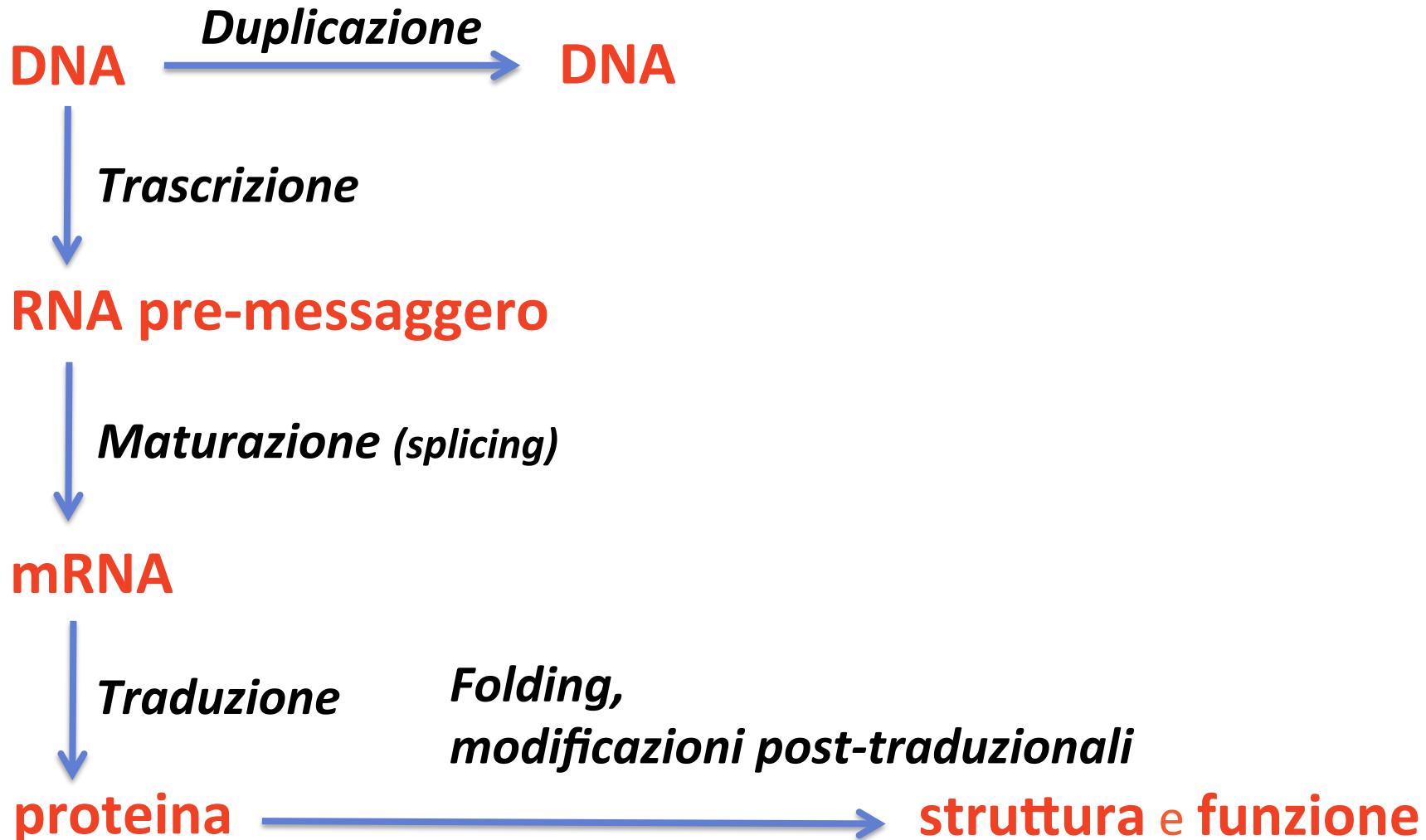


Flusso dell'informazione genetica

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'-.....TACCTTACGTCTCCTAACGAACT.....-5'

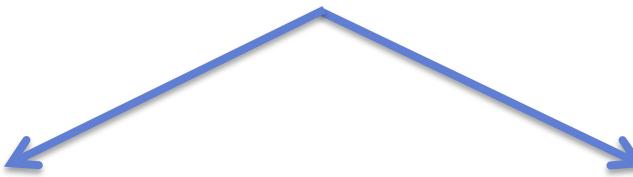
Flusso dell'informazione genetica

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

3' -..... TACCTTTACGTCTCCTAACGAACT.....-5'

Trascrizione

Duplicazione



5' -.. ATGGAAATGCAGAGGATTGCTTGA..-3'

3' -.. TACCTTTACGTCTCCTAACGAACT..-5'

Flusso dell'informazione genetica

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

3' -..... TACCTTACGTCTCCTAACGAACT.....-5'

Trascrizione

Duplicazione

5' -.. AUGGAAAUGCAGAGGAUUGCUGA..-3'

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

Traduzione

NH₂-.. MetGluMetGlnArgIleAla..-COOH

	T	C	A	G	
T	TTT <i>Phe (F)</i> TTC " TTA <i>Leu (L)</i> TTG "	TCT <i>Ser (S)</i> TCC " TCA " TCG "	TAT <i>Tyr (Y)</i> TAC " TAA stop TAG stop	TGT <i>Cys (C)</i> TGC " TGA stop TGG <i>Trp (W)</i>	T C A G
C	CTT <i>Leu (L)</i> CTC " CTA " CTG "	CCT <i>Pro (P)</i> CCC " CCA " CCG "	CAT <i>His (H)</i> CAC " CAA <i>Gln (Q)</i> CAG "	CGT <i>Arg (R)</i> CGC " CGA " CGG "	T C A G
A	ATT <i>Ile (I)</i> ATC " ATA " ATG <i>Met (M)</i>	ACT <i>Thr (T)</i> ACC " ACA " ACG "	AAT <i>Asn (N)</i> AAC " AAA <i>Lys (K)</i> AAG "	AGT <i>Ser (S)</i> AGC " AGA <i>Arg (R)</i> AGG "	T C A G
G	GTT <i>Val (V)</i> GTC " GTA " GTG "	GTC <i>Ala (A)</i> GCC " GCA " GCG "	GAT <i>Asp (D)</i> GAC " GAA <i>Glu (E)</i> GAG "	GGT <i>Gly (G)</i> GGC " GGA " GGG "	T C A 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' -..... TACCTTACGTCTCCTAACGAACT.....-5'

