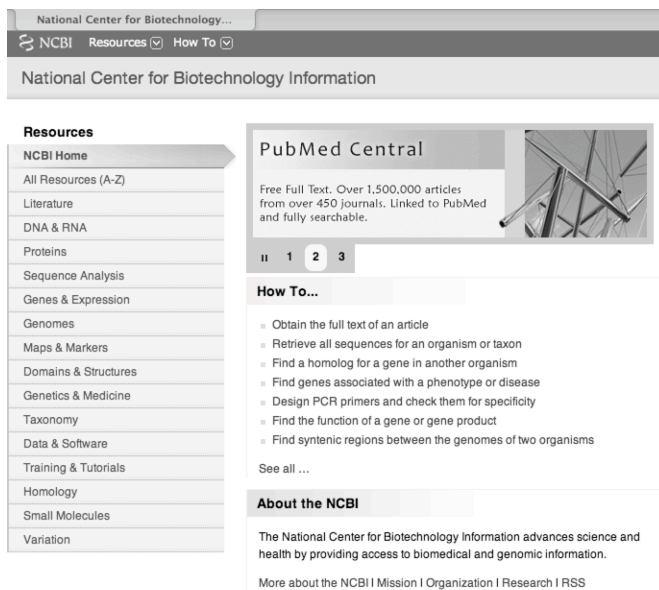


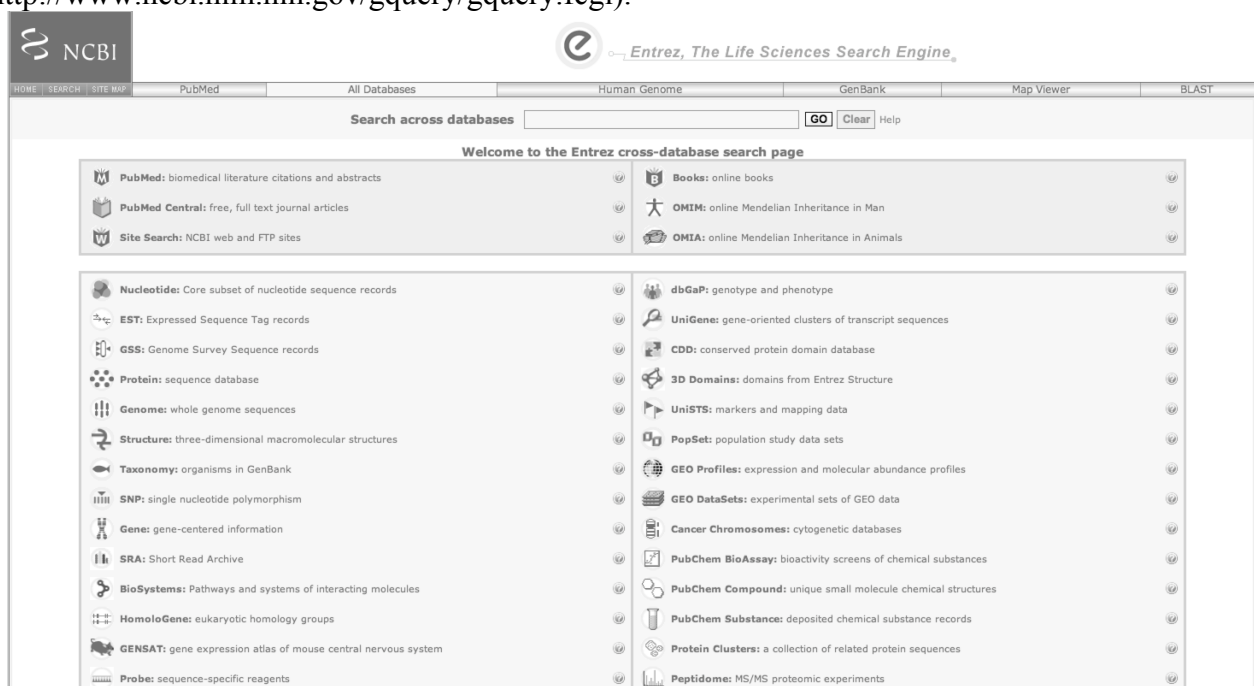
1. Le banche dati biologiche

NCBI, ENTREZ, MEDLINE, PUBMED

Il National Center for Biotechnology Information (NCBI), un ente governativo americano, gestisce un sito web (www.ncbi.nlm.nih.gov) attraverso il quale è possibile accedere a una serie di banche di dati genomiche, proteiche e così via.

