Giornate sugli sbocchi professionali
Del corso di Laurea in
Biotecnologie Industriali
(BIOTIN)

**Oristano 23/24 Aprile 2013**

BIOCLOUD
SEARCH ENGENE

URL *http://biocloud.unica.it*
email *biocloud@unica.it*

**Emanuele Pascariello**
emanuele.pascariello@gmail.com

CONSORZIO UNO
L'Università a Oristano

REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA

# Scenario attuale

Negli ultimi 15 anni si è assistito ad una rivoluzione in campo Biomedico, che ha determinato la generazione di una enorme mole di dati. **Introduzione di tecnologie ad alto flusso**

## Nuovi metodi di indagine

- ✓Gene Expression profiling by array
- ✓Expression profiling by high throughput sequencing
- ✓SNP genotyping by SNP array
- ✓Next Generation Sequencing

Generazione di una enorme mole di dati in tempi sempre più brevi ed a costi sempre inferiori.

## Cosa è cambiato?

**Scenario precedente:**
La quantità di dati generati dalle tecnologie disponibili era inferiore alle nostre capacità di gestirli

High – Throughput Technologies

A T T C G C G A T T A C G T A A T C G A A
T A A G C G C T A A A T G C A T T A G C T T

MERMLPLLALGLLAAGFCPAVLCHPNSPLDE
ENLTQENQDRGTHVDLGLASANVDFAFS

Annotazioni da processi manuali
Ed automatizzati

**Scenario attuale:**
La quantità di dati generati dalle tecnologie ad elevato flusso superano o mettono costantemente alla prova la nostra capacità di gestirli

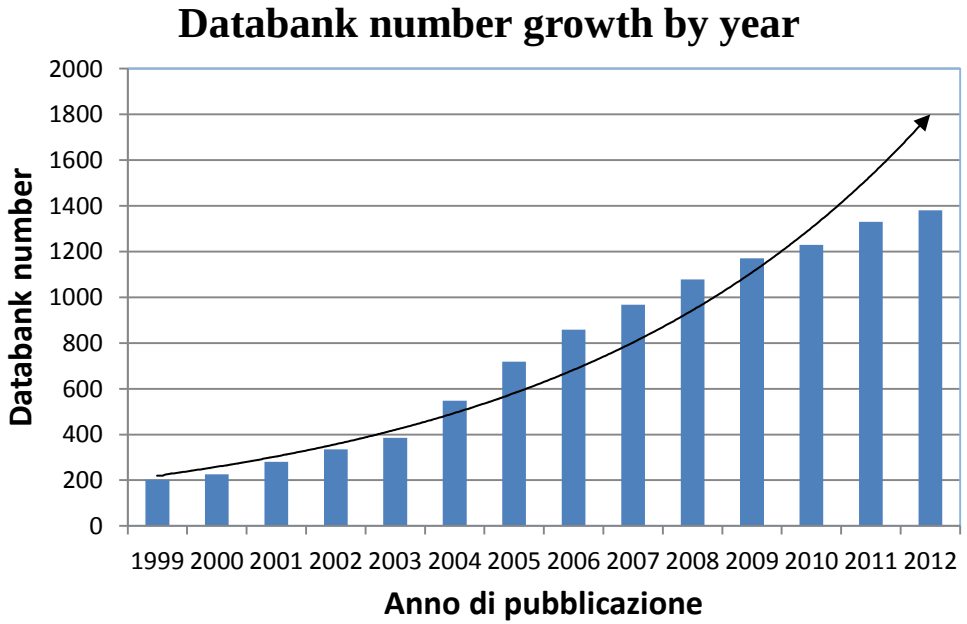
# Opportunità e conseguenze

Crescita del numero di Banche dati di tipo:

✓**Primario** (DNA, RNA, Proteine)
✓**Derivato** o specializzato (EST, SNP, Genomi, Microarray data, pathways, genetic disorders

✓I dati generati presentano elevata dimensionalità – struttura
✓Sono quantitativamente molto consistenti
✓Obsolescenza del dato
✓Presentano fenomeni di ridondanza esterna e interna.

**Problema**: Molte banche dati diverse, molti Accession e riferimenti diversi.
Necessità di consultare differenti servizi/ banche dati per ottenere informazioni specifiche.



## Scopo del progetto

✓Creazione di un servizio di integrazione di una parte consistente di queste informazioni provenienti da banche dati primarie e derivate

✓Riorganizzazione con un approccio “modulare” del dato di origine

✓Accesso mirato alle banche dati esterne

✓Web usability

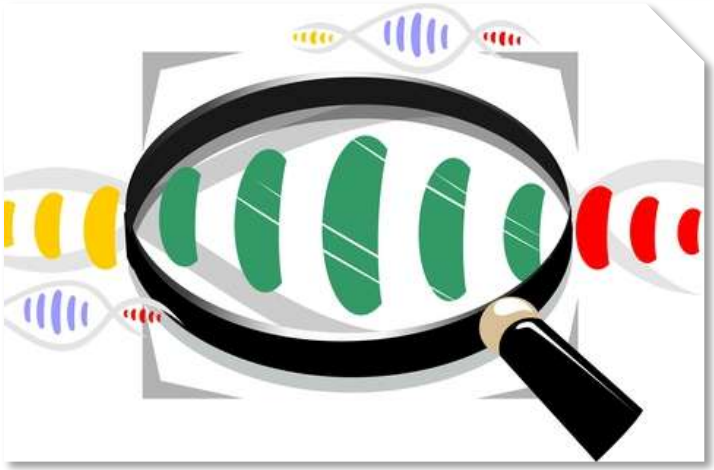
✓Approccio *Pay as you go*

## Intento

✓Fornire un servizio per la ricerca di informazioni geniche, farmacologiche e fenotipiche nell'uomo, più facile e intuitivo

✓Raccolta di grossa parte delle informazioni esterne in unico punto.

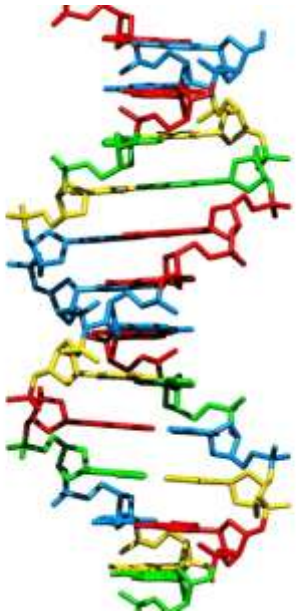
✓Un **Hub** verso le maggiori fonti e Web-services a livello mondiale per ulteriori approfondimenti



## Su cosa si basa Biocloud search enGene

**Biocloud search enGene** si basa sui dati contenuti nei principali database pubblici di Biomedicina.

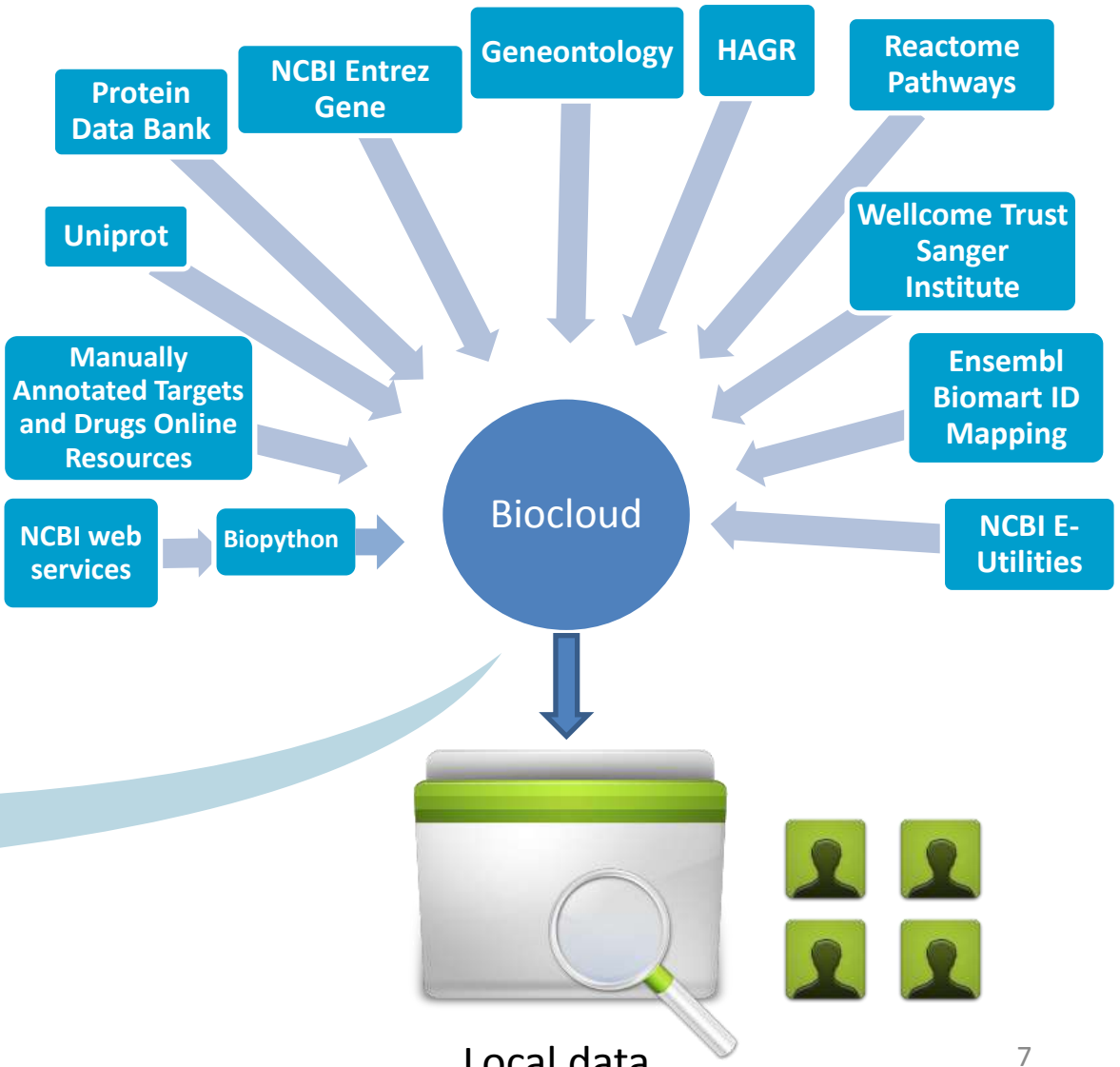
Nella versione attuale comprende annotazioni relative all'***Homo Sapiens***



