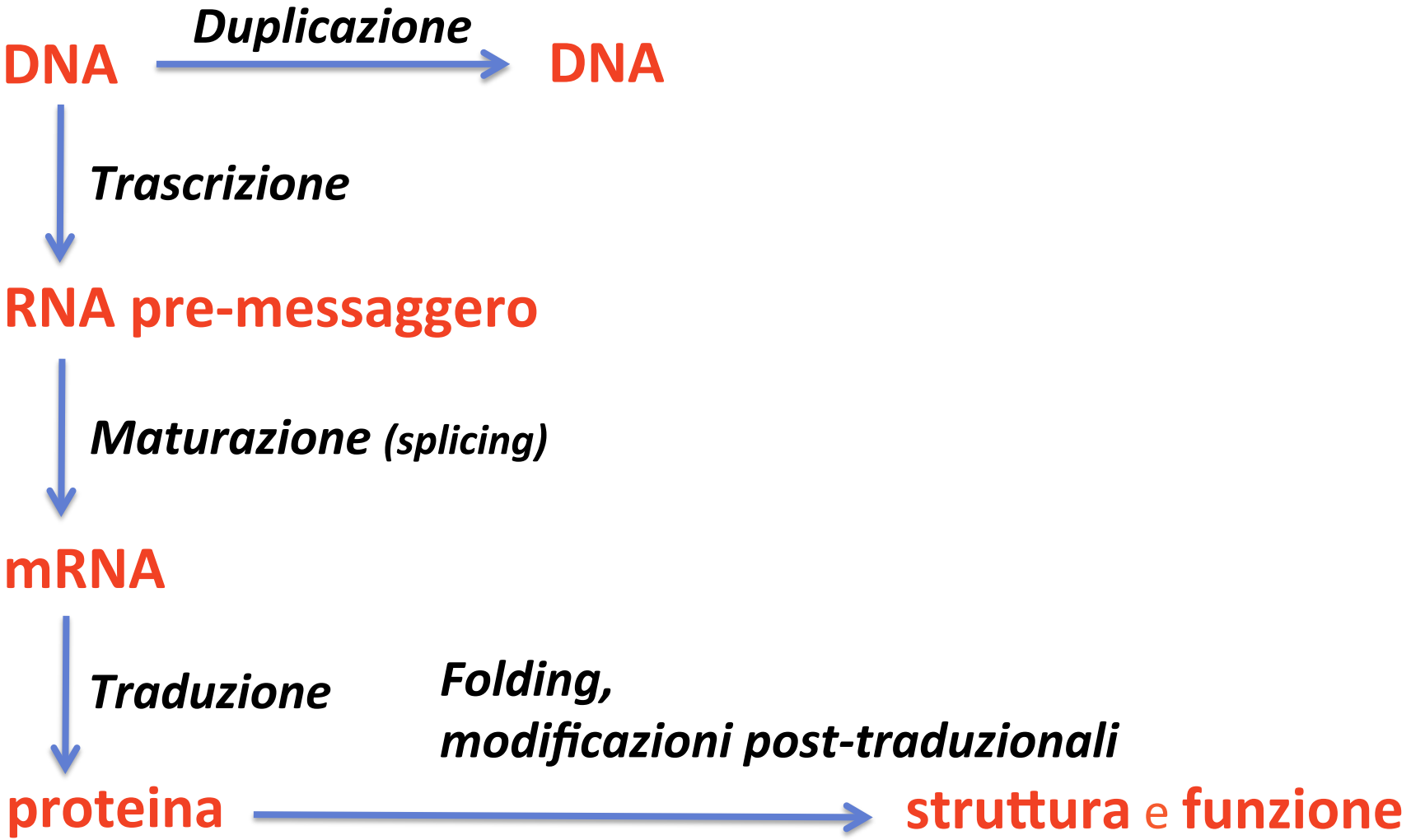
Flusso dell'informazione genetica

Un tempo: "un gene, una proteina"...



Flusso dell'informazione genetica

5'-.....ATGGAAATGCAGAGGATTGCTTGA.....-3'

Open Reading Frame (ORF)

gene

N.B.: questa è una semplificazione estrema!

Flusso dell'informazione genetica

```
5'-......ATGGAAATGCAGAGGATTGCTTGA.......-3'
3'-......TACCTTTACGTCTCCTAACGAACT.......-5'
```

Flusso dell'informazione genetica

5′-.....ATGGAAATGCAGAGGATTGCTTGA.....-3′
3′-.....TACCTTTACGTCTCCTAACGAACT.....-5′

Trascrizione

Duplicazione

5′-..ATGGAAATGCAGAGGATTGCTTGA..-3′
3′-..TACCTTTACGTCTCCTAACGAACT..-5′

Flusso dell'informazione genetica

5′-.....ATGGAAATGCAGAGGATTGCTTGA.....-3′
3′-.....TACCTTTACGTCTCCTAACGAACT.....-5′

Trascrizione

Duplicazione

5′-..AUGGAAAUGCAGAGGAUUGCUUGA..-3′

5′-..ATGGAAATGCAGAGGATTGCTTGA..-3′
3′-..TACCTTTACGTCTCCTAACGAACT..-5′

Traduzione

NH_2-..MetGluMetGlnArgIleAla..-COOH

	T	C	A	G	
T	TTT *Phe (F)*	TCT *Ser (S)*	TAT *Tyr (Y)*	TGT *Cys (C)*	T
	TTC "	TCC "	TAC "	TGC "	C
	TTA *Leu (L)*	TCA "	TAA ***stop***	TGA ***stop***	A
	TTG "	TCG "	TAG ***stop***	TGG *Trp (W)*	G
C	CTT *Leu (L)*	CCT *Pro (P)*	CAT *His (H)*	CGT *Arg (R)*	T
	CTC "	CCC "	CAC "	CGC "	C
	CTA "	CCA "	CAA *Gln (Q)*	CGA "	A
	CTG "	CCG "	CAG "	CGG "	G
A	ATT *Ile (I)*	ACT *Thr (T)*	AAT *Asn (N)*	AGT *Ser (S)*	T
	ATC "	ACC "	AAC "	AGC "	C
	ATA "	ACA "	AAA *Lys (K)*	AGA *Arg (R)*	A
	ATG *Met (M)*	ACG "	AAG "	AGG "	G
G	GTT *Val (V)*	GTC *Ala (A)*	GAT *Asp (D)*	GGT *Gly (G)*	T
	GTC "	GCC "	GAC "	GGC "	C
	GTA "	GCA "	GAA *Glu (E)*	GGA "	A
	GTG "	GCG "	GAG "	GGG "	G

Tabella di conversione del codice genetico in aminoacidi. Le basi gialle si riferiscono al primo nucleotide del codone, quelle rosa al secondo e quelle verdi al terzo.

Flusso dell'informazione genetica

5′-.....ATGGAAATGCAGAGGATTGCTTGA.....-3′
3′-.....TACCTTTACGTCTCCTAACGAACT.....-5′

Trascrizione

Duplicazione

5′-..AUGGAAAUGCAGAGGAUUGCUUGA..-3′

5′-..ATGGAAATGCAGAGGATTGCTTGA..-3′
3′-..TACCTTTACGTCTCCTAACGAACT..-5′

Traduzione

NH_2-..MetGluMetGlnArgIleAla..-COOH

NH_2-.. M E M Q R I A ..-COOH

Flusso dell'informazione genetica

5′-.....ATGGAAATGCAGAGGATTGCTTGA.....-3′

3′-.....TACCTTTACGTCTCCTAACGAACT.....-5′

Trascrizione
Traduzione

NH_2-..MetGluMetGlnArgIleAla..-COOH

NH_2-.. M E M Q R I A ..-COOH

Folding, modificazioni post-traduzionali

NH_2-.. M E M O R I A ..-COOH

Flusso dell'informazione genetica

5′-.....ATGGAAATGCAGAGGATTGCTTGA.....-3′
3′-.....TACCTTTACGTCTCCTAACGAACT.....-5′

Trascrizione
Traduzione

NH_2-..MetGluMetGlnArgIleAla..-COOH
NH_2-.. M E M Q R I A ..-COOH

Folding, modificazioni post-traduzionali

NH_2-.. M E M O R I A ..-COOH

struttura e funzione

fenotipo

DATABASE o **banca dati** (generica)

- “Individuare tutte le pubblicazioni rilevanti;
- analizzarne criticamente i dati e risolvere i risultati conflittuali;
- presentare i dati in un formato che possa riflettere quegli aspetti della struttura che sono stati determinati sperimentalmente e quelli che possono ragionevolmente essere ricavati dell’omologia;
- identificare il materiale inerente alla funzione chimica, alla sorgente biologica, al controllo genetico e all’origine evolutiva…”