Trascrizione

Duplicazione

5' -.. AUGGAAAUGCAGAGGAUUGCUGA..-3'

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

Traduzione

NH₂-.. **Met**Glu**Met**Gln**Arg**Ile**Ala**..-COOH

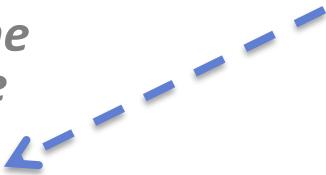
NH₂-.. M E M **Q** R I A ..-COOH

Flusso dell'informazione genetica

5' -.....ATGGAAATGCAGAGGATTGCT**TGA**.....-3'

3' -.....TACCTTACGTCTCCTAACGAACT.....-5'

*Trascrizione
Traduzione*



NH₂-... **Met** Glu Met Gln Arg Ile Ala ... -COOH

NH₂-... M E M **Q** R I A ... -COOH

Folding, modificazioni post-traduzionali



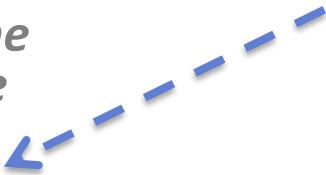
NH₂-... M E M **O** R I A ... -COOH

Flusso dell'informazione genetica

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

3' -.....TACCTTACGTCTCCTAACGAACT.....-5'

*Trascrizione
Traduzione*



NH₂-... Met Glu Met Gln Arg Ile Ala ... -COOH

NH₂-... M E M Q R I A ... -COOH

Folding, modificazioni post-traduzionali



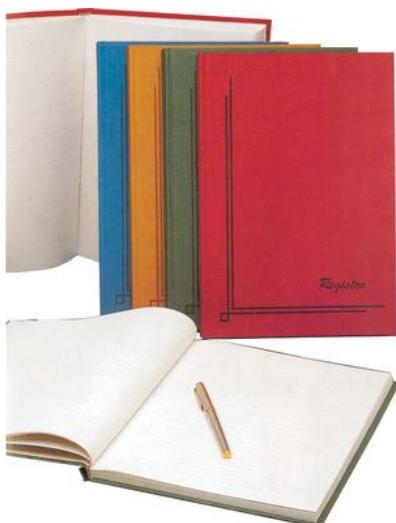
NH₂-... M E M O R I A ... -COOH

struttura e funzione
fenotipo

DATABASE o banca dati (generica)

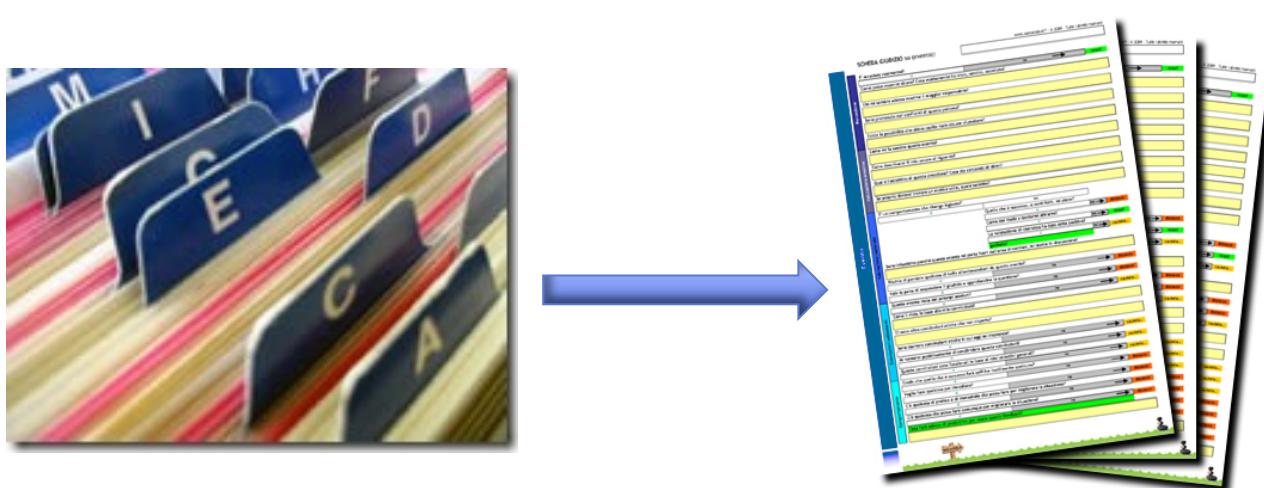
- “Individuare tutte le pubblicazioni rilevanti;
- analizzarne criticamente i dati e risolvere i risultati conflittuali;
- presentare i dati in un formato che possa riflettere quegli aspetti della struttura che sono stati determinati sperimentalmente e quelli che possono ragionevolmente essere ricavati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