# Come funziona

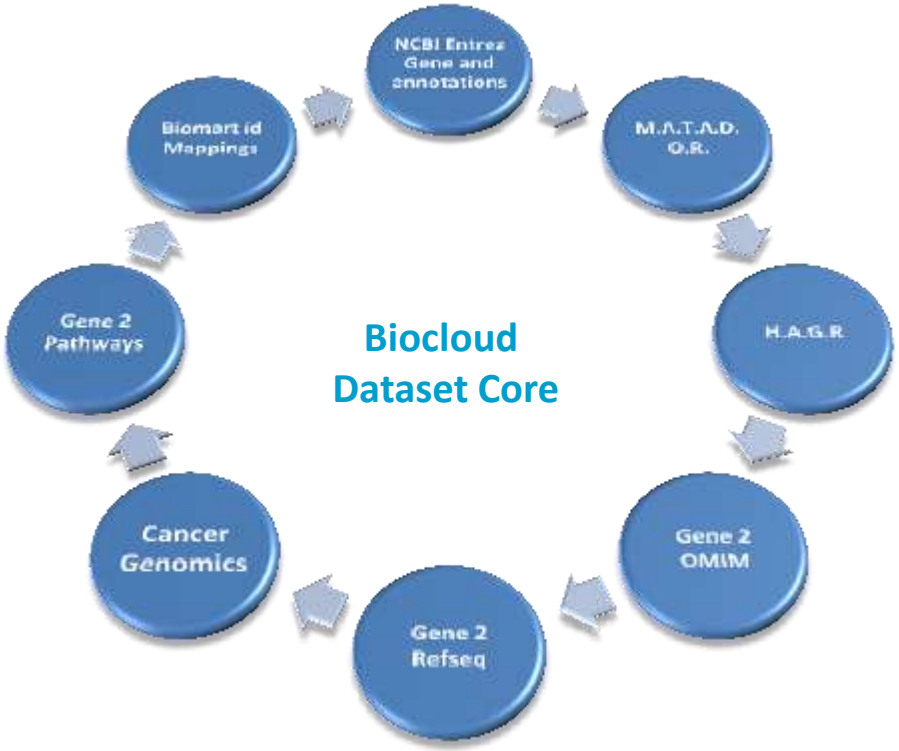
✓**Redirect coerente con la query** verso le principali banche dati, primarie e derivate per ulteriori informazioni specifiche

✓Eliminazione della necessità di orientarsi all'interno di tali servizi.



# Come è fatto – Local data

L'applicazione è fortemente ***Gene – Centric*** e colleziona le annotazioni sui geni umani provenienti dalle banche dati tra le principali e maggiormente accreditate



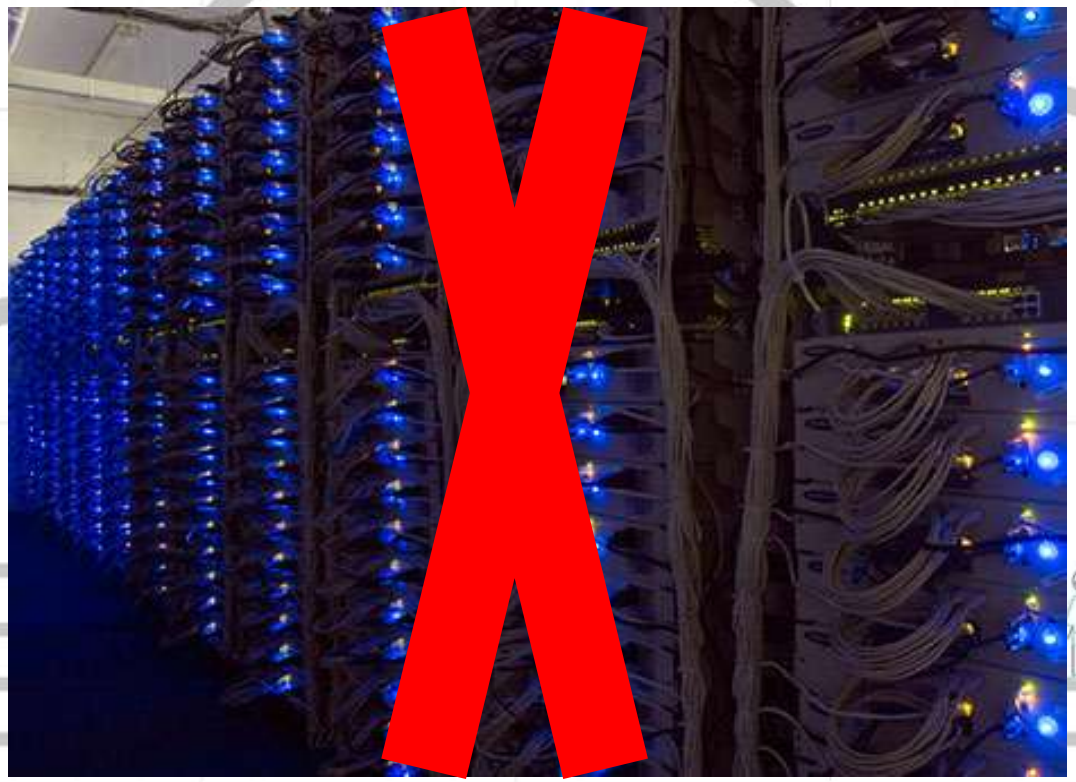

**Analisi struttura Datasets**

## Relazioni tra i dataset

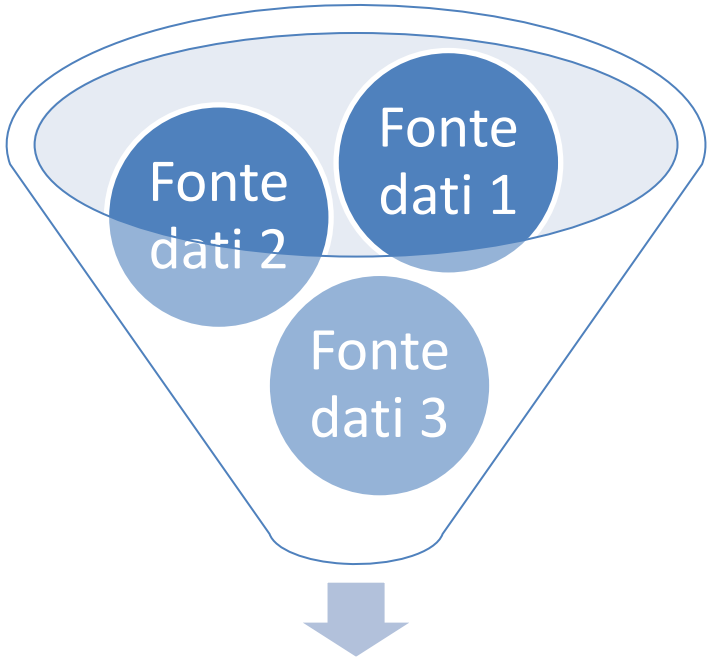
- ✓*Aging*
- ✓*Fenotipo patologico*
- ✓*Sequenze di riferimento nucleotidiche e aminoacidiche*
- ✓*Strutture proteiche 3D*
- ✓*Networks di interazioni*
- ✓*Interazione farmaco – prodotto di espressione*
- ✓*Definizione termini ontologici*
- ✓*Pathways*
- ✓*Aspetti genomici della sensibilità a chemioterapici oncologici*
- ✓*Tipologia di relazione tra geni*
- ✓*Relazioni tra geni omologhi*

# Vantaggi

L’utilizzo di tecnologie cloud e di database non relazionali, permette di scalare con minori problemi di natura sistemistica l’enorme mole di dati in costante crescita.



## Nella realtà




**Biocloud search enGene**

Viene fatta la cattura “al volo” di una notevole quantità di dati mentre altri sono ritenuti all’interno del datastore dell’applicazione ed usati come indice.

In questo modo è possibile by-passare il problema della enorme mole di dati con una efficienza maggiore.

# Come è possibile interrogare Biocloud - Contesti

## Queries sui geni

Nella sua versione attuale Biocloud permette di eseguire delle queries sulla base dei seguenti criteri:

✓Simbolo ufficiale del gene **HGNC** o identificativo univoco **Entrez gene ID**

✓Identificativo dei prodotti di espressione **Uniprot ID**

✓*bulk queries* Ricerca in base a criteri quali: **natura del gene** - **Cromosoma di appartenenza** – **Annotazioni relative ai processi *Aging related* - *Annotazioni farmacogenomiche relative a sensibilità a chemioterapici oncologici***

## Queries sui farmaci

✓Queries su molecole farmacologicamente attive tramite **nome ufficiale** della molecola o tramite **Pubchem ID**

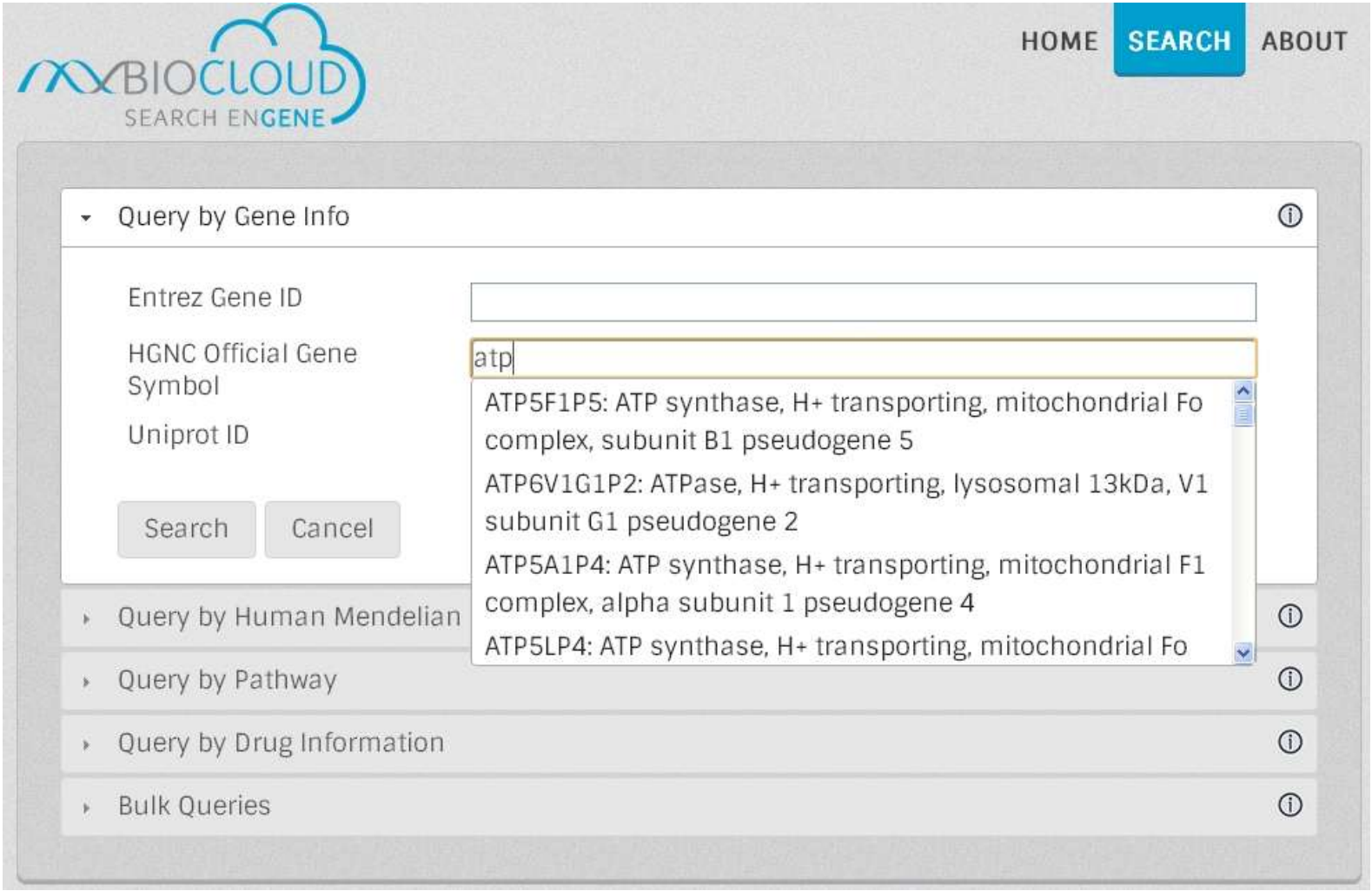
## Queries sui fenotipi

✓Queries basate su **fenotipi** associati a disordini genetici con trasmissione di tipo Mendeliano

## Queries sui Pathways

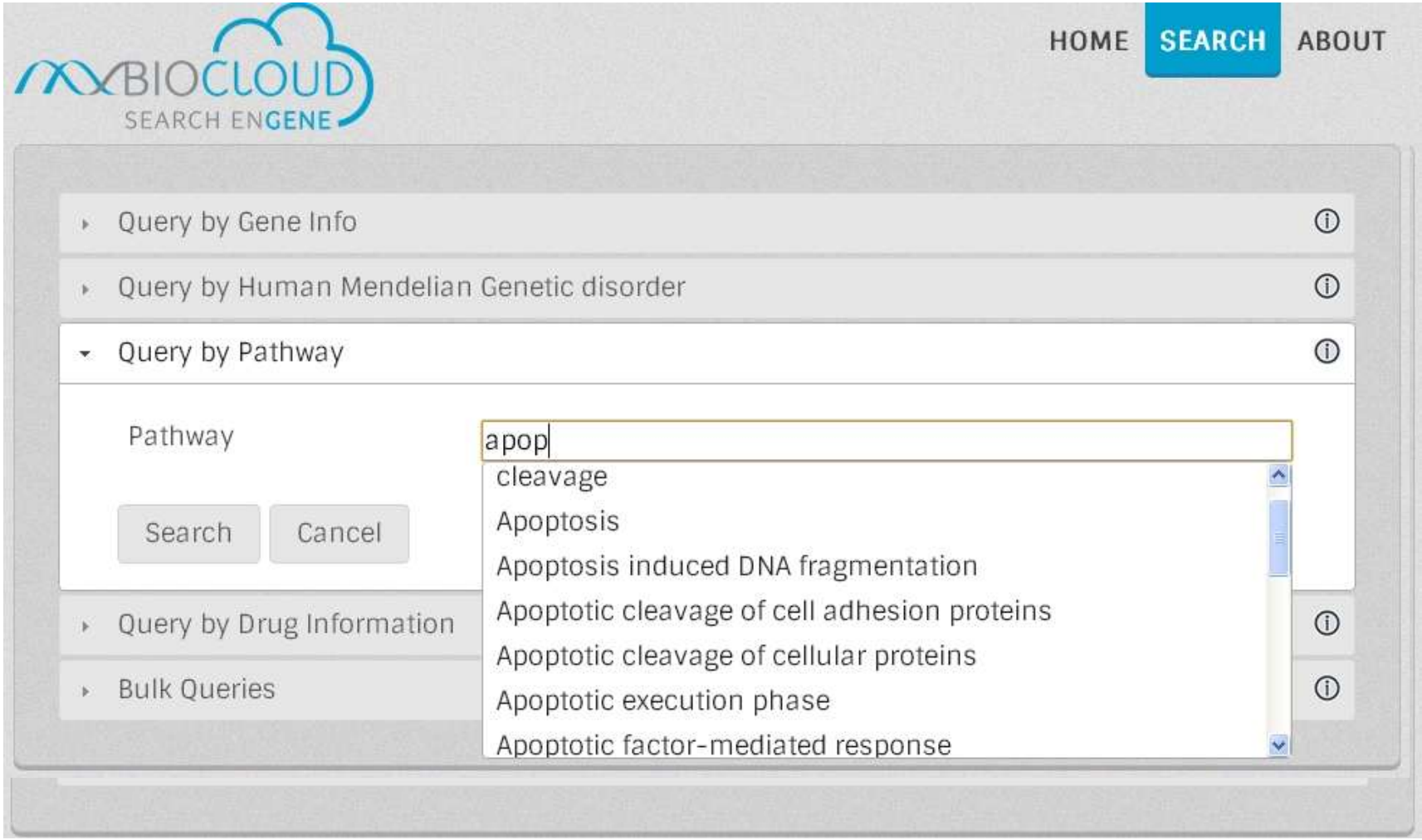
✓Queries basate su nomi descrittivi di pathways metabolici in cui sono coinvolti i geni

**Query tramite l'utilizzo del simbolo ufficiale del gene**



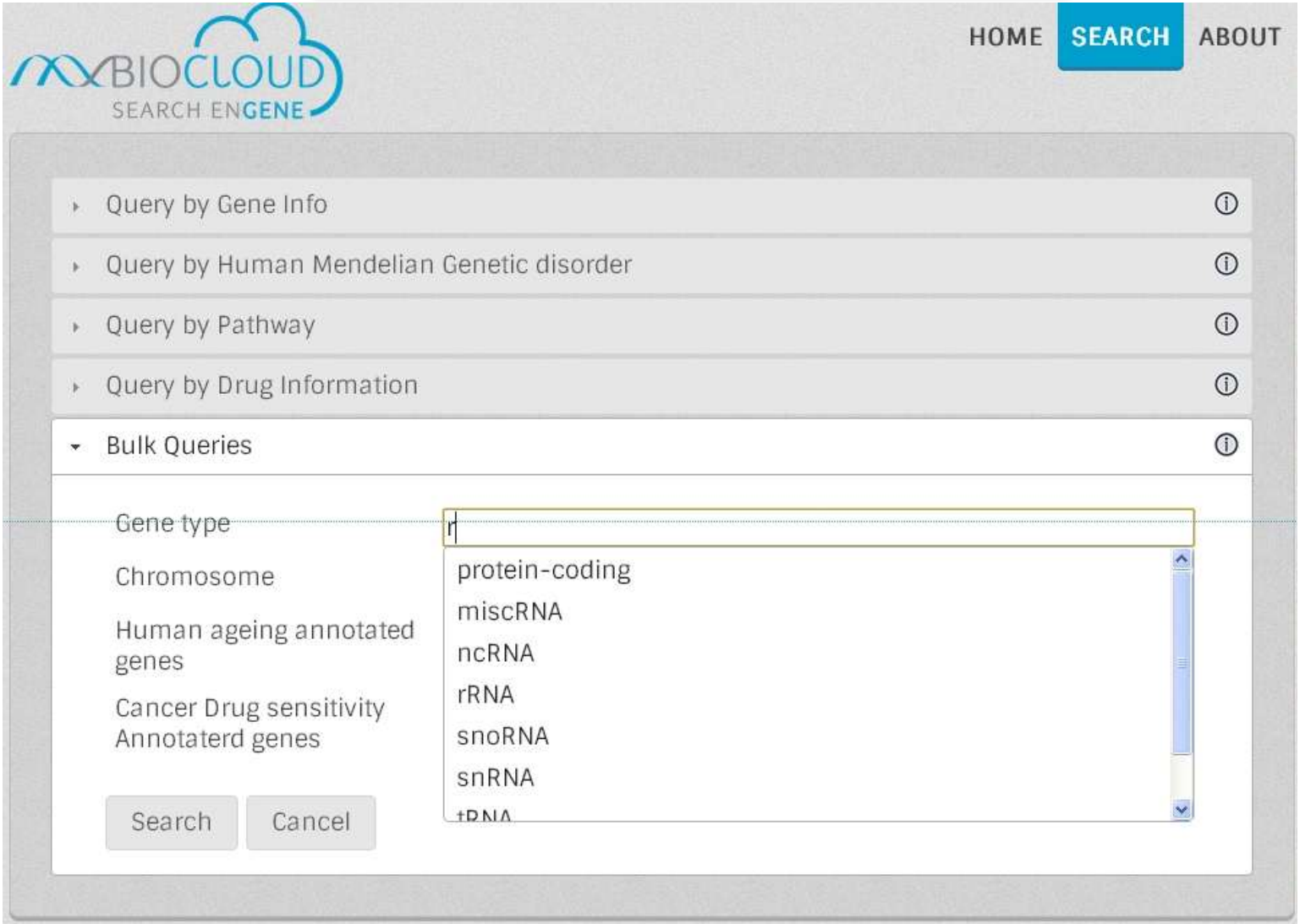
**Query tramite l'utilizzo del Pathway metabolico**