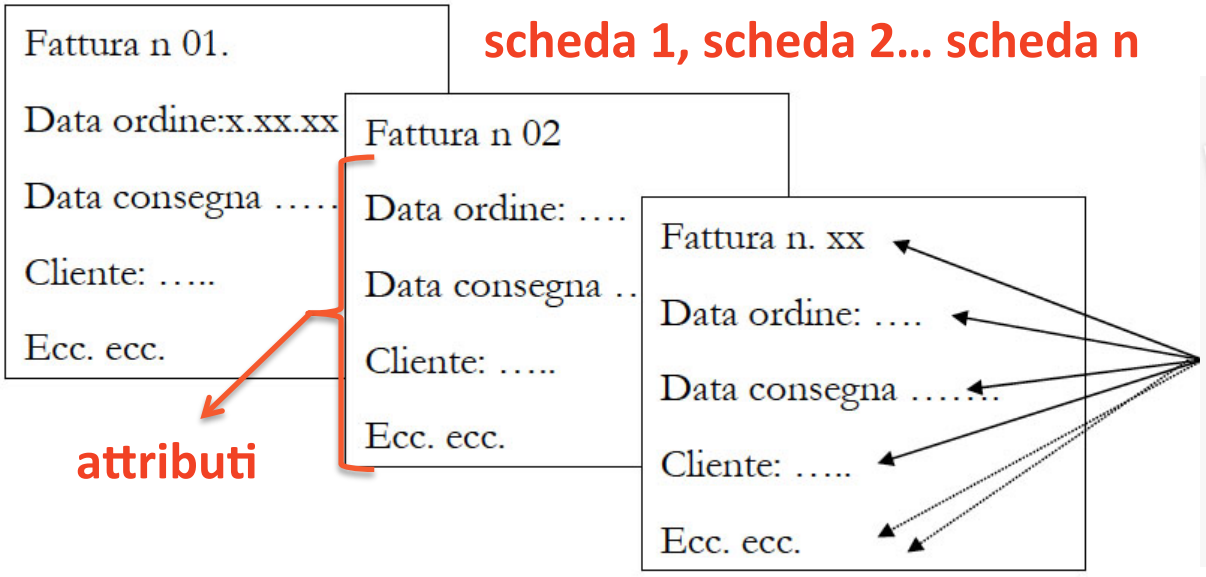
M.Dayhoff, Atlas of protein sequence and structure

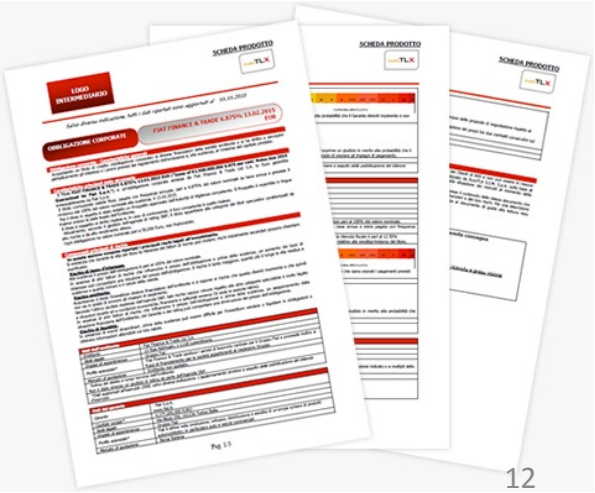
DATABASE o banca dati

- Un **archivio** (o **schedario**) non può contenere 'tutto' ; deve essere costruito esattamente per lo scopo a cui serve.
- La **scheda** è l'elemento principale dello schedario, ma quello che caratterizza l'archivio è il contenuto delle schede.
- Ogni scheda deve contenere le informazioni cioè gli **attributi** (chiamati anche categorie di informazioni) che caratterizzano l'elemento

Esempio: **schedario di fatturazioni di una ditta commerciale**.

- ✓ Ogni fattura è rappresentata da una (e una sola) scheda ognuna di questa deve contenere le informazioni correlate (attributi).
- ✓ Esempi di attributi possibili: data dell'ordine, data della consegna, importo, nome del cliente, indirizzo del cliente, telefono del cliente, ecc...
- ✓ **N.B.**: le informazioni contenute su una scheda possono anche essere ripetute su altre schede (uno stesso cliente può essere associato a differenti fatture, differenti fatture possono essere emesse nello stesso giorno), la fattura però deve essere univoca e rappresentare solo quella specifica transizione





DATABASE o **banca dati** (generica)

In informatica, il termine *database*, tradotto in italiano con **banca dati**, base di dati o anche base dati, indica un archivio di dati, riguardanti uno stesso argomento o più argomenti correlati tra loro, strutturato in modo tale da consentire la gestione dei dati stessi (l' inserimento, la ricerca, la cancellazione ed il loro aggiornamento) da parte di applicazioni software gestite da un elaboratore.

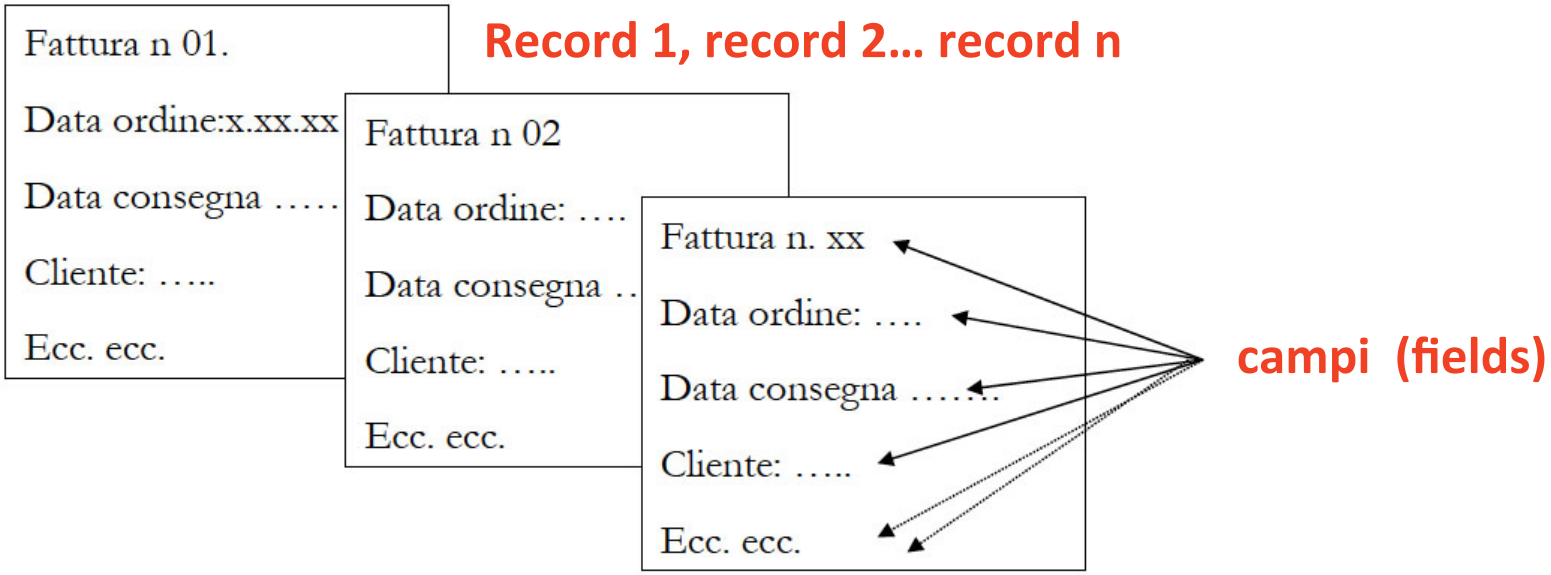
Altre definizioni:

- un *database* può essere definito come un insieme di informazioni strettamente correlate e memorizzate su un supporto di memoria di massa, costituenti un tutt'uno, che possono essere manipolate da più programmi applicativi;
- si definiscono strutture dati i modi di organizzare secondo regole precise una certa quantità, detta anche "base", di informazioni; tali regole possono definirsi sia in forma teorica, sia in forma pratica, riferendo quest'ultima all'effettivo ordinamento fisico dei dati sulla memoria di un elaboratore elettronico. La parte di memoria di un elaboratore elettronico destinato a contenere le informazioni così organizzate prende il nome di *database*.

DATABASE o **banca dati** (generica)

Nei *database* l'organizzazione è simile, con qualche differenza.

- Definizioni:
  - le schede prendono il nome di RECORDS;
  - gli attributi prendono il nome di CAMPI (FIELDS);
- Modalità di registrazione: gli schedari su supporto cartaceo, i database nelle memorie fisiche dei computer;
- Gestione dei dati: (manuale negli schedari) gestita da software negli elaboratori.



DATABASE o **banca dati** (generica)

Nei *database* valgono le stesse proprietà degli schedari.