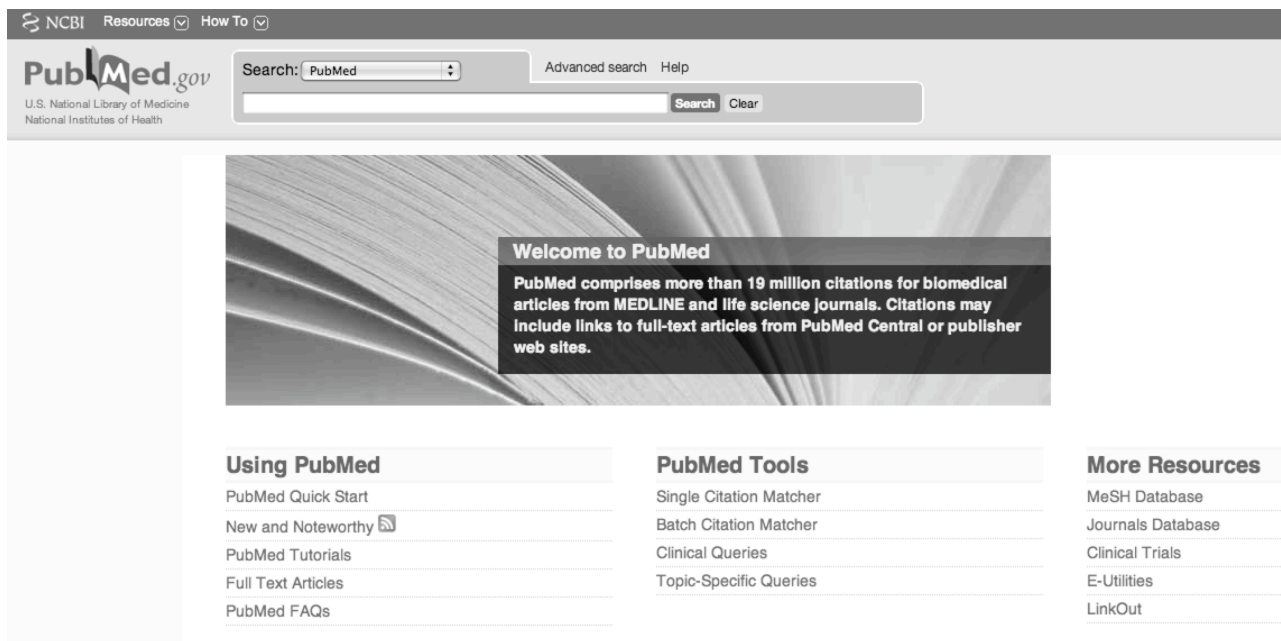


L'interfaccia che permette di effettuare l'accesso alle banche dati si chiama ENTREZ (<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>).



Per esempio, l'accesso a MEDLINE (*Medical Literature Analysis and Retrieval System Online*), la più autorevole banca di dati bibliografici in campo medico, tramite ENTREZ è sul sito dell'NCBI all'indirizzo www.ncbi.nlm.nih.gov/entrez/.

PUBMED raccoglie le pubblicazioni in campo medico dal 1966 fino a oggi e comprende materiale anche precedente (più di 100 000 schede fra il 1963 e il 1965). La banca dati mette a disposizione le schede di ogni articolo di circa 4800 riviste internazionali in 30 lingue. Dal 2002 viene aggiornata 5 volte alla settimana con 1500-3500 nuove schede al giorno.



Cercare informazioni bibliografiche in PUBMED

Ricordiamo che le interrogazioni su una base di dati strutturati si effettuano utilizzando gli operatori **booleani**. Il termine *booleano*, che ricorre spesso nell'ambito delle ricerche in database sia locali sia nel web, si riferisce a un sistema logico sviluppato dal matematico inglese George Boole (1815-64). La logica booleana consiste di tre operatori logici:

- OR
- AND
- NOT

Nelle ricerche booleane, l'operatore AND tra due parole, campi o valori (p. e. «pera AND mela», oppure «Editore: Zanichelli AND Autore: Rossi») significa che si stanno ricercando documenti contenenti entrambe le parole o valori e non uno solo di essi. L'operatore OR («pera OR mela») significa che si stanno cercando documenti contenenti almeno una delle parole o valori. Infine, l'operatore NOT («pera NOT mela») significa che si stanno cercando documenti contenenti la prima parola (o valore) e non la seconda.

Una base di dati bibliografici come PUBMED è fatta di schede come quella riportata di seguito (notate i nomi sintetici dei campi, sulla sinistra, che sono al massimo 4 lettere; sulla destra dopo il trattino compare il contenuto relativo):

PMID - 870973
 OWN - NLM
 STAT- MEDLINE
 DA - 19770630
 DCOM - 19770630
 LR - 20041117
 PUBM - Print
 IS - 0036-8075
 VI - 196
 IP - 4294
 DP - 1977 Jun 3
 TI - *Erythema chronicum migrans and Lyme arthritis: cryoimmunoglobulins and clinical activity of skin and joints.*
 PG - 1121-2
 AB - *We report the presence of serum cryoimmunoglobulins in patients with attacks of a newly described epidemic arthritis--Lyme arthritis--and in some patients with a characteristic skin lesion--erythema chronicum migrans--.....*
 AU - Steere AC
 AU - Hardin JA
 AU - Malawista SE
 LA - eng
 PT - Journal Article
 PL - UNITED STATES
 TA - Science
 JID - 0404511
 EDAT - 1977/06/03
 MHDA - 1977/06/03 00:01
 PST - ppublsh
 SO - Science 1977 Jun 3;196(4294):1121-2.

Impostando una ricerca con la stringa «HOMO AND SHMT», verranno visualizzati gli articoli in cui sono presenti sia il termine HOMO sia il termine SHMT in uno qualsiasi dei campi delle schede conservate in PubMed:

The screenshot shows the PubMed search interface. The search bar contains 'SHMT AND HOMO'. The results page displays a list of articles, with the first one being 'Serine hydroxymethyltransferase from Plasmodium vivax is different in substrate specificity from its homologues.' The interface includes navigation options like 'First', 'Prev', 'Page 1', 'Next', and 'Last'. On the right side, there are filters for 'Filter your results' (All (72), Review (5), Free Full Text (27)) and 'Find related data' (Database: Select, Find items). The search details section shows the query: 'SHMT[All Fields] AND "hominidae"[MeSH Terms] OR "hominidae"[All Fields] OR "homo"[All Fields]'. The recent activity section shows 'Turn Off' and 'Clear' buttons.

In realtà, la nostra ricerca viene «tradotta» nella seguente espressione:

SHMT[All Fields] AND ("hominidae"[MeSH Terms] OR "hominidae"[All Fields] OR "homo"[All Fields])

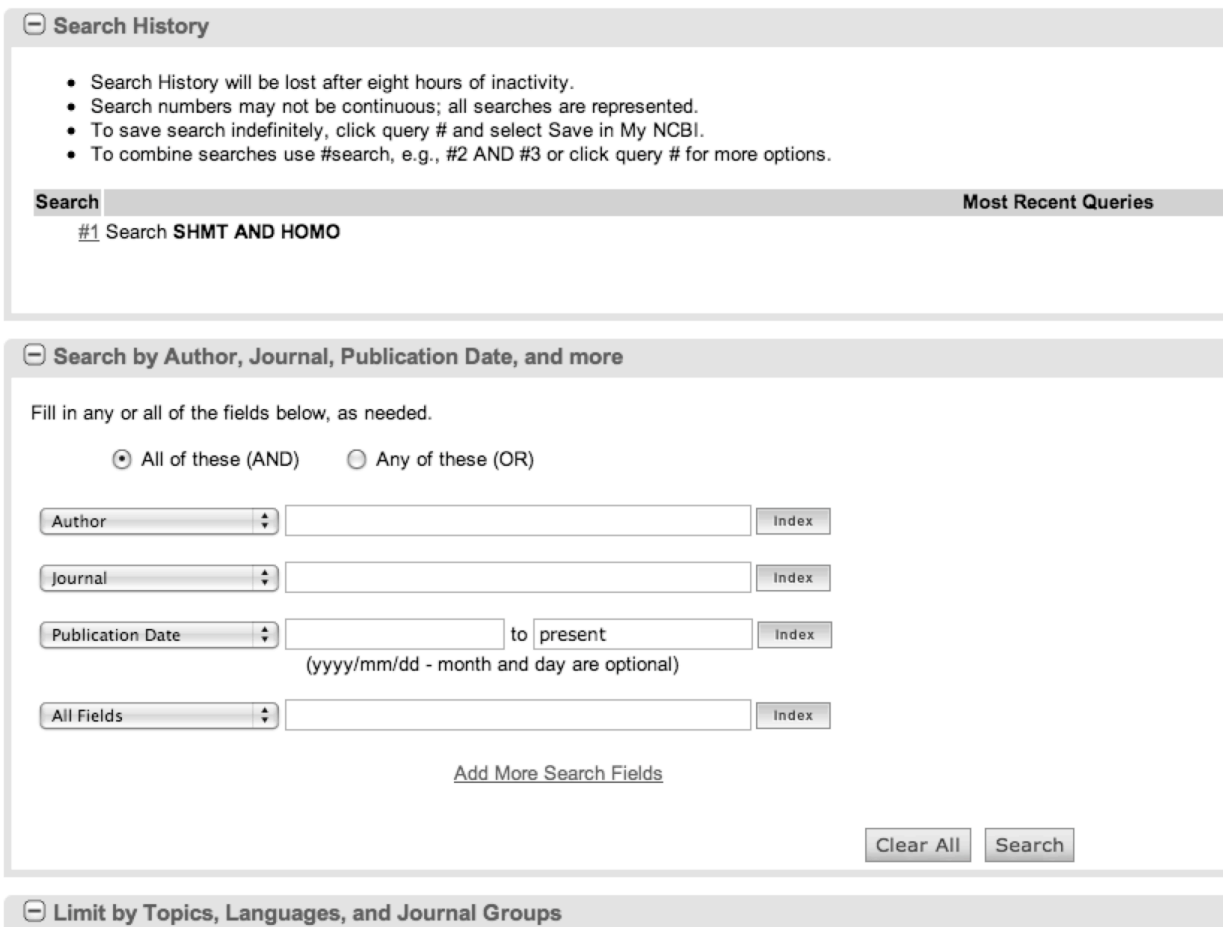
Il termine *All Fields*, chiuso tra parentesi quadre, indica che la ricerca sarà effettuata in tutti i campi. *MeSH Terms* si riferisce invece ai termini biomedici indicizzati in un vocabolario curato dall'NCBI.

Tipi di ricerca

L'interfaccia di ENTREZ permette di effettuare ricerche semplici o avanzate. Quando si inseriscono i termini nella casella di ricerca, il sistema automaticamente crea una interrogazione booleana. I parametri della query possono essere modificati attraverso la ricerca avanzata (*advanced search*), accessibile attraverso cliccando sul link accanto al campo di input della ricerca.



Advanced Search



I settori *Search by Author, Journal, Publication Date and more* e *Limit by Topics, Languages and Journal Groups* consentono di raffinare la ricerca specificando i valori di determinati campi. Per esempio, è possibile limitare la ricerca ad articoli pubblicati solo da un certo autore o in un determinato intervallo di tempo. Il settore *Index of Fields and Field Values* consente di formare un'interrogazione booleana selezionando i campi su cui fare la ricerca e combinandone i termini. *Search History* elenca le ricerche già fatte e consente di combinarle fra loro.

ESERCITAZIONE

Rispondete alle domande dopo aver effettuato le seguenti ricerche:

- Impostate una ricerca che consenta di ottenere tutti gli articoli sulla *serine hydroxymethyltransferase* pubblicati dal 1990 al 2000. Quanti risultati si ottengono? (*Suggerimento*: utilizzate i campi *Publication date*.)
- Impostate una ricerca che consenta di ottenere tutti gli articoli che presentino nel titolo le parole *serine* e *hydroxymethyltransferase* pubblicati dal 1990 al 2000. Quanti risultati si ottengono? (*Suggerimento*: utilizzate i campi *Publication date* e *Title*.)
- Impostate una ricerca che consenta di ottenere tutti gli articoli pubblicati da Bat sul DNA. Quanti risultati si ottengono? (*Suggerimento*: utilizzate i campi *Title* e *MeSH terms*, accessibile dal menù a tendina *All fields*.)

Cercare sequenze proteiche con ENTREZ

Oltre alla possibilità di effettuare ricerche bibliografiche, ENTREZ permette di consultare ed eventualmente prