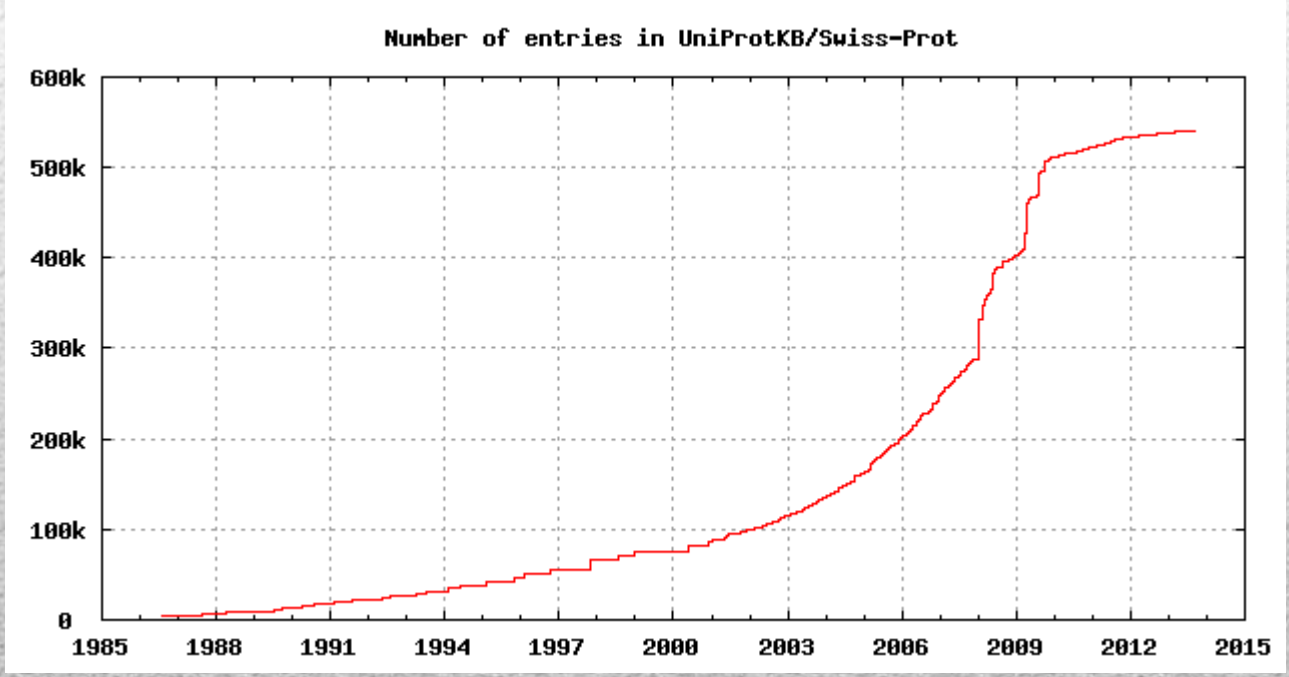


Manipolazione informazione: es. sintesi proteina

- Copia informazione: un tratto di DNA (**gene**) viene copiato in una molecola (**RNA**)in grado di trasportare l'informazione altrove.
- Traduzione: l'RNA che ha portato l'informazione altrove (**messaggero**) viene **tradotto**: viene letto a gruppi di tre lettere, ad ogni tripletta corrisponde uno di 20 aminoacidi. mRNA viene letto e viene sintetizzata una catena di aminoacidi (la **proteina** codificata dal messaggero)

Cosa c'entra tutto questo con la lezione di oggi?

Tutte le biosequenze (DNA, geni, mRNA, proteine,…) vengono immagazzinate in apposite banche dati. La banca dati di riferimento per le sequenze di proteine (**Uniprot** : http://www.uniprot.org/) contiene, al momento, quasi 600.000 proteine.



Dato il numero estremamente elevato di sequenze non è possibile accedere a queste informazioni senza utilizzare strumenti dedicati. Oggi ci occupiamo di questi strmenti.

BANCHE DATI BIOLOGICHE

Le prime banche dati biologiche sono state create nel 1982. In quel periodo i calcolatori erano poco potenti. La possibilità di scambiarsi informazioni utilizzando internet non era diffusa come al giorno d'oggi.

Tuttavia questa possibilità esisteva già nelle università. In quel periodo serviva molto tempo e molto denaro per ottenere informazioni su una biomolecola.

Ogni volta che il progetto dedicato alla caratterizzazione di una biomolecola terminava, tutti i risultati venivano inseriti in semplici **file di testo** e inviati ad una banca dati che rendeva queste «schede informative» pubblicamente disponibili.

BANCHE DATI BIOLOGICHE

Quindi le prime banche dati biologiche erano semplici pagine web che fornivano l'accesso a collezioni di file di testo.

Ogni file di testo aveva un nome che corrispondeva ad un codice identificativo della molecola descritta al suo interno. Ad esempio se il file conteneva informazioni su una proteina avente codice **PR001** allora il file di testo contenente la sua scheda aveva nome **PR001.txt**.

La pratica di assegnare un identificativo ad ogni molecola inserita in una banca dati biologica è utilizzata ancora adesso. L'identificativo viene detto **accession number** ed è una parola contenente **lettere, numeri e simboli** (spazi vuoti non ammessi)

BANCHE DATI BIOLOGICHE

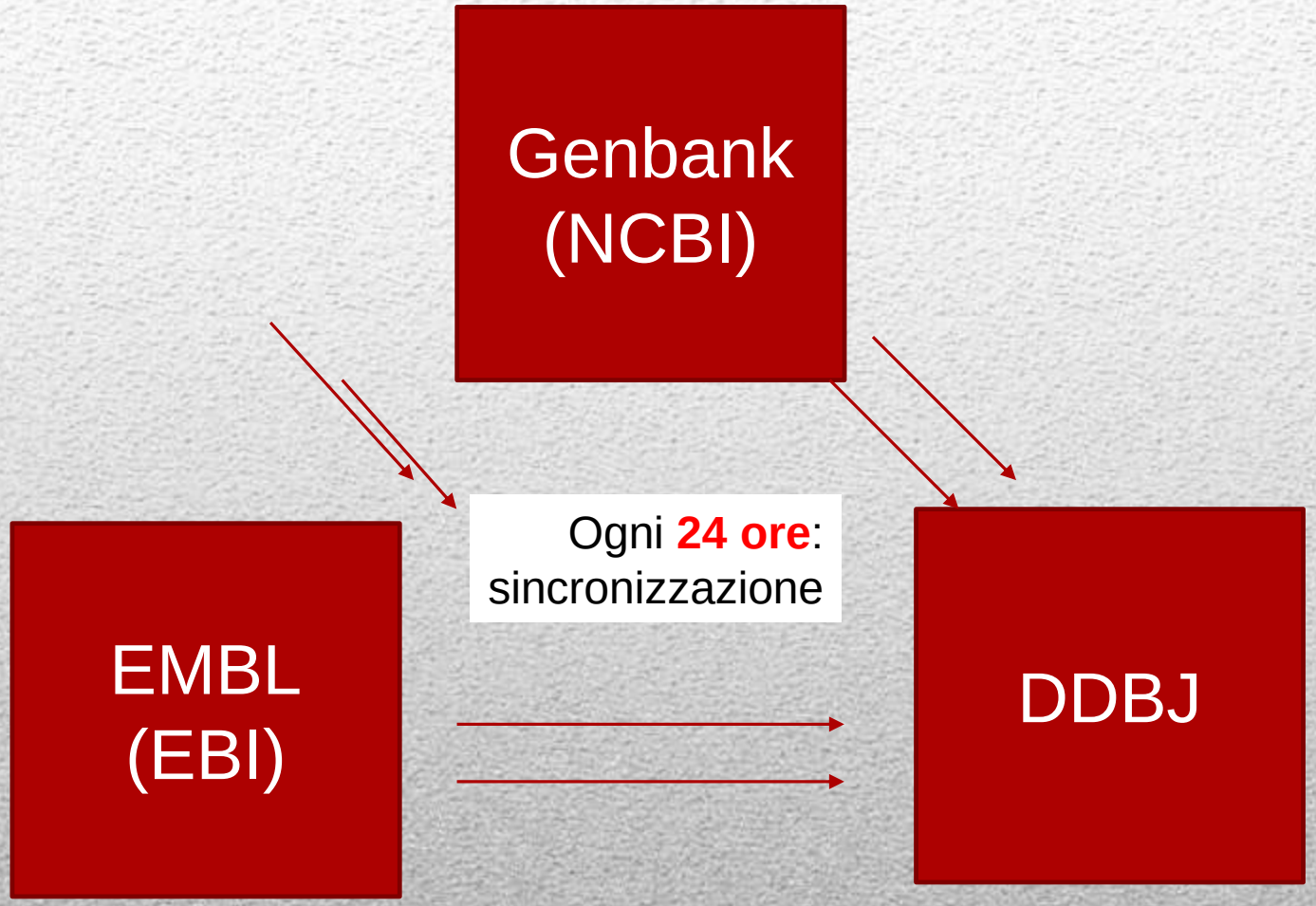
Esistono diversi tipi di banche dati biologiche. Esse si possono classificare in base a vari criteri. Uno di questi riguarda la **qualità delle informazioni** in esse contenute.

Da questo punto di vista possiamo suddividere la banche dati bio in 2 classi principali:

•Banche dati **PRIMARIE** : dette anche collettori primari. Il loro ruolo è quello di raccogliere, giornalmente, tutte le informazioni che riguardano biomolecole prodotte in tutti i laboratori del mondo e renderle disponibili.

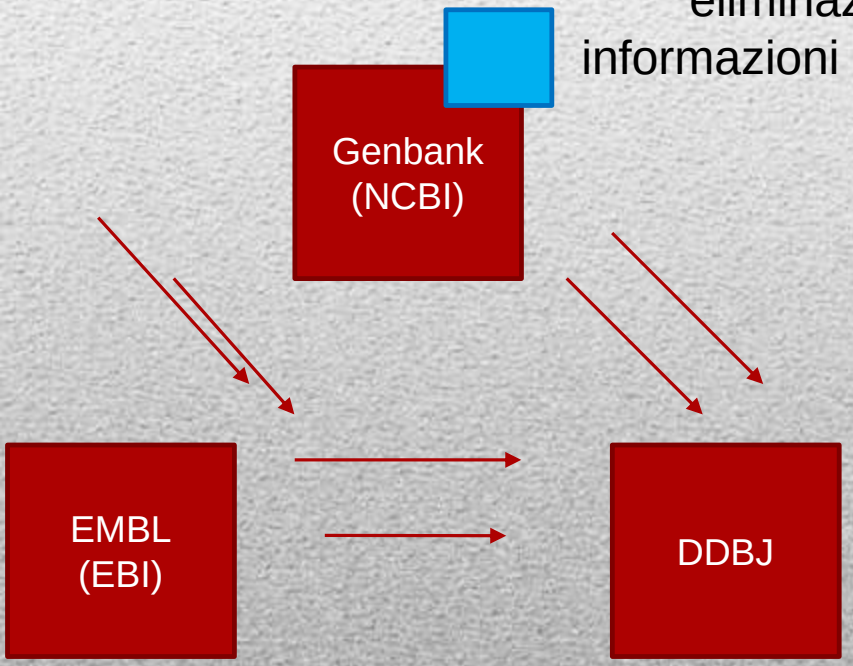
•Banche dati **SECONDARIE**: dato che l'informazione contenuta nei collettori primari è sporca e ridondante esse esaminano i dati dei collettori, correggono eventuali errori, includono informazioni aggiuntive e rendono disponibili i risultati di questo processo di raffinamento.

BANCHE DATI PRIMARIE



… va avanti così dagli anni ’80 … secondo voi quanti dati contengono ?

BANCHE DATI SECONDARIE



Intervento umano: correzione errori ed eliminazione informazioni ridondanti

RefSeq : reference sequences (NCBI)

Uniprot : Universal proteins resource (EBI)

Aggiunta di informazioni:
- Letteratura
- Integrazione di informazioni da diverse banche dati

… va avanti così dagli anni ’80 … secondo voi quanti dati contengono ?

BANCHE DATI PRIMARIE

Queste banche dati contengono, letteralmente, **miliardi di schede**. Sarebbe impossibile trovare quello di cui abbiamo bisogno in assenza di strumenti che permettano di cercare le informazioni a cui siamo interessati.

Questo ci aiuta a capire il motivo per cui le banche dati biologiche non sono **mai** costituite **solamente dalla collezione di dati** che contengono ma anche da un insieme di strumenti progettati per rendere possibile estrazione e manipolazione delle informazioni in esse contente.

… va avanti così dagli anni ’80 … secondo voi quanti dati contengono ?

Banca dati biologica: definizione

Obiettivi:

1. Disseminare dati ed informazioni biologiche

1. Strutturare l'informazione in modo che essa sia leggibile/modificabile da parte di un calcolatore

Una banca dati biologica **DEVE** avere **almeno uno strumento specifico** per la ricerca ed estrazione dei dati.

- Pagine web, libri, articoli scientifici, tabelle, file di testo, e fogli di calcolo **non possono** essere considerati banche dati biologiche.

Liste pubbliche di banche dati biologiche

- Wikipedia (lista di banche dati biologiche)
https://en.wikipedia.org/wiki/List_of_biological_databases
- Nucleic Acids Research Database Listing
http://nar.oupjournals.org/cgi/content/full/30/1/1/DC1 (esempio di pubblicazione in cui è presente una lista di database biologici "storica" … articolo del 2002)

  - Sono un buon punto di partenza per farsi un'idea sul numero e varietà delle banche dati biologiche.
  - **Più di 500** banche dati esistenti sono state catalogate fino ad oggi. E sono costantemente in crescita...

Esempio di accesso a banca dati biologica:

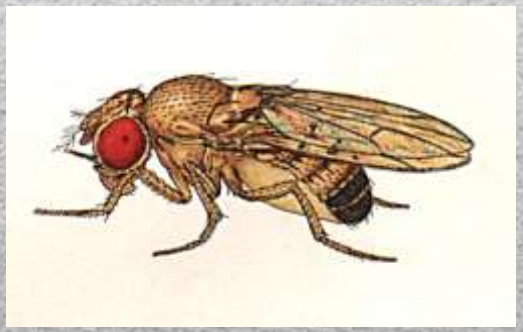
Supponiamo di conoscere il “nome” di un **gene**: **INDY** (ebbene sì … ogni gene ha un suo nome). E di voler cercare informazioni su di esso in una banca dati.

Da dove iniziamo?

Prima di iniziare …

L'effetto delle mutazioni nei geni viene studiato utilizzando organismi facili da manipolare in laboratorio.
In uno di questi organismi, il moscerino della frutta (nome scientifico: Drosophila *melanogaster*) è stato identificato un gene che,se mutato, **raddoppia la durata della vita media dei moscerini**. A questo gene è stato dato il nome di **INDY**:

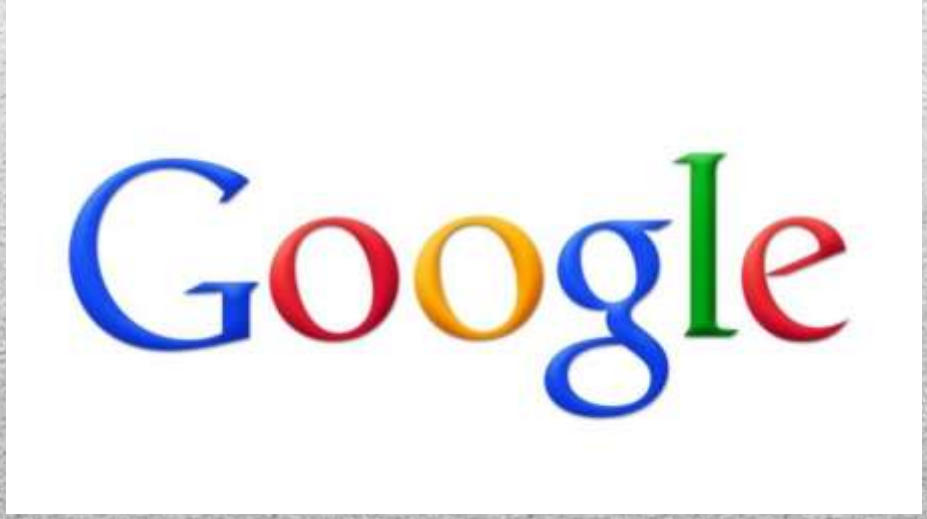
"I'm Not Dead Yet"



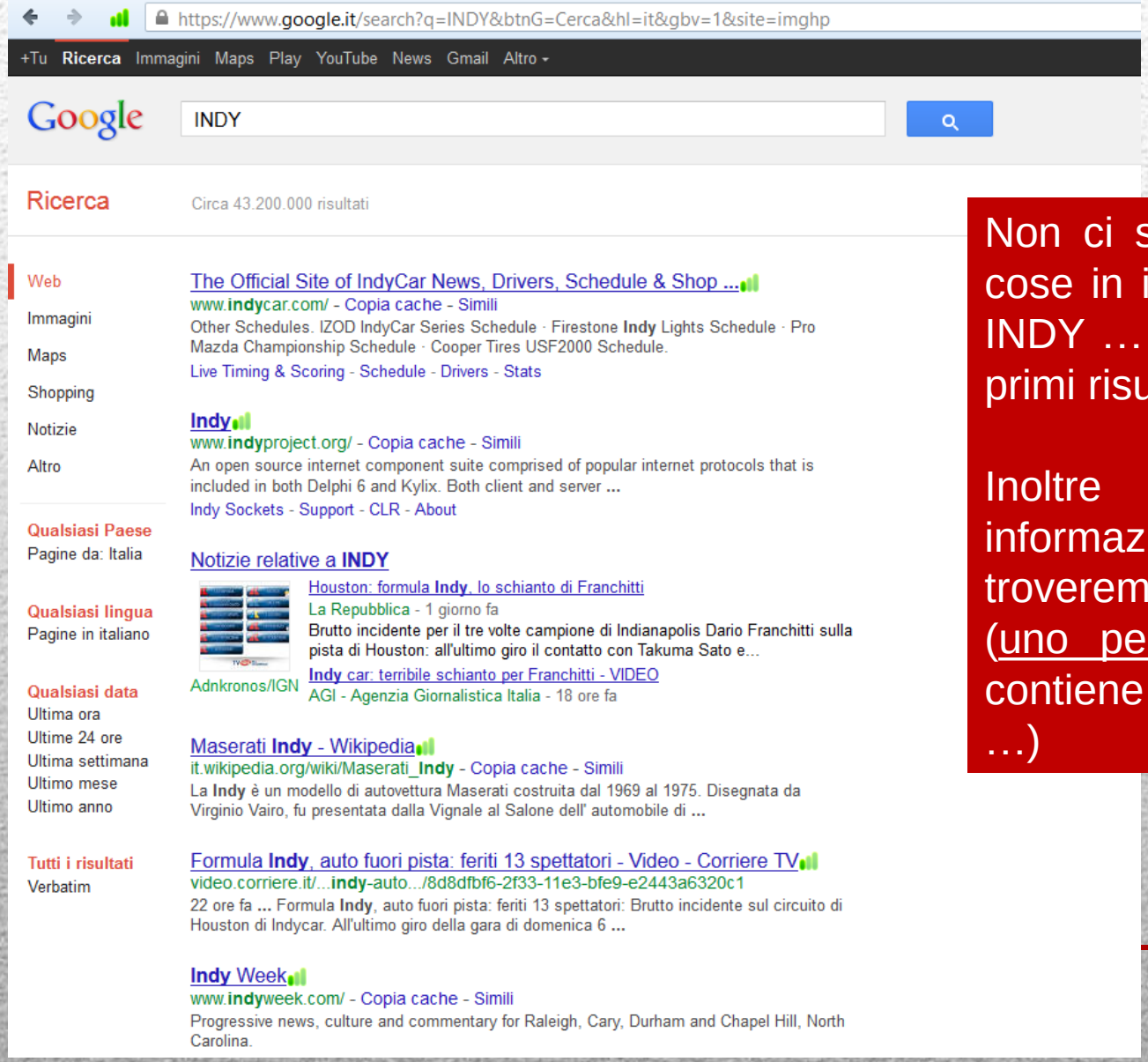
Esempio di accesso a banca dati biologica:

Ricerca di informazioni sul gene: **INDY**

Prima prova: usiamo strumenti “classici” per cercare informazioni …



Ricerca di informazioni sul gene: INDY



Non ci siamo … ci sono troppe cose in internet che si chiamano INDY … e il gene non è uno dei primi risultati riportati.

Inoltre anche se trovassimo informazioni relative al gene troveremmo **molti collegamenti** (uno per ogni banca dati che contiene informazioni sul gene …)

Ricerca di informazioni sul gene: INDY

RICERCA PER PAROLA CHIAVE

Apriamo il web browser e colleghiamoci alla divisione **NUCLEOTIDE** delle banche dati gestite da NCBI: http://www.ncbi.nlm.nih.gov/nuccore

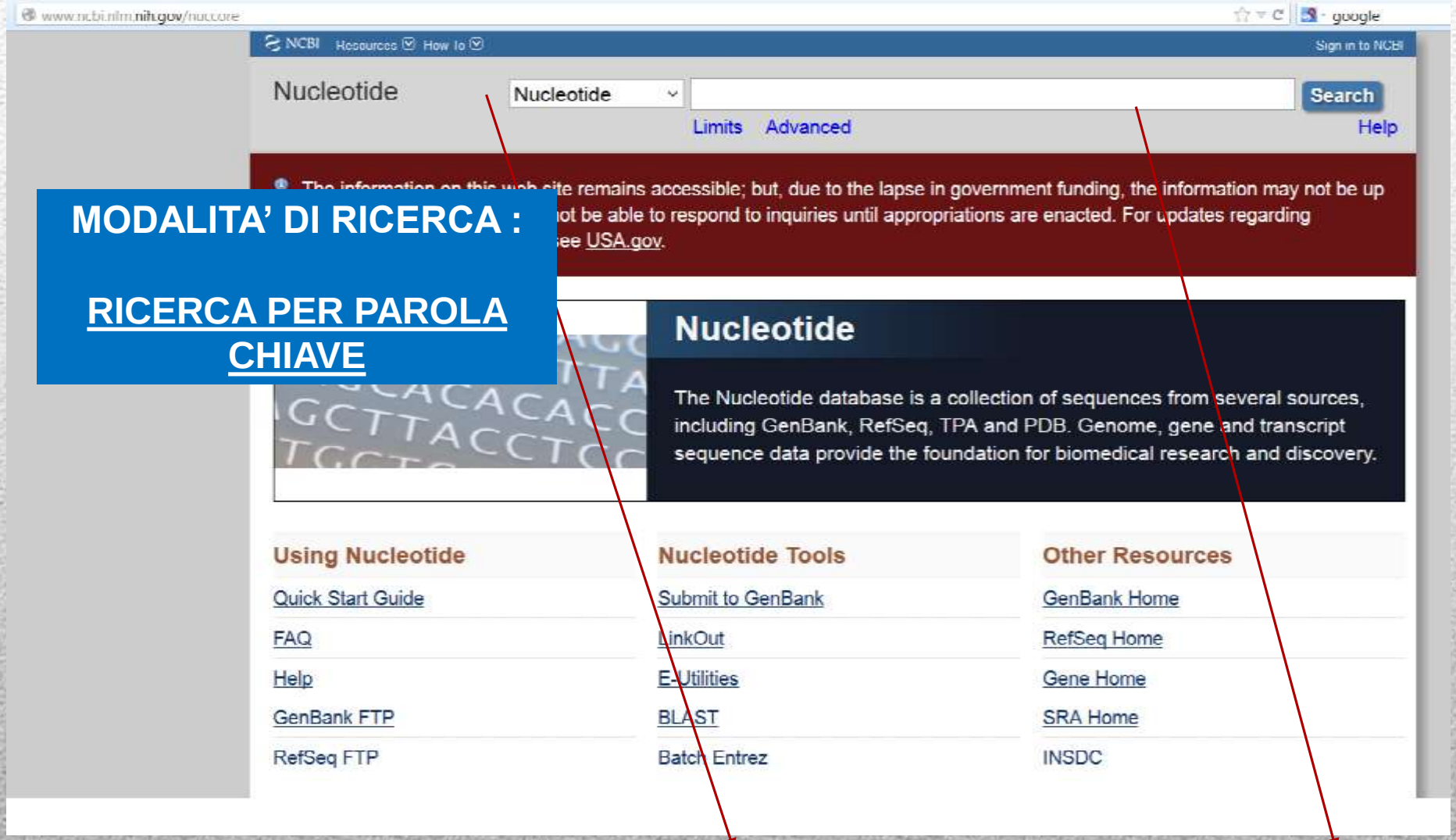
Scegliendo la “sezione” Nucleotide di Genbank otterremo solo risultati riguardanti molecole composte da nucleotidi (DNA o RNA). Non otterremo risultati riguardanti le schede delle proteine.

Tipo di sequenze contenute: DNA e RNA … *NON* Proteine

Ricerca di informazioni sul gene: INDY

MODALITA' DI RICERCA :

RICERCA PER PAROLA CHIAVE

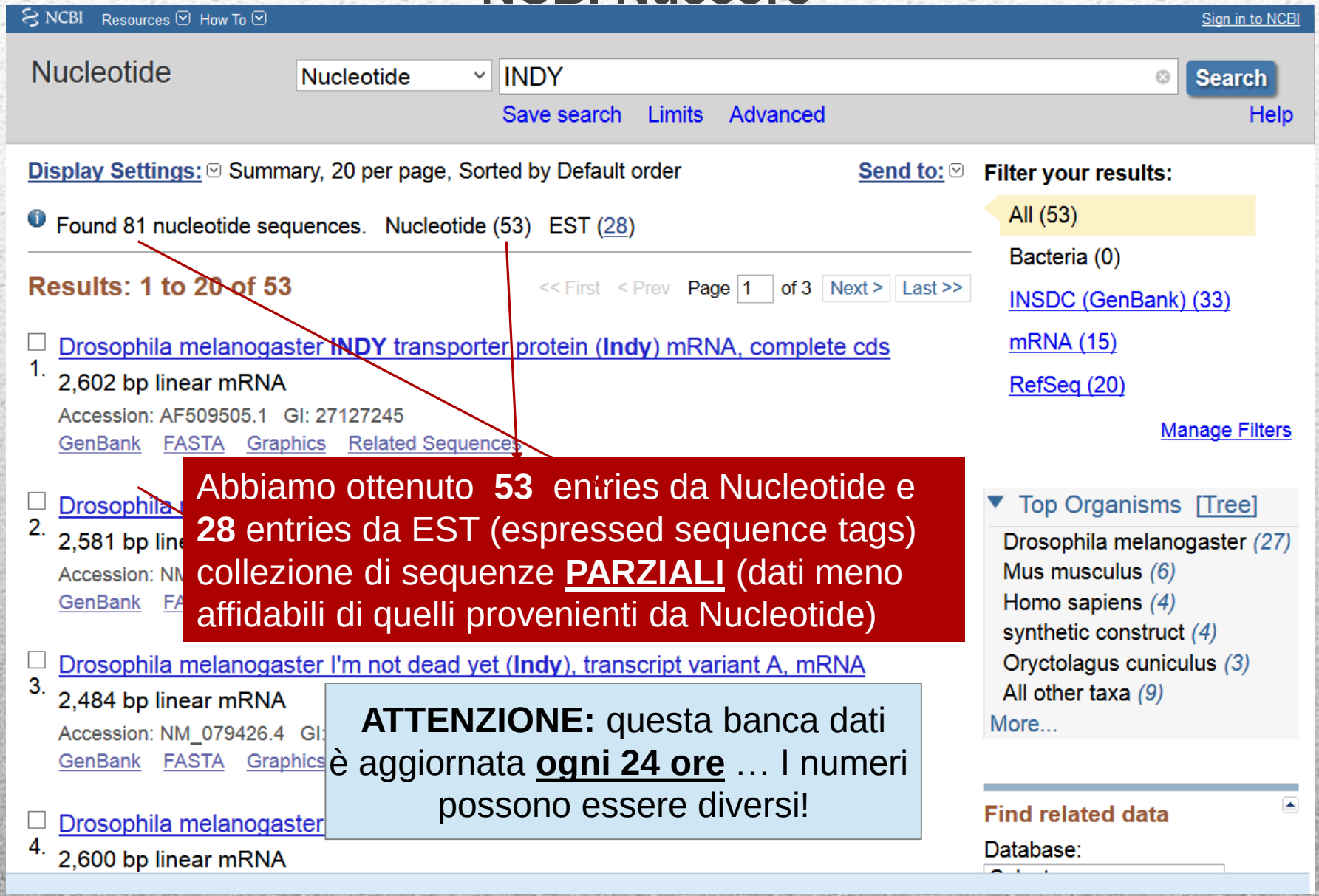


Scrivete **qui** il nome del gene e, poi, premete **search**

gene: INDY, ricerca per parola chiave
NCBI Nuccore

Otteniamo lista di «schede» informative che contengono la parola INDY. **Ogni elemento è un link** che porta ad una singola scheda (entry)

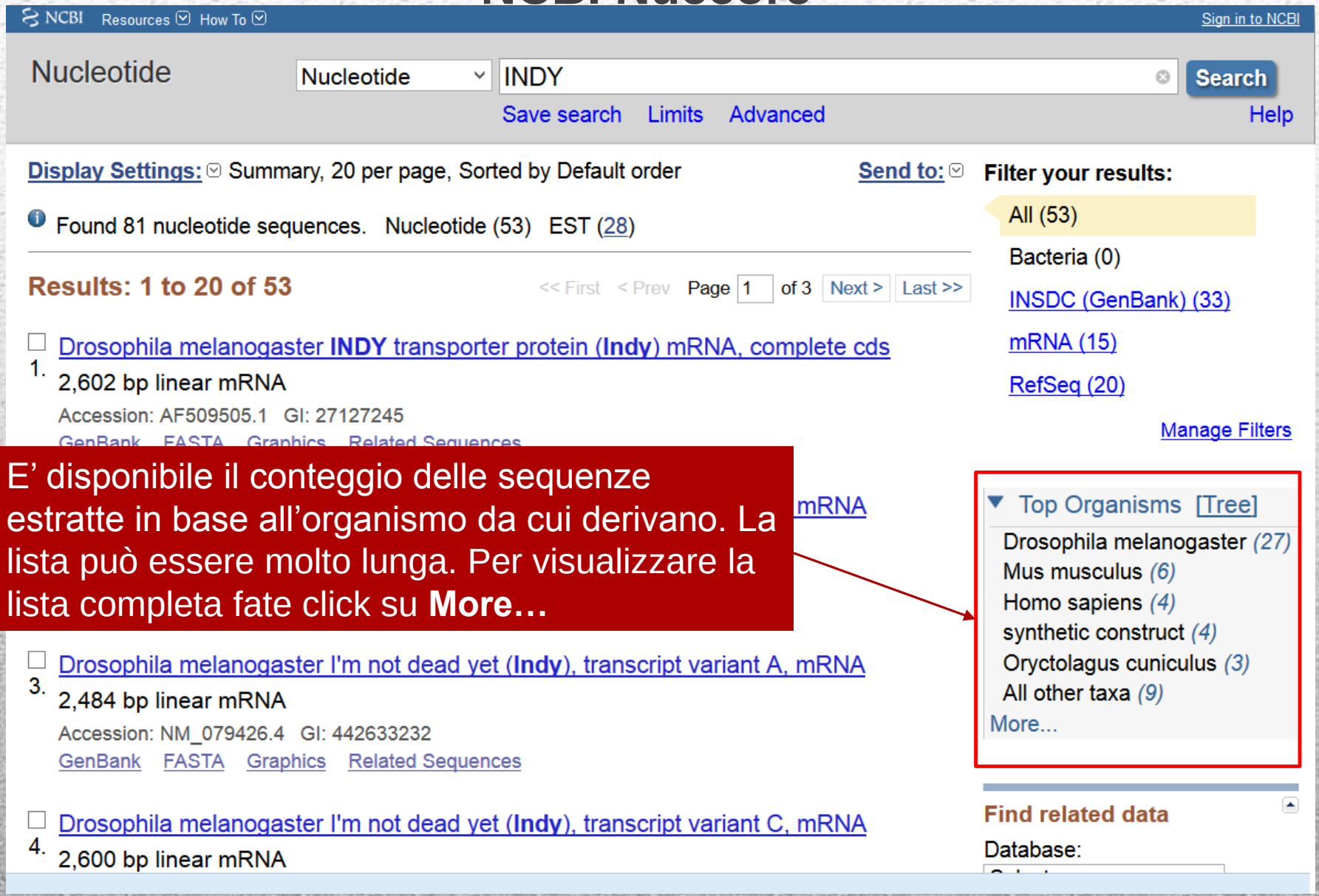
gene: INDY, ricerca per parola chiave
NCBI Nuccore



Abbiamo ottenuto **53** entries da Nucleotide e **28** entries da EST (espressed sequence tags) collezione di sequenze **PARZIALI** (dati meno affidabili di quelli provenienti da Nucleotide)

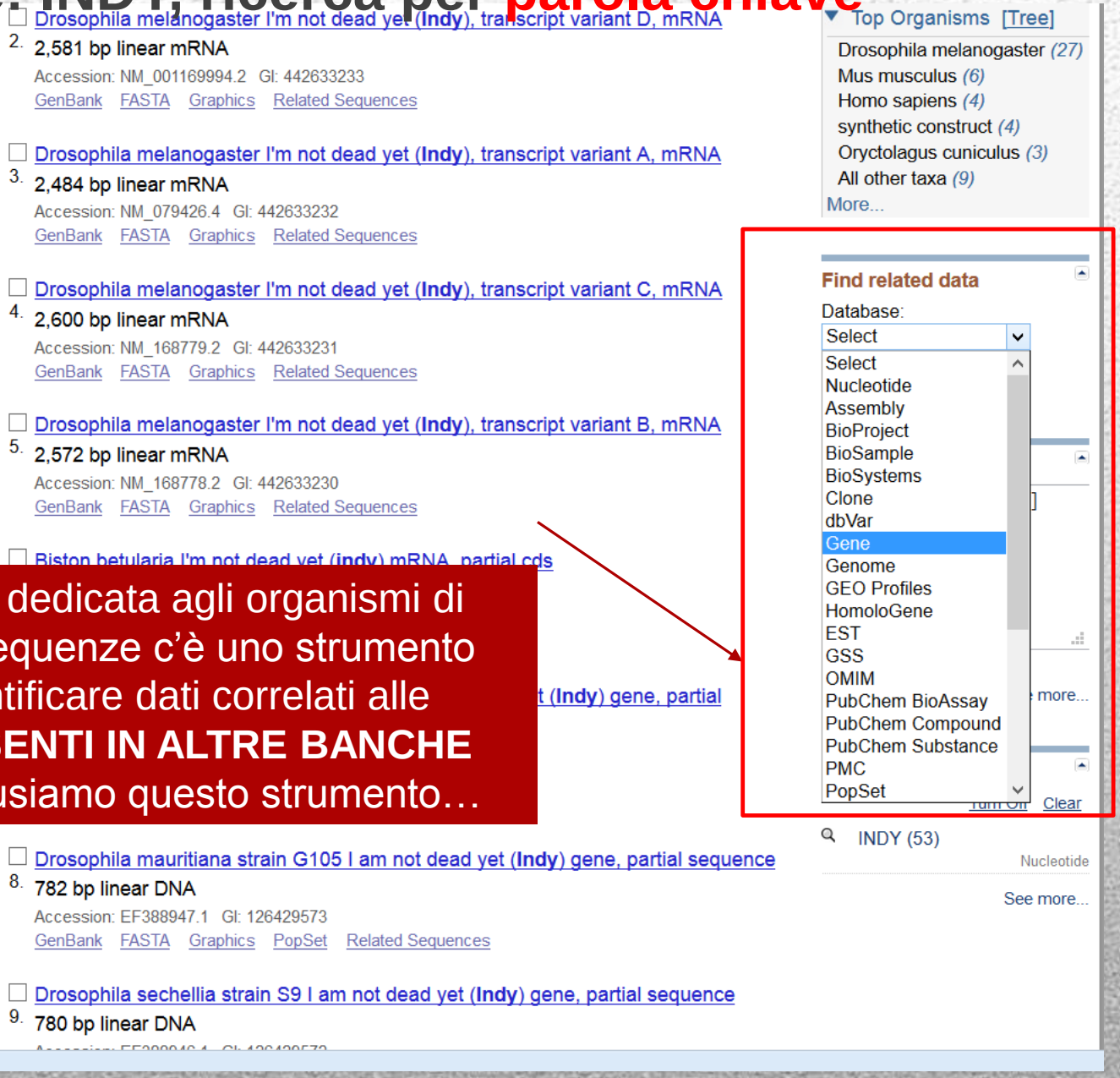
ATTENZIONE: questa banca dati è aggiornata **ogni 24 ore** … I numeri possono essere diversi!

gene: INDY, ricerca per parola chiave
NCBI Nuccore

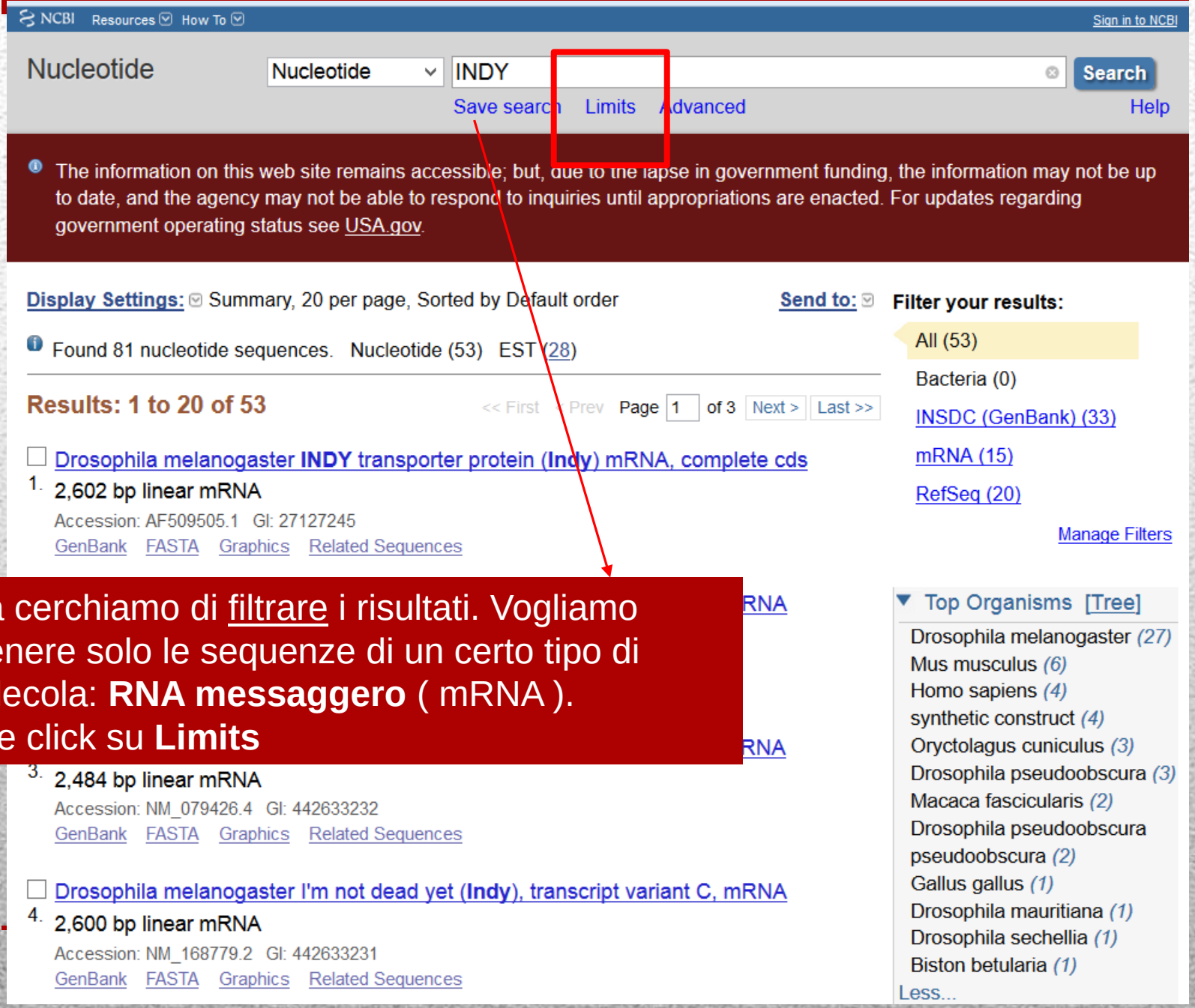


E' disponibile il conteggio delle sequenze estratte in base all'organismo da cui derivano. La lista può essere molto lunga. Per visualizzare la lista completa fate click su **More…**

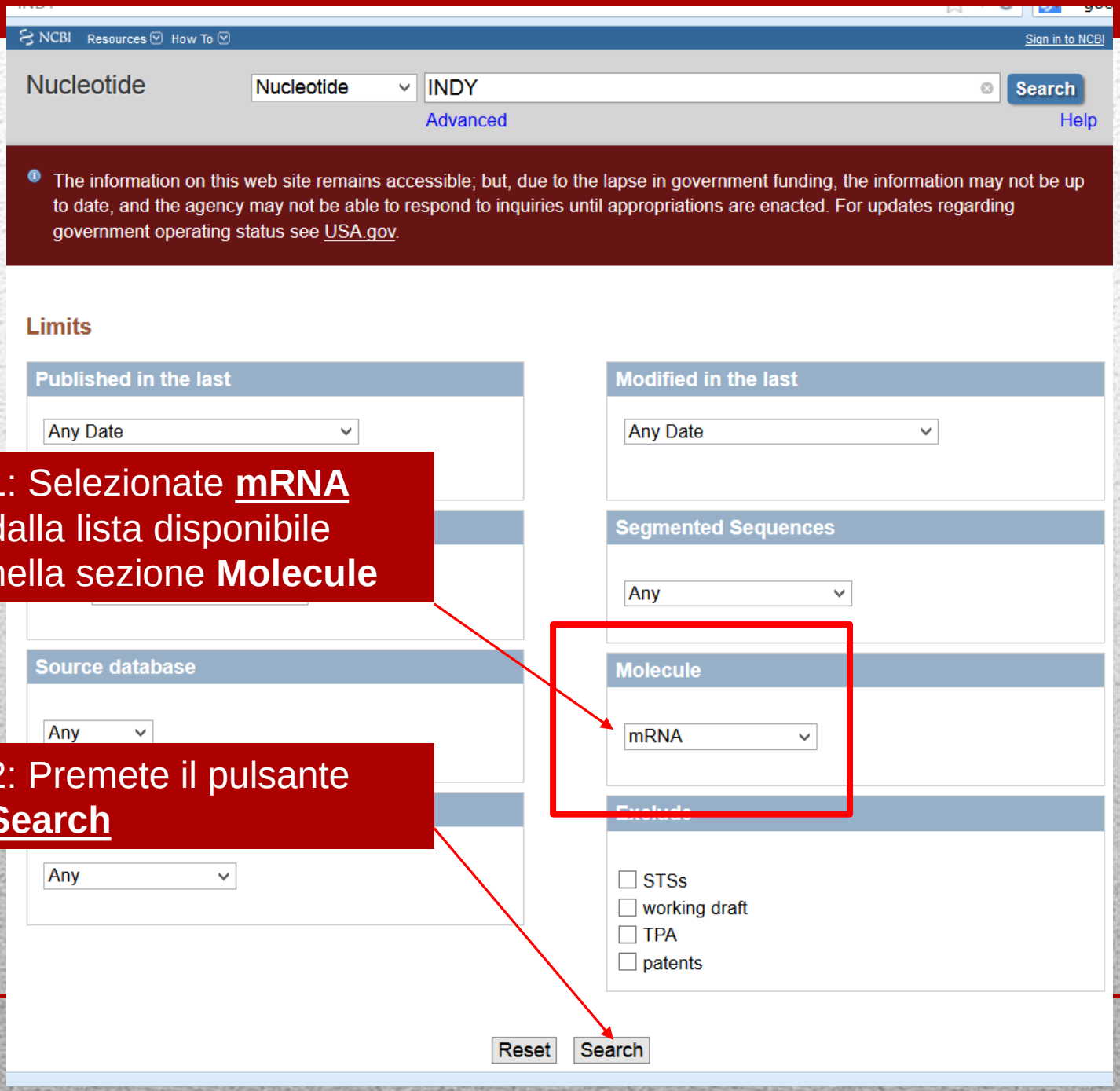
gene: INDY, ricerca per parola chiave

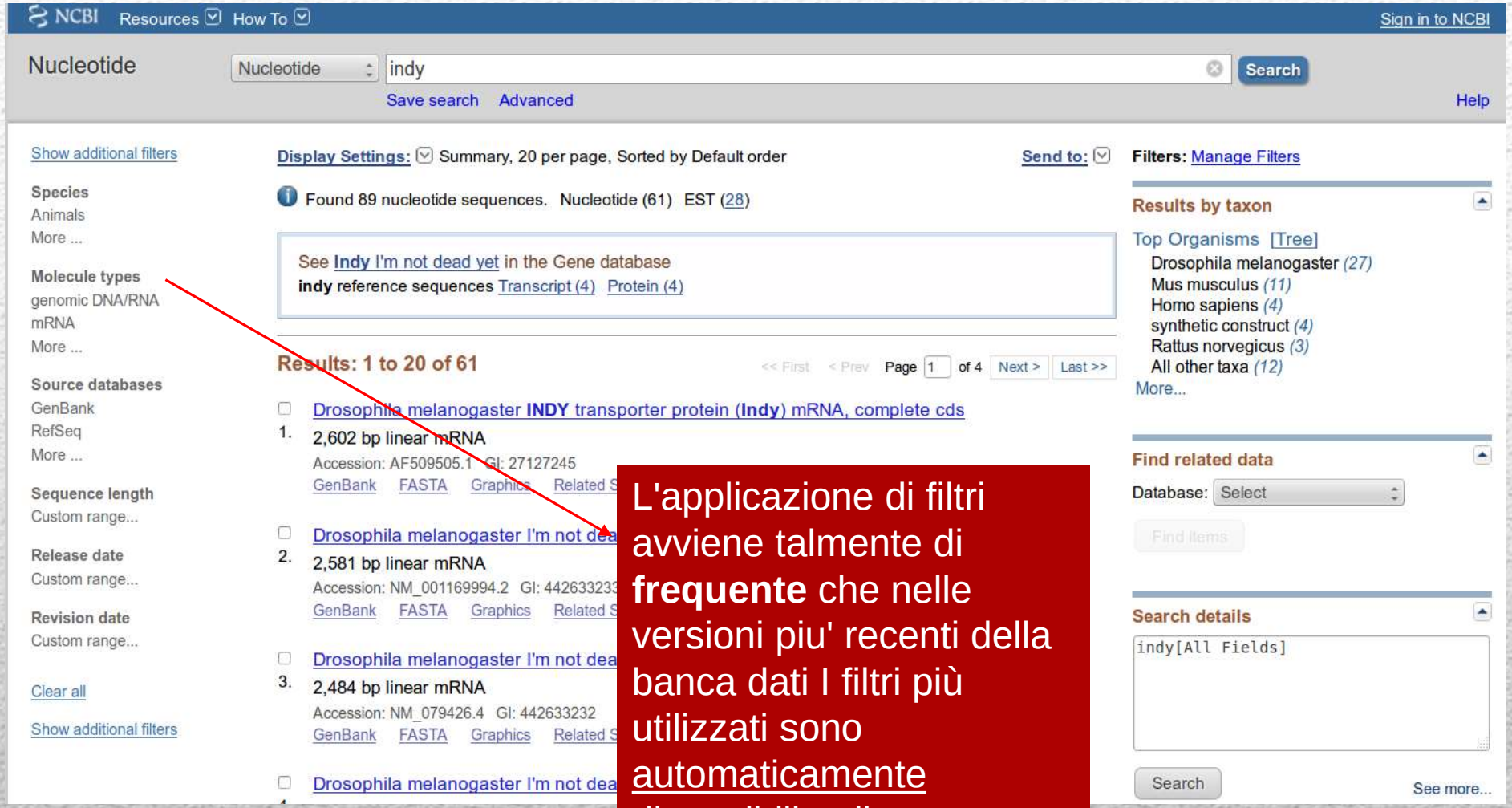


Al di sotto della lista dedicata agli organismi di provenienza delle sequenze c'è uno strumento che permette di identificare dati correlati alle sequenze ma **PRESENTI IN ALTRE BANCHE DATI**. Per ora non usiamo questo strumento…



Ora cerchiamo di filtrare i risultati. Vogliamo ottenere solo le sequenze di un certo tipo di molecola: **RNA messaggero** (mRNA).
Fate click su **Limits**


NCBI Resources How To
Sign in to NCBI
Nucleotide
Nucleotide
INDY
Search
Advanced
Help
The information on this web site remains accessible; but, due to the lapse in government funding, the information may not be up to date, and the agency may not be able to respond to inquiries until appropriations are enacted. For updates regarding government operating status see USA.gov.
Limits
Published in the last
Any Date
Modified in the last
Any Date
Segmented Sequences
Any
Source database
Any
Molecule
mRNA
Any
STSs
working draft
TPA
patents
Reset
Search
1: Selezionate mRNA dalla lista disponibile nella sezione Molecule
2: Premete il pulsante Search



L'applicazione di filtri avviene talmente di **frequente** che nelle versioni piu' recenti della banca dati I filtri più utilizzati sono automaticamente disponibili nella parte sinistra dell'interfaccia web.


Nucleotide
Nucleotide
INDY
Search
Save search
Limits
Advanced
Help
The information on this web site remains accessible; but, due to the lapse in government funding, the information may not be up to date, and the agency may not be able to respond to inquiries until appropriations are enacted. For updates regarding government operating status see USA.gov.
Display Settings: Summary, 20 per page, Sorted by Default order
Limits Activated: Molecule: mRNA
Change | Remove
Questa scrtta ci ricorda che in Limits ci sono elementi attivi
Results: 15
Drosophila melanogaster INDY transporter protein (Indy) mRNA, complete cds
1. 2,602 bp linear mRNA
Accession: AF509505.
GenBank FASTA G
Drosophila melanc
2. 2,581 bp linear mR
Accession: NM_00116
GenBank FASTA Graphics Related Sequences
Nelle versioni + recenti dell'interfaccia web I filtri attivi sono evidenziati in grassetto
Drosophila melanogaster I'm not dead yet (Indy), transcript variant A, mRNA
3. 2,484 bp linear mRNA
Accession: NM_079426.4 GI: 442633232
GenBank FASTA Graphics Related Sequences
Drosophila melanogaster I'm not dead yet (Indy), transcript variant C,
4. 2,600 bp linear mRNA
Accession: NM_168779.2 GI: 442633231
GenBank FASTA Graphics Related Sequences
Drosophila melanogaster I'm not dead yet (Indy), transcript variant B, mRNA
5. 2,572 bp linear mRNA
Accession: NM_168778.2 GI: 442633230
INSDC (GenBank) (5)
mRNA (15)
RefSeq (10)
Manage Filters
Top Organisms [Tree]
Drosophila melanogaster (6)
Homo sapiens (4)
Macaca fascicularis (2)
Mus musculus (1)
Oryctolagus cuniculus (1)
All other taxa (1)
More...
Il numero di risultati è diminuito. Tutte le sequenze ottenute sono di tipo: mRNA
Find related data
Database:
Select

Nucleotide | Nucleotide | INDY | Search

Save search | Limits | Advanced | Help

The information on this web site remains accessible; but, due to the lapse in government funding, the information may not be up to date, and the agency may not be able to respond to inquiries until appropriations are enacted. For updates regarding government operating status see USA.gov.

Display Settings: Summary, 20 per page, Sorted by Default order | Send to:

Limits Activated: Molecule: mRNA Change | Remove

Results: 15

Drosophila melanogaster **INDY** transporter protein (**Indy**) mRNA, complete cds

Nonostante il filtro i risultati ottenuti corrispondono a sequenze appartenenti a più organismi. Siamo interessati solo alle sequenze di **Drosophila melanogaster** (il moscerino della frutta … un noto organismo utilizzato in laboratorio).
Selezionamo l'organismo da questa lista

, mRNA

, mRNA

Accession: NM_079426.4 GI: 442633232
GenBank FASTA Graphics Related Sequences

4. Drosophila melanogaster I'm not dead yet (**Indy**), transcript variant C, mRNA
2,600 bp linear mRNA
Accession: NM_168779.2 GI: 442633231
GenBank FASTA Graphics Related Sequences

5. Drosophila melanogaster I'm not dead yet (**Indy**), transcript variant B, mRNA
2,572 bp linear mRNA
Accession: NM_168778.2 GI: 442633230

Filter your results:

All (15)
Bacteria (0)
INSDC (GenBank) (5)
mRNA (15)
RefSeq (10)
Manage Filters

Top Organisms [Tree]
Drosophila melanogaster *(6)*
Homo sapiens *(4)*
Macaca fascicularis *(2)*
Mus musculus *(1)*
Oryctolagus cuniculus *(1)*
All other taxa *(1)*
More...

Analyze these sequences
Run BLAST

Find related data
Database:
Select

I risultati ottenuti derivano da più banche dati. Vogliamo che i risultati provengano da **una sola** banca dati.

Selezioniamo RefSeq nella left bar dedicata ai filtri (nella sezione **source databases**)

Ora, tutti i risultati ottenuti corrispondono a sequenze che appartengono ad un solo organismo.

gene: INDY, ricerca per parola chiave

ESERCIZIO 1 :

Al momento abbiamo ottenuto la lista di sequenze presenti in NCBI Nucleotide che :

- Contengono **INDY** nella loro sceda descrittiva
- Corrispondono a molecole di mRNA
- Corrispondono a sequenze del moscerino della frutta

Rispondete a queste domande:

- **Quante sono** le sequenze ottenute?
- Quante di esse sono presenti nella banca dati secondaria **RefSeq**?

gene: INDY, ricerca per parola chiave

ESERCIZIO 2 :

Limitare i risultati ottenuti alle sole sequenze che appartengono alla banca dati secondaria **RefSeq.**

Suggerimento: cercate qualcosa che vi permetta di restringere il numero di banche dati sulle quali viene effettuata l'estrazione.

Domanda (a cui rispondere dopo aver risolto l'esercizio) :

•Quante sequenze avete ottenuto?

gene: INDY, ricerca per parola chiave

ESERCIZIO 3 :

Dopo aver risolto l'esercizio 2 cercate informazioni collegate alle sequenze ottenute **in un'altra banca dati NCBI.** Per questo esercizio cercate informazioni provenienti dalla banca dati **Gene**.

Suggerimento: cercate in una delle slide precedenti il modo di raggiungere la lista dei risultati collegati alla lista di sequenze corrente ma presenti nella banca dati **Gene**. Prima di procedere selezionate alcune delle sequenze RefSeq di partenza. **EVITATE** di selezionare sequenze RefSeq la cui descrizione contiene la parola "chromosome" o la parola PREDICTED.

Domande (a cui rispondere dopo aver risolto l'esercizio) :

- Quanti geni ottenete?
- A che organismo appartengono?
- Che nome ha il/i geni ottenuti?

Modalità di visualizzazione dei risultati

Chi è stato attento ha notato che, nelle slide precedenti, a volte descrivevamo i risultati riferendoci ad essi come «sequenze». Questo sembra suggerire che per ogni risultato ottenuto il link che punta ad esso restituisca la sequenza della molecola.

Sarà davvero così?

Selezionate il link che punta a :

Drosophila melanogaster I'm not dead yet (Indy), transcript variant A, mRNA

Modalità di visualizzazione dei risultati : ENTRY

Drosophila melanogaster I'm not dead yet (Indy), transcript variant A, mRNA

NCBI Reference Sequence: NM_079426.4

FASTA Graphics

Go to:

```
LOCUS       NM_079426               2484 bp    mRNA    linear   INV 16-JAN-2013
DEFINITION  Drosophila melanogaster I'm not dead yet (Indy), transcript variant
            A, mRNA.
ACCESSION   NM_079426
VERSION     NM_079426.4  GI:442633232
KEYWORDS    RefSeq.
SOURCE      Drosophila melanogaster (fruit fly)
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1  (bases 1 to 2484)
  AUTHORS   Hoskins,R.A., Carlson,J.W., Kennedy,C., Acevedo,D., Evans-Holm,M.,
            Frise,E., Wan,K.H., Park,S., Mendez-Lago,M., Rossi,F.,
            Villasante,A., Dimitri,P., Karpen,G.H. and Celniker,S.E.
  TITLE     Sequence finishing and mapping of Drosophila melanogaster
            heterochromatin
  JOURNAL   Science 316 (5831), 1625-1628 (2007)
   PUBMED   17569867
```

Titolo , banca dati (RefSeq) e identificativo nella banca dati

Lunghezza molecola, tipo molecola, data ultimo aggiornamento

Identificativo sequenza (ACCESSION), organismo di provenienza

Lista pubblicazioni che parlano di questa sequenza (elenco può essere lungo …)

Modalità di visualizzazione ENTRY :

La entries sono simili a schede informative.

Le entries sono composte da **1 sezione** principale (descritta nella slide precedente) e **3 sezioni aggiuntive**:

- **Comment** : contiene commenti inseriti dai curatori della banca dati
- **Features** : Caratteristiche della sequenza, informazioni generali (che valgono per l'intera sequenza, e informazioni su **tratti specifici della sequenza** (ad es. note riguardanti il tratto di sequenza che va dal nucleotide 10 al nucleotide 20).
- **Sequence** : La sequenza vera e propria

Potete scegliere di andare in una sezione specifica della entry utilizzando il menù **Go to** posto immediatamente all'inizio della entry stessa.

Sezione ENTRY : Feature table

```
FEATURES             Location/Qualifiers
     source          1..2484
                     /organism="Drosophila melanogaster"
                     /mol_type="mRNA"
                     /db_xref="taxon:7227"
                     /chromosome="3L"
                     /genotype="y[1]; cn[1] bw[1] sp[1]; Rh6[1]"
     gene            1..2484
                     /gene="Indy"
                     /locus_tag="Dmel_CG3979"
                     /gene_synonym="anon-EST:fe3A7; anon-WO0172774.70;
                     BEST:LP01220; CG3979; Dmel\CG3979; drIndy; indy; INDY"
                     /note="I'm not dead yet"
                     /map="75E1-75E2"
                     /db_xref="FLYBASE:FBgn0036816"
                     /db_xref="GeneID:40049"
     CDS             175..1893
                     /gene="Indy"
                     /locus_tag="Dmel_CG3979"
                     /gene_synonym="anon-EST:fe3A7; anon-WO0172774.70;
                     BEST:LP01220; CG3979; Dmel\CG3979; drIndy; indy; INDY"
                     /note="CG3979 gene product from transcript CG3979-RA;
```

intera sequenza

Coordinate

intera sequenza

gene

nome feature

parte di sequenza

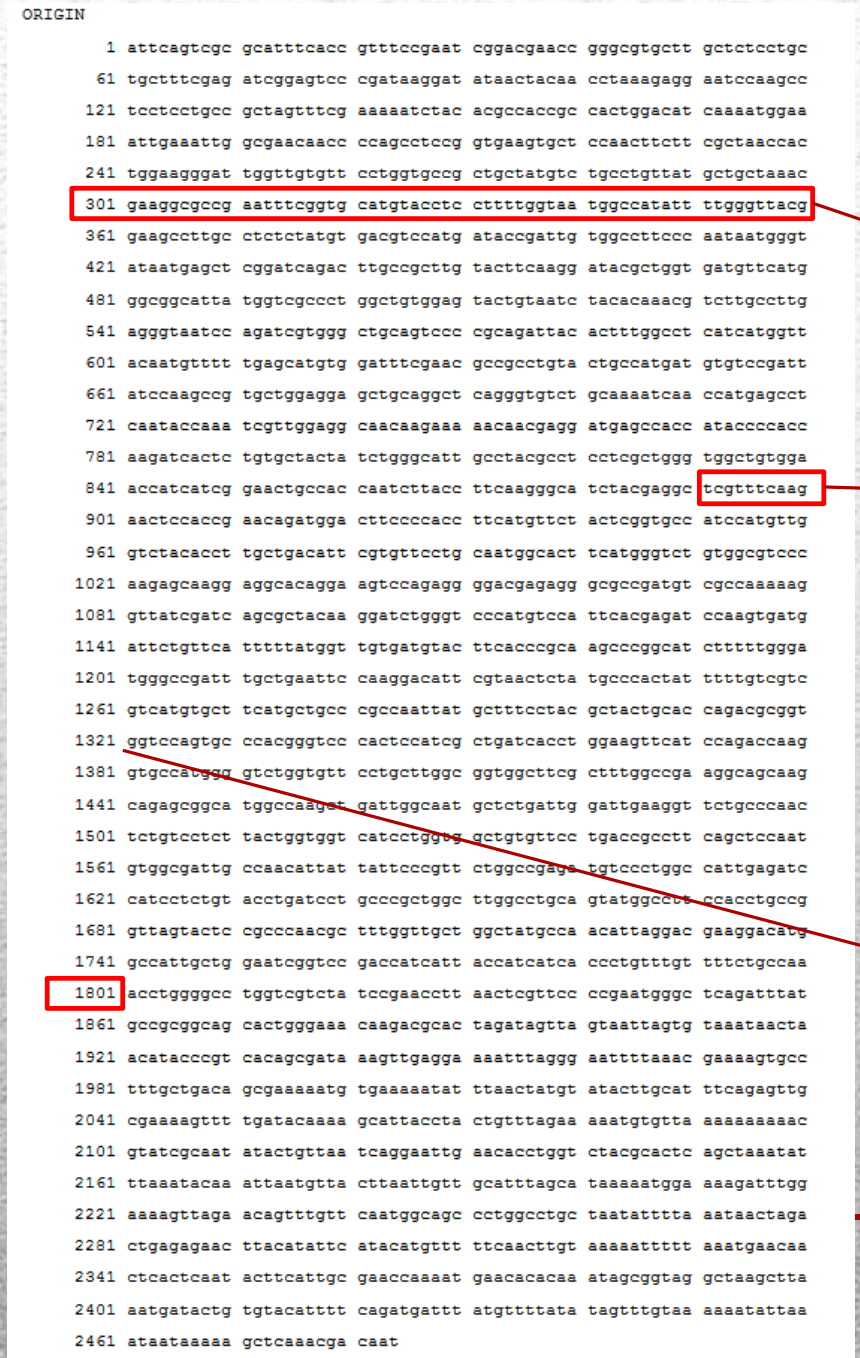
CDS

hyperlink

CDS (coding sequence):

È la parte di RNA che viene tradotta in proteina.

Sezione ENTRY : Sequence



```
ORIGIN
        1 attcagtcgc gcatttcacc gtttccgaat cggacgaacc gggcgtgctt gctctcctgc
       61 tgctttcgag atcggagtcc cgataaggat ataactacaa cctaaagagg aatccaagcc
      121 tcctcctgcc gctagtttcg aaaaatctac acgccaccgc cactggacat caaaatggaa
      181 attgaaattg gcgaacaacc ccagcctccg gtgaagtgct ccaacttctt cgctaaccac
      241 tggaagggat tggttgtgtt cctggtgccg ctgctatgtc tgcctgttat gctgctaaac
      301 gaaggcgccg aatttcggtg catgtacctc cttttggtaa tggccatatt ttgggttacg
      361 gaagccttgc ctctctatgt gacgtccatg ataccgattg tggccttccc aataatgggt
      421 ataatgagct cggatcagac ttgccgcttg tacttcaagg atacgctggt gatgttcatg
      481 ggcggcatta tggtcgccct ggctgtggag tactgtaatc tacacaaacg tcttgccttg
      541 agggtaatcc agatcgtggg ctgcagtccc cgcagattac actttggcct catcatggtt
      601 acaatgtttt tgagcatgtg gatttcgaac gccgcctgta ctgccatgat gtgtccgatt
      661 atccaagccg tgctggagga gctgcaggct cagggtgtct gcaaaatcaa ccatgagcct
      721 caataccaaa tcgttggagg caacaagaaa aacaacgagg atgagccacc ataccccacc
      781 aagatcactc tgtgctacta tctgggcatt gcctacgcct cctcgctggg tggctgtgga
      841 accatcatcg gaactgccac caatcttacc ttcaagggca tctacgaggc tcgtttcaag
      901 aactccaccg aacagatgga cttccccacc ttcatgttct actcggtgcc atccatgttg
      961 gtctacacct tgctgacatt cgtgttcctg caatggcact tcatgggtct gtggcgtccc
     1021 aagagcaagg aggcacagga agtccagagg ggacgagagg gcgccgatgt cgccaaaaag
     1081 gttatcgatc agcgctacaa ggatctgggt cccatgtcca ttcacgagat ccaagtgatg
     1141 attctgttca tttttatggt tgtgatgtac ttcacccgca agcccggcat ctttttggga
     1201 tgggccgatt tgctgaattc caaggacatt cgtaactcta tgcccactat ttttgtcgtc
     1261 gtcatgtgct tcatgctgcc cgccaattat gctttcctac gctactgcac cagacgcggt
     1321 ggtccagtgc ccacgggtcc cactccatcg ctgatcacct ggaagttcat ccagaccaag
     1381 gtgccatggg gtctggtgtt cctgcttggc ggtggcttcg ctttggccga aggcagcaag
     1441 cagagcggca tggccaagat gattggcaat gctctgattg gattgaaggt tctgcccaac
     1501 tctgtcctct tactggtggt catcctgggg gctgtgttcc tgaccgcctt cagctccaat
     1561 gtggcgattg ccaacattat tatcccgtt ctggccgaga tgtccctggc cattgagatc
     1621 catcctctgt acctgatcct gcccgctggc ttggcctgca gtatggcctt ccacctgccg
     1681 gttagtactc cgcccaacgc tttggttgct ggctatgcca acattaggac gaaggacatg
     1741 gccattgctg gaatcggtcc gaccatcatt accatcatca ccctgtttgt tttctgccaa
     1801 acctggggcc tggtcgtcta tccgaacctt aactcgttcc ccgaatgggc tcagatttat
     1861 gccgcggcag cactgggaaa caagacgcac tagatagtta gtaattagtg taaataacta
     1921 acatacccgt cacagcgata aagttgagga aaatttaggg aattttaaac gaaaagtgcc
     1981 tttgctgaca gcgaaaaatg tgaaaaatat ttaactatgt atacttgcat ttcagagttg
     2041 cgaaaagttt tgatacaaaa gcattaccta ctgtttagaa aaatgtgtta aaaaaaaaac
     2101 gtatcgcaat atactgttaa tcaggaattg aacacctggt ctacgcactc agctaaatat
     2161 ttaaatacaa attaatgtta cttaattgtt gcatttagca taaaaatgga aaagatttgg
     2221 aaaagttaga acagtttgtt caatggcagc cctggcctgc taatatttta aataactaga
     2281 ctgagagaac ttacatattc atacatgttt ttcaacttgt aaaaatttttt aaatgaacaa
     2341 ctcactcaat acttcattgc gaaccaaaat gaacacacaa atagcggtag gctaagctta
     2401 aatgatactg tgtacatttt cagatgattt atgttttata tagtttgtaa aaaatattaa
     2461 ataataaaaa gctcaaacga caat
```