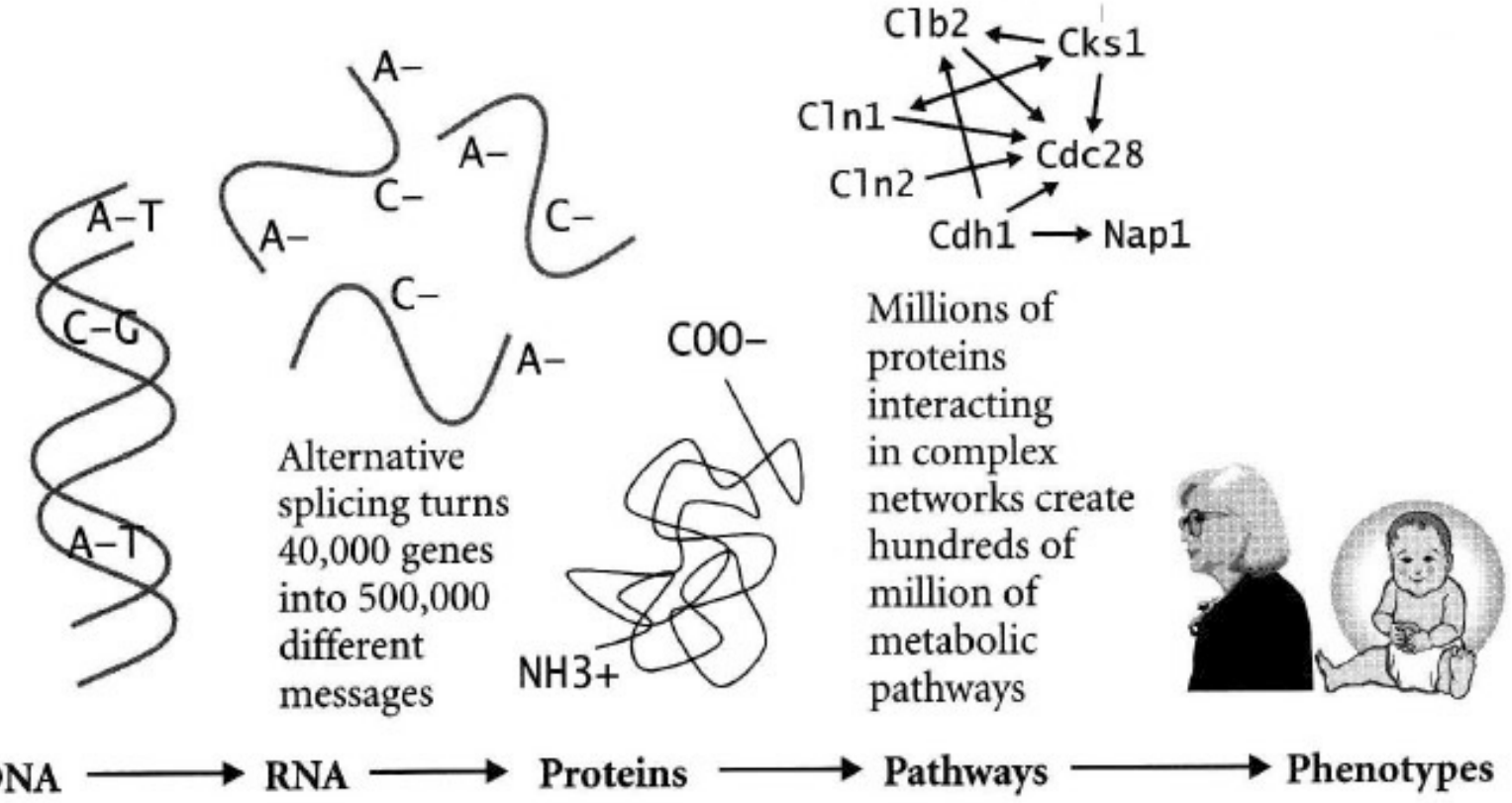
Se il database rappresenta un'entità del mondo reale, allora ogni record rappresenta un'istanza di quella entità e quindi non possono esistere più record per la stessa istanza.

Identificatore di record

È importante avere un contrassegno che identifica in modo univoco il record. Deve quindi esistere un campo speciale chiamato 'chiave' che deve essere diverso per ogni record.

Il campo chiave può essere rappresentato da un numero progressivo, oppure da una sigla, o anche da un nome, comunque sia, è essenziale che sia unico. Nei *database* biologici il campo chiave è chiamato "**ID**" oppure "**AC**" (**Accession number**).

Banche dati biologiche



DNA → **RNA** → **Proteins** → **Pathways** → **Phenotypes**

40,000 genes (approx 100 million bases) represent less that 3% of the genome. The remaining 97% is involved in creating structure and regulatinng gene expression.

Post-translational modifications turn 500,000 messages into millions of proteins

Hundreds of millions of metabolic pathways and environmental effects create billions of different individuals

Banche dati biologiche

Acquisizione dei dati

- Le **banche dati** sono dei contenitori costruiti per immagazzinare grandi quantità di dati biologici in modo efficiente e razionale;

- Le banche dati biologiche raccolgono informazioni e dati derivati da:
  - Letteratura;
  - Analisi di laboratorio (*in vitro* e *in vivo*);
  - Analisi bioinformatiche (*in silico*).

- Ogni banca dati è caratterizzata da un elemento biologico centrale che costituisce l'oggetto intorno al quale viene costruita la **ENTRY** principale della banca dati;

- La maggior parte delle banche dati sono fruibili in **formato Flat-file**: Ogni entry è memorizzata in un file di testo generalmente strutturato, contenente le informazioni;

```
ID   EKV49554; SV 1; linear; genomic DNA; CON; FUN; 1671 BP.
XX
PA   JH931607.1
XX
PR   Project:PRJNA61005;
XX
DT   02-DEC-2012 (Rel. 115, Created)
DT   02-DEC-2012 (Rel. 115, Last updated, Version 1)
XX
DE   Agaricus bisporus var. bisporus H97 tyrosinase
XX
KW   .
XX
OS   Agaricus bisporus var. bisporus H97
OC   Eukaryota; Fungi; Dikarya; Basidiomycota; Agaricomycotina; Agaricomycetes;
OC   Agaricomycetidae; Agaricales; Agaricaceae; Agaricus.
OX   NCBI_TaxID=936046;
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1671
FT                   /organism="Agaricus bisporus var. bisporus H97"
FT                   /chromosome="Unknown"
FT                   /strain="H97"
FT                   /variety="bisporus"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:936046"
FT   CDS             join(JH931607.1:2251601..2251709,
FT                   JH931607.1:2251769..2251953,JH931607.1:2252015..2252337,
FT                   JH931607.1:2252387..2252586,JH931607.1:2252640..2252792,
FT                   JH931607.1:2252851..2253407,JH931607.1:2253459..2253602)
FT                   /codon_start=1
FT                   /locus_tag="AGABI2DRAFT_191532"
FT                   /product="tyrosinase"
FT                   /note="GO_function: GO:16491 - oxidoreductase activity;
FT                   GO_process: GO:8152 - metabolic process"
FT                   /db_xref="GOA:K9HSW6"
FT                   /db_xref="InterPro:IPR002227"
FT                   /db_xref="InterPro:IPR008922"
FT                   /db_xref="InterPro:IPR016216"
FT                   /db_xref="UniProtKB/TrEMBL:K9HSW6"
FT                   /protein_id="EKV49554.1"
FT                   /translation="MSLIATVGPTGGVKNRLNIVDFVKNEKFFTLYVRSLELLQAKEQH
FT                   DYSSFFQLAGIHGLPFTEWAKERPSMNLYKAGYCTHGQVLFPTWHRTYLSVFEQILQGA
FT                   AIEVANKFTSNQTDWIQAAQDLRQPYWDWGFELMPPDEVIKNEEVNITNYDGKKISVKN
FT                   PILRYHFHPIDPSFKPYGDFATWRTTVRNPDRNRREDIPGLIKKMRLEEGQIREKTYNM
FT                   LKFNDAWERFSNHGISDDQHANSLESVHDDIHVMVGYGKIEGHMDHPFFAAFDPIFWLH
FT                   HTNVDRLLSLWKAINPDVWVTSGRNRDGTMGIAPNAQINDETPLEPFYQSEDKVWTSAS
FT                   LADTARLGYSYPDFDKLVGGTKELIRDAIDDLIDERYGSKPSSGARNTAFDLLADFKGI
FT                   TKEHKEDLKMYDWTIHVAFKKFELKESFSLLFYFASDGGDYDQENCFVGSINAFRGTTP
FT                   ETCANCQDNENLIQEGFIHLNHYLARDLESFEPQDVHKFLKEKGLSYKLYSREDKSLTS
FT                   LSVKIEGRPLHLPPGEHRPKYDHTQDRVVFDDVAVHVIN"
XX
SQ   Sequence 1671 BP; 446 A; 401 C; 390 G; 434 T; 0 other; 1471638968 CRC32;
     atgtcgctga ttgctactgt cggacctact ggcggagtta agaaccgttt gaacatcgtt        60
     gattttgtga agaatgaaaa gtttttcacg ctttatgtac gctccctcga acttctacaa       120
```

Formato flat file

- ID
- AC
- DT
- DE
- GN
- OC
- R(X) referenze bibliografiche
- DR riferimenti incrociati
- KW key words
- FT features
- SQ intestazione sequenza

```
     gccaaggaac agcatgacta ctcgtctttc ttccaactcg ccgggattca tggtctaccc       180
     tttactgagt gggccaaaga gcggccttcc atgaacctat acaaggctgg ttattgtacc       240
     catgggcagg ttctgttccc gacttggcat agaacgtacc tttctgtgtt cgagcaaata       300
     cttcaaggag ctgccatcga agttgctaac aagttcactt ctaatcaaac cgattggatc       360
     caggcggcgc aggatctacg ccagccctac tgggattggg gtttcgaact tatgcctcct       420
     gatgaggtta tcaagaacga agaggtcaac attacgaact acgatggaaa gaagatttcc       480
     gtcaagaacc ctatcctccg ctatcacttc catccgatcg atccttcttt caagccatac       540
     ggagactttg caacctggcg aacaacagtc cgaaaccccg atcgtaatag gcgagaggat       600
     atccccggtc taatcaaaaa aatgagactc gaggaaggtc agattcgtga gaagacctac       660
     aatatgttga agttcaacga tgcttgggag agatttagta accacggcat atctgatgat       720
     cagcatgcta acagcttgga gtctgttcac gatgacattc atgttatggt tggatacggc       780
     aaaatcgaag gacatatgga ccacccttc tttgctgcct tcgacccgat tttctggtta       840
     catcatacca acgtcgaccg tctactatcc ctttggaaag caatcaatcc agatgtgtgg       900
     gttacatcgg gacgtaacag ggatggtacc atgggcatcg caccaaacgc tcagatcaac       960
     gacgagactc ctcttgagcc attctatcaa tctgaggata aagtgtggac ctcggcctct      1020
     ctcgctgata ctgctcggct cggctactcc taccccgatt tcgacaagtt ggttggagga      1080
     acaaaggagt tgattcgcga cgctatcgac gacctcatcg atgagcggta tggaagcaaa      1140
     ccttcgagtg gggctcgcaa tactgccttt gatctcctcg ccgatttcaa gggcattacc      1200
     aaggagcaca aggaggatct caaaatgtac gactggacca tccatgttgc cttcaagaag      1260
     ttcgagttga aagagagttt cagtcttctc ttctactttg cgagtgatgg tggcgattat      1320
     gatcaggaga attgctttgt tggatcaatt aacgccttcc gtgggactac tcccgaaact      1380
     tgcgcgaact gtcaagataa cgagaacttg attcaagaag gctttattca cttgaatcat      1440
     tatcttgctc gtgaccttga atctttcgag ccgcaggacg tgcacaagtt cttaaaggaa      1500
     aaaggactgt catacaaact ctacagcagg gaagataaat ctttgacatc gttgtcagtc      1560
     aagattgaag gacgtcccct tcatttgcca cccggagaac atcgtccgaa gtacgatcac      1620
     actcaggacc gagtagtgtt tgatgatgtc gcggtgcatg ttatcaactg a               1671
//
```

// linea di terminazione

Informazioni associate

Ontologia: una formale descrizione delle entità e delle relazioni intercorrenti fra esse

Banche dati biologiche

Acquisizione dei dati

- Banche dati in **formato XML** (eXtensible Markup Language); non è propriamente un linguaggio di programmazione, ma una **sintassi di descrizione dei dati** ➔ è sempre un file di testo ma ne rende possibile l'utilizzo indipendentemente dall'aplicazione adottata, senza preoccuparsi della formattazione

Formato XML

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT request="EKV49554&amp;display=xml">
<entry accession="EKV49554" version="1" entryVersion="1" dataClass="CON" taxonomicDivision="FUN" moleculeType="genomic DNA" sequenceLength="1671" topology="linear"
firstPublicRelease="115" lastUpdated="2012-12-02" lastUpdatedRelease="115">
        <description>Agaricus bisporus var. bisporus H97 tyrosinase</description>
        <xref db="EMBL" id="JH931607.1"/>
        <feature name="source" location="1..1671">
                <taxon scientificName="Agaricus bisporus var. bisporus H97" taxId="936046">
                        <lineage>
                                <taxon scientificName="Eukaryota"/>
                                <taxon scientificName="Fungi"/>
                                <taxon scientificName="Dikarya"/>
                                <taxon scientificName="Basidiomycota"/>
                                <taxon scientificName="Agaricomycotina"/>
                                <taxon scientificName="Agaricomycetes"/>
                                <taxon scientificName="Agaricomycetidae"/>
                                <taxon scientificName="Agaricales"/>
                                <taxon scientificName="Agaricaceae"/>
                                <taxon scientificName="Agaricus"/>
                        </lineage>
                </taxon>
                <qualifier name="organism">
                        <value>
Agaricus bisporus var. bisporus H97
                        </value>
                </qualifier>
                <qualifier name="chromosome">
                        <value>
Unknown
                        </value>
                </qualifier>
                <qualifier name="strain">
                        <value>
H97
                        </value>
                </qualifier>
                <qualifier name="variety">
                        <value>
bisporus
                        </value>
                </qualifier>
        </feature>
        <feature name="CDS"
location="join(JH931607.1:2251601..2251709,JH931607.1:2251769..2251953,JH931607.1:2252015..2252337,JH931607.1:2252387..2252586,JH931607.1:2252640..2252792,JH931607.
53602)">
                <xref db="GOA" id="K9HSW6"/>
                <xref db="InterPro" id="IPR002227"/>
                <xref db="InterPro" id="IPR008922"/>
                <xref db="InterPro" id="IPR016216"/>
                <xref db="UniProtKB/TrEMBL" id="K9HSW6"/>
                <qualifier name="codon_start">
                        <value>
1
```

Caratterizzato dalla presenza di *tag* , simboli racchiusi da parentesi angolari < e > , organizzati gerarchicamente.

Il dato risulta “strutturato”.

Moltissime applicazioni informatiche sono in grado di interpretarlo ➔ formato standard per la trasmissione di dati.

```
                        </value>
                </qualifier>
                <qualifier name="locus_tag">
                        <value>
AGABI2DRAFT_191532
                        </value>
                </qualifier>
                <qualifier name="product">
                        <value>
tyrosinase
                        </value>
                </qualifier>
                <qualifier name="note">
                        <value>
GO_function: GO:16491 - oxidoreductase activity; GO_process: GO:8152 - metabolic
process
                        </value>
                </qualifier>
                <qualifier name="protein_id">
                        <value>
EKV49554.1
                        </value>
                </qualifier>
                <qualifier name="translation">
                        <value>
MSLIATVGPTGGVKNRLNIVDFVKNEKFFTLYVRSLELLQAKEQHDYSSFFQLAGIHGLPFTEWAKERPSMNLYKAGYCT
HGQVLFPTWHRTYLSVFEQILQGAAIEVANKFTSNQTDWIQAAQDLRQPYWDWGFELMPPDEVIKNEEVNITNYDGKKIS
VKNPILRYHFHPIDPSFKPYGDFATWRTTVRNPDRNRREDIPGLIKKMRLEEGQIREKTYNMLKFNDAWERFSNHGISDD
QHANSLESVHDDIHVMVGYGKIEGHMDHPFFAAFDPIFWLHHTNVDRLLSLWKAINPDVWVTSGRNRDGTMGIAPNAQIN
DETPLEPFYQSEDKVWTSASLADTARLGYSYPDFDKLVGGTKELIRDAIDDLIDERYGSKPSSGARNTAFDLLADFKGIT
KEHKEDLKMYDWTIHVAFKKFELKESFSLLFYFASDGGDYDQENCFVGSINAFRGTTPETCANCQDNENLIQEGFIHLNH
YLARDLESFEPQDVHKFLKEKGLSYKLYSREDKSLTSLSVKIEGRPLHLPPGEHRPKYDHTQDRVVFDDVAVHVIN
                        </value>
                </qualifier>
        </feature>
        <sequence>
atgtcgctgattgctactgtcggacctactggcggagttaagaaccgtttgaacatcgtt
gattttgtgaagaatgaaaagtttttcacgctttatgtacgctccctcgaacttctacaa
gccaaggaacagcatgactactcgtctttcttccaactcgccgggattcatggtctaccc
tttactgagtgggccaaagagcggccttccatgaacctatacaaggctggttattgtacc
catgggcaggttctgttcccgacttggcatagaacgtacctttctgtgttcgagcaaata
cttcaaggagctgccatcgaagttgctaacaagttcacttctaatcaaaccgattggatc
caggcggcgcaggatctacgccagccctactgggattggggtttcgaacttatgcctcct
gatgaggttatcaagaacgaagaggtcaacattacgaactacgatggaaagaagatttcc
gtcaagaaccctatcctccgctatcacttccatccgatcgatccttctttcaagccatac
ggagactttgcaacctggcgaacaacagtccgaaaccccgatcgtaataggcgagaggat
atccccggtctaatcaaaaaaatgagactcgaggaaggtcagattcgtgagaagacctac
aatatgttgaagttcaacgatgcttgggagagatttagtaaccacggcatatctgatgat
cagcatgctaacagcttggagtctgttcacgatgacattcatgttatggttggatacggc
aaaatcgaaggacatatggaccacccttttctttgctgccttcgacccgattttctggtta
catcataccaacgtcgaccgtctactatccctttggaaagcaatcaatccagatgtgtgg
gttacatcgggacgtaacagggatggtaccatgggcatcgcaccaaacgctcagatcaac
gacgagactcctcttgagccattctatcaatctgaggataaagtgtggacctcggcctct
```

Banche dati biologiche

Acquisizione dei dati

- Banche dati in **formato TABELLA**

campo A, campo B, … campo N

record 1,

record 2,

… …

record n

```
>Feature Seq1
<1     >1050    gene
                            gene            ATH1
<1     1009     CDS
                            product         acid trehalase
                            product         Athlp
                            codon_start     2
<1     >1050    mRNA
                            product         acid trehalase

>Feature Seq2
2626   2590     tRNA
2570   2535
                            product         tRNA-Phe