**Con attività farmacologica**

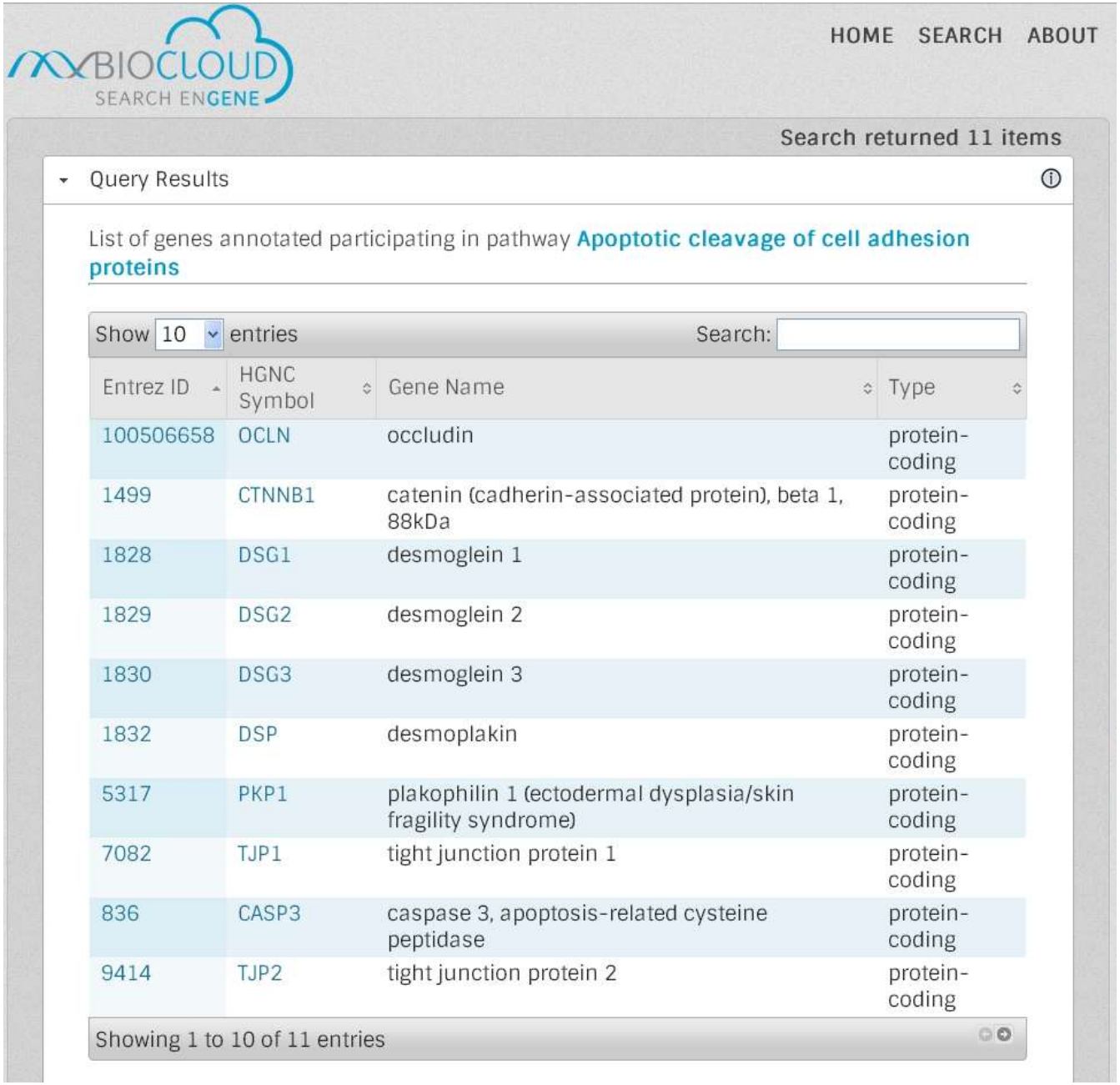


## Query tramite la selezione di criteri quali:

- ✓ Gene biotype
- ✓ Chromosome belonging
- ✓ Aging Annotation
- ✓ Cancer Drug sensitivity



# Risultato query su Pathway metabolico



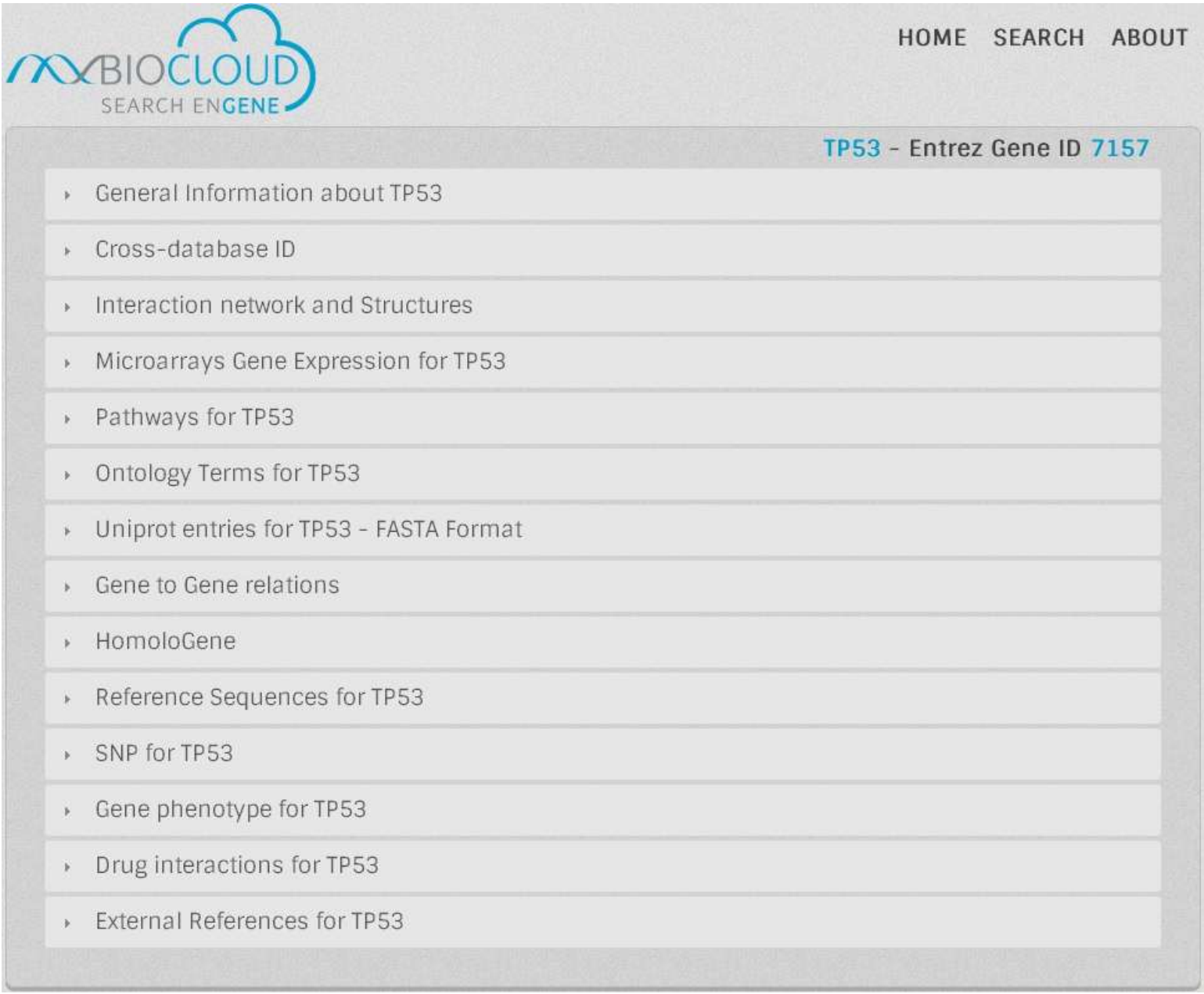
# Cosa restituisce Biocloud

## Queries eseguite sui geni

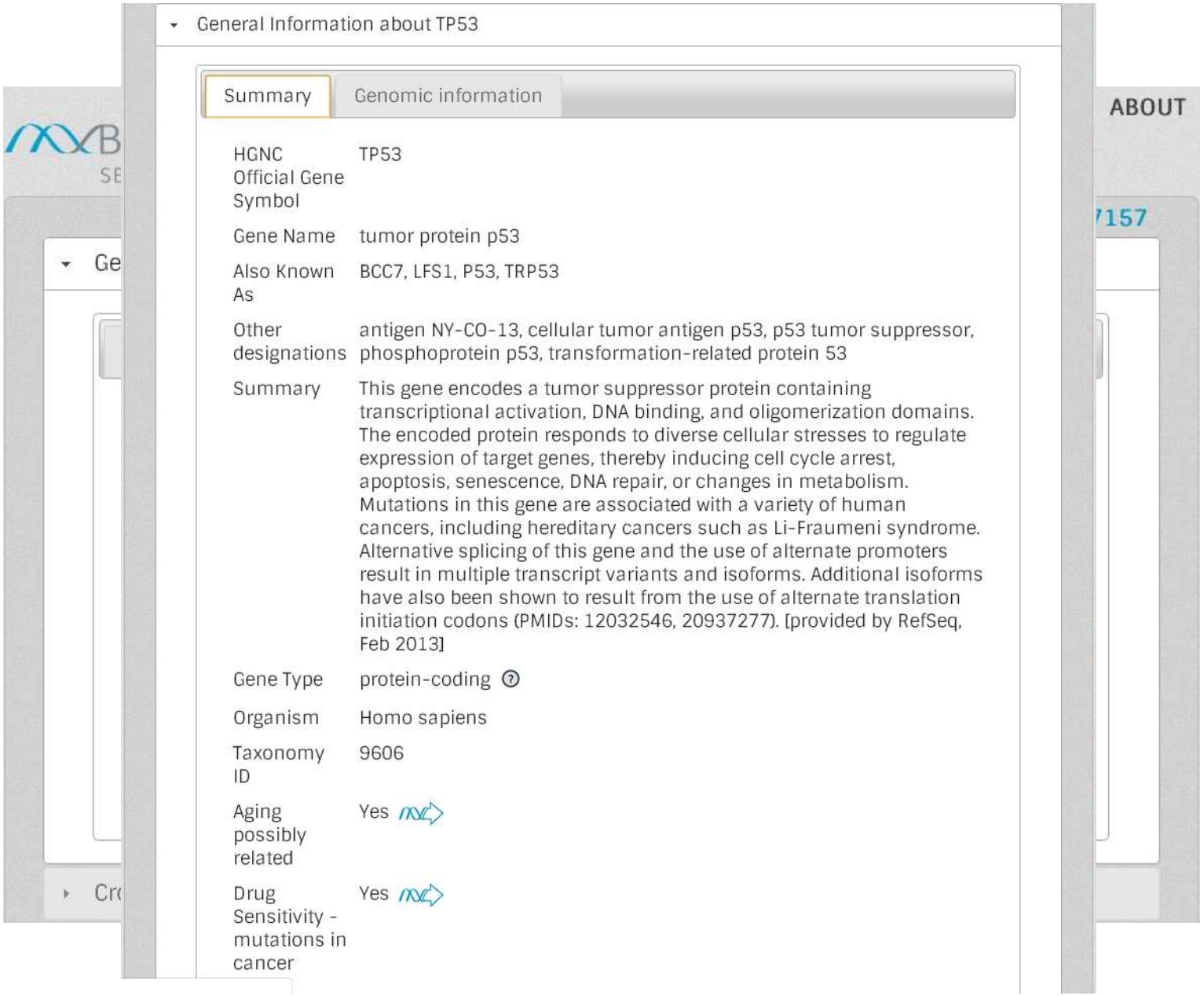
- ✓Summary
- ✓ Full gene name
- ✓ Alieases gene name
- ✓ Taxonomy ID
- ✓ Posizione di **Start** e **Stop** del gene
- ✓Numero degli **Esoni** del gene
- ✓ **Gene type**
- ✓ **HGNC** symbol
- ✓Informazioni legate ad "**Aging**"
- ✓ **Mappa** completa del **Cromosoma** di appartenenza e riferimento cytoband del gene
- ✓ **Mappa** della posizione del **gene** sul cromosoma
- ✓ Informazioni **epigenomiche**
- ✓ Strutture proteiche **3D** relative al gene oggetto della query
- ✓ **Network di interazioni** note e presunte del prodotto di espressione del gene con altri prodotti di espressione
- ✓ Drug sensitivity – mutations in **Cancer**
- ✓ **Sequenza Aminoacidica** del prodotto di espressione in formato FASTA

- ✓Tipologia delle relazioni tra il gene della query e altri geni presenti nel dataset delle annotazioni
- ✓Elenco dei **termini ontologici** che annotano il gene oggetto della query
- ✓Grafo di ciascun termine ontologico
- ✓ **Accession** Nucleotidiche e Aminoacidiche di riferimento
- ✓Relazione con molecole farmacologicamente attive
- ✓ **Pathways** metabolici.
- ✓Elenco dei **fenotipi** patologici annotati in O.M.I.M
- ✓**Datasets e profili di espressione** correlati, da Gene Expression Omnibus relativi ad esperimenti di Microarray gene expression
- ✓**SNP**
- ✓**Homologene**
- ✓Bibliografia Pubmed relativa al gene della query
- ✓Link **diretti** a web services esterni