M.Dayhoff, *Atlas of protein sequence and structure*



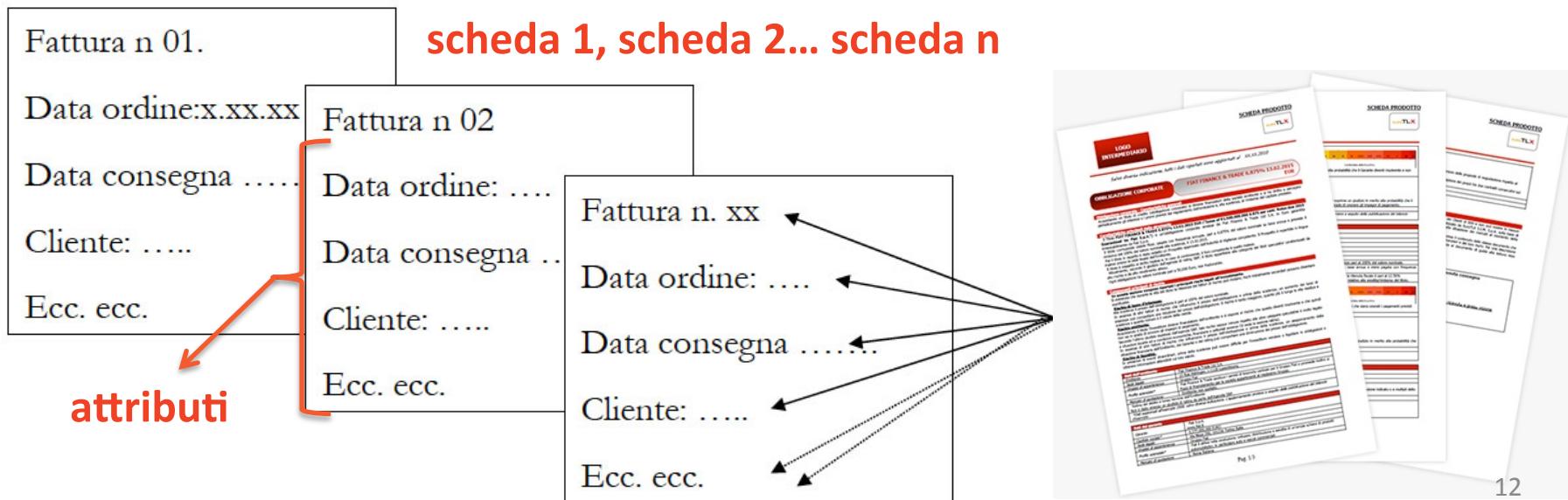
DATABASE o banca dati

- 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 queste 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.

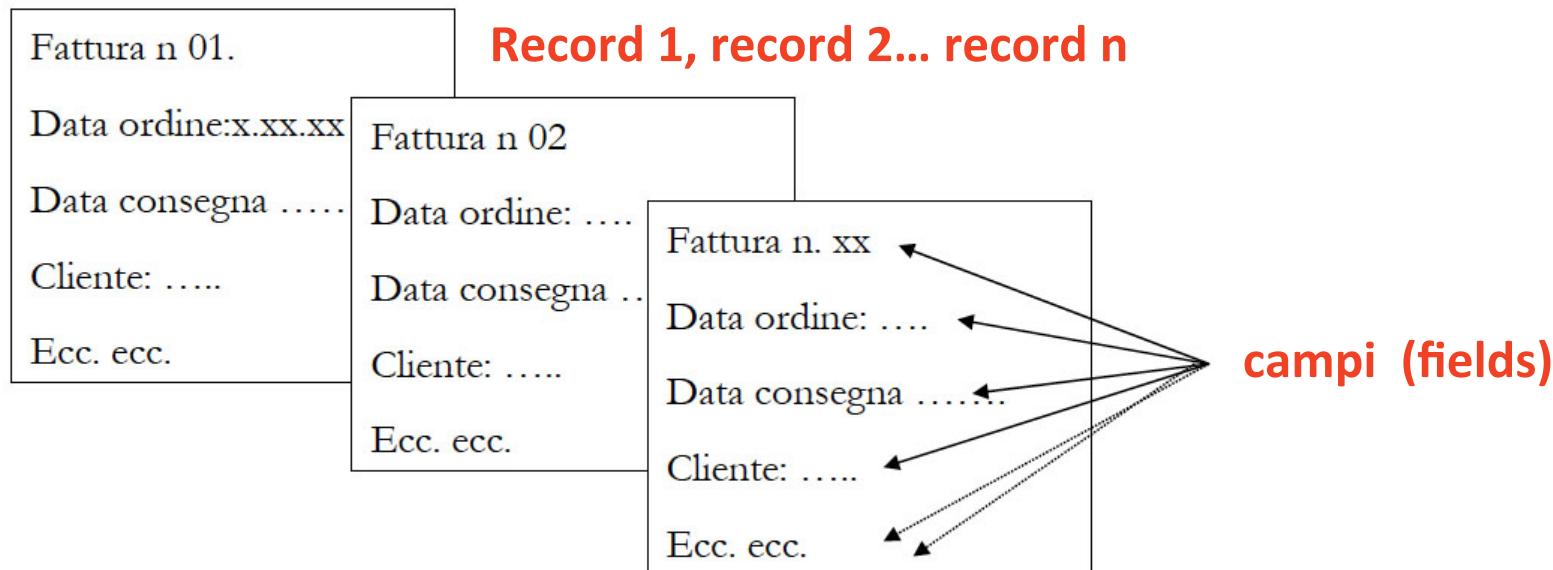
Altre definizioni:

- 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.