>Feature Seq3
1080   1210   CDS
1275   1315
                            product         actin
                            note            alternatively spliced
1055   1210   mRNA
1275   1340
                            product         actin
1055   1340   gene
                            gene            ACT
1055   1079   5'UTR
1316   1340   3'UTR
```

- Questo tipo di formato è utilizzato soprattutto per costruire ➔ database relazionali

Banche dati biologiche

Tipologie

- **Database ordinati**: i records sono disposti fisicamente (o virtualmente) in ordine secondo il contenuto di uno o più particolari campi (flat files e indicizzazione).
- Un archivio ordinato permette l'applicazione di particolari algoritmi che velocizzano la ricerca (per es. applicando un algoritmo di ricerca dicotomica -detta anche ricerca binaria-)
- **Database relazionali**: sono insieme di DB (generalmente in formato tabella) correlati tra loro attraverso relazioni che fanno riferimento ai campi 'chiave.'

DB clienti

ID cliente	Nome	Indirizzo	Ecc.ecc.
01	Rossi Guido	Padova, via ….	..
02	Barca Mario	Milano, ….	..
03	Milan Lino	….	..
ecc	..	..	..

DB prodotti

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

Relazione

Relazione

DB ordini

ID ordine	ID cliente	Data ordine	ID Prodotto	Ecc ecc
01	03	24/12/2005	02	---
02	---	---	---	---
---	---	---	---	---

Banche dati biologiche

Gestione e utilizzo dei dati

- Con il crescere dei dati si è reso necessario adottare **DBMS** (sistemi di gestione delle basi di dati, **Database Management Systems**);

- La consultazione e il recupero dei dati da parte dell'utente raramente è diretta ma si avvale di **linguaggi di interrogazione delle basi di dati**, il più diffuso dei quali è **SQL** (**Structured Query Language**);

SELECT	nome del gene / della proteina
FROM	genoma / proteoma del tal organismo
WHERE	condizione imposta alla ricerca
ORDER BY	nome / dimensione / ecc.

- **Parole chiave** da introdurre nei campi di ricerca e utilizzo degli **operatori booleani** (**AND, OR** e **NOT**);

- Uso del web per accedere a informazioni tra loro correlate (**cross-referencing**, **riferimenti incrociati**) attraverso link ipertestuali.

- Uso del web per accedere a informazioni tra loro correlate (meccanismi di **cross-referencing**, **riferimenti incrociati**) attraverso link ipertestuali ➔ fondamentali: permettono consentono di navigare fra i database anche se dislocati su siti tra di loro remoti

Names and origin

Protein names	*Submitted name:* Tyrosinase EMBL EKV49554.1
Gene names	ORF Names:AGABI2DRAFT_191532 EMBL EKV49554.1
Organism	**Agaricus bisporus var. bisporus (strain H97 / ATCC MYA-4626 / FGSC 10389) (White button mushroom)** [Complete proteome]
Taxonomic identifier	936046 [NCBI]
Taxonomic lineage	Eukaryota › Fungi › Dikarya › Basidiomycota › Agaricomycotina › Agaricomycetes › Agaricomycetidae › Agaricales › Agaricaceae › A

Protein attributes

Sequence length	556 AA.
Sequence status	Complete.
Protein existence	Predicted

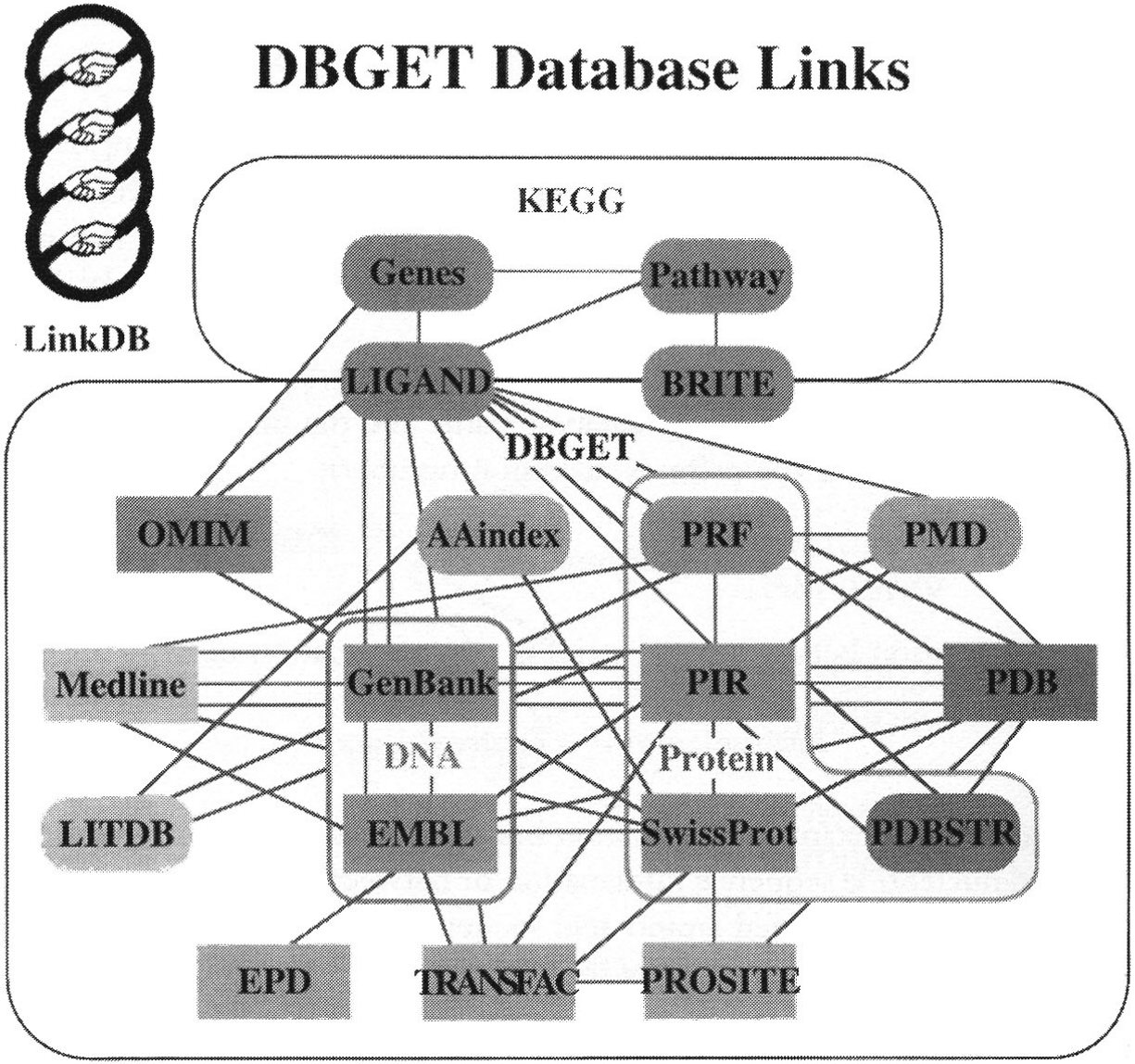
Ontologies

Keywords

Ligand	Metal-binding SAAS SAAS002227
Technical term	Complete proteome

Gene Ontology (GO)

Molecular_function	metal ion binding Inferred from electronic annotation. Source: UniProtKB-KW monophenol monooxygenase activity Inferred from electronic annotation. Source: InterPro



A link-based integration of molecular biology databases in the DBGET/LinkDB system at GenomeNet (http://www.genome.ad.jp/). The lines indicate thet the cross-references are given by the original databases.

Banche dati biologiche

Gestione dei dati

- Ridondanze e Errori:
  - Errori durante l'estrazione delle sequenze;
  - Algoritmi per la previsione di strutture imperfetti;
  - Inserimento erroneo di duplicati nei DB;
  - Diversi nomi per la stessa sequenza;
  - Non vi è un'unica struttura per un gene (splicing alternativi). Lo stesso gene può essere rappresentato da numerose sequenze nei vari DB;

- Nonostante diversi accorgimenti (strumenti per l'inserimento accessibili dalla rete, controllo automatizzato di correttezza e validità del dato inserito, controllo manuale), i centri di raccolta eseguono periodicamente controlli e aggiornamenti; invitano gli utenti a segnalare errori o discrepanze

- NCBI accetta tutte le sequenze ma le eleva al rango di **REFSEQ** (sequenza di riferimento) e assegna un refseqID solo dopo numerosi controlli anche manuali.

Banche dati primarie

- Acidi nucleici (DNA, RNA):
  - ✓ **EMBL datalibrary** (EMBL – European Molecular Biology Laboratory - 1980);
  - ✓ **GenBank** (NCBI – National Center for Biotechnology Information - 1982);
  - ✓ **DDBJ** (DNA Database of Japan - 1986).
- EMBL adotta un formato diverso dalle altre banche di acidi nucleici;

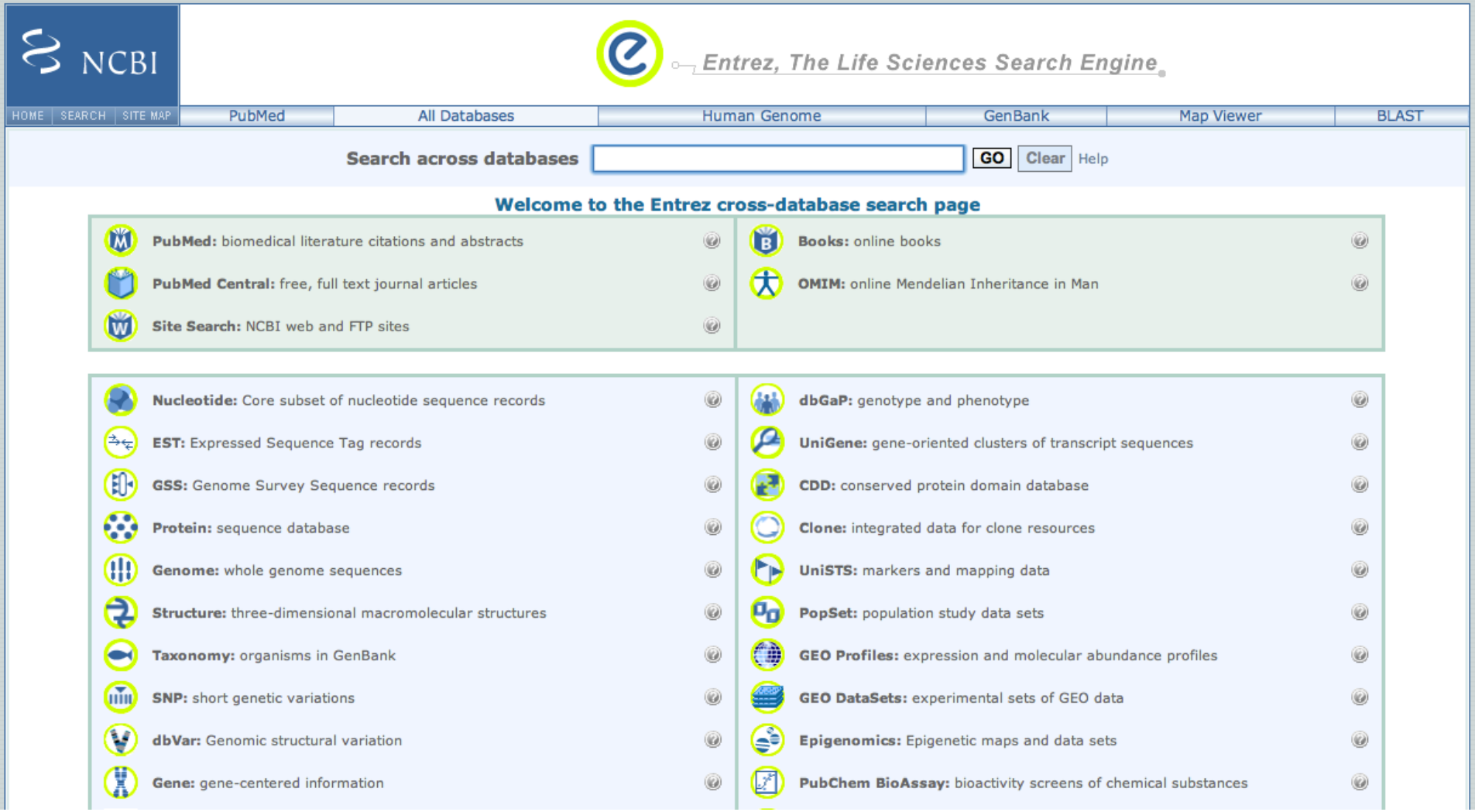
- Proteine:
  - ✓ **UniProt** (2002) dall'unione di PIR, Swiss-Prot e TrEMBL.

- Esiste un accordo tra le tre banche per cui l'inserimento di dati in una, comporta l'automatico inserimento nelle altre.

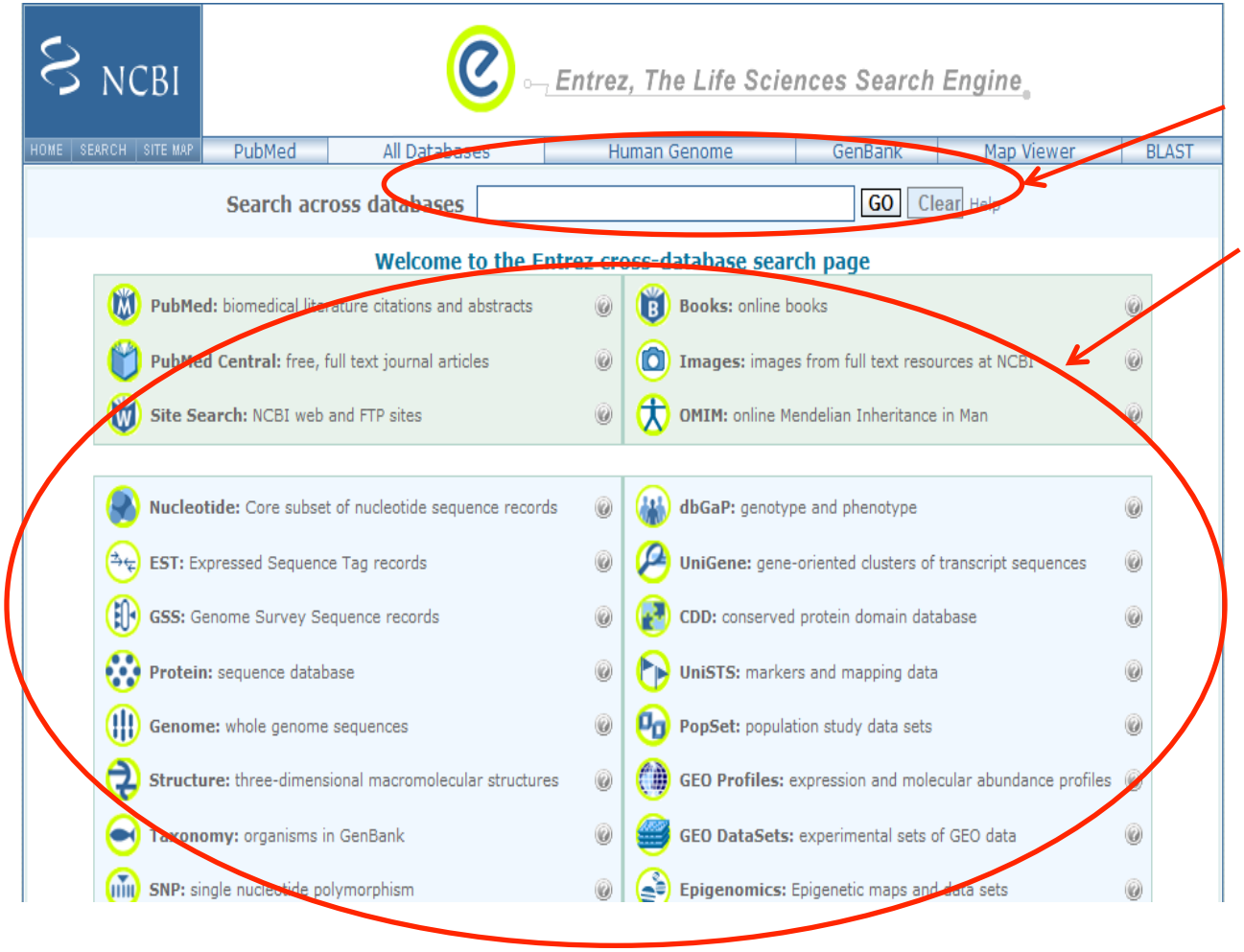
Interrogazione di banche dati

- I sistemi più utilizzati per interrogare le banche dati sono:
  - **Entrez** (Sviluppato da NCBI): Permette di accedere a numerose banche dati (anche contemporaneamente) attraverso una interfaccia web.
    - Permette di effettuare ricerche testuali sui DB utilizzando diverse sintassi per i vari DB.
  - **SRS** - Sequence Retrieval System (Sviluppato da EBI – European Bioinformatics Institute);
  - Anche DDBJ offre un metodo di ricerca e analisi dei dati via WEB (ma in pratica si tratta delle stesse cose che vedremo per Entrez e SRS);

Entrez - http://www.ncbi.nlm.nih.gov/Entrez



Entrez - http://www.ncbi.nlm.nih.gov/Entrez

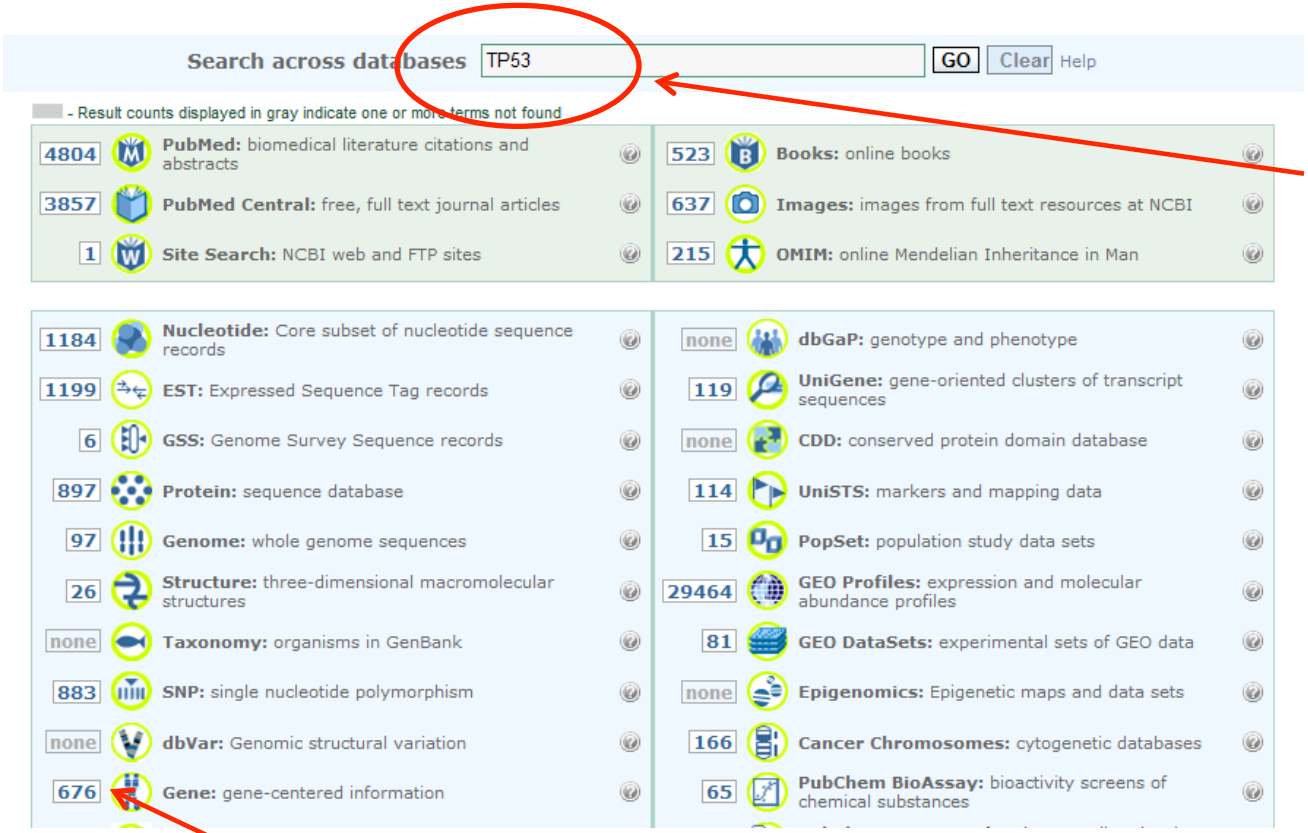


Ricerca in tutti i database

Risorse principali:

- Nucleotide;
- Protein;
- Genome;
- Gene;
- Taxonomy;
- Pubmed;

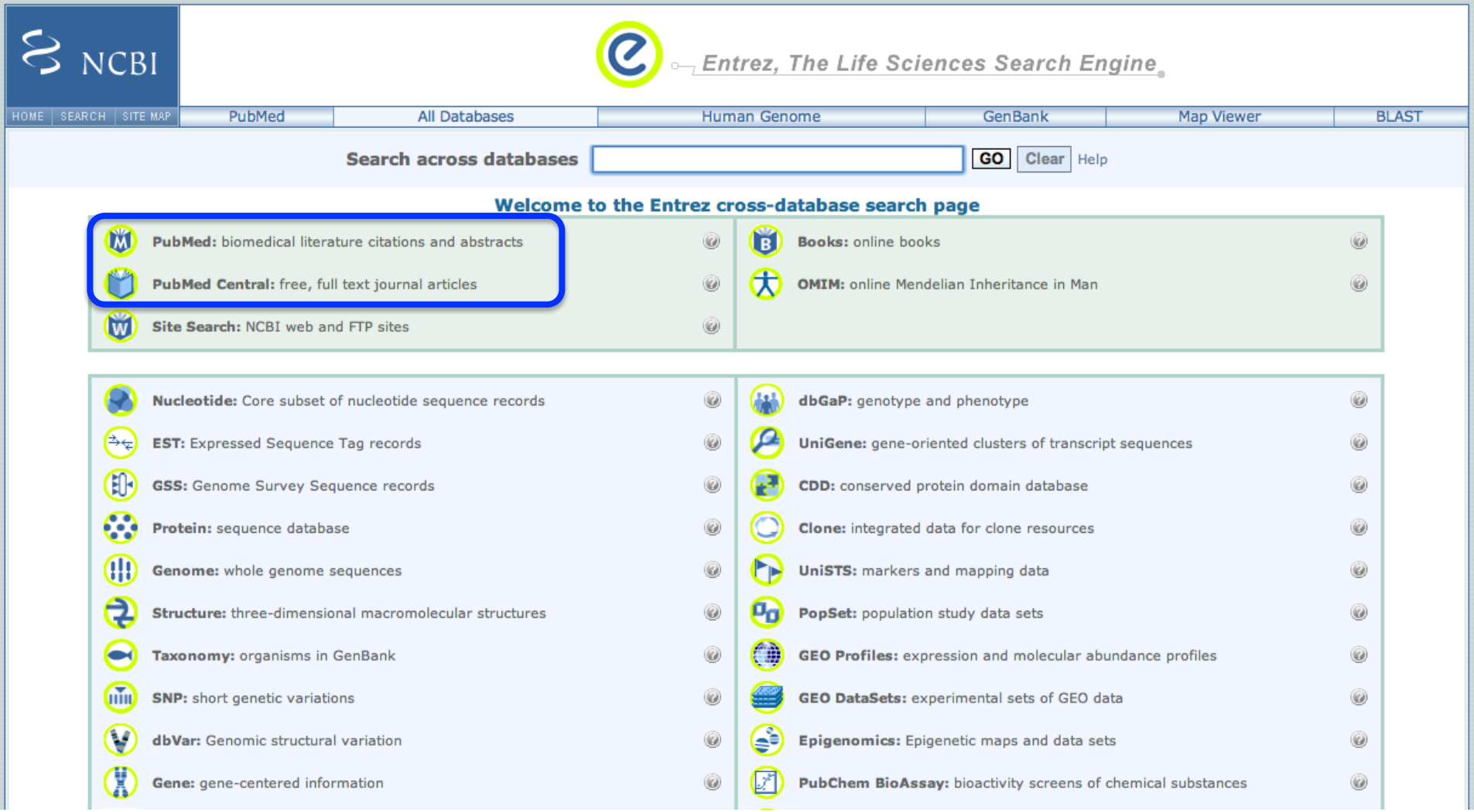
Entrez - http://www.ncbi.nlm.nih.gov/Entrez



Cerchiamo informazioni relativamente al gene umano TP53

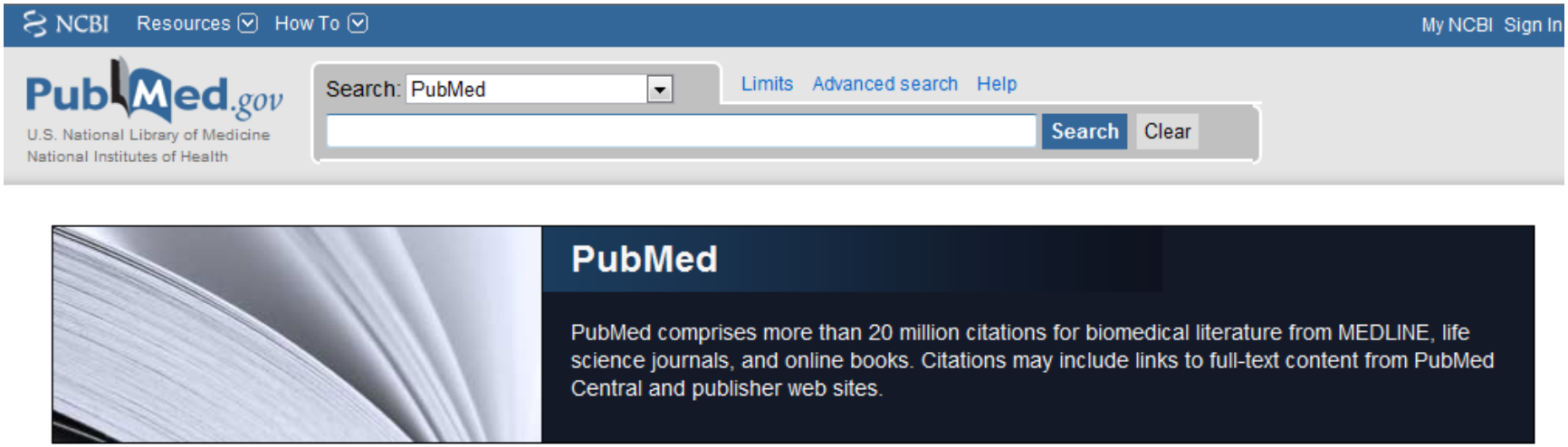
Clicchiamo in corrispondenza di Gene

- … Ricerca bibliografica nella letteratura scientifica



Entrez - Pubmed

http://www.ncbi.nlm.nih.gov/pubmed



Entrez - Pubmed

Entrez, The Life Sciences Search Engine

PubMed contiene al suo interno 4 database:

- **MEDLINE**

  citazioni dal 1966 ad oggi; abstract; MESH; aggiornamento settimanale;

- **OLDMEDLINE**

  con citazioni dal 1951 al 1965 , no abstract, no MESH

- **PREMEDLINE** (In Process citations)

  per citazioni non ancora indicizzate; no MeSH ; aggiornamento giornaliero

- **PUBLISHER SUPPLIED CITATIONS**

  per citazioni ricevute per via elettronica direttamente dall'editore. Non ancora pubblicate in cartaceo.

Entrez - Pubmed

Entrez, The Life Sciences Search Engine

Anche PubMed ha il suo formato Flat file:

[AU] campo autore

[TI] campo titolo

[TA] nome della rivista

[LA] lingua di pubblicazione dell'articolo

[MH] Mesh terms (soggetti)

[DP] data di pubblicazione(A/M/G)

[EDAT] data di inserimento nel pubmed (A/M/G)

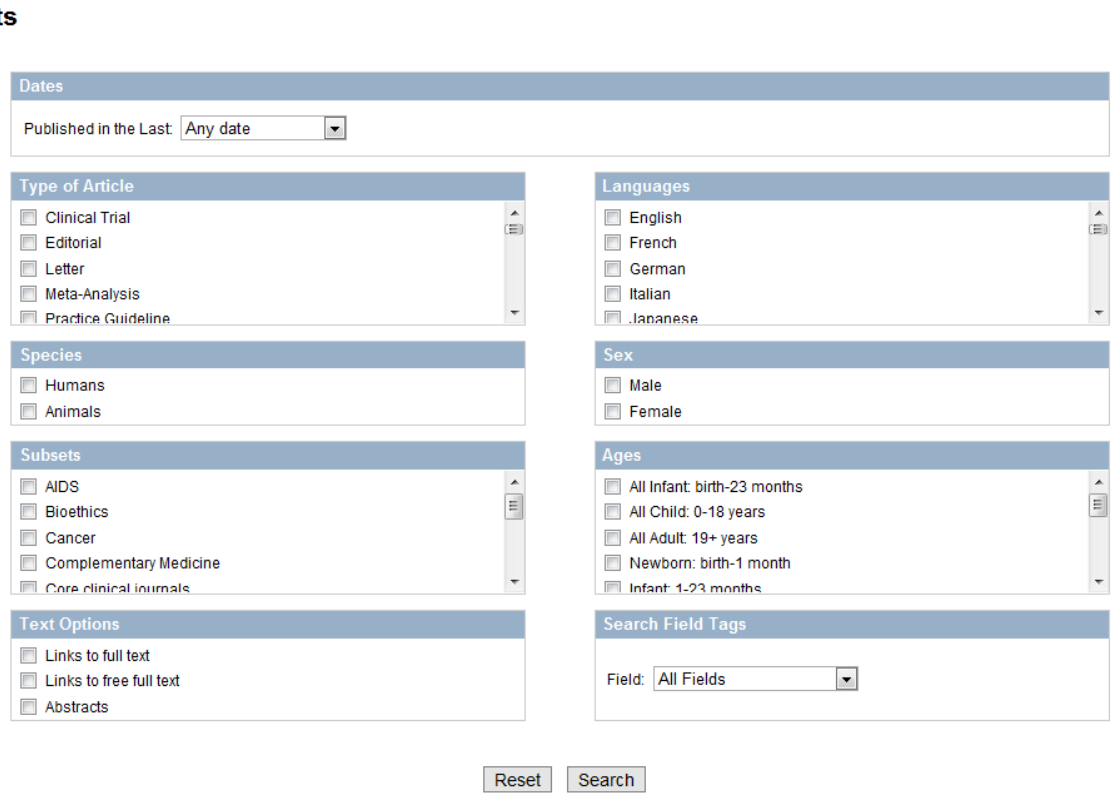
[AB] abstract

```
PMID- 21390126
OWN - NLM
STAT- In-Data-Review
DA  - 20110310
IS  - 1476-4687 (Electronic)
IS  - 0028-0836 (Linking)
VI  - 471
IP  - 7337
DP  - 2011 Mar 10
TI  - Inactivating mutations of acetyltra
PG  - 189-95
AB  - B-cell non-Hodgkin's lymphoma compr
      diseases the pathogenesis of which
      oncogenes and tumour-suppressor gen
      types-follicular lymphoma and diffu
      structural alterations inactivating
      related histone and non-histone ace
      transcriptional co-activators in mu
      of diffuse large B-cell lymphoma an
      genomic deletions and/or somatic mu
      coding domain of these two genes. T
      suggesting that reduction in HAT do
      demonstrate specific defects in ace
      oncoprotein and activation of the p
      CREBBP/EP300 mutations as a major p
      of B-cell non-Hodgkin's lymphoma, w
      targeting acetylation/deacetylation
AD  - 1] Institute for Cancer Genetics an
      Center, Columbia University, New Yo
      Pathology and Cell Biology, Columbi
FAU - Pasqualucci, Laura
AU  - Pasqualucci L
FAU - Dominguez-Sola, David
AU  - Dominguez-Sola D
FAU - Chiarenza, Annalisa
AU  - Chiarenza A
FAU - Fabbri, Giulia
AU  - Fabbri G
```

Entrez - Pubmed

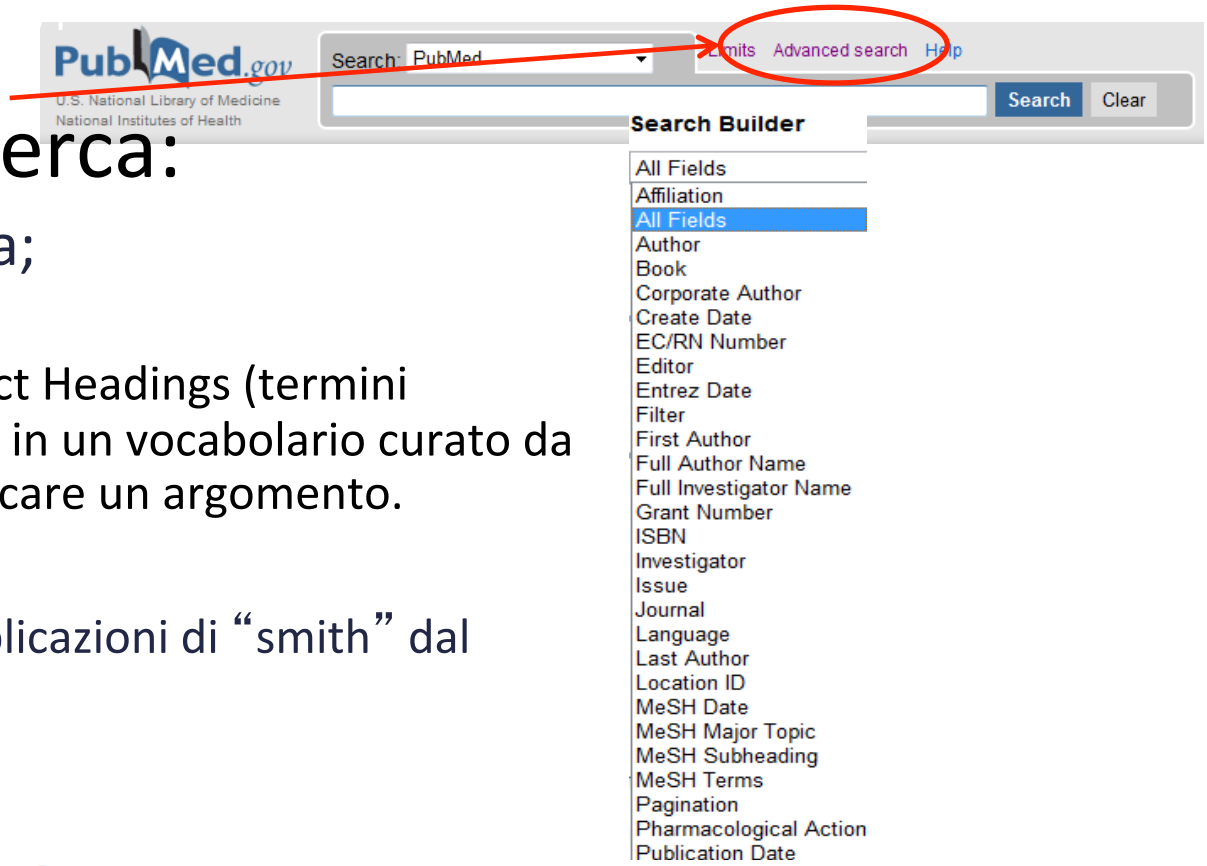
- Metodi di ricerca:
  - Usare i Limits;

Cliccare su limits nella pagina principale di entrez Pubmed



  - Data di pubblicazione;
  - Nome dell'autore;
  - Tipo di articolo;
  - Linguaggio;
  - Specie;
  - Sesso;

Entrez - Pubmed

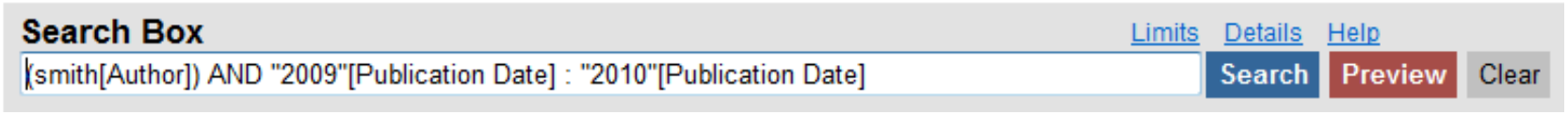


- Metodi di ricerca:
  - Ricerca avanzata;

[mesh] Medical Subject Headings (termini biomedici indicizzati in un vocabolario curato da NCBI). Usati per indicare un argomento.

Esempio: tutte le pubblicazioni di “smith” dal 2009 al 2010

PubMed Advanced Search



Entrez - Pubmed

Entrez, The Life Sciences Search Engine

Usare il tag **MeSH - Medical Subject Headings.**
Dalla Pagina della ricerca avanzata è possibile accedere al vocabolario di termini medici utili alla ricerca.

MeSH è un dizionario dei sinonimi e contrari (thesaurus) controllato dalla NML (National Library of Medicine's).