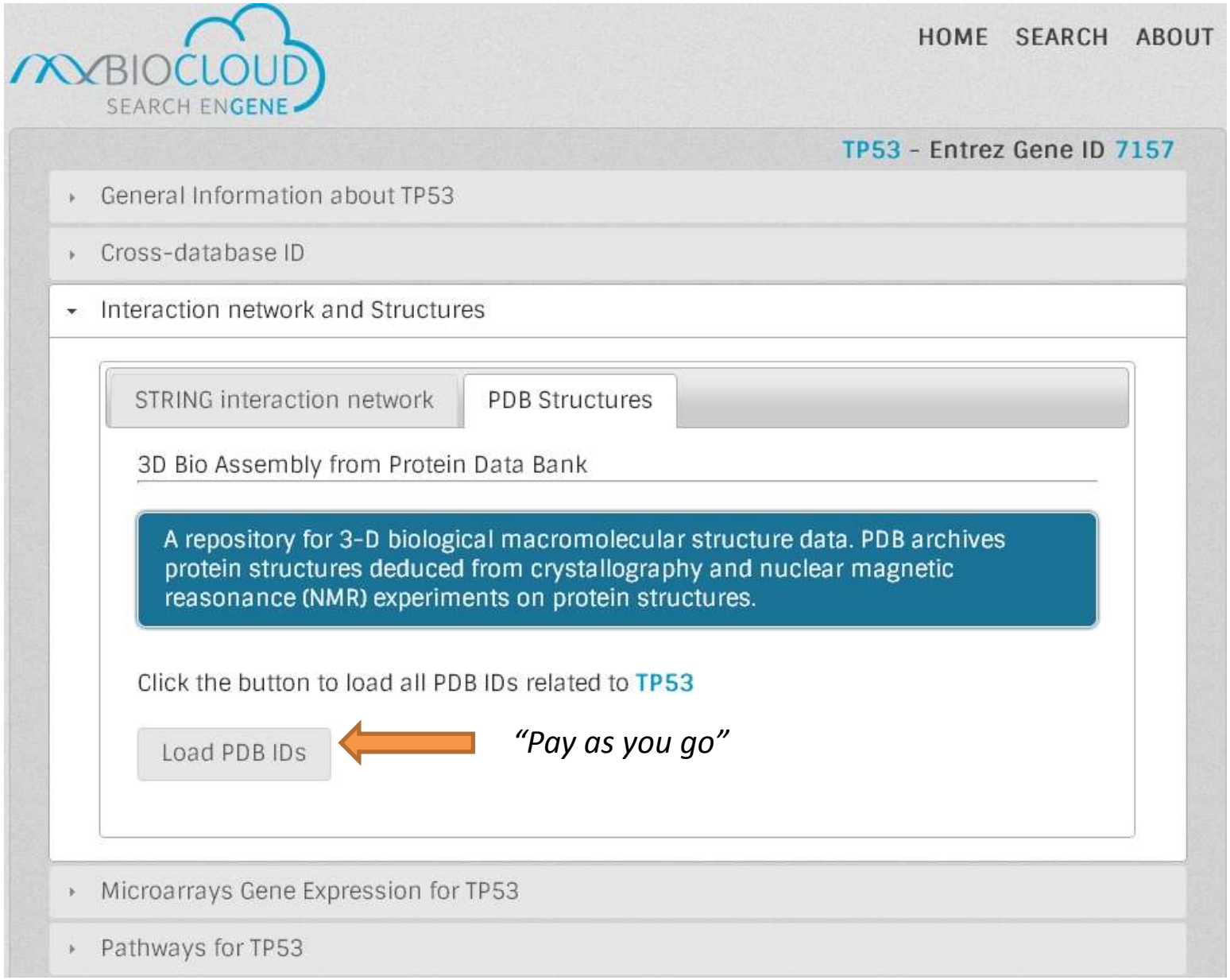
# La struttura dei menu



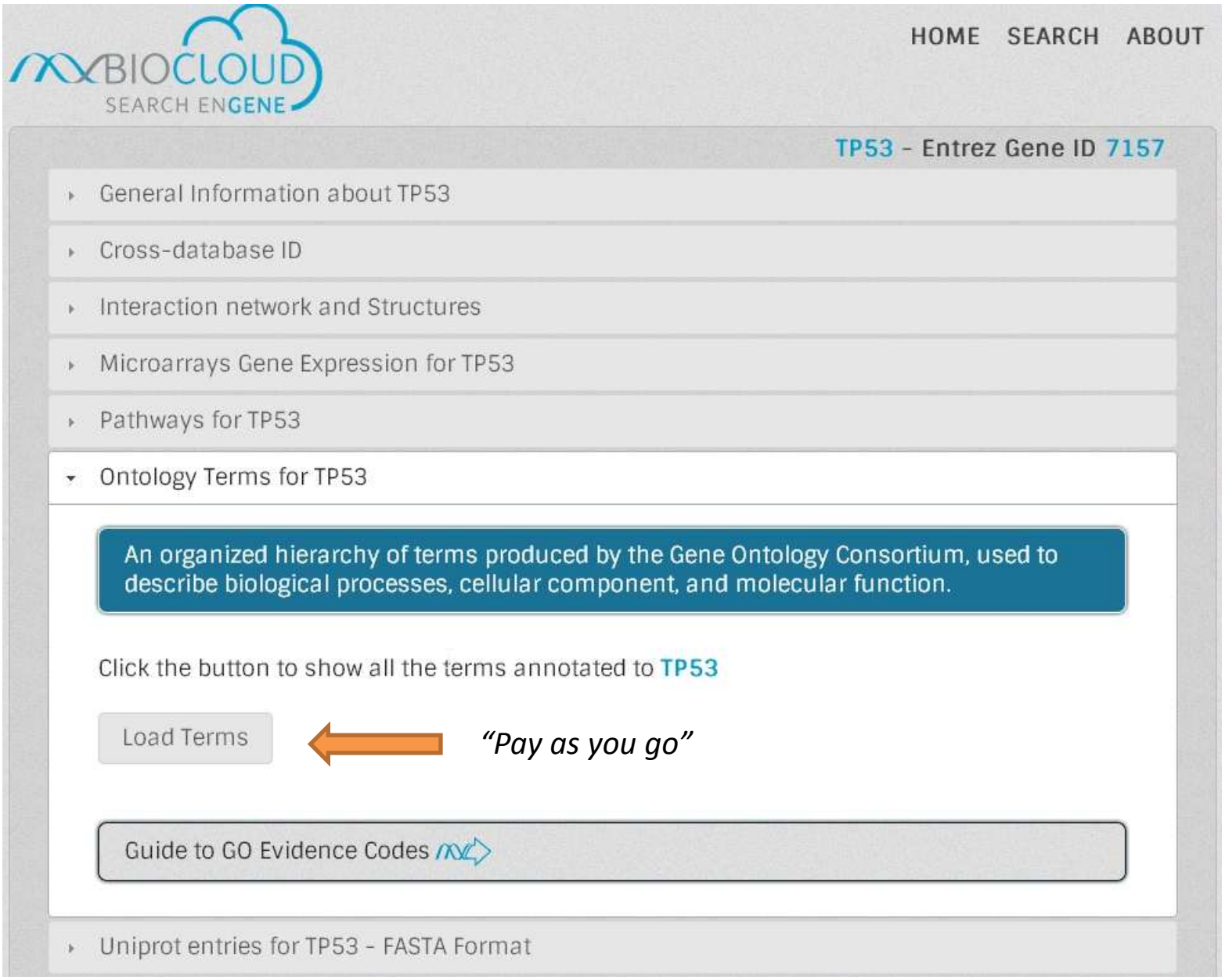
# Menu: general info



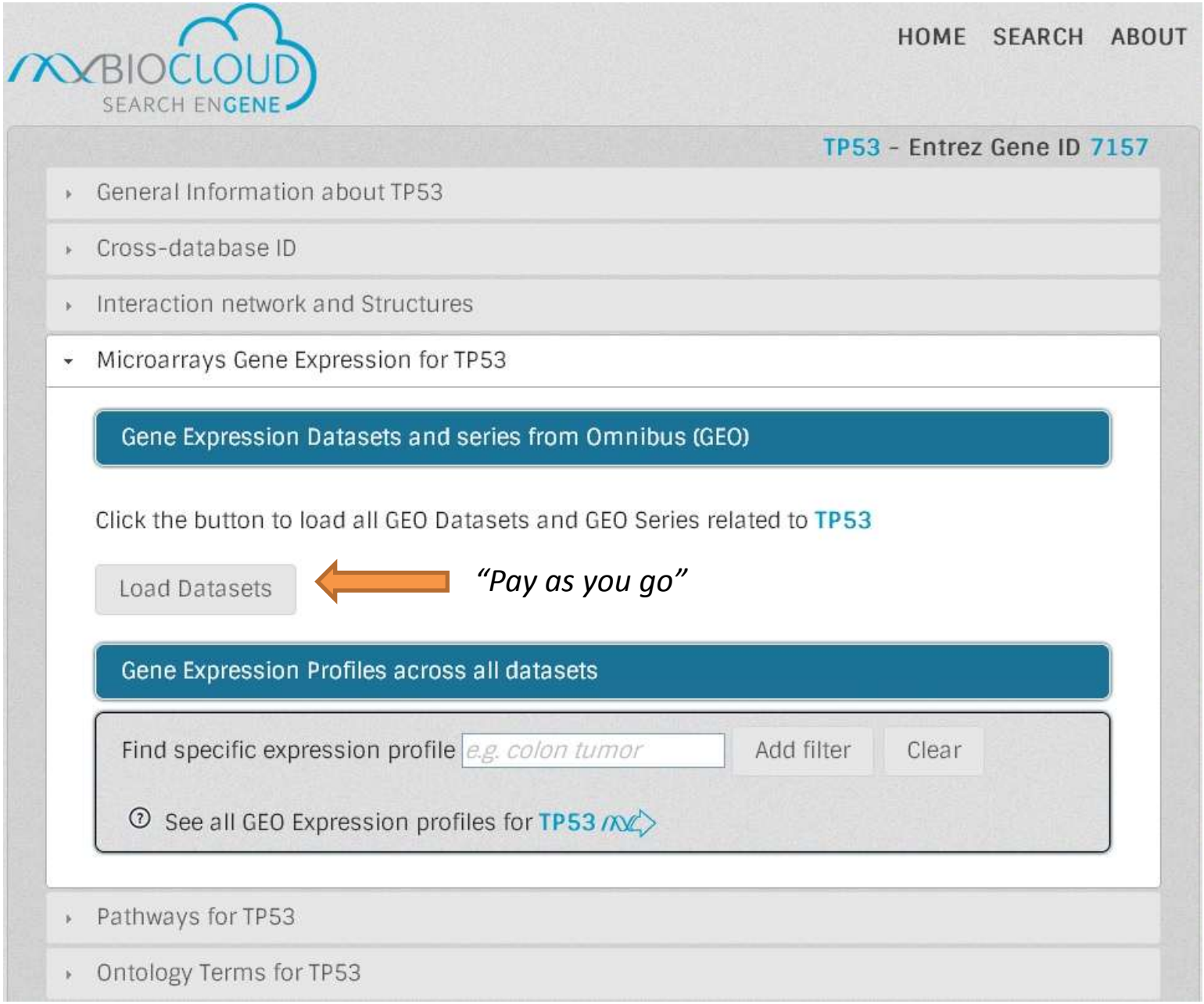
# Menu: Structures and interactions



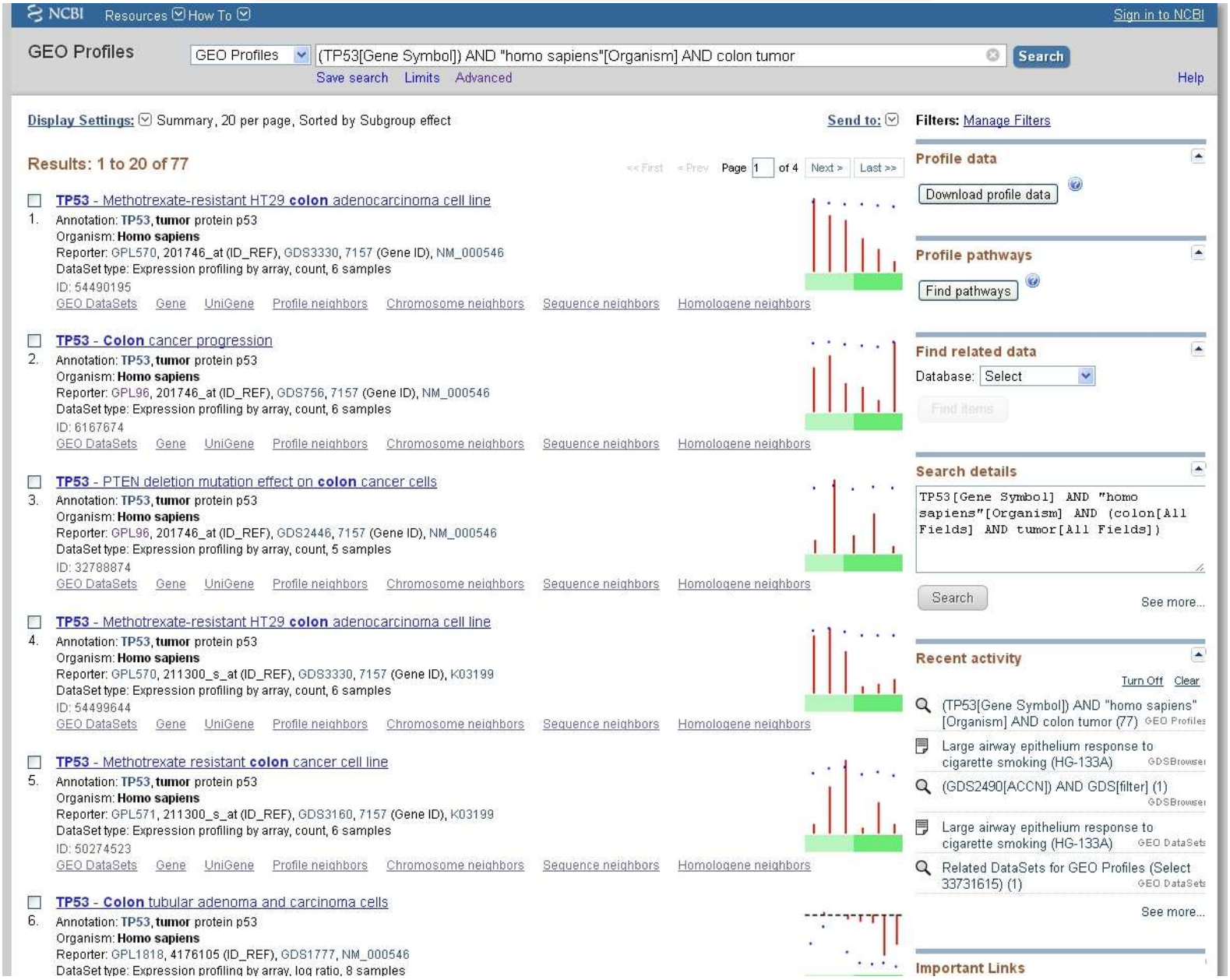
# Menu: Ontology terms



# Menu: Microarray datasets



# Microarray datasets – Pre built query



# Menu: gene phenotype

- HomoloGene
- Reference Sequences for TP53
- SNP for TP53
- Gene phenotype for TP53

**Gene - Morbid associations Data**

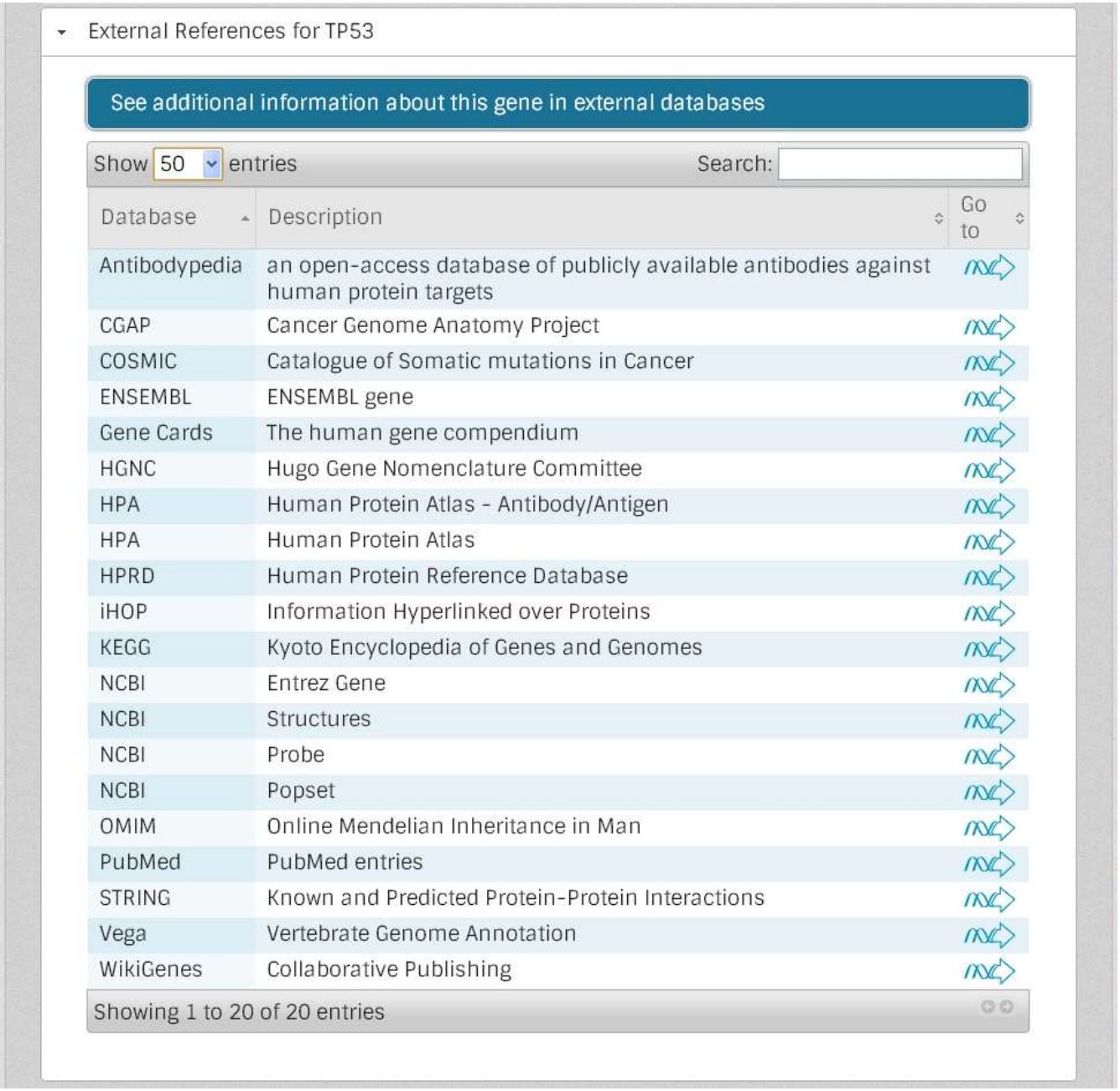
Show 10 entries  Search:

| Phenotype accession | MIM Morbid Description |
|---|---|
| 133239 | ESOPHAGEAL CANCER |
| 151623 | LI-FRAUMENI SYNDROME 1; LFS1 |
| 202300 | ADRENOCORTICAL CARCINOMA, HEREDITARY; ADCC |
| 211980 | LUNG CANCER |
| 260500 | PAPILLOMA OF CHOROID PLEXUS |
| 275355 | SQUAMOUS CELL CARCINOMA, HEAD AND NECK; HNSCC |

Showing 1 to 6 of 6 entries

- Drug interactions for TP53
- External References for TP53

# External references



# Cosa restituisce Biocloud

## Queries su molecole Farmacologicamente attive

✓Drug name

✓Pubchem Id e link diretti alla banca dati NCBI Pubchem

✓Molecular weight

✓IUPAC name

✓Molecular formula

✓Anatomical Therapeutic Chemical (ATC) Code

✓ Lista attività farmacologiche

✓Struttura 2D della molecola

✓Un elenco di prodotti di espressione con i quali interagisce la molecola e la tipologia di tali interazioni

✓ Pubmed Health entries

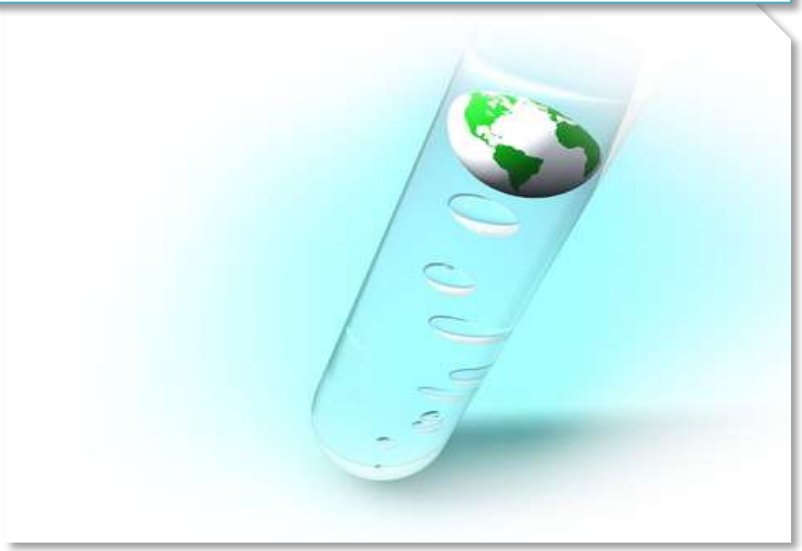
✓STITCH – Network di interazioni note e predette tra la molecola e altri prodotti di espressione
✓Pubchem Bioassay entries

✓Sinonimi

✓Side effects

✓MeSH terms

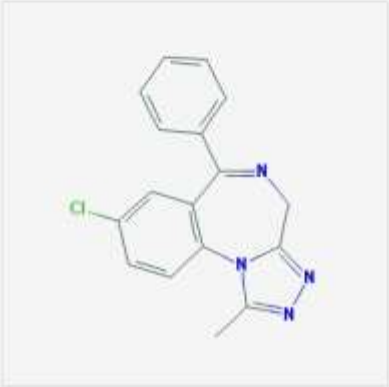
✓Bibliografia Pubmed correlata alla molecola oggetto della query