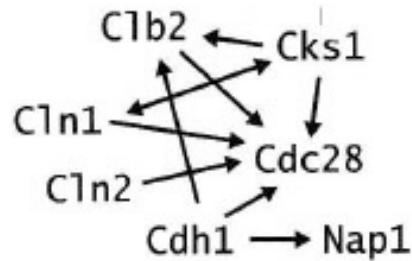
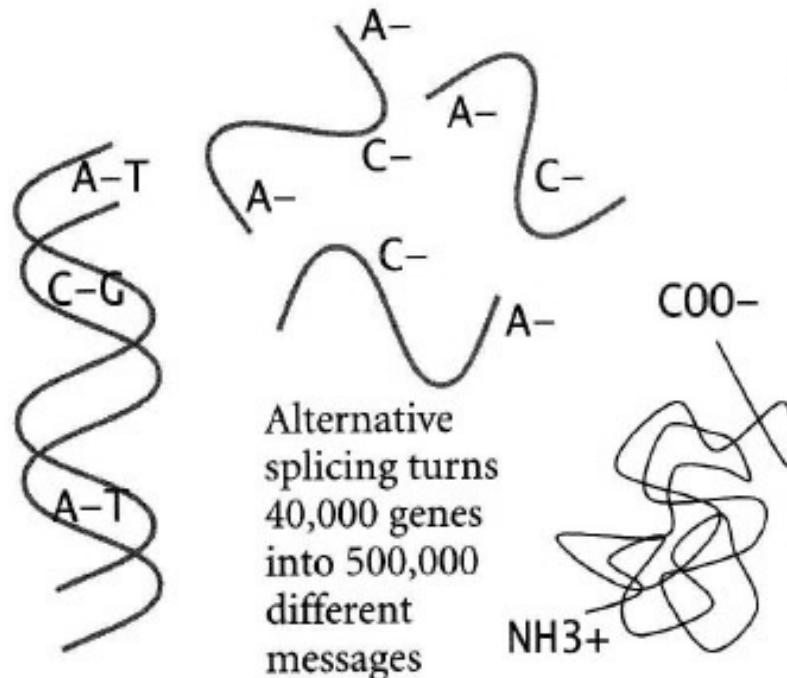
Se il database rappresenta un'entità del mondo reale, allora ogni record rappresenta un'istanza di quella entità e quindi non possono esistere più record per la stessa istanza.

Identificatore di record

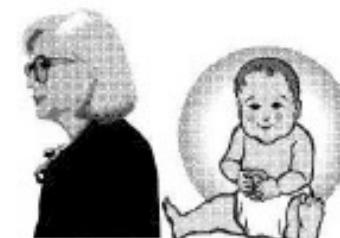
È importante avere un contrassegno che identifica in modo univoco il record. Deve quindi esistere un campo speciale chiamato ‘chiave’ che deve essere diverso per ogni record.

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



Millions of proteins interacting in complex networks create hundreds of millions of metabolic pathways



DNA → RNA → Proteins → Pathways → Phenotypes

40,000 genes (approx 100 million bases) represent less than 3% of the genome. The remaining 97% is involved in creating structure and regulating 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
/organism="Agaricus bisporus var. bisporus H97"
/chromosome="Unknown"
/strain="H97"
/variety="bisporus"
/mol_type="genomic DNA"
/db_xref="taxon:936046"
FT CDS join(JH931607.1:2251601..2251709,
JH931607.1:2251769..2251953,JH931607.1:2252015..2252337,
JH931607.1:2252387..2252586,JH931607.1:2252640..2252792,
JH931607.1:2252851..2253407,JH931607.1:2253459..2253602)
/codon_start=1
/locus_tag="AGABI2DRAFT_191532"
/product="tyrosinase"
/note="GO_function: GO:16491 - oxidoreductase activity;
GO_process: GO:8152 - metabolic process"
/db_xref="GOA:K9HSW6"
/db_xref="InterPro:IPR002227"
/db_xref="InterPro:IPR008922"
/db_xref="InterPro:IPR016216"
/db_xref="UniProtKB/TrEMBL:K9HSW6"
/protein_id="EKV49554.1"
/translation="MSLIATVGPTGGVKNRLNIVDFVKNEKFTLYVRSLELLQAKEQH
DYSSFFQLAGIHLCLPFTEWAKERPSMNLYKAGYCTHGQVLFPTWHRTYLSVFEQILQGA
AIEVANKFTSNQTDWIQAAQDLRQPYWDWGPFELMPDDEVIKNEEVNITNYDGKKISVKN
PILRYHFPIPDPSFKPYGDFATWRRTTVRNPDNRREDIPGLIKKMRLEEGQIREKTYNM
LKFNDAWERFSNHGISDDQHANSLESVHDDIHVMVGYGKIEGHMDHPFFAAFDPIFWLH
HTNVDRLLSLWKAINPDVWVTSGRNRDGTMGIAFPNAQINDETPLEPFYQSEDKVWTSAS
LADTARLGYSYPDFDKLVGGTKELIRDAIDDLLIDERYSKPSGARNNTAFDILLADFKGI
TKEHKEDLKMVDWTIHVAFKKFELKESFSLLFYFASDGGDYDQENCFCVGSINAFRGTTP
ETCANCQDNENLIQEGFIHLNHYLARDLESFEPQDVHKFLKEKGLSYKLYSREDKSLTS
LSVKIEGRPLHLPPGEHRPKYDHTQDRVVFDDVAVHVIN"
XX
SQ Sequence 1671 BP; 446 A; 401 C; 390 G; 434 T; 0 other; 1471638968 CRC32;
atgtcgctga ttgctactgt cggaccaact ggcggaggtta agaacatcggtt 60
gattttgtga agaatgaaaa gttttcacg ctttatgtac gtcctccctcga 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	gggccaAAAGA	gcggcCTTC	atgaacctat	acaaggctgg	ttattgtacc	240
catggcagg	ttctgttccc	gacttggcat	agaacgtacc	tttctgtgtt	cgagcaaata	300
cttcaggAG	ctgccccatcgA	agttgtctaaC	aagttcacttC	ctaatacAAAC	cgttggatC	360
caggcggcgc	aggatctacg	ccagccctac	tgggatttggg	gttgcgaACT	tatgcctcCT	420
gatgaggTTA	tcaagaacGA	agaggctcaAC	attacgaACT	acgatggAAA	gaagatttCC	480
gtcaagaACC	ctatcCTCCG	ctatcacttC	catccgatCG	atccttCTtta	caagccatac	540
ggagactttG	caacctgCG	aacaacAGTC	cgaaacCCCG	atcgtaatAG	gcgagaggat	600
atccccggTC	taatcaaaaa	aatgagactC	gaggaaggTC	agattcgtGA	gaagacctAC	660
aatatgttGA	agttcaacGA	tgcttgggag	agatttagtA	accacggCAT	atctgtatGAT	720
cagcatgCTA	acagcttggA	gtctgttCAC	gatgacattC	atgttatgtt	tggatACGGC	780
aaaatcgaAG	gacatatgGA	ccacccttTC	tttgcgtcCT	tcgaccCGAT	tttctggTTA	840
catcataCCA	acgtcgaccG	tctactatCC	cttggaaAG	caatcaatCC	agatgtgtGG	900
gttacatCGG	gacgtaacAG	ggatgttacc	atgggcatCG	caccAAACGC	tcagatcaAC	960
gacgagactC	ctcttgAGCC	attctatCAA	tctgaggATA	aagtgtggAC	ctcgccCTCT	1020
ctcgctgATA	ctgctcgCTG	cggtctactCC	taccccGATT	tcgacaAGTT	gttggagGA	1080
acaaaggAGT	tgatTCGCGA	cgctatcgAC	gacctcatCG	atgagcggTA	tggaaAGCAA	1140
ccttcgagtG	gggctcgCAA	tactgcTTT	gatcttcTG	ccgatttCAA	gggcatttACC	1200
aaggagcACA	aggaggatCT	caaaaatgtAC	gactggacCA	tccatgttGC	cttcaagaAG	1260
ttcgagttGA	aagagaggTT	cagtcttCTC	tttactttG	cgagtgtatGG	tggcatttAT	1320
gatcaggAGA	attgcttTGT	tggatcaATT	aacgccttCC	gtgggactAC	tcccggAAact	1380
tgcgcgaACT	gtcaagataAA	cgagaactTG	attcaagaAG	gttttattCA	cttgaatCAT	1440
tatcttgCTC	gtgaccttGA	atctttcgAG	ccgcaggACG	tgcacaAGTT	cttAAaggAA	1500
aaaggactGT	catacaaACT	ctacagcagg	gaagataAAAT	cttgacatC	gttgtcAGTC	1560
aagattgAAAG	gacgtccctCT	tcatttgCCA	ccccggagaAC	atcgccgAA	gtacgatcac	1620
actcaggACC	gagtagtGTT	tgtatgttGC	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'applicazione 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>
```

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>
MSLIATVGPTGGVKNRNLNIVDFVKNEKFTLYVRSLELLQAKEQHDYSSFFQLAGIHGLPTEWAKERPSMNLYKACYCT
HCGVLFPTWHRTYLSVFEQILQGAAIEVANKFTSNQTDWIQAQDLRQPYWDWGFELMPPDEVIKNEEVNITNYDGKKIS
VKNPILRYHFHPIDPSFKPYGDFATWRTVRNPDRNRREDIPGLIKKMRLEEGQIREKTYNMLKFNDAWERFSNHGISDD
QHANSLESVHDDIHVMVGYKGKIEGHMDHPFFAAFDPFIPFWLHHTNVDRLLSLWKAINPDVWVTSGRNRDGTMGIAAPNAQIN
DETPLEPFYQSEDKVWTSASLADTARLGYSYPDFDKLVGGTKELIRDAIDLIDERYGSKPSSGARNATAFDLLADFKGIT
KEHKEDLKMYYDWTHIHVAFKKFELKESFSLLFYFASDGGDYDQENCFCVGSINAFRGTTPETCANCQDNENLIQEGFIHLNH
YLARDLESFEPQDVHKFLKEKGKLSYKLYSREDKSLSVSKIEGRPLHLPPGEHRPKYDHQTDRVVFDDVAHVVIN
        </value>
    </qualifier>
</feature>
<sequence>
atgtcgctgattgtactgtcgaccaactggcgaggatgttgcacatcgtt
gattttgtgaagaatgaaatggatttttcacgcgtttatgtacgcgtccctcgaaccttctacaa
gccaaggAACAGCATGACTACTCGTCTTCACCTCGCCGGATTATGGCTACCC
tttactgagtgcccggaaagagcggccttccatgaacctatacaaggctgttattgtacc
catggcagggtttgttcccgacttgcatacgtacccgttctgtgttcgagcaata
cttcaaggAGCTGCCATGCAAGGTGTAACAAGTTCACTCTAAACCCGATTGATC
caggccgcgcaggatctacgcgcggccatgtggattgggttcgaaactatgcctcc
gtatgggttatcaagaacgaaaggtaacattacgcactacgtggaaagatttcc
gtcaagaaccctatcctccgcatactccatccgtatgcacgcgttctttcaagccatac
ggagactttgcacccgtggcaacaacagtccgaaaccccgatcgtataaggcgagaggat
atccccgggtcaatcaaaaaaatggactcgaggaaaggctcgatccgtgagaagacccatc
aatatgttgcaggatcaacgtgttggggagagatttagtaaccacgcgcatactgtatgtat
cagcatgtcaacagcttggagtctgttgcacgtacgtacattcatgttatgttggatacggc
aaaatcgaaaggacatatggaccacccttctttgtgcctcgaccggatttctgttgc
catcataccacgtcgaccgtctactatcccttggaaagcaatcaatccagatgtgtgg
gttacatcgggacgttaacaggatgttgcacatggcatcgaccacaaacgctcagatcaac
gacgagactcccttgcggcattctatcaatctgaggataaagtgtggacccctcgccct

```

Banche dati biologiche

Acquisizione dei dati

- Banche dati in **formato TABELLA**

campo A, campo B, ... campo N

record 1,

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

record 2,

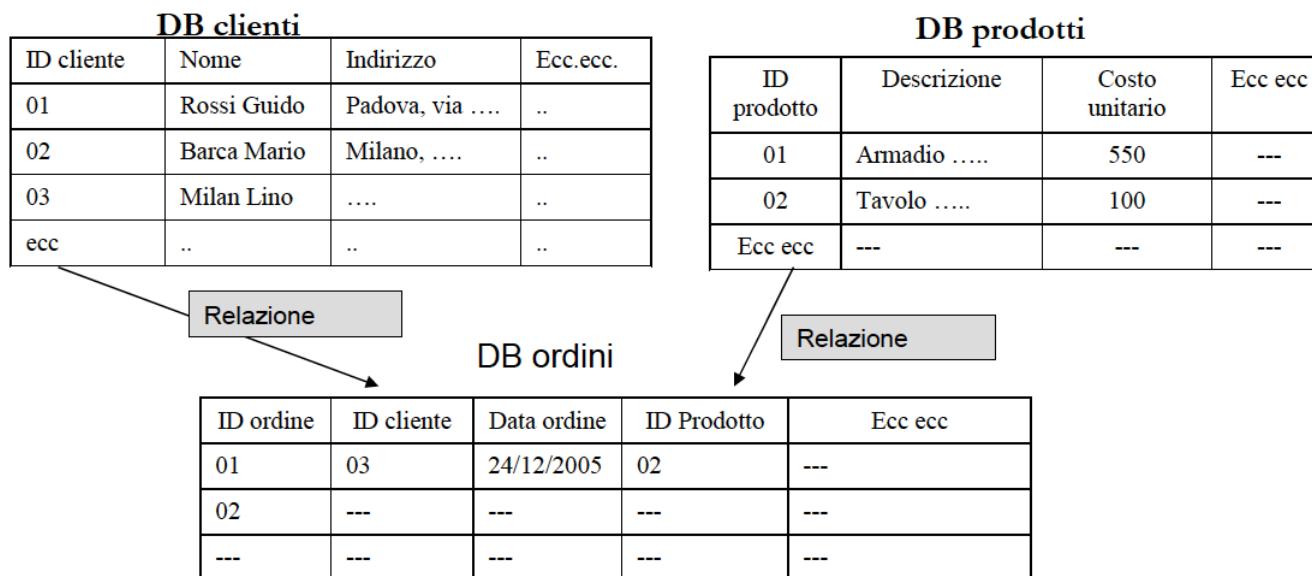
>Feature Seq2			
2626	2590	tRNA	
2570	2535		
			product
			tRNA-Phe
... ...			
>Feature Seq3			
1080	1210	CDS	
1275	1315		
			product
			note
			actin
1055	1210	mRNA	alternatively spliced
1275	1340		
			product
			actin
1055	1340	gene	
			gene
			ACT
1055	1079	5'UTR	
1316	1340	3'UTR	

- 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.'



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 di consentire di navigare fra i database anche se dislocati su siti tra di loro remoti

Names and origin

Protein names	<i>Submitted name:</i> 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

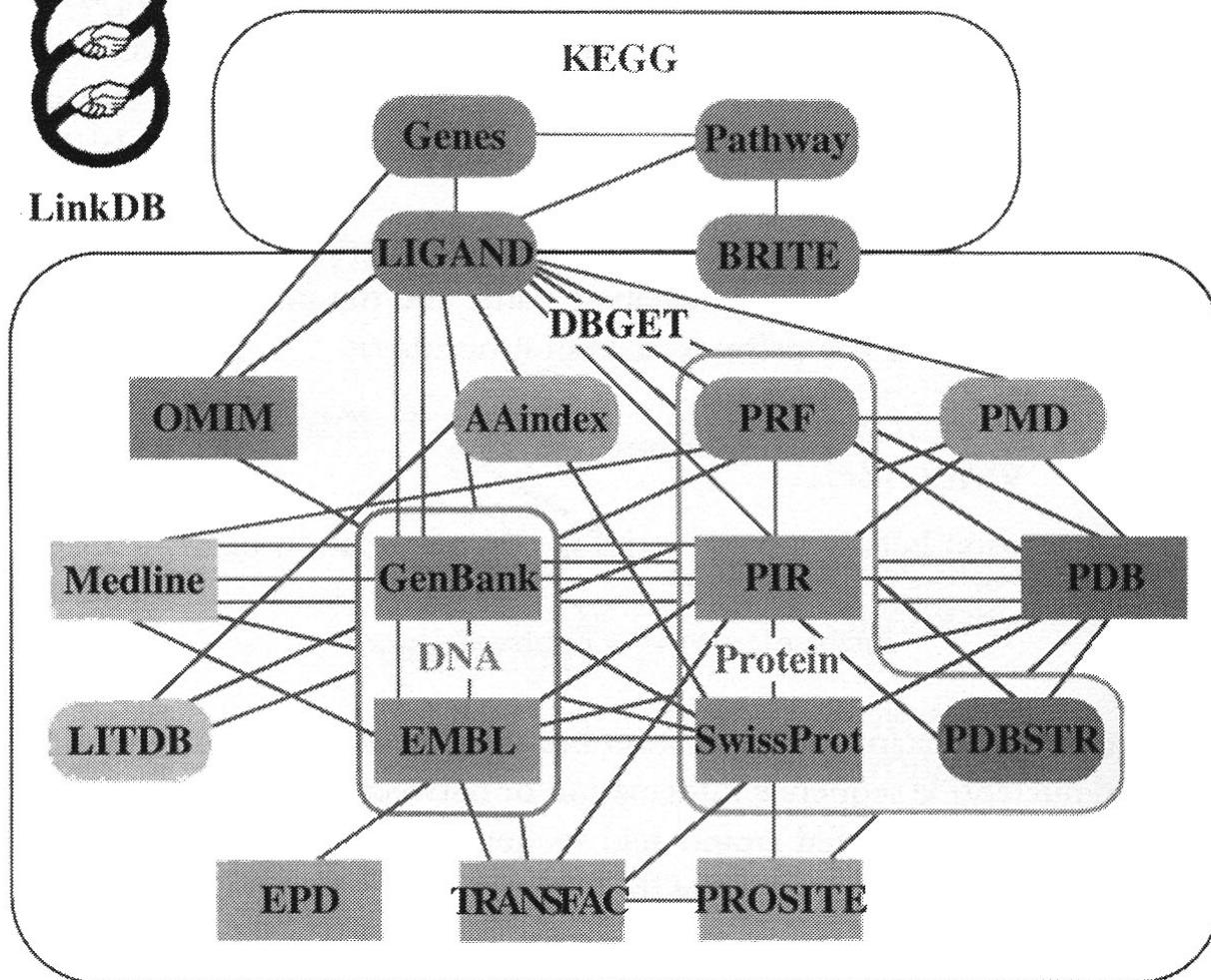
Ligand	Metal-binding SAAS SAAS002227
Technical term	Complete proteome

Gene Ontology (GO)

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



DBGET Database Links



A link-based integration of molecular biology databases in the DBGET/LinkDB system at GenomeNet (<http://www.genome.ad.jp/>). The lines indicate that 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, The Life Sciences Search Engine

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

NCBI

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases Help

Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts	Books: online books
PubMed Central: free, full text journal articles	OMIM: online Mendelian Inheritance in Man
Site Search: NCBI web and FTP sites	

Nucleotide: Core subset of nucleotide sequence records	dbGaP: genotype and phenotype
EST: Expressed Sequence Tag records	UniGene: gene-oriented clusters of transcript sequences
GSS: Genome Survey Sequence records	CDD: conserved protein domain database
Protein: sequence database	Clone: integrated data for clone resources
Genome: whole genome sequences	UniSTS: markers and mapping data
Structure: three-dimensional macromolecular structures	PopSet: population study data sets
Taxonomy: organisms in GenBank	GEO Profiles: expression and molecular abundance profiles
SNP: short genetic variations	GEO DataSets: experimental sets of GEO data
dbVar: Genomic structural variation	Epigenomics: Epigenetic maps and data sets
Gene: gene-centered information	PubChem BioAssay: bioactivity screens of chemical substances



Entrez, The Life Sciences Search Engine

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

The screenshot shows the NCBI Entrez search interface. At the top, there's a navigation bar with links for HOME, SEARCH, SITE MAP, PubMed, All Databases, Human Genome, GenBank, Map Viewer, and BLAST. Below the navigation bar is a search bar labeled "Search across databases" with a "GO" button and "Clear Help" links. A large red circle highlights this search area, and a red arrow points from it to the main content area. The main content area is titled "Welcome to the Entrez cross-database search page". It lists various resources in two columns:

Resource	Description
PubMed	biomedical literature citations and abstracts
PubMed Central	free, full text journal articles
Site Search	NCBI web and FTP sites
Nucleotide	Core subset of nucleotide sequence records
EST	Expressed Sequence Tag records
GSS	Genome Survey Sequence records
Protein	sequence database
Genome	whole genome sequences
Structure	three-dimensional macromolecular structures
Taxonomy	organisms in GenBank
SNP	single nucleotide polymorphism
Books	online books
Images	images from full text resources at NCBI
OMIM	online Mendelian Inheritance in Man
dbGaP	genotype and phenotype
UniGene	gene-oriented clusters of transcript sequences
CDD	conserved protein domain database
UniSTS	markers and mapping data
PopSet	population study data sets
GEO Profiles	expression and molecular abundance profiles
GEO DataSets	experimental sets of GEO data
Epigenomics	Epigenetic maps and data sets

Ricerca in tutti i database

Risorse principali:

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



Entrez, The Life Sciences Search Engine

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

Search across databases **TP53**

- Result counts displayed in gray indicate one or more terms not found

4804 PubMed: biomedical literature citations and abstracts	523 Books: online books
3857 PubMed Central: free, full text journal articles	637 Images: images from full text resources at NCBI
1 Site Search: NCBI web and FTP sites	215 OMIM: online Mendelian Inheritance in Man

1184 Nucleotide: Core subset of nucleotide sequence records	none dbGaP: genotype and phenotype
1199 EST: Expressed Sequence Tag records	119 UniGene: gene-oriented clusters of transcript sequences
6 GSS: Genome Survey Sequence records	none CDD: conserved protein domain database
897 Protein: sequence database	114 UniSTS: markers and mapping data
97 Genome: whole genome sequences	15 PopSet: population study data sets
26 Structure: three-dimensional macromolecular structures	29464 GEO Profiles: expression and molecular abundance profiles
none Taxonomy: organisms in GenBank	81 GEO DataSets: experimental sets of GEO data
883 SNP: single nucleotide polymorphism	none Epigenomics: Epigenetic maps and data sets
none dbVar: Genomic structural variation	166 Cancer Chromosomes: cytogenetic databases
676 Gene: gene-centered information	65 PubChem BioAssay: bioactivity screens of chemical substances

Cerchiamo informazioni relativamente al gene umano TP53

Clicchiamo in corrispondenza di Gene

● ... Ricerca bibliografica nella letteratura scientifica

NCBI

Entrez, The Life Sciences Search Engine.

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases Help

Welcome to the Entrez cross-database search page

 PubMed: biomedical literature citations and abstracts	 Books: online books
 PubMed Central: free, full text journal articles	 OMIM: online Mendelian Inheritance in Man
 Site Search: NCBI web and FTP sites	

 Nucleotide: Core subset of nucleotide sequence records	 dbGaP: genotype and phenotype
 EST: Expressed Sequence Tag records	 UniGene: gene-oriented clusters of transcript sequences
 GSS: Genome Survey Sequence records	 CDD: conserved protein domain database
 Protein: sequence database	 Clone: integrated data for clone resources
 Genome: whole genome sequences	 UniSTS: markers and mapping data
 Structure: three-dimensional macromolecular structures	 PopSet: population study data sets
 Taxonomy: organisms in GenBank	 GEO Profiles: expression and molecular abundance profiles
 SNP: short genetic variations	 GEO DataSets: experimental sets of GEO data
 dbVar: Genomic structural variation	 Epigenomics: Epigenetic maps and data sets
 Gene: gene-centered information	 PubChem BioAssay: bioactivity screens of chemical substances



Entrez, The Life Sciences Search Engine

Entrez - Pubmed

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

NCBI Resources ▾ How To ▾ My NCBI Sign In

PubMed.gov U.S. National Library of Medicine National Institutes of Health

Search: PubMed Limits Advanced search Help

Search Clear

PubMed

PubMed comprises more than 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

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 acetyltransferase genes in B-cell non-Hodgkin's lymphoma
PG - 189-95
AB - B-cell non-Hodgkin's lymphoma comprises a heterogeneous group of diseases the pathogenesis of which is believed to involve oncogenes and tumour-suppressor genes. Follicular lymphoma and diffuse large B-cell lymphoma are characterized by structural alterations inactivating related histone and non-histone acetyltransferases. Transcriptional co-activators in most cases of diffuse large B-cell lymphoma and genomic deletions and/or somatic mutations in the coding domain of these two genes. This suggesting that reduction in HAT does not demonstrate specific defects in acetyltransferase genes. Oncoprotein and activation of the p300/CREBBP/EP300 mutations as a major pathogenic mechanism in the development of B-cell non-Hodgkin's lymphoma, where targeting acetylation/deacetylation
AD - 1] Institute for Cancer Genetics and Genomic Medicine, Department of Pathology and Cell Biology, Columbia University, New York, NY, USA
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



Entrez, The Life Sciences Search Engine

- ▶ 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;

The screenshot shows the PubMed search interface. At the top, there is a search bar with "PubMed" and a dropdown menu. To the right of the search bar are links for "Limits", "Advanced search", and "Help". Below the search bar is a large button labeled "Search" and a smaller "Clear" button. A red arrow points from the "Usare i Limits;" bullet point to the "Limits" link in the top navigation. The main area is titled "Limits" and contains several filter sections: "Dates" (Published in the Last: Any date), "Type of Article" (Clinical Trial, Editorial, Letter, Meta-Analysis, Practice Guideline), "Species" (Humans, Animals), "Subsets" (AIDS, Bioethics, Cancer, Complementary Medicine, Core clinical journals), "Text Options" (Links to full text, Links to free full text, Abstracts), "Languages" (English, French, German, Italian, Japanese), "Sex" (Male, Female), "Ages" (All Infant: birth-23 months, All Child: 0-18 years, All Adult: 19+ years, Newborn: birth-1 month, Infant: 1-23 months), and "Search Field Tags" (Field: All Fields). At the bottom of the "Limits" section are "Reset" and "Search" buttons.

Entrez - Pubmed



Entrez, The Life Sciences Search Engine

- ▶ 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

The screenshot shows the PubMed Advanced Search interface. At the top, there's a navigation bar with links for 'Limits', 'Advanced search', and 'Help'. Below the navigation is a search bar containing 'PubMed'. To the right of the search bar is a 'Search Builder' sidebar with a list of search fields. The 'All Fields' option is selected and highlighted with a blue background. Other options listed include Affiliation, Author, Book, Corporate Author, Create Date, EC/RN Number, Editor, Entrez Date, Filter, First Author, Full Author Name, Full Investigator Name, Grant Number, ISBN, Investigator, Issue, Journal, Language, Last Author, Location ID, MeSH Date, MeSH Major Topic, MeSH Subheading, MeSH Terms, Pagination, Pharmacological Action, and Publication Date.

PubMed Advanced Search

Search Box

(smith[Author]) AND "2009"[Publication Date] : "2010"[Publication Date]

[Limits](#) [Details](#) [Help](#)

Search **Preview** **Clear**

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 NLM (National Library of Medicine's).

NCBI Resources How To

MeSH
NLM Controlled Vocabulary

Search: MeSH

Limits Advanced search Help

Search Clear

Search History

Search #27 Search p53

Clear History

Search History Instructions

More Resources

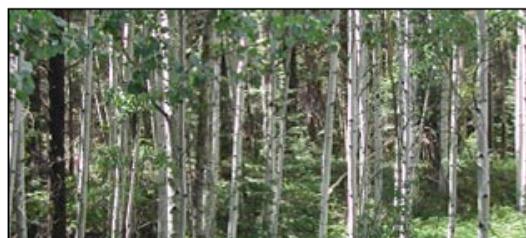
MeSH Database (circled)

Journals in NCBI Databases

Single Citation Matcher

Clinical Queries

Topic-Specific Queries



MeSH

MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus for PubMed.

Using MeSH

[Help](#)

[Tutorials](#)

More Resources

[E-Utilities](#)

[NLM MeSH Homepage](#)