Ogni riga contiene 60 caratteri (in questo caso nucleotidi)…

Divisi in gruppi di 10 caratteri (per facilitare conteggi)

Ogni riga inizia con il numero del primo carattere (nucleotide) della riga stessa

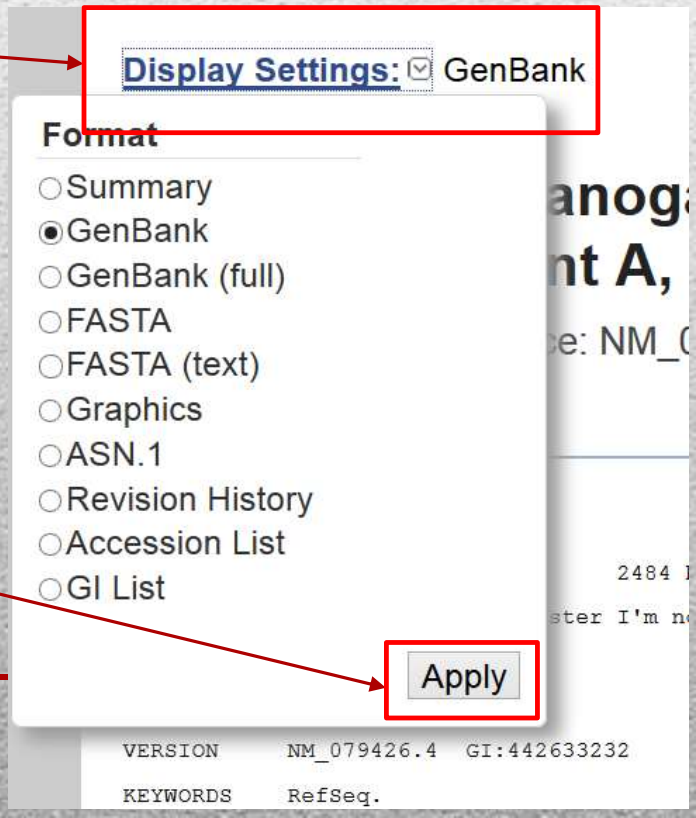
Modalità di visualizzazione alternative :

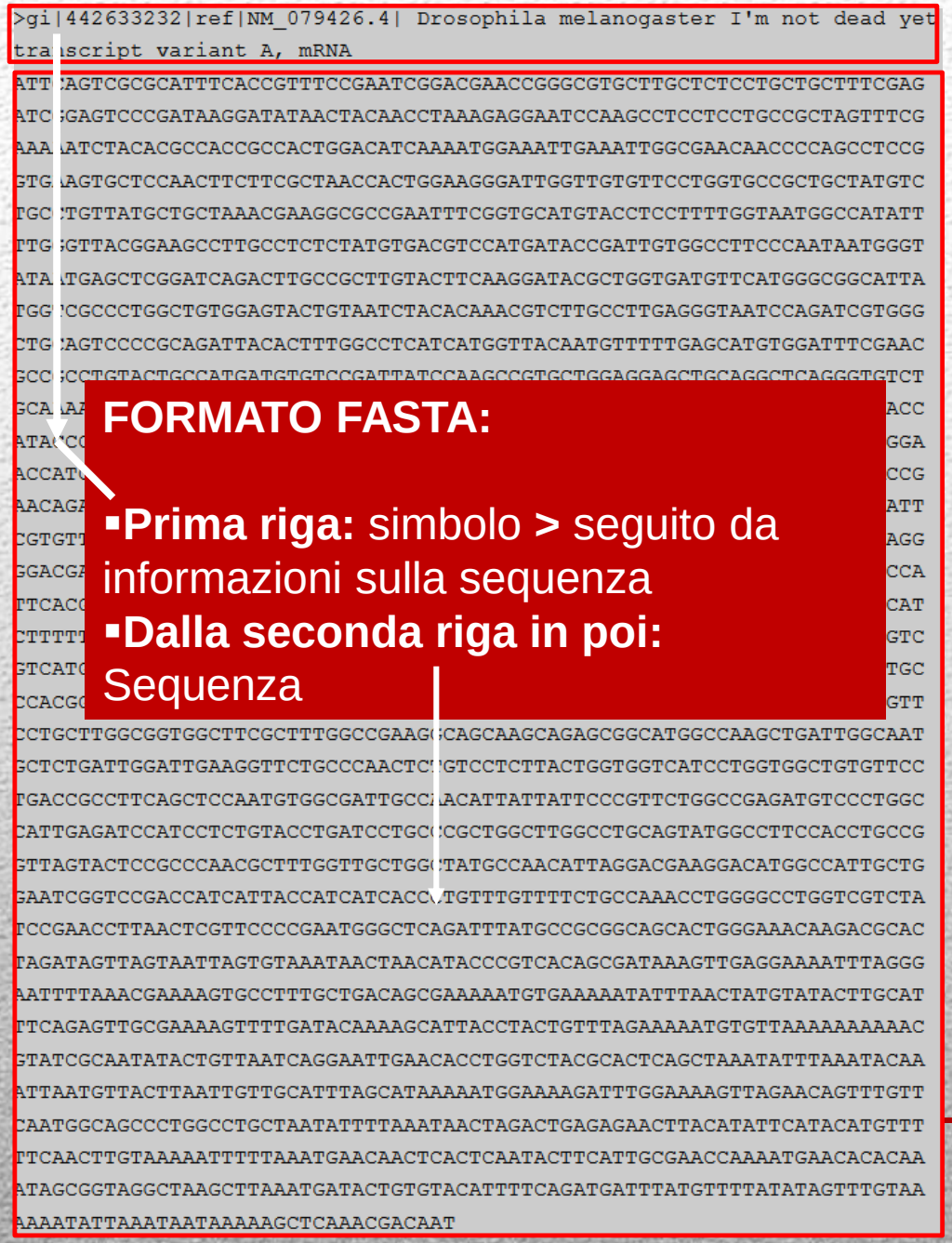
Oltre a visualizzare le informazioni sulla sequenza in modalità ENTRY (a volte detta modalità GenBank) è possibile visualizzare le informazioni in modo diverso.

E' possibile selezionare la modalità di visualizzazione grazie al menù **Display Settings** posto immediatamente all'inizio della entry.

- Summary: solo informazioni principali
- FASTA / FASTA(text): solo sequenza
- GenBank (full) informazioni estese

Dopo aver selezionato la modalità di display premete il pulsante **Apply**. Provate Fasta(text).





Display mode: FASTA (text)

Il formato di visualizzazione FASTA (text) mostra solo la sequenza (senza spazi vuoti e senza numeri). Inoltre **non contiene caratteri invisibili** (a differenza delle altre modalità di visualizzazione della sequenza).

In questo formato la sequenza può essere usata come input per programmi che effettuano analisi su di essa (ad esempio composizione: frequenza caratteri A,C,G e T in questo caso)

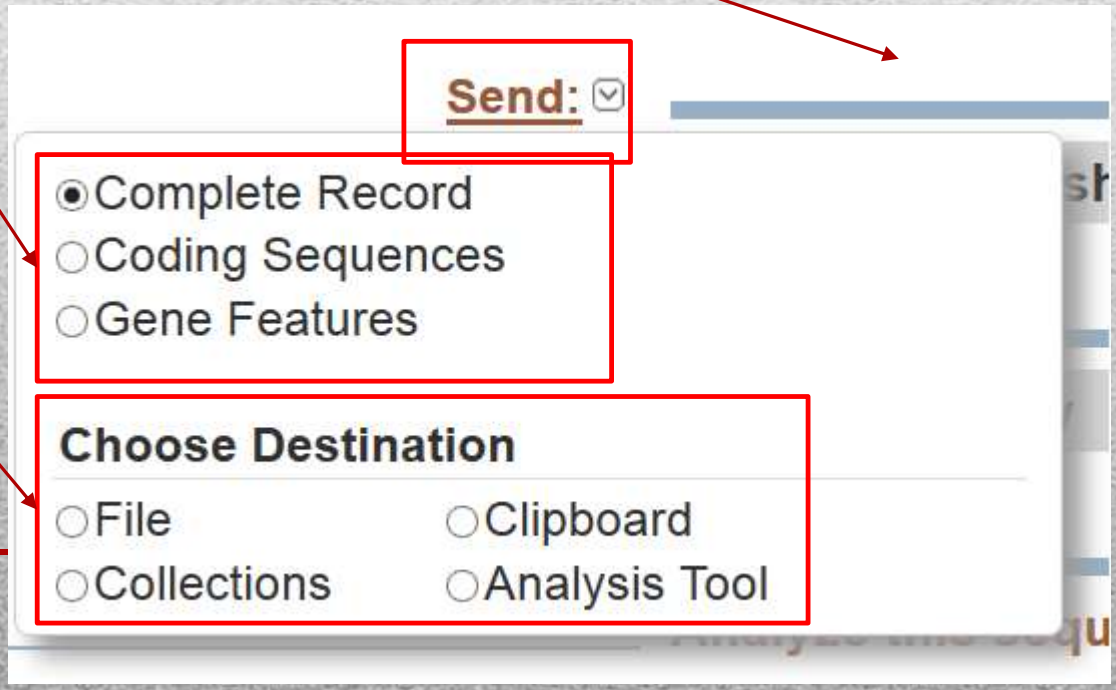
Una volta che abbiamo reperito le informazioni che cerchiamo come le utilizziamo?

Abbiamo varie opzioni … la più banale è fare una stampa ma questa via presenta diverse limitazioni e serve solo quando abbiamo necessità, ad esempio, di aggiungere delle note manuali. Esiste un modo migliore per **esportare e/o salvare in un file** solo le informazioni di cui abbiamo bisogno.

E' possibile esportare dati (operazione in **2 passaggi**) tramite menù **Send** :

1. **Selezionate** le informazioni a cui siete interessati (ad es. **Complete Record**)

2. **Selezionate** la destinazione di queste informazioni (appunti, file, strumenti di analisi, …)

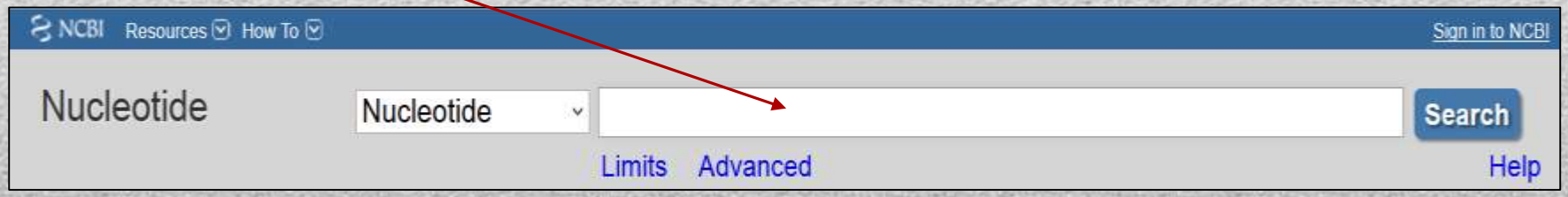


Riepilogo : ricerca per parola chiave

Quella che abbiamo visto è la modalità classica di ricerca di in formazione nelle banche dati biologiche. Essa si definisce «ricerca per parola chiave».

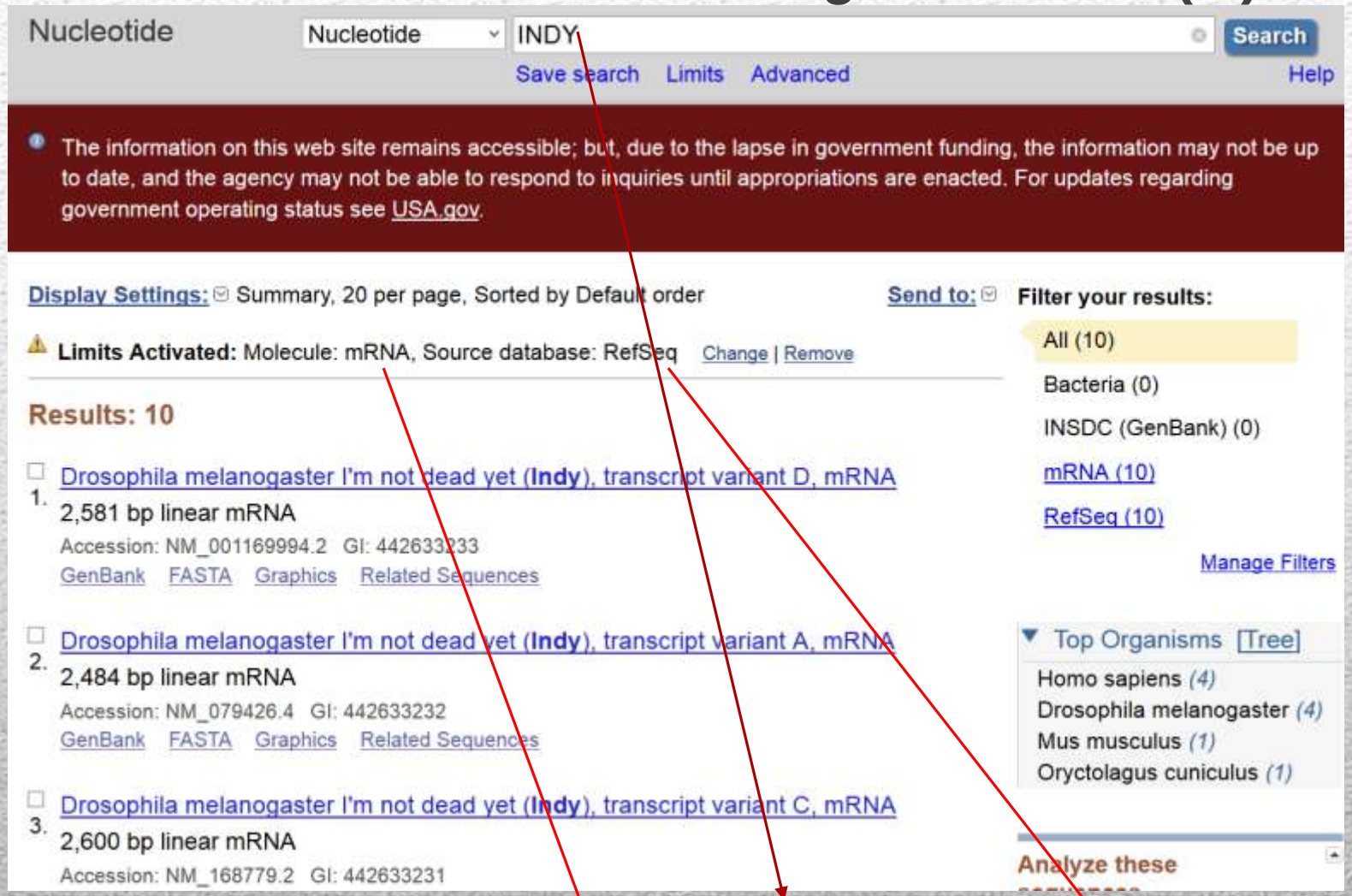
Nel nostro esempio la parola chiave era il **nome del gene** (INDY) .

In origine non esistevano tutti gli strumenti che abbiamo visto (i menù nell'interfaccia web). L'unica modalità con cui era possibile raffinare la ricerca era quella di costruire complesse stringhe di testo da inserire nella casella di ricerca dove avevamo inserito la parola **INDY**.



Anche se in maniera non evidente … **il sistema funziona ancora così** … solo che la stringa di ricerca viene costruita dinamicamente in base alle scelte che operiamo sugli strumenti dell'interfaccia web

Costruzione dinamica stringa di ricerca (I)



Quando cerchiamo la parola chiave INDY ed impostiamo nei limiti (Limits) tipo di molecola: mRNA e source database: RefSeq ….

Costruzione dinamica stringa di ricerca (II)

Mano a mano che aggiungiamo dei filtri il sistema aggiorna una stringa di testo che specifica tutti i passaggi della nostra ricerca …

E li utilizza per aggiornare una stringa di testo che, se salvata ci permetterà di ritornare al sito e rieffettuare la ricerca incollando la stringa nella casella posta di lato al pulsante **Search** (e premendo **Search**)

INDY[All Fields] AND (biomol_mRNA[PROP] AND srcdb_refseq[PROP])

Quando cerchiamo la parola chiave INDY ed impostiamo nei limiti (Limits) tipo di molecola: mRNA e source database: RefSeq ….

Ricerca mediante similarità di sequenza

Abbiamo potuto trovare informazioni sul gene INDY secondo le modalità descritte nelle slide precedenti solo perchè avevamo a disposizione una parola chiave da utilizzare (**INDY**) … ma come possiamo cercare informazioni per una molecola **di cui non sapiamo nulla**?

Questa situazione si verifica quando otteniamo la sequenza in laboratorio. L'unica cosa che otteniamo è, appunto, la sequenza. Nient'alto. In questo caso una ricerca per parola chiave è impossibile da realizzare!

Ricerca mediante similarità di sequenza

Supponiamo di aver ottenuto, in laboratorio, questa sequenza di cui non sappiamo nulla:

>Sequenza_sconosciuta
CTCGCAGGCTCCAGGGGCGGGGCGTGGCCGGGGCGCAGCGACGGGCGCGGAGGTCCGGCCGGGCGCGCGC
GCCCCCGCCACACGCACGCCGGGCGTGCCAGTTTATAAAGGGAGAGAGCAAGCAGCGAGTCTTGAAGCTC
TGTTTGGTGCTTTGGATCCATTTCCATCGGTCCTTACAGCCGCTCGTCAGACTCCAGCAGCCAAGATGGT
GAAGCAGATCGAGAGCAAGACTGCTTTTCAGGAAGCCTTGGACGCTGCAGGTGATAAACTTGTAGTAGTT
GACTTCTCAGCCACGTGGTGTGGGCCTTGCAAAATGATCAAGCCTTTCTTTCATGATGTTGCTTCAGAGT
GTGAAGTCAAATGCATGCCAACATTCCAGTTTTTTAAGAAGGGACAAAAGGTGGGTGAATTTTCTGGAGC
CAATAAGGAAAAGCTTGAAGCCACCATTAATGAATTAGTCTAATCATGTTTTCTGAAAATATAACCAGCC
ATTGGCTATTTAAAACTTGTAATTTTTTTAATTTACAAAAATATAAAATATGAAGACATAAACCCAGTTG
CCATCTGCGTGACAATAAAACATTAATGCTAACACTTTTTAAAACCGTCTCATGTCTGAATAGCTTTCAA
AATAAATGTGAAATGGTCATTTAATGTATTTTCCTATATTCTCAATCACTTTTTAGTAACCTTGTAGGCC
ACTGATTATTTTAAGATTTTAAAAATTATTATTGCTACCTTAATGTATTGCTACAAAAATCTCTTGTTGG
GGGCAATGCAGGTAATAAAGTAGTATGTTGTTATTTGTAAAAAAAAAAAAAAAAAA

E’ una semplice sequenza **FASTA (text)** … potete scaricarla (sottoforma di file di testo) dalla sezione dedicata al materiale didattico presente sul sito del laboratorio.

Aprite il file, selezionate tutto e copiate negli appunti di Windows

Ricerca mediante similarità di sequenza

>seq_sconosciuta
ATTCGATCTAGCGATCTA
CTAATTCGAGGCGATCTA
TCAGCGACTAGCTAGCAT
CGACTACGATCACC…

Ricerca in una **collezione** di sequenze di **sequenze simili** a quella che si usa per effettuare l'interrogazione (**QUERY**)

>seq_1
ATTCGGATCTAGGCTATC
TAGCGATCGACTGACTAG
CTAGCTAGCATCGATCAC

>seq_2
ATTCGAGCGATCTTTTTA
TTATATCGGATTTCGATCG
ATCGATCGACTAAAAAA

>seq_3
ATTCGGATCTAGGCTATC
TAGCGATCGACTGACTAG
CTAGCTAGCATCGATCAC

Punti da ricordare:

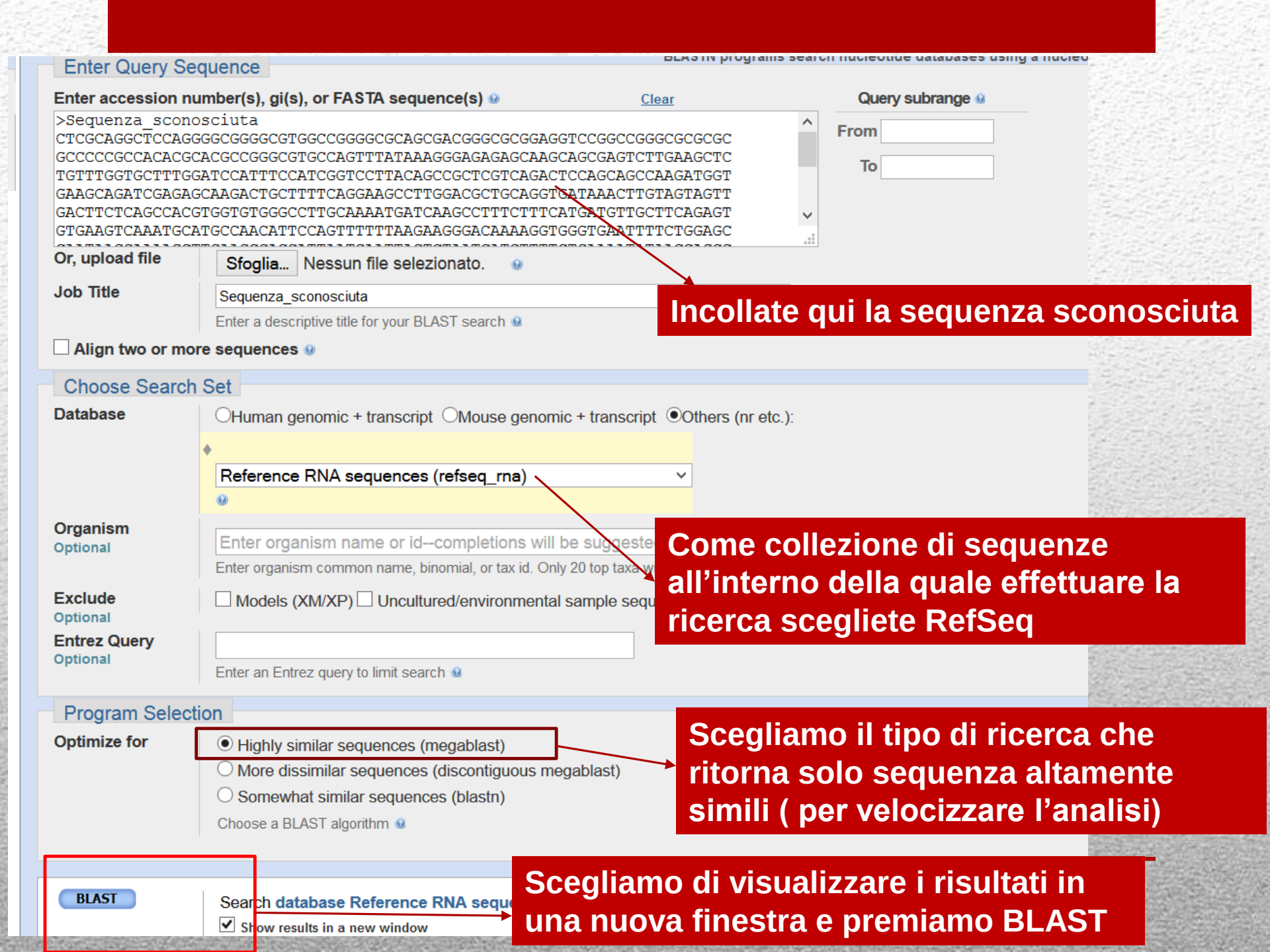
- Il risultato è una **lista di sequenze** ordinate dalla più simile alla nostra sequenza QUERY alla meno simile.
- Al posto di una parola chiave utilizziamo una **sequenza**.

Ricerca tramite similarità di sequenza

Apriamo il web browser e colleghiamoci al sito dello strumento di ricerca per similarità di sequenza **BLAST,** **B**asic **L**ocal **A**lignment **S**earch **T**ool (NCBI):

http://blast.ncbi.nlm.nih.gov/Blast.cgi

.. nella sezione **Basic Blast**, seguite il link **nucleotide blast**


Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>Sequenza_sconosciuta
CTCGCAGGCTCCAGGGGCGGGGCGTGGCCGGGGCGCAGCGACGGGCGCGGAGGTCCGGCCGGGCGCGCGC
GCCCCCGCCACACGCACGCCGGGCGTGCCAGTTTATAAAGGGAGAGAGCAAGCAGCGAGTCTTGAAGCTC
TGTTTGGTGCTTTGGATCCATTTCCATCGGTCCTTACAGCCGCTCGTCAGACTCCAGCAGCCAAGATGGT
GAAGCAGATCGAGAGCAAGACTGCTTTTCAGGAAGCCTTGGACGCTGCAGGTGATAAACTTGTAGTAGTT
GACTTCTCAGCCACGTGGTGTGGGCCTTGCAAAATGATCAAGCCTTTCTTTCATGATGTTGCTTCAGAGT
GTGAAGTCAAATGCATGCCAACATTCCAGTTTTTTAAGAAGGGACAAAAGGTGGGTGAATTTTCTGGAGC
Or, upload file
Sfoglia...
Nessun file selezionato.
Job Title
Sequenza_sconosciuta
Enter a descriptive title for your BLAST search
Align two or more sequences
Incollate qui la sequenza sconosciuta
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Reference RNA sequences (refseq_rna)
Come collezione di sequenze all'interno della quale effettuare la ricerca scegliete RefSeq
Organism
Optional
Enter organism name or id--completions will be suggeste
Enter organism common name, binomial, or tax id. Only 20 top taxa w
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequ
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
Scegliamo il tipo di ricerca che ritorna solo sequenza altamente simili (per velocizzare l'analisi)
BLAST
Search database Reference RNA sequ
Show results in a new window
Scegliamo di visualizzare i risultati in una nuova finestra e premiamo BLAST

Sequenza_sconosciuta

RID 5DTJ5PTT016 (Expires on 10-12 01:39 am)

Query ID lcl|10343
Description Sequenza_sconosciuta
Molecule type nucleic acid
Query Length 826

Database Name refseq_rna
Description NCBI Transcript Reference Sequences
Program BLASTN 2.2.28+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results]

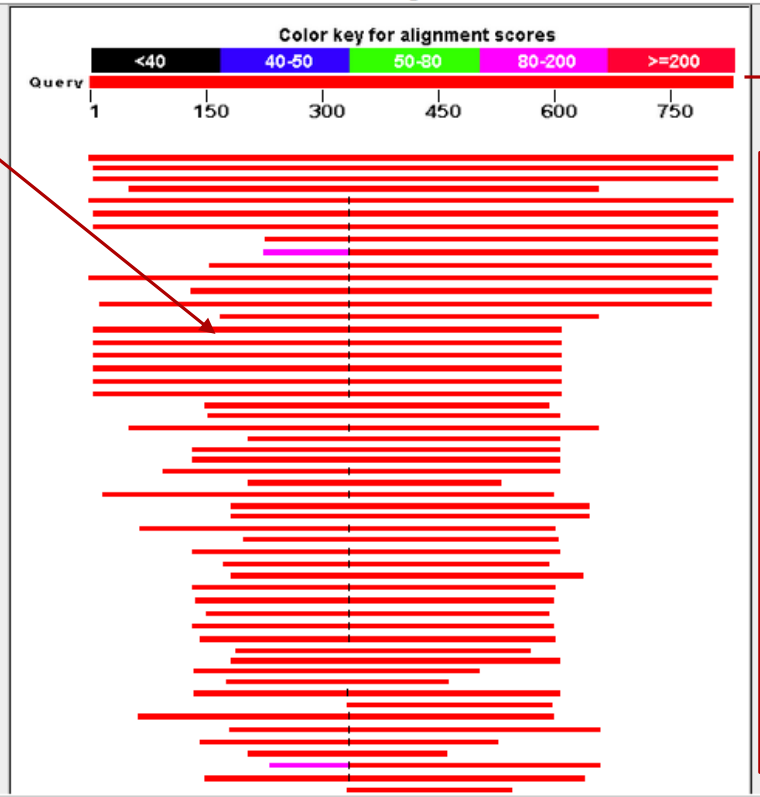
Otteniamo un output composto da varie parti

Graphic Summary

Sommario (in forma grafica)

Distribution of 154 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Questa è la **sequenza QUERY** (in italiano sequenza sonda)

Abbiamo ottenuto **154** corrispondenze

Queste sono le sequenze restituite dalla ricerca…
La PRIMA (sequenza più simile alla query) ha la **stessa lunghezza** della query

Descrizione sequenze estratte (lista risultati)

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens thioredoxin (TXN), transcript variant 2, mRNA	1526	1526	100%	0.0	100%	NM_001244938.1
PREDICTED: Gorilla gorilla gorilla thioredoxin, transcript variant 3 (TXN), mRNA	1443	1443	97%	0.0	99%	XM_004048428.1
PREDICTED: Pan troglodytes thioredoxin (TXN), mRNA	1437	1437	97%	0.0	99%	XM_003951491.1
PREDICTED: Macaca fascicularis thioredoxin (TXN), transcript variant X2, mRNA	950	950	73%	0.0	95%	XM_005581101.1
Homo sapiens thioredoxin (TXN), transcript variant 1, mRNA	911	1529	100%	0.0	99%	NM_003329.3
PREDICTED: Gorilla gorilla gorilla thioredoxin, transcript variant 1 (TXN), mRNA	856	1446	97%	0.0	99%	XM_004048426.1
PREDICTED: Pan troglodytes thioredoxin, transcript variant 1 (TXN), mRNA	850	1440	97%	0.0	99%	XM_001142154.2
PREDICTED: Nomascus leucogenys thioredoxin, transcript variant 2 (TXN), mRNA	828	1028	70%	0.0	98%	XM_004087002.1
PREDICTED: Nomascus leucogenys thioredoxin, transcript variant 1 (TXN), mRNA	828	1027	70%	0.0	98%	XM_003260469.2

La prima sequenza della lista copre l'intera lunghezza della sequenza Query ed è anche identica alla sequenza Query.

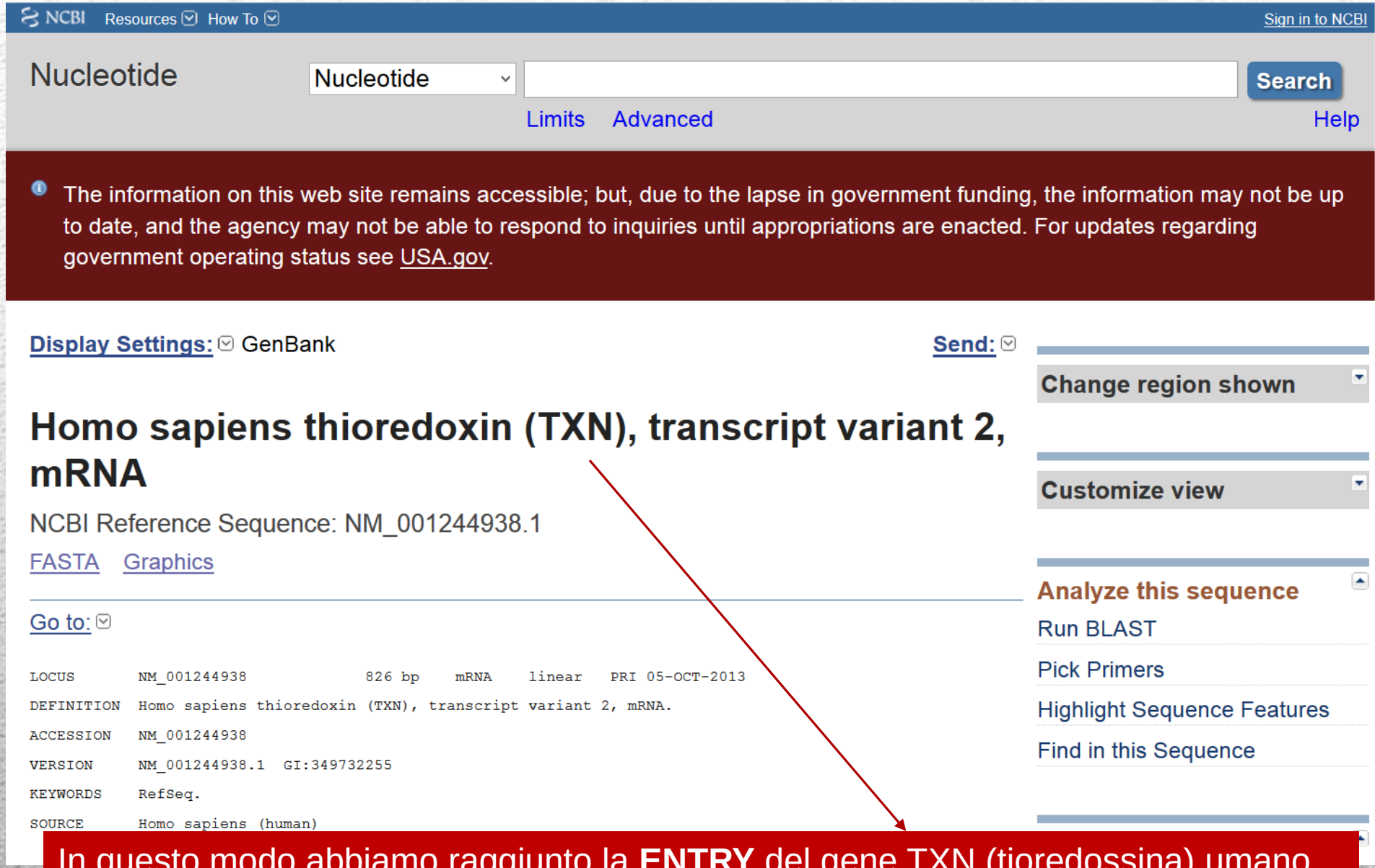
Abbiamo identificato la sequenza sconosciuta!

Descrizione sequenze estratte (lista risultati)

Link ai risultati del confronto tra la query e la prima sequenza ottenuta

Link alla **ENTRY** della prima sequenza ottenuta

Ora abbiamo a disposizione due collegamenti che ci permettono di ottenere ulteriori informazioni. Seguiamo il link che riporta l'accession (codice identificativo) della sequenza. In questo esempio è : NM_001244938.1



In questo modo abbiamo raggiunto la **ENTRY** del gene TXN (tioredossina) umano. De qui in poi vale tutto quello che abbiamo già visto nella sezione dedicata alla ricerca per parola chiave.

Ricerca per similarità di sequenza

Domanda 1 :

•a) A che banca dati appartiene la entry che abbiamo appena estratto (TXN human) ?
•b) Pensate che ci sia una relazione tra la risposta della domanda 1.a e le scelte che avete fatto **PRIMA** di effettuare la ricerca BLAST (suggerimento: riguardate la slide n. 53)? Se si quale?

Esercizio 1 :

Scoprite tutto quello che potete sulla sequenza **Sequenza_sconosciuta_2** presente nello stesso file da cui avete copiato la sequenza, in formato FASTA (text), di Sequenza_Sconosciuta.

•Che tipo di molecola è (DNA o mRNA)?
•Come si chiama il gene da cui deriva?
•A quale organismo appartiene questa sequenza?

Riepilogo

Le banche dati biologiche sono collezioni di informazioni riguardanti molecole presenti nei viventi. Esse hanno dimensioni considerevoli e quindi vengono rese disponibili al pubblico **unitamente a strumenti specializzati** per svolgere ricerche al loro interno.

Esistono principalmente due tipi di ricerca all'interno di una banca dati biologica:

•Ricerca per **parola chiave** : permette di estrarre una o più sequenze fornendo una seria di parole chiave opportunamente combinate. La stringa di interrogazione può essere costruita dinamicamente grazie ad una serie di filtri progressivi. Non può essere utilizzata in assenza di informazioni sull'obiettivo della nostra ricerca.

•Ricerca per **similarità di sequenza** : al posto delle parole chiave si utilizza una sequenza sonda che serve per trovare le sequenze più simili ad essa in banca dati.

Indipendentemente dalla modalità di ricerca adottata è possibile raggiungere delle schede (entries) che contengono molti collegamenti ad altre informazioni sulla sequenza presenti in altre banche dati

Docente: **Matteo Re**
re@di.unimi.it
matteo.re@unimi.it

Università degli studi di milano

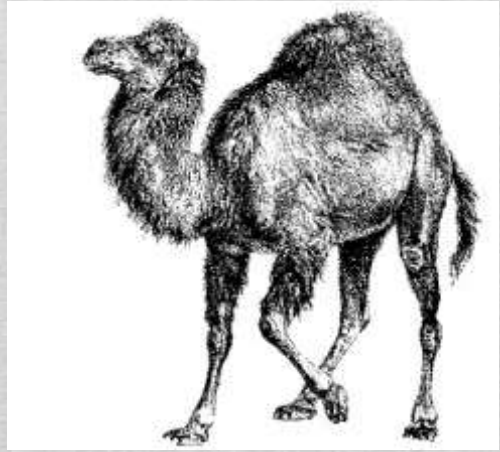
Banche dati biologiche

A.A. 2014-2015 semestre I

5 Accesso diretto

•Obiettivi

- **Interrogazione diretta di banche dati biologiche**
  - Accesso mediante Perl
  - Linguaggio SQL
- **Biologia computazionale**
  - Struttura (db schema) Ensembl database
  - API Ensembl
  - Estrazione di annotazioni
  - Estrazione di sequenze



•Linee guida

- **Il livello di complessità di questa esercitazione è medio-alto**
  - Cercate di risolvere il problema dopo aver compreso gli schemi dai database presentati
  - I template script di questa esercitazione sono estremamente semplici … non fatevi ingennare da questa apparente semplicità **la difficoltà dell'esercizio risiede nella necessità di costruire le interrogazioni in linguaggio SQL** e di itegrarle in maniera opportuna negli script. Come sempre il codice che mi invierete DEVE essere commentato (in questo caso il commento riguarderà principalmente le query SQL).

- **Modalità di svolgimento dell'esercitazione:**
  - Nessun file da scaricare questa volta … lo script di base per effettuare le query SQL è molto contenuto ed è riportato in queste slide.
  - Lo stesso vale per gli esercizi sulle API Ensembl core (trovate molti più esempi risolti mediante le API che mediante SQL… Questo dipende dal fatto che la difficoltà intrinseca degli esercizi SQL sta nella necessità **di dover esplorare lo schema della banca dati Ensembl**).

•Tipi di banche dati biologiche:

•Collettori primari:

- Sequenze sottomesse direttamente dai laboratori di ricerca alle banche dati Genbank, DDBJ ed EMBL. Qualità bassa, a volte contengono errori di annotazione.

•Banche dati secondarie:

•Le informazioni contenute in queste banche dati sono curate manualmente: qualità superiore. Spesso sono banche dati specializzate nel senso che contengono un solo tipo di informazione (seq. proteiche, seq. di trascritti, …).

•Banche dati associate a progetti genomici:

•Le sequenze genomiche sono un tipo di dato molto particolare. Esse si prestano ad essere annotate a diversi livelli. A causa di questa caratteristica la loro annotazione richiede l'utilizzo di informazioni derivanti da un numero consistente di banche dati esterne. Come conseguenza le banche dati associate a progetti di annotazione genomica sono gli strumenti di elezione per **INTEGRARE** il contenuto di altre banche dati in modo da ottenerne una **rappresentazione unitaria**.

•Tipi di dati biologici (solo alcuni)

•**Livello dei trascritti** misurati in particolari condizioni: esistono siti dedicati a collezioni di esperimenti microarray (es. NCBI Gene Expression Omnibus (NCBI GEO), http://www.ncbi.nlm.nih.gov/geo/)

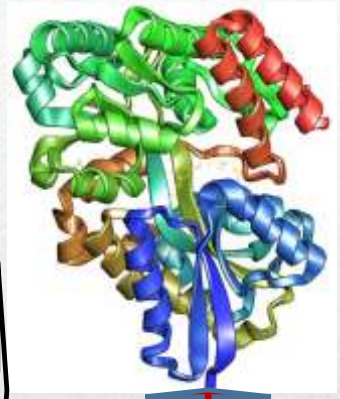
•**Annotazione funzionale di proteine**: «funzionale» viene utilizzato come termine a «**basso**» livello, annotazione di una sequenza proteica **residuo per residuo**. Molti tipi di annotazione: siti di fosforilazione, presenza di ponti disolfuro, struttura secondaria della proteina, struttura 3D della proteina. Sito di riferimento è una banca dati che integra le informazioni di diverse banche dati: Uniprot (Universal Protein Resource, http://www.uniprot.org/) .

•**Annotazione funzionale di geni:** «funzionale» viene utilizzato come termine ad «**alto**» livello. Creazione di vocabolari controllati a partire da materiale reperibile in **LETTERATURA**. Team di curatori assegnano ogni gene ai termini dei vocabolari (**ontologie**). Sito di riferimento: Gene Ontology (http://www.geneontology.org/) .

•**Variabilità genetica:** Database dedicati a SNP (es. NCBI dbSNP) e a progetti su vasta scala (HapMap). Esistono inoltre databases dedicati a studi di associazione genome-wide (es GWAS central) http://www.gwascentral.org/index.

•E MOLTI ALTRI …

•Esistono molti tipi di banche dati … perché quelle associate a progetti di annotazione genomica dovrebbero essere considerate più «importanti» di altre?



•fuctional annotation

•Proteins DB

•cross-links

•post tanslational editing

•folding

•Transcripts DB

•translation

•post transcriptional editing

•transcription

•alignment

•alignment

•genome

•regulatory

•region

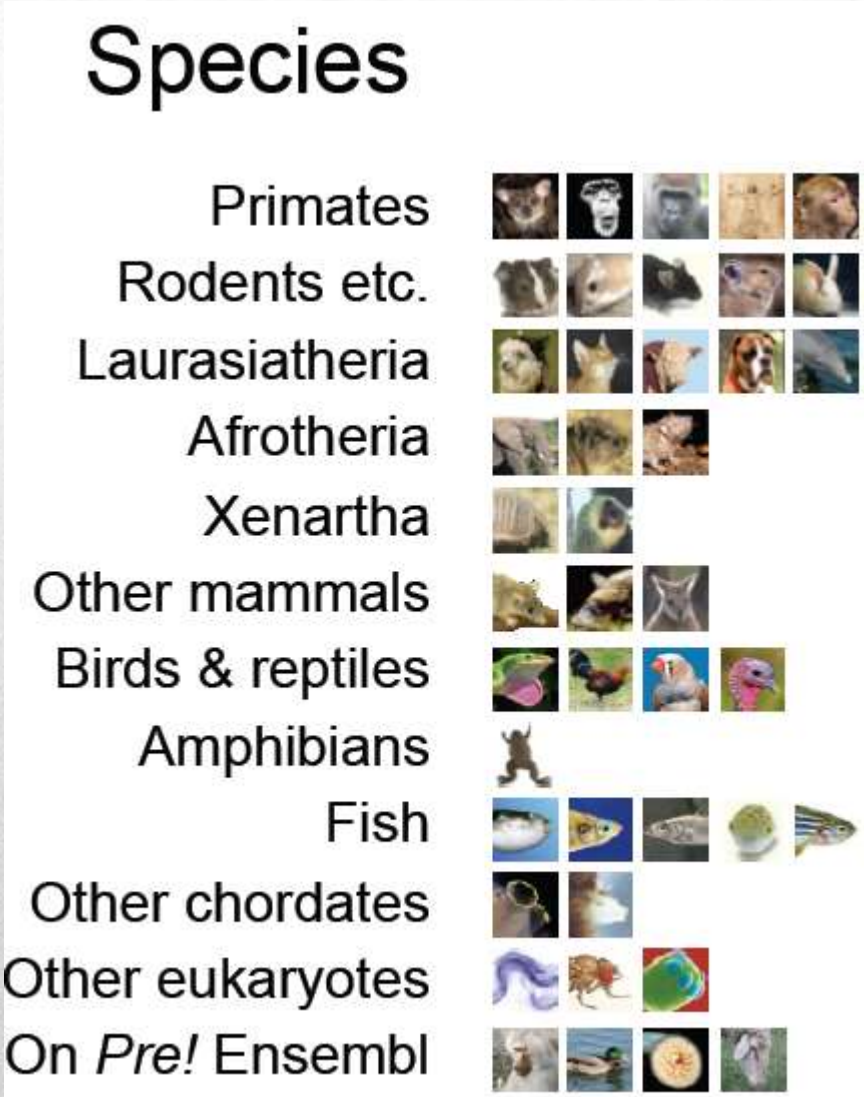
•regulatory motifs

•Annotazione in coordinate genomiche

•BROWSERS GENOMICI

•**Ne esistono diversi:** Principalmente 3:
•NCBI map viewer (http://www.ncbi.nlm.nih.gov/projects/mapview/)
•Ensembl (http://www.ensembl.org/index.html)
•UCSC genome browser (http://genome.ucsc.edu/)
•**Presentano le stesse informazioni, ma in modo diverso**: tutti e tre permettono di trovare la posizione genomica di una sequenza (mediante allineamento o ricerca per parola chiave) e di visualizzare la regione genomica associata.
•**I dati contenuti nei browser genomici dipendono dal contenuto di altre banche dati:** necessità di aggiornare i dati molto spesso. Ensembl viene aggiornato mensilmente .
•**Produzione dei dati di annotazione genomica:** E' un processo cosatoso dal punto di vista delle risorse di calcolo (allineamento di intere banche dati di sequenze al genoma). I principali browser genomici contengono più di un genoma (in realtà contengono molti genomi). E' un processo basato su **pipeline di annotazione automatizzate**.

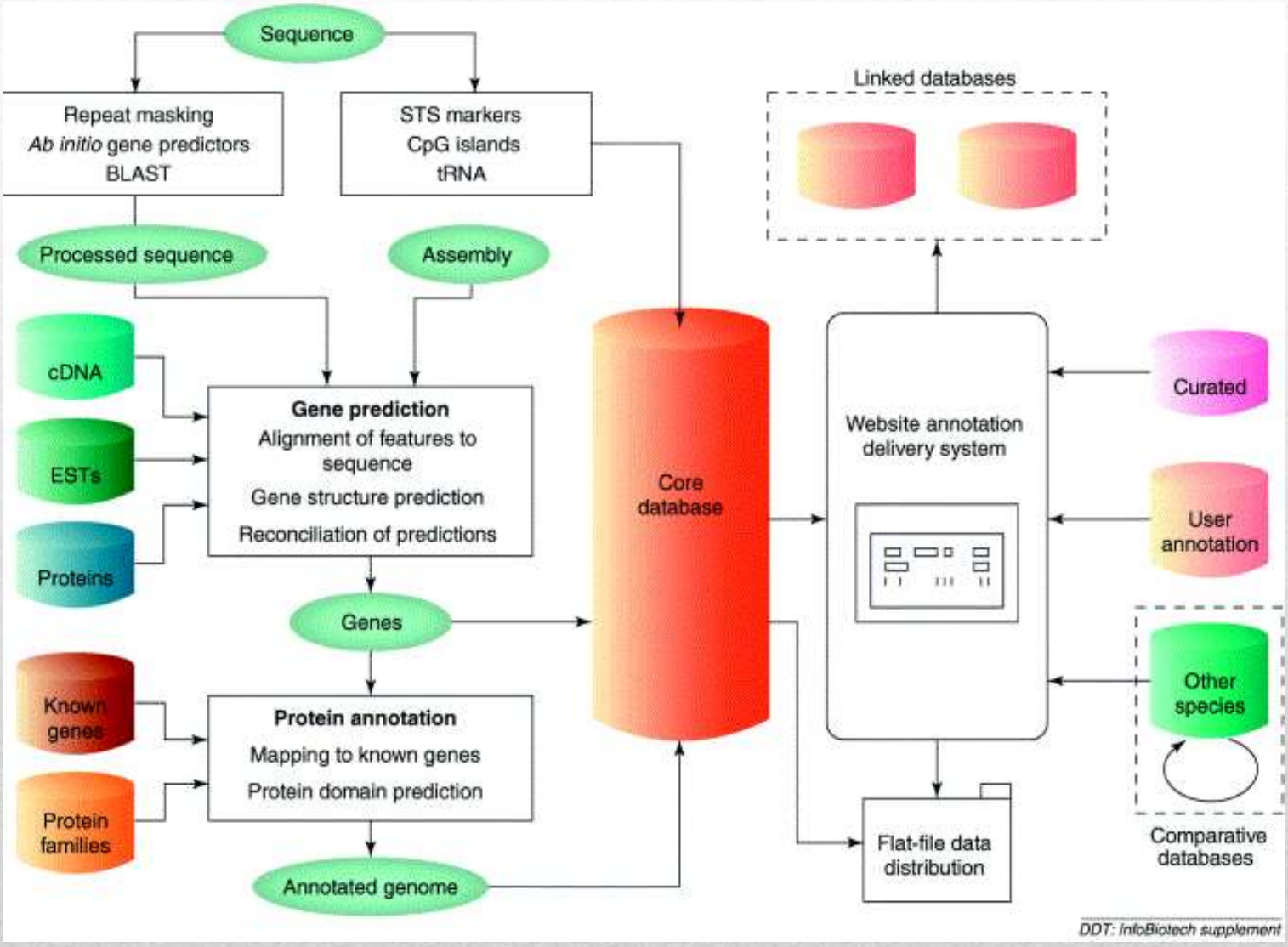
•**Ensembl**



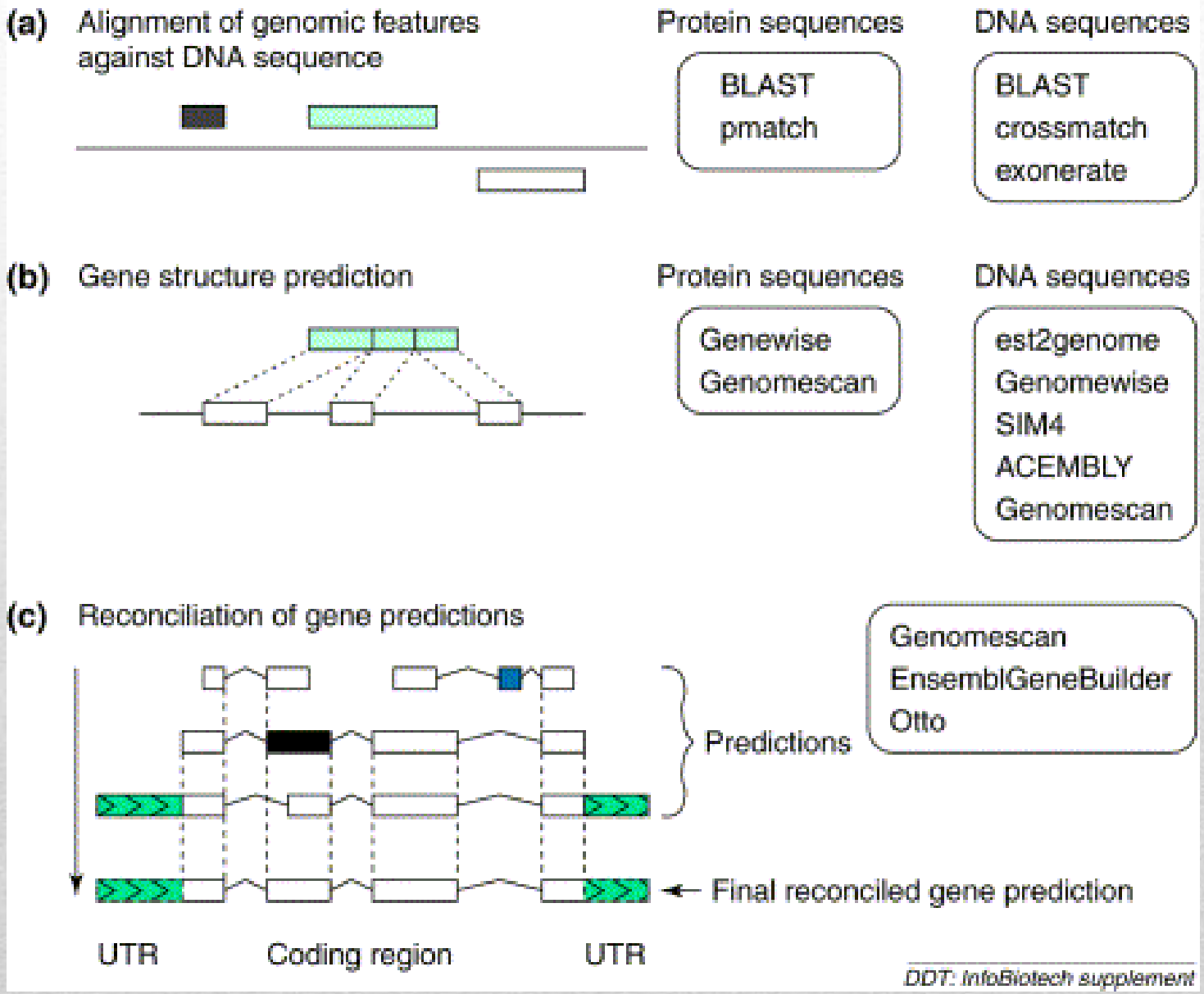
•**Caratteristica specifica di Ensembl :**

•contiene **modelli** di <u>geni</u> (altri browser utilizzano come entità fondamentale il trascritto o, comunque, la «sequenza allineata al genoma»)

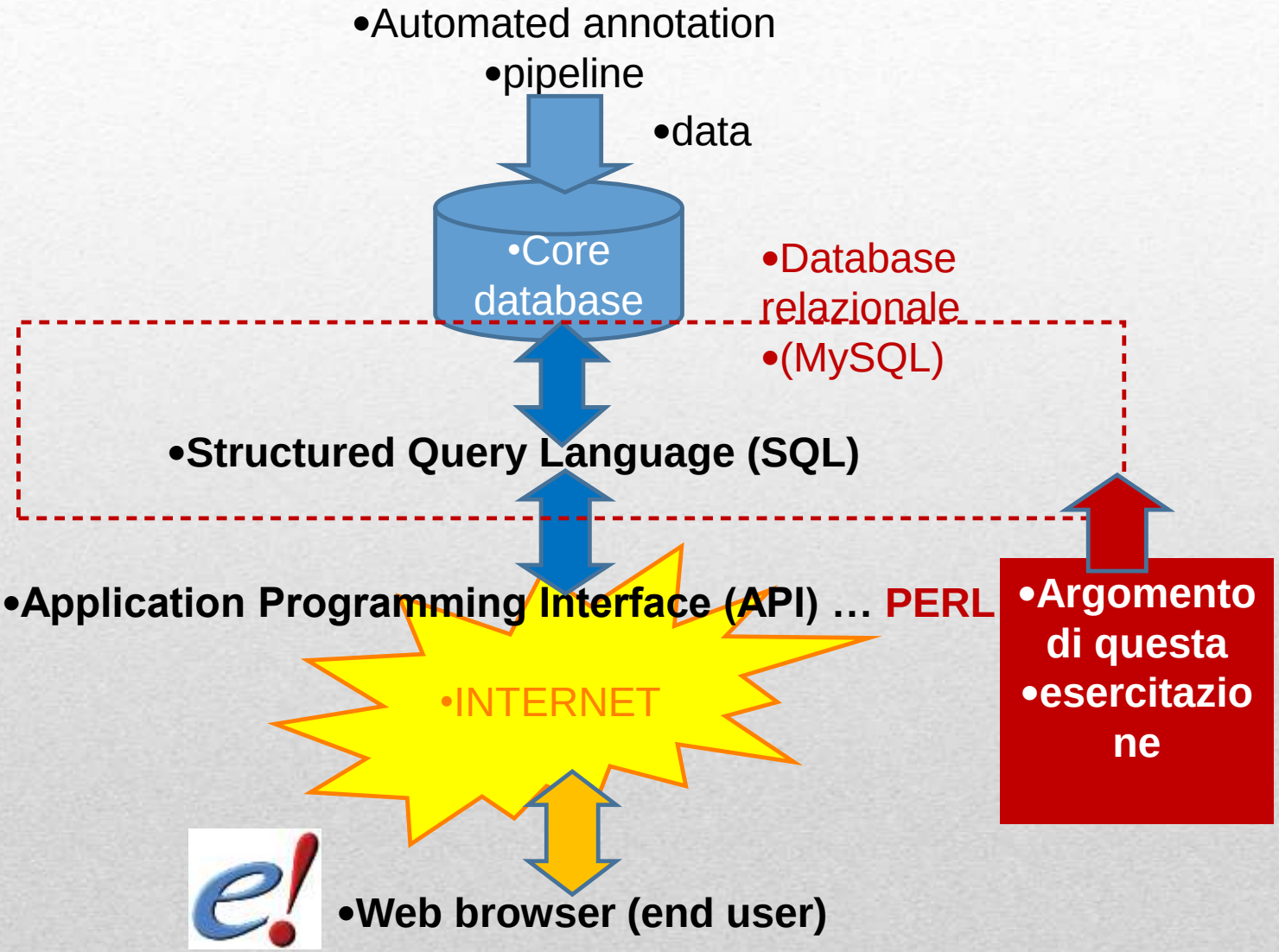
- **Automatizzazione del processo di annotazione di una sequenza genomica**



•Creazione di «modelli» di geni



•Architettura del browser genomico Ensembl



•Structured Query Language (SQL) e (R)DBMS

•I database sono estremamente eterogenei per quanto riguarda la loro struttura e la quantità di dati contenuta. Essi possono essere costituiti da file di testo ASCII o file che rappresentano complesse strutture composte da alberi binari (ad es. Oracle o Sybase). In ogni caso un database è un contenitore di dati.

•PROBLEMA:

•Se un database è una semplice collezione di dati … chi tiene traccia del cambiamento dei dati stessi?

•Questo è il ruolo dei sistemi di gestione delle basi di dati (database management systems o DBMS). Alcuni DBMS sono relazionali. In tal caso ci si riferisce ad essi come relational DBMS o RDBMS. Le relazioni su cui si basano i sistemi RDBMS assicurano che diverse colelzioni di dati (ad es. tabelle) possano essere interrogate "all'unisono". Le relazioni, di fatto, rappresentano delle regole di integrità referenziale tra collezioni di dati. Supponiamo di avere un RDBMS che contiene I dati di tutti gli impiegati di un azienza e di avere 2 tabelle: reparto e impiegato. Tra di esse potrebbe esistere una relazione che permette l'inserimento di un nuovo impiegato SOLO se esso è assegnato ad un reparto esistente.

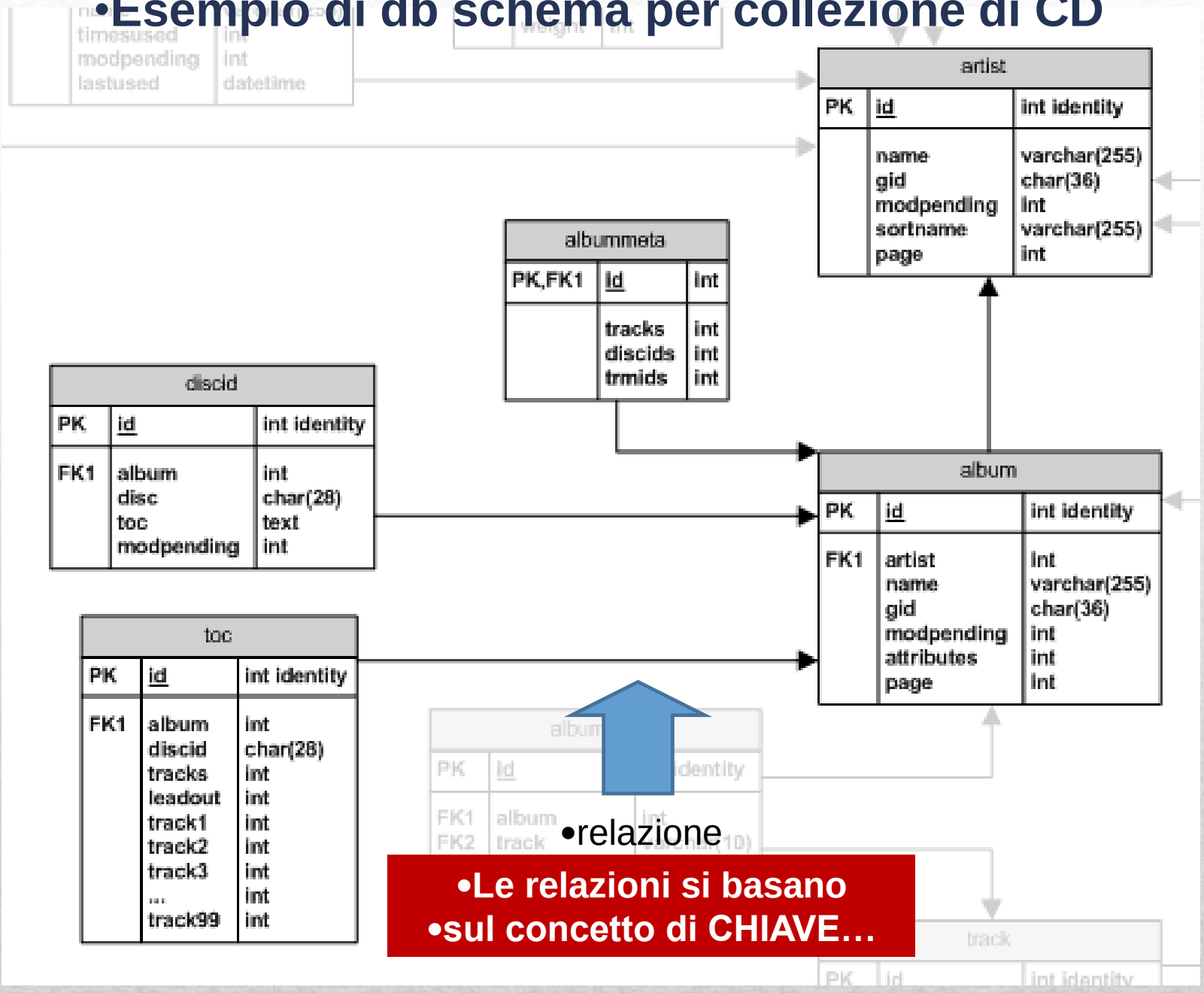
•Structured Query Language (SQL) e (R)DBMS

•Un database relazionale (come quello associato alla maggioranza delle banche dati genomiche) è costituito da :

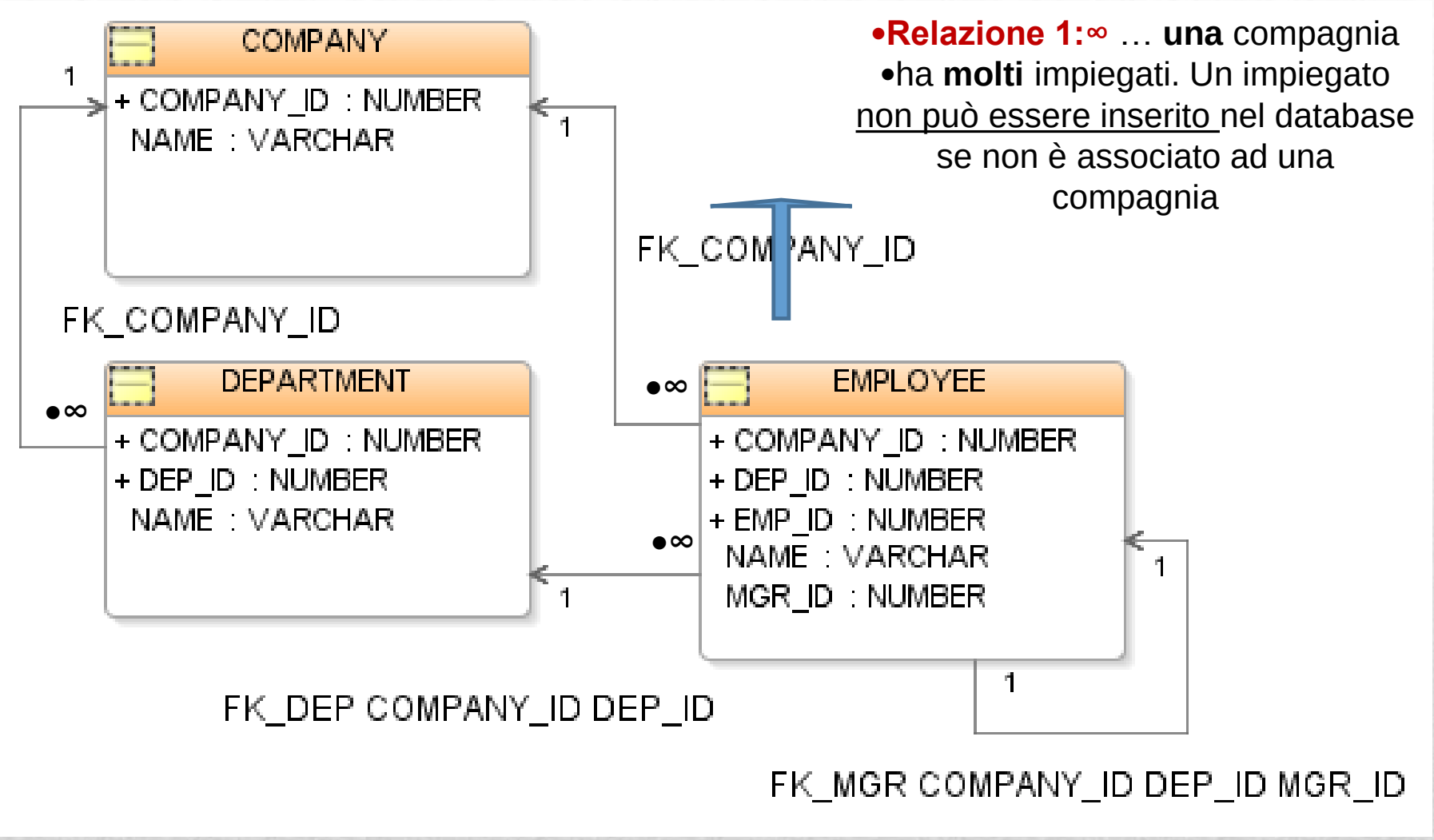
1.Una parte INVARIANTE nel tempo detta database schema. Essa definisce la struttura logica delle unità di memorizzazione delle informazioni. Tale struttura, di solito, viene rappresentata sottoforma di tabella. La rappresentazione tabulare permette di nascondere I dettagli del formato reale di memorizzazione su disco.

1.I dati veri e propri: ad essi ci si riferisce con il termine generico di istanze. Per ogni tabella presente nella banca dati è disponibile una DEFINIZIONE composta da numero e nomi dei campi (colonne) della tabella, tipo di dato ammesso in ogni campo e altre caratteristiche (che descrivono ad esempio, il coinvolgimento di una relazione di integrità associata ad un dato campo). Prima di inserire una nuova riga in una tabella il sistema RDBMS verifica che la colelzione di dati (la riga della tabella) tispetti tutte le specifiche della tabella stessa. Un altro modo comune di riferirsi alle righe delle tabelle è il termine RECORD.

•Esempio di db schema per collezione di CD



•Esempio di db schema per collezione di CD



•**Relazione 1:∞** … **una** compagnia •ha **molti** impiegati. Un impiegato non può essere inserito nel database se non è associato ad una compagnia

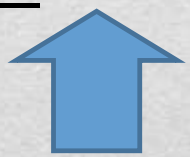
•Le relazioni si basano •sul concetto di CHIAVE…

•Interazione con RDBMS e ruolo di SQL

•E’ necessaro uno strumento che permetta di interagire con la banca dati. Questo ruolo è svolto da un linguaggio standardizzato detto Structured Query Language (SQL). SQL permette non solo l’estrazione dei dati ma anche la creazione/modifica di database e tabelle nonchè la definizione di vincoli relazionali. SQL si divide in:
•

•**DATA DEFINITION LANGUAGE (DDL): linguaggio di definizione dei dati, serve per creare databases, definizioni di tabelle e vincoli di integrità referenziale. Permette inoltre di modificare la struttura di tabelle esistenti.**

•**DATA MANIPULATION LANGUAGE (DML): insieme di enunciati che permettono, principalmente, di estrarre informazioni da una banca dati.**



•A noi interessa DML (DDL non verrà trattato)

•Operazioni realizzabili mediante SQL DML

•SQL DML permette di realizzare diverse operazioni che possono essere attribuite a tre grandi macrocategorie:
•

•**PROIEZIONE:** **Estrazione di attributi (valori contenuti in un sottoinsieme di colonne di una tabella specificate dall'utente)**

•**ESTRAZIONE:** **Selezione di alcune righe (record) da una tabella nel caso in cui queste corrispondano ad alcuni criteri specificati dall'utente**

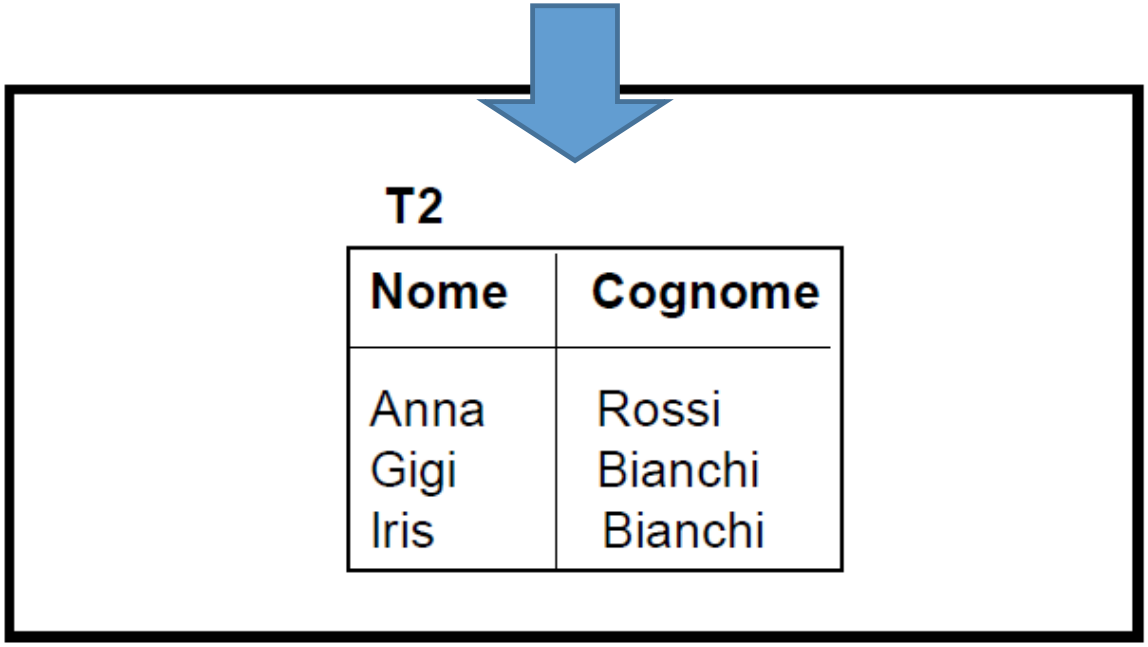
•**JOIN:** **Interrogazione simultanea di più tabelle basata su relazioni. Concettualmente equivale a creare in memoria una macrotabella costituita dai dati contenuti in più tabelle. Solitamente dopo il join viene effettuata un'estrazione.**

•Esempio di PROIEZIONE

T1

Nome	Cognome	Nato il	Nato a
Anna	Rossi	2/2/71	TO
Gigi	Bianchi	23/4/80	Ivrea
Iris	Bianchi	15/9/45	CN

La **proiezione** di T1 sugli attributi Nome e Cognome restituisce



T2

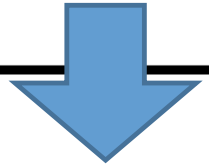
Nome	Cognome
Anna	Rossi
Gigi	Bianchi
Iris	Bianchi

•Esempio di ESTRAZIONE (o selezione)

T1

Nome	Cognome	Nato il	Nato a
Anna	Rossi	2/2/71	TO
Gigi	Bianchi	23/4/80	Ivrea
Iris	Bianchi	15/9/45	CN

La **selezione** dei record di T1 tali che **"Nato il >= 1/1/1960"** restituisce



T2

Nome	Cognome	Nato il	Nato a
Anna	Rossi	2/2/71	TO
Gigi	Bianchi	23/4/80	Ivrea

•Esempio di utilizzo di enunciato SELECT

•Il database **Ensembl core** contiene una tabella **gene**:

Field	Type		Null	Key	Default		Extra
gene_id	int(10) unsigned	16B	NO	PRI	(NULL)	0K	auto_increment
biotype	varchar(40)	11B	NO		(NULL)	0K	
analysis_id	smallint(5) unsigned	20B	NO	MUL	(NULL)	0K	
seq_region_id	int(10) unsigned	16B	NO	MUL	(NULL)	0K	
seq_region_start	int(10) unsigned	16B	NO		(NULL)	0K	
seq_region_end	int(10) unsigned	16B	NO		(NULL)	0K	
seq_region_strand	tinyint(2)	10B	NO		(NULL)	0K	
display_xref_id	int(10) unsigned	16B	YES	MUL	(NULL)	0K	
source	varchar(20)	11B	NO		(NULL)	0K	
status	enum('KNOWN','NOVEL','PUTATIVE','PREDICTED','KNOWN_BY_PRO...	76B	YES		(NULL)	0K	
description	text	4B	YES		(NULL)	0K	
is_current	tinyint(1)	10B	NO		1	1B	
canonical_transcript_id	int(10) unsigned	16B	NO		(NULL)	0K	
canonical_annotation	varchar(255)	12B	YES		(NULL)	0K	

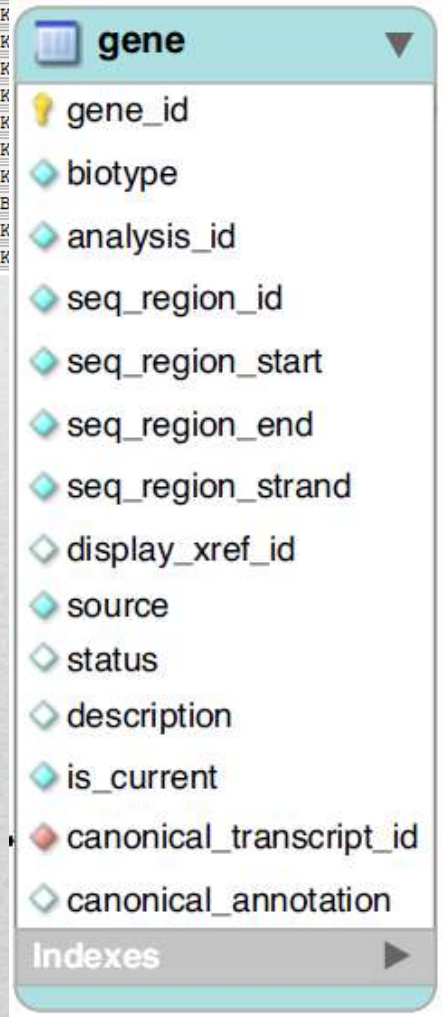
•chiave primaria

•chiavi esterne

•descrizione dettagliata (nomi campi, tipi di dato …)

•**SQL:** DESCRIBE gene;

•Rappresentazione **semplificata** (nomi campo + simboli ma non tipo di dato). Comune in molti strumenti ad interfaccia grafica ed **estremamente comune** nei diagrammi che descrivono gli schemi delle banche dati



•Strumenti free per l'accesso a banche dati relazionali

•Proveremo ad effettuare alcuni esperimenti pratici utilizzando uno strumento free: **SQLyog**

•Scaricatelo da questo sito:
•http://code.google.com/p/sqlyog/downloads/list
•(scaricate l'ultima versione : SQLyog-9.0.1-1Community.exe)

•Provate ad installarlo in una directory in cui **avete i permessi di scrittura** (es. Documenti).

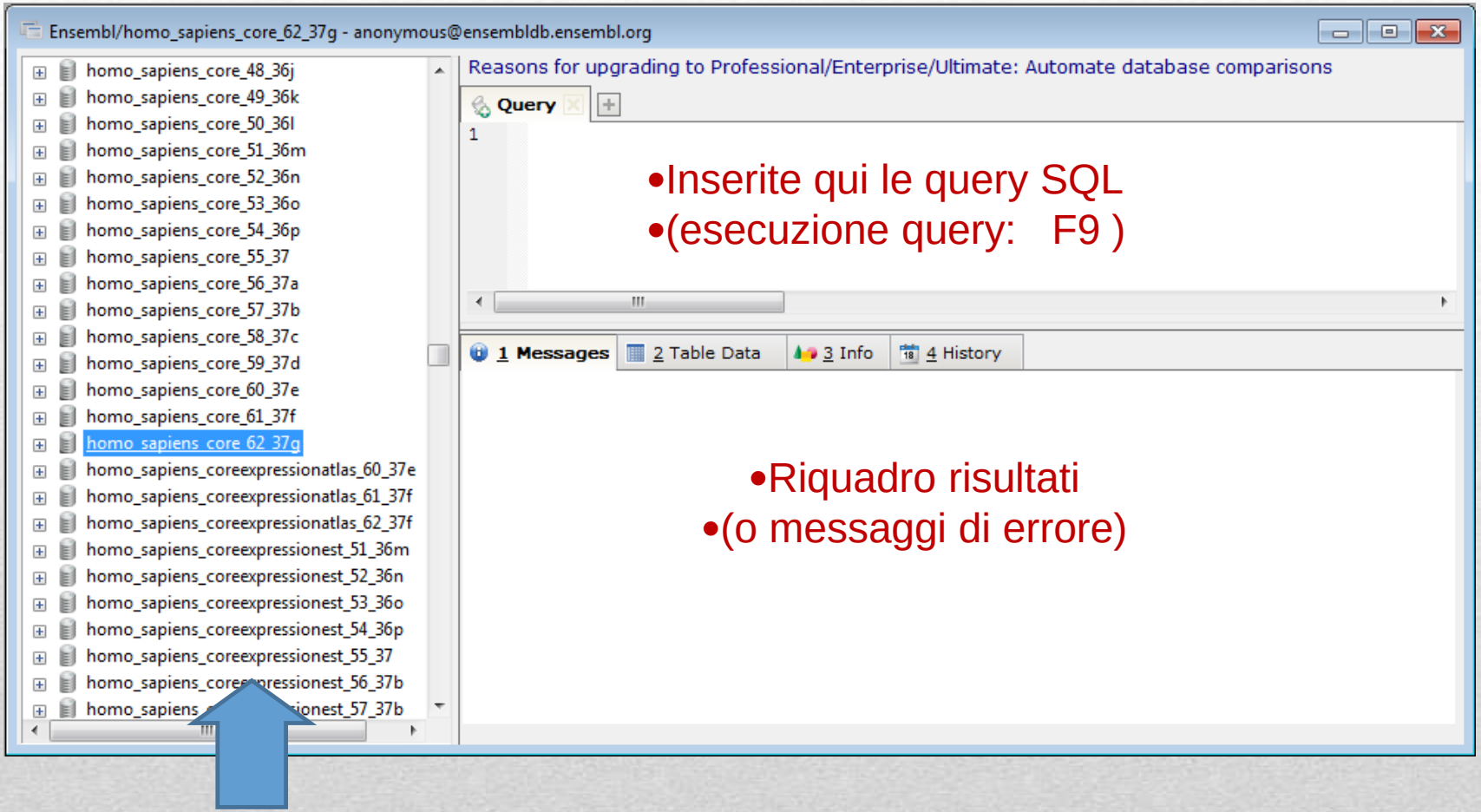
•Una volta installato definite i parametri per una nuova connessione:

•File -> New **connection**
•**Valori**:
•Nome connessione EnsEMBL
•MySQL Host Address ensembldb.ensembl.org
•Username anonymous
•Port 5306
•(lasciate vuota la password che non serve)

•Accesso in sola lettura

(alternativa) **mysql client** command line :
mysql -u anonymous -h ensembldb.ensembl.org -P 5306

•SQLyog: finestra principale



•Database disponibili: a noi interessa **homo_sapiens_core_62_37g**

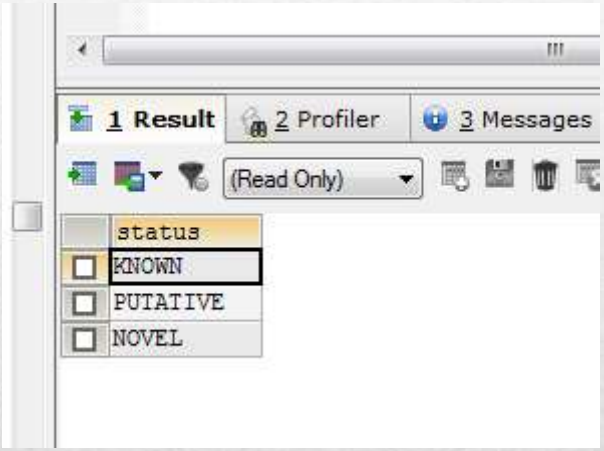
• (click sx per selezionarlo)

•Interrogazione diretta di Ensembl

•Quali sono i possibili valori presenti un una data colonna?

•SELECT **DISTINCT(**status**)** FROM gene;

•Quanti record ottenete?



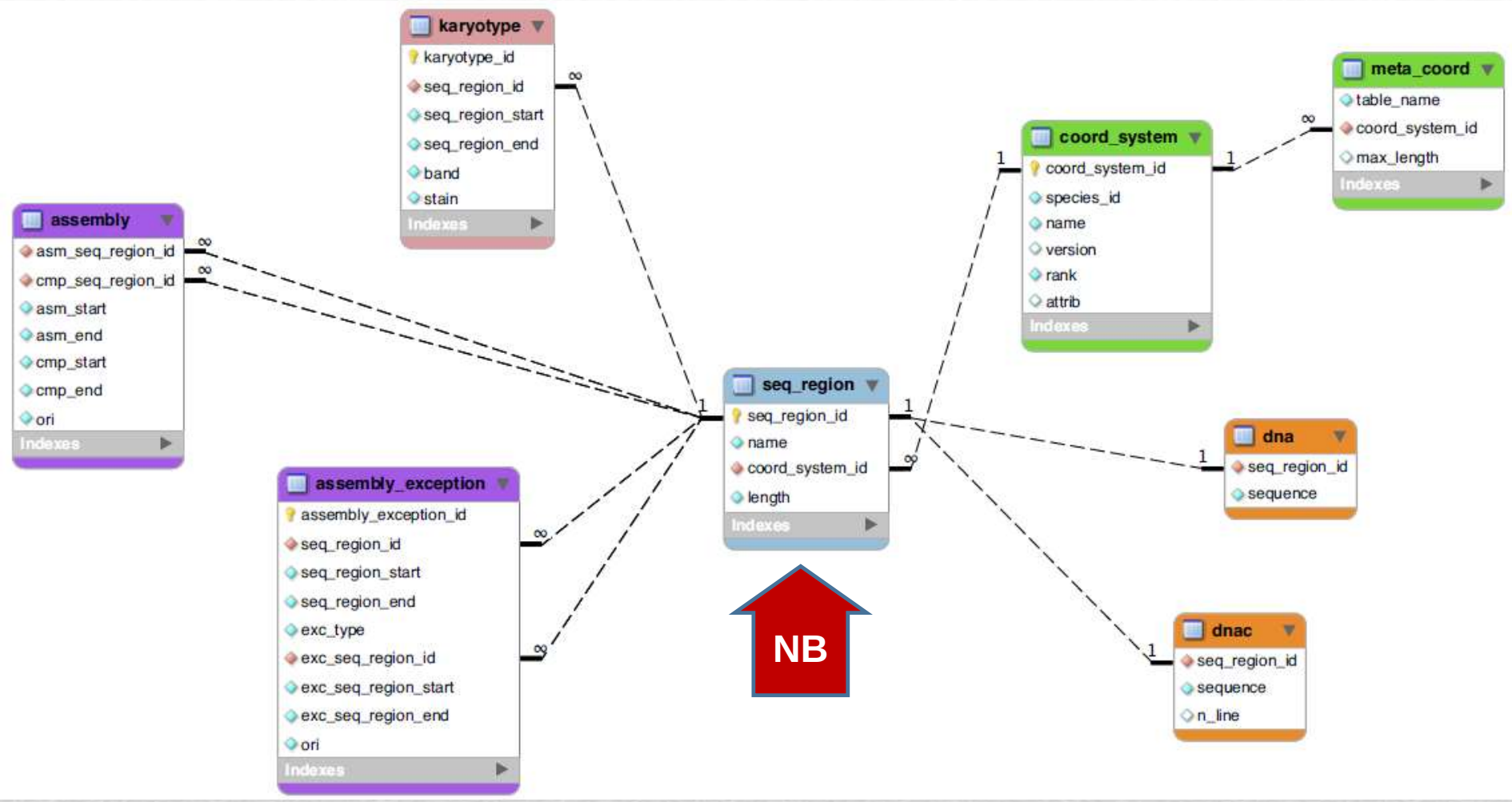
•**NON** utilizzeremo SQLyog per realizzare i nostri accessi diretti ad Ensembl (utilizzeremo Perl), ma esso è uno strumento molto comodo per testare le query SQL **prima** di inserirle in uno script, in modo da essere sicuri che si comportino secondo le attese.

•Esecuzione di query SQL da remoto in Perl

```
•# PERL MODULES
•use DBI;
•use DBD::mysql;


•# CONFIG VARIABLES
•$platform = "mysql";
•$host = "ensembldb.ensembl.org";
•$port = "5306";
•$user = "anonymous";
•$pw = "";
•$database="homo_sapiens_core_62_37g";


•#DATA SOURCE NAME ##################################
•$dsn = "dbi:mysql:$database:$host:5306";


•#CONNECTION ########################################
•$DBIconnect = DBI->connect($dsn, $user, $pw);


•#Query #############################################

•$sqlquery = "select * from gene limit 10";

•$sth = $DBIconnect->prepare($sqlquery);

•$sth->execute;


•#PRINT RESULTS #####################################
•while (@row = $sth->fetchrow_array) {
•print "@row\n";
•}
```

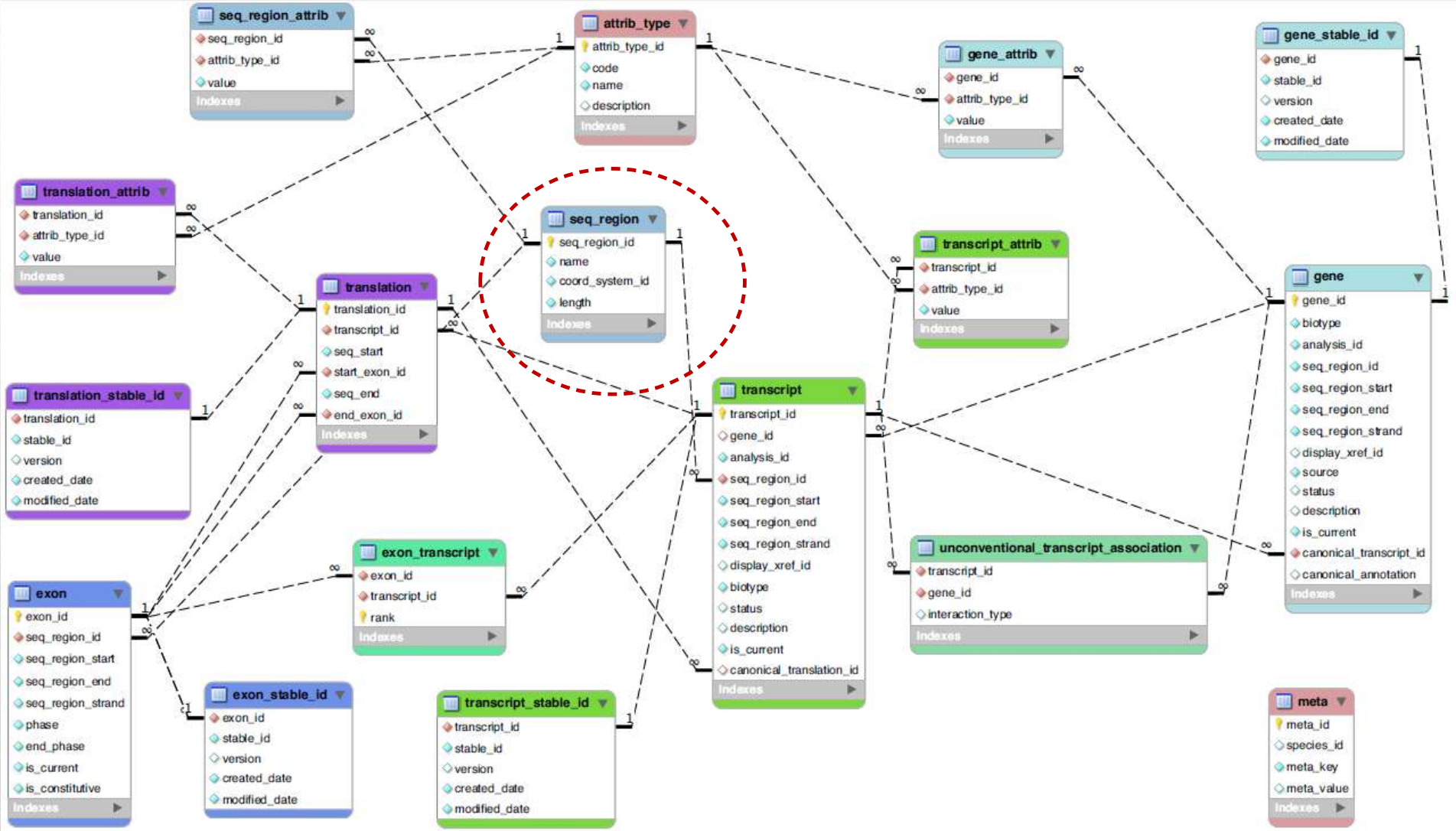
- Librerie SQL
- Parametri di connessione
- Connessione
- Interrogazione
- Stampa risultati

•Ensembl core db schema (I)



•**NB:** il processo di annotazione genomica **non viene effettuato <u>unicamente</u> sulle sequenze genomiche assemblate**. Parte di esso viene effettuato su cloni, contigui, supercontigui ecc.. OGNI ANNOTAZIONE esiste in uno specifico sistema di coordinate

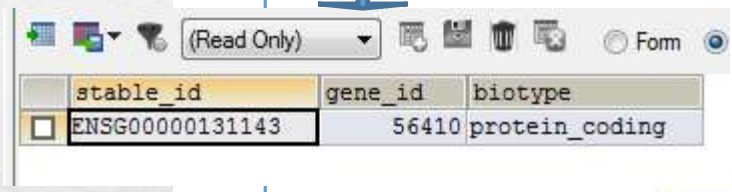
•Ensembl core db schema (II)



•Ensembl core db schema (II) : JOIN

•nometabella.nomecampo

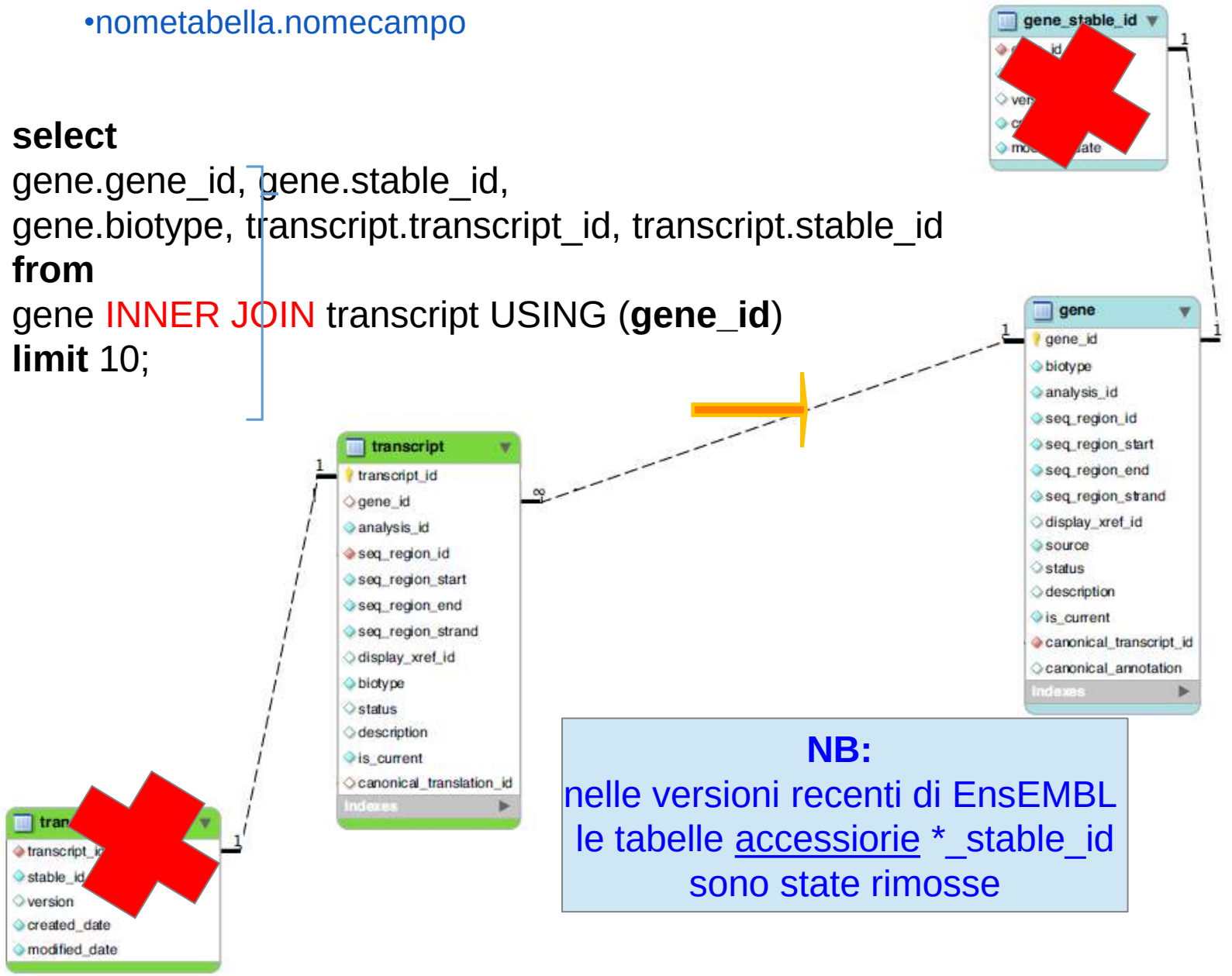
- •SELECT gene_stable_id.stable_id, gene.gene_id, gene.biotype
- •FROM gene_stable_id INNER JOIN gene USING (gene_id)
- •WHERE gene_stable_id.stable_id = 'ENSG00000131143';



•Ensembl core db schema (II) : JOIN

•nometabella.nomecampo

```
select
gene.gene_id, gene.stable_id,
gene.biotype, transcript.transcript_id, transcript.stable_id
from
gene INNER JOIN transcript USING (gene_id)
limit 10;
```

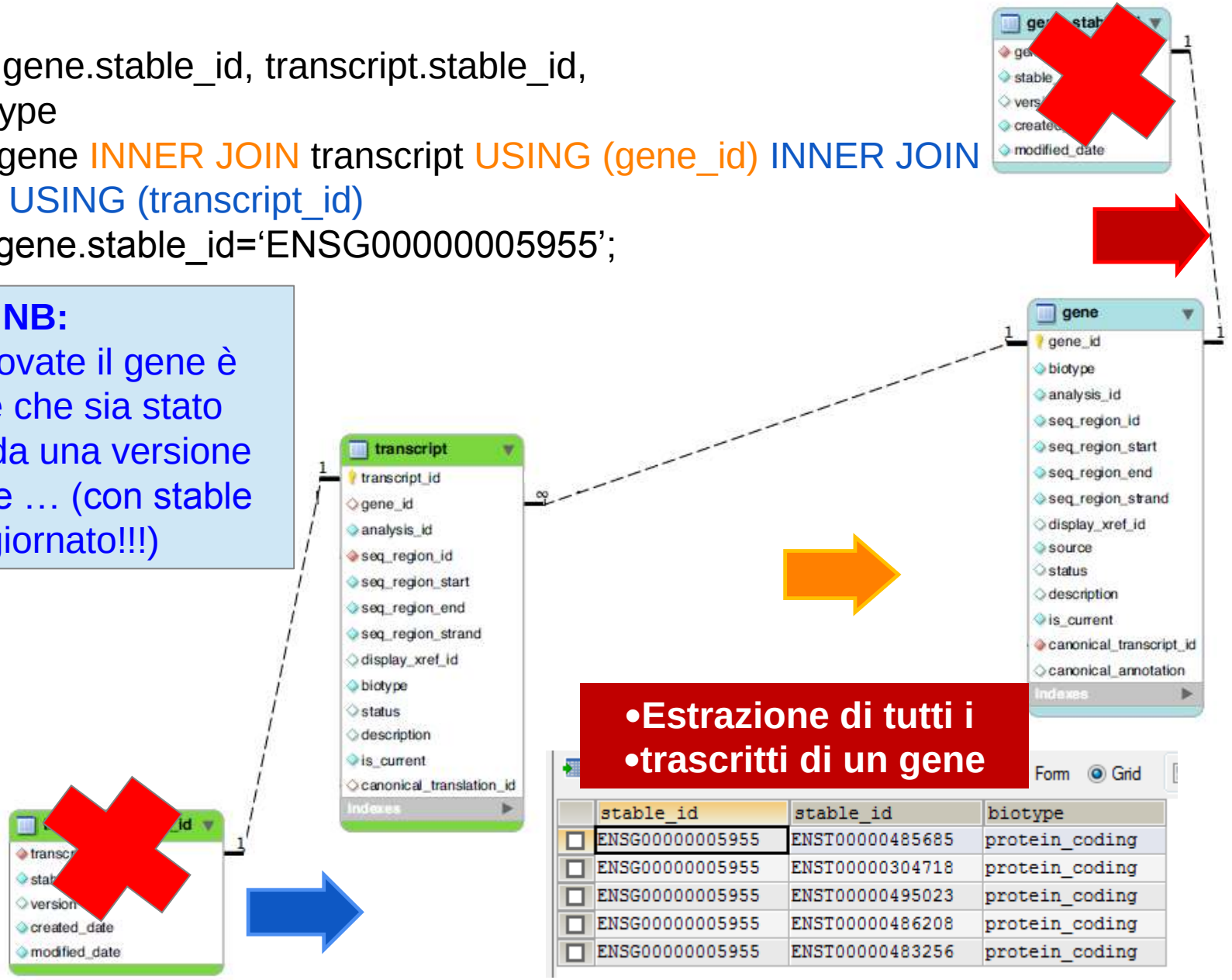


NB:
nelle versioni recenti di EnsEMBL le tabelle accessiorie *_stable_id sono state rimosse

•Ensembl core db schema (II) : JOIN

- **SELECT** gene.stable_id, transcript.stable_id,
- gene.biotype
- **FROM** gene INNER JOIN transcript USING (gene_id) INNER JOIN
- transcript USING (transcript_id)
- **WHERE** gene.stable_id='ENSG00000005955';

NB:
Se non trovate il gene è possibile che sia stato sostituito da una versione più recente … (con stable Id aggiornato!!!)



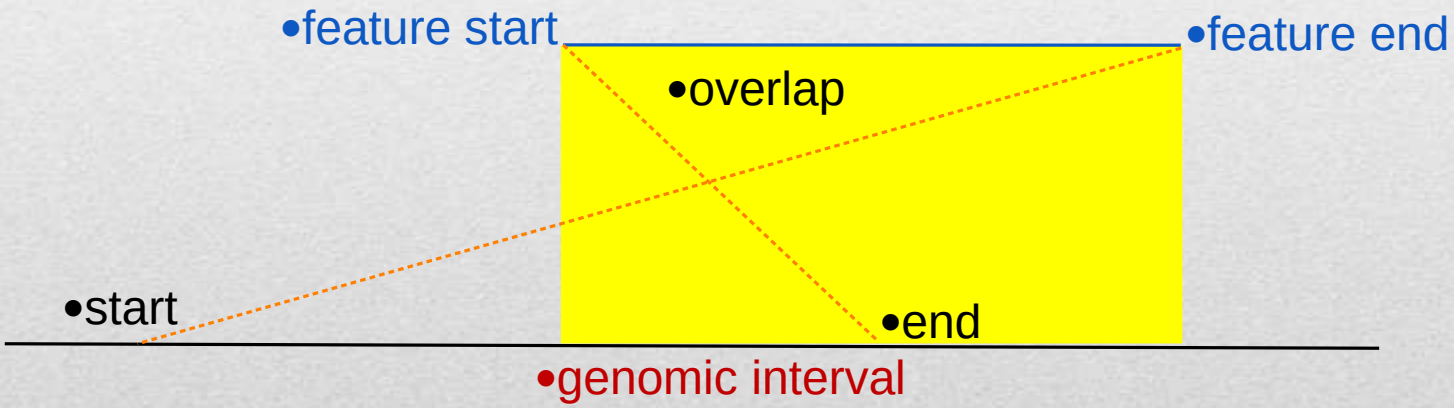
- **Estrazione di tutti i**
- **trascritti di un gene**

Form Grid

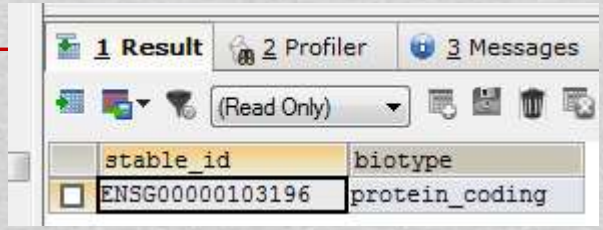
stable_id	stable_id	biotype
ENSG00000005955	ENST00000485685	protein_coding
ENSG00000005955	ENST00000304718	protein_coding
ENSG00000005955	ENST00000495023	protein_coding
ENSG00000005955	ENST00000486208	protein_coding
ENSG00000005955	ENST00000483256	protein_coding

•Ensembl core db schema (II) :
•estrazione su base *posizionale*

•**SELECT** gene_stable_id.stable_id, gene.biotype
•**FROM** seq_region INNER JOIN gene USING (seq_region_id) INNER JOIN gene_stable_id USING (gene_id)
•**WHERE NOT(**gene.seq_region_start>**84966302** OR gene.seq_region_end<**84826528)** AND seq_region.name = '**16**';



•Con queste coordinate trova solo il gene **ENSG00000103196** … sarà vero?



•Ensembl SQL :

•estrazione basata su pattern di testo

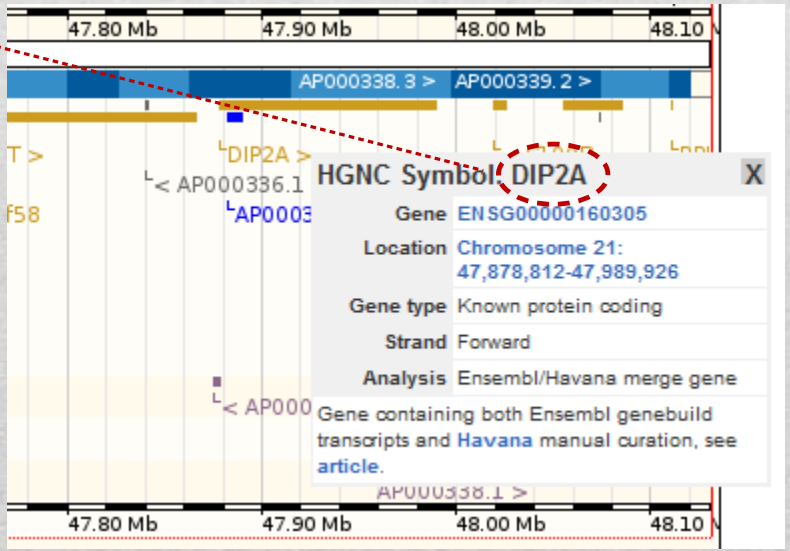
•**SELECT** gene_stable_id.stable_id, gene.biotype, xref.display_label **FROM** gene_stable_id INNER JOIN gene USING (gene_id) INNER JOIN xref ON (gene.display_xref_id = xref.xref_id) **WHERE** xref.display_label **LIKE 'DIP2%'**;

1 Result | 2 Profiler | 3 Messages | 4 Table Data

(Read Only) Form Grid

stable_id	biotype	display_label
ENSG00000160305	protein_coding	DIP2A
ENSG00000066084	protein_coding	DIP2B
ENSG00000151240	protein_coding	DIP2C

•**NBB**: Tabella **xref** (**external reference**) fa riferimento ad un dato mappato sul genoma ma **ORIGINARIO** di un'altra banca dati. Nel caso (ad es.) di **DIP2A** si tratta di un gene symbol definito da **HGNC** (HUGO Gene nomenclature committee)



•Esercizi (SQL)

•Scrivete una query SQL che restituisca tutti i trascritti di un gene a vostra scelta (**3 pt**)

•Scrivete una query che restituisca **IL NUMERO** degli pseudogeni umani annotati in Ensembl (**3 pt**)

•Scrivete una query che restituisca tutti i geni del cromosoma 1 di tipo **diverso** da protein_coding (**3 pt**)

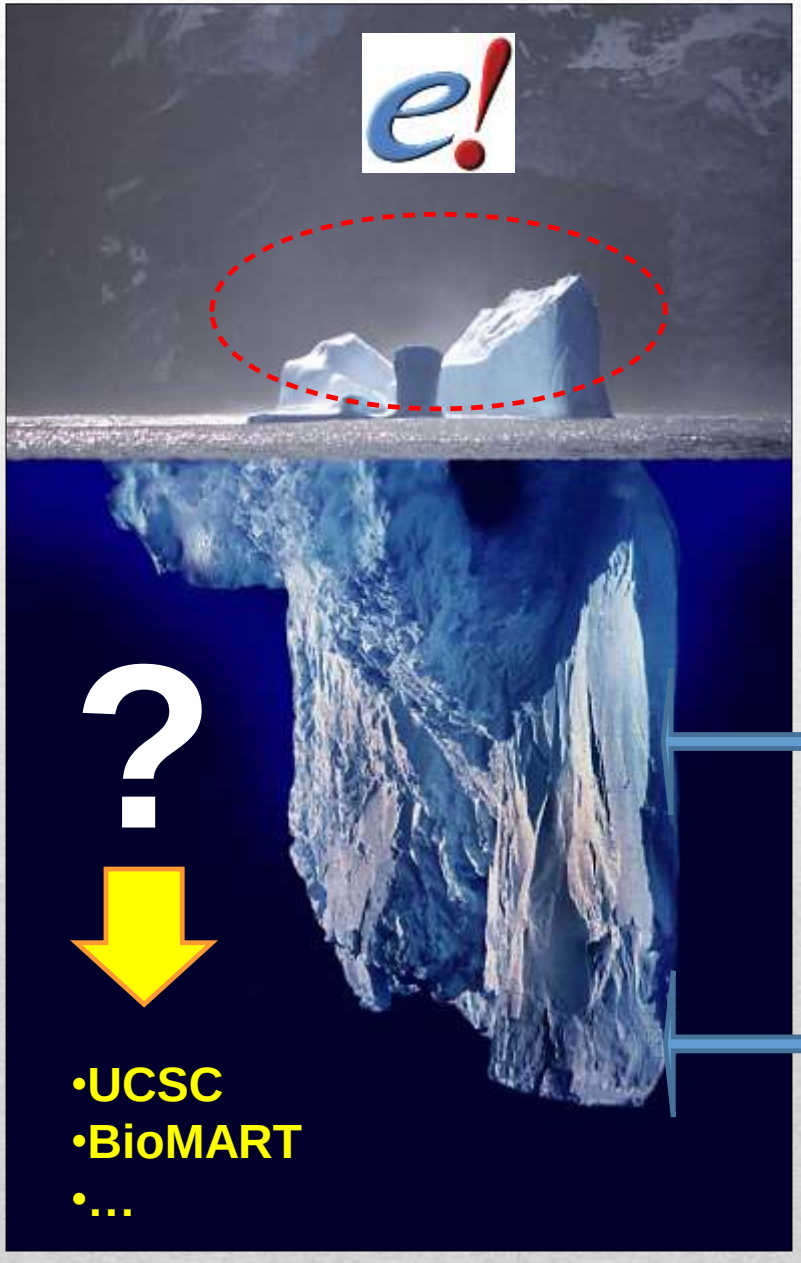
•Scrivete **uno script** a cui passare come parametro il nome di un gene e che restituisca il numero dei suoi trascritti e, per ciascun trascritto, I nomi e le posizioni dei suoi esoni . Potete realizzare l'esercizio mediante **più query** successive (gene → trascritti, per ogni trascritto → esoni) (6 **pt**).

•Scrivete **uno script** a cui passare come parametro delle coordinate genomiche e che restituisca le simple_feature annotate nella regione genomica, la loro posizione ed il loro tipo . NB: questo esercizio **richiede l'utilizzo di INNER JOIN da una tabella che dovete identificare ad altre due tabelle: seq_region e analysis** (**6 pt**).

•RIEPILOGO (I):

•Alcune banche dati (non tutte) sono disponibili sottoforma di database relazionali pubblici
•Possiamo interrogarle in vari modi (SQL/API dedicate)
•Il **linguaggio di programmazione** delle eventuali API disponibili varia da banca dati a banca dati
•Abbiamo visto una modalità d'accesso (SQL) alla componente core di un browser genomico (Ensembl) tra i più utilizzati (ma ne esistono altri!!!)
•Nel caso in cui una banca dati non permetta l'accesso diretto (SQL o API) solitamente permette scaricare I suoi dati sottoforma di file di testo… I dati sono sempre dati, ma così la loro gestione è **molto più complicata** (inoltre i files possono essere molto grandi (Gb)).

•RIEPILOGO (II):



•Anche se abbiamo visto «da vicino» una banca dati (Ensembl), abbiamo visto solo la «**punta dell'iceberg**» …

•Per quanto riguarda **Ensembl**:

•**Altre API:**

•API **Variation** (variabilità genetica)
•API **Compara** (Genomica comparata)
•API **FunctionalGenomics** (il nome dice tutto)

•**SQL:**
•Databases per genomica comparata, genomica funzionale, variabilità genetica ecc … per **OGNI SPECIE !**

•RIEPILOGO (II):

•«INDIRIZZI UTILI» (warning: possono cambiare da release a release … controllate sempre il sito della banca dati.):

•UCSC genome browser:

- host: genome-mysql.cse.ucsc.edu
- user: genome
- access type: SQL

•Vedere anche:

- UCSC Table browser (interfaccia web per «costruire» query):
- http://genome.ucsc.edu/cgi-bin/hgTables

•BIOMART:

- host: martdb.ensembl.org
- user: anonymous
- port: 5316
- access type: SQL (API disponibile)