# Drug detail

alprazolam – Pubchem ID 2118

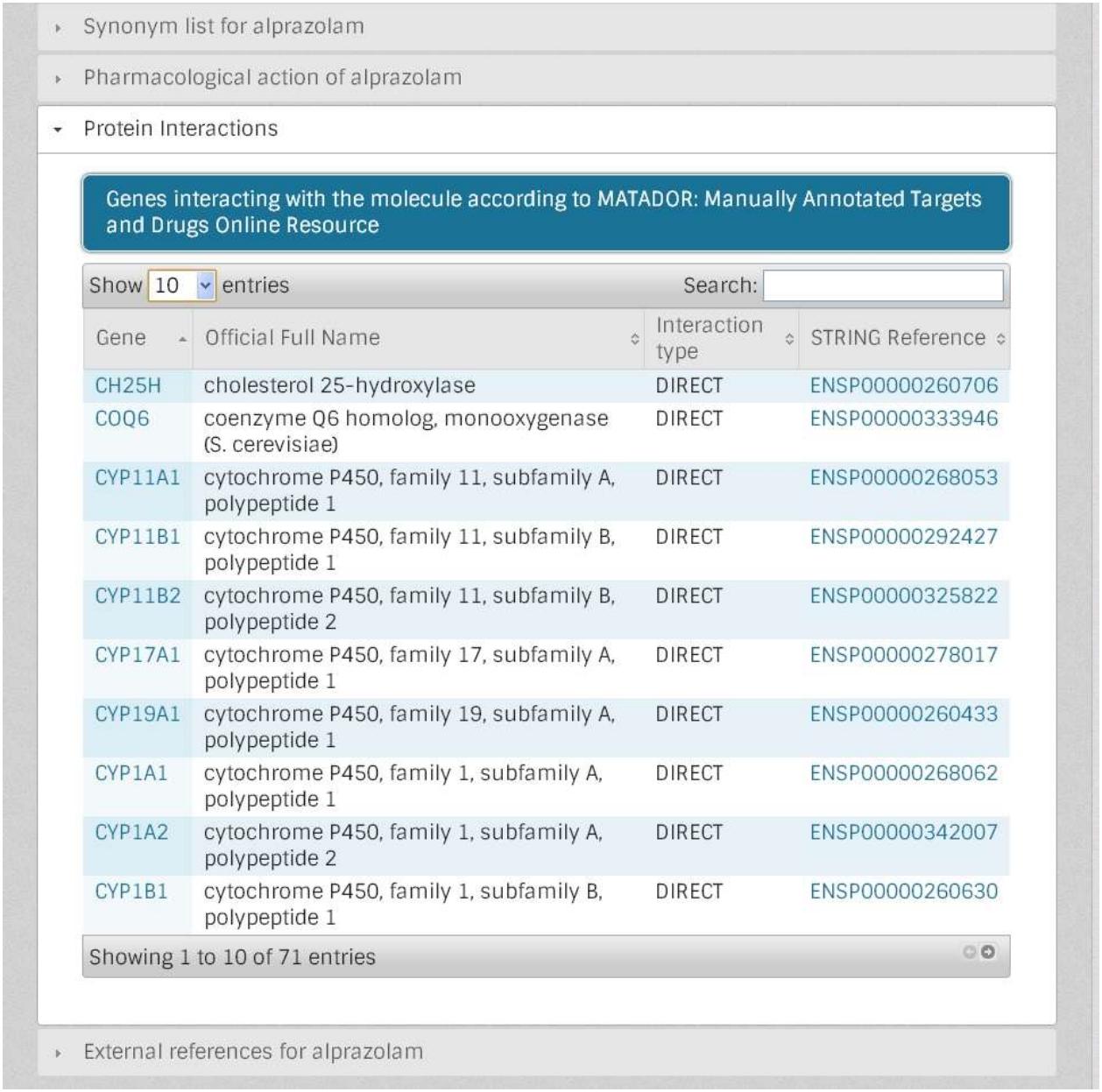
## General Information about alprazolam

| | |
|---|---|
| Drug Name | alprazolam |
| Pubchem Compound ID | 2118 |
| ATC Classification system | N05BA12 |
| Molecular Weight | 308.764920 |
| Molecular Formula | C17H13ClN4 |
| 2D Structure |  |



- Compound Descriptors
- Compound IDs
- MeSH term List for alprazolam
- Synonym list for alprazolam
- Pharmacological action of alprazolam
- Protein Interactions
- External references for alprazolam

# Drug - Protein interactions



Synonym list for alprazolam

Pharmacological action of alprazolam

Protein Interactions

Genes interacting with the molecule according to MATADOR: Manually Annotated Targets and Drugs Online Resource

Show 10 entries Search:

| Gene | Official Full Name | Interaction type | STRING Reference |
|---|---|---|---|
| CH25H | cholesterol 25-hydroxylase | DIRECT | ENSP00000260706 |
| COQ6 | coenzyme Q6 homolog, monooxygenase (S. cerevisiae) | DIRECT | ENSP00000333946 |
| CYP11A1 | cytochrome P450, family 11, subfamily A, polypeptide 1 | DIRECT | ENSP00000268053 |
| CYP11B1 | cytochrome P450, family 11, subfamily B, polypeptide 1 | DIRECT | ENSP00000292427 |
| CYP11B2 | cytochrome P450, family 11, subfamily B, polypeptide 2 | DIRECT | ENSP00000325822 |
| CYP17A1 | cytochrome P450, family 17, subfamily A, polypeptide 1 | DIRECT | ENSP00000278017 |
| CYP19A1 | cytochrome P450, family 19, subfamily A, polypeptide 1 | DIRECT | ENSP00000260433 |
| CYP1A1 | cytochrome P450, family 1, subfamily A, polypeptide 1 | DIRECT | ENSP00000268062 |
| CYP1A2 | cytochrome P450, family 1, subfamily A, polypeptide 2 | DIRECT | ENSP00000342007 |
| CYP1B1 | cytochrome P450, family 1, subfamily B, polypeptide 1 | DIRECT | ENSP00000260630 |

Showing 1 to 10 of 71 entries

External references for alprazolam

# Conclusioni

## Scenario prossimo futuro

E' lecito prevedere una crescita sempre maggiore di dati ad elevata *dimensionalità.*

✓Un numero sempre maggiore di banche dati e servizi è assolutamente atteso

✓**Ridondanza** dei dati in un numero sempre maggiore di repositories pubblici

✓Numero sempre crescente di accession identificativi delle *entries* nei database pubblici.

✓Aumento delle attività di **Data Mining** su grandi volumi di dati collezionati su banche dati differenti con scopi e finalità diverse renderà sempre più difficile orientarsi e ottenere una sintesi di ciò che realmente si sta cercando

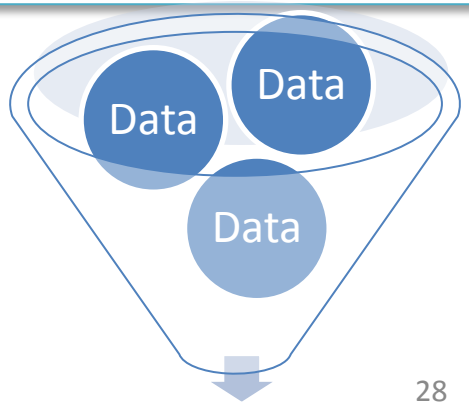
## Sfide da raccogliere

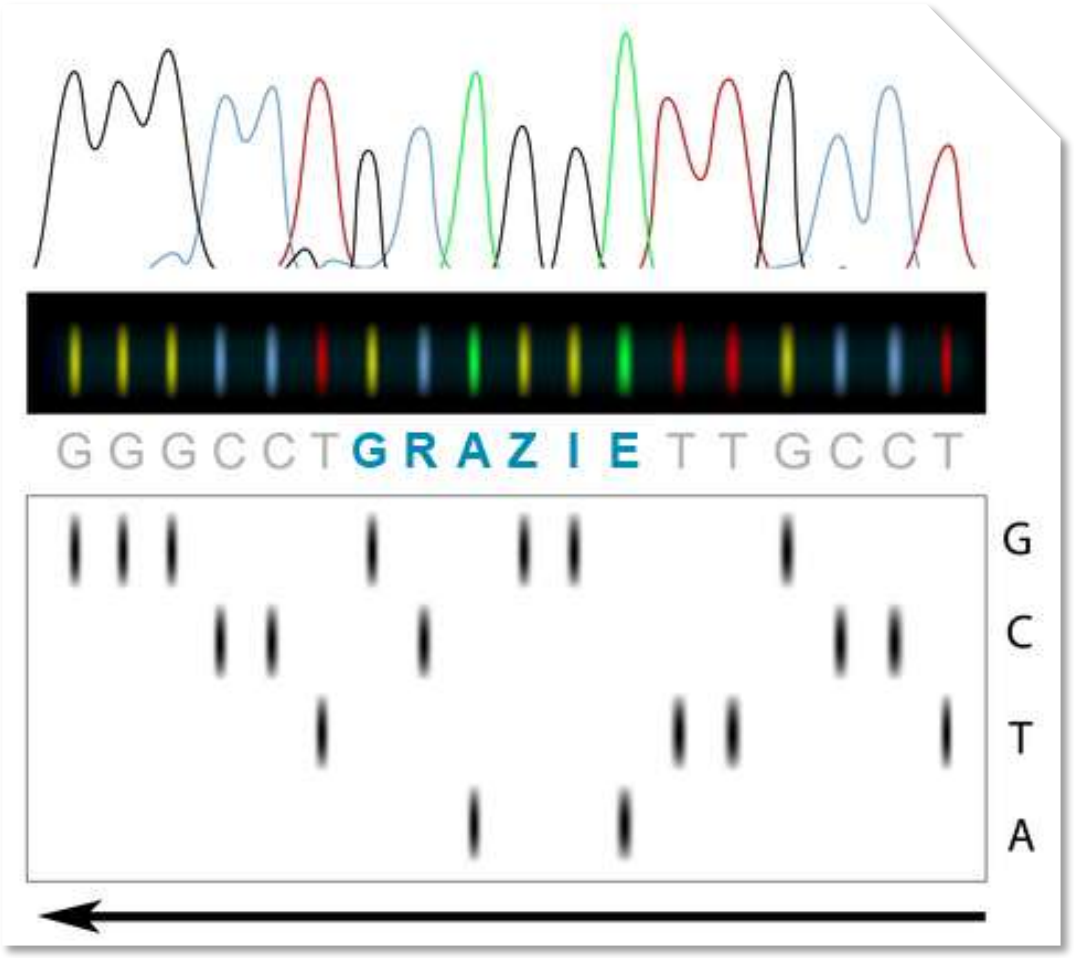
✓Integrazione di questa mole di informazioni costantemente in crescita; **riorganizzazione** in un contesto che sia quanto più possibile vicino alla natura dell'uomo di recepirle.

✓Creazione di **servizi di integrazione** delle numerose fonti – intento del progetto **Biocloud search enGene**

✓Trasformazione del dato in informazione

✓ **Trasformazione delle informazioni in conoscenza**






Un ringraziamento particolare a:

CONSORZIO UNO
L'Università a Oristano

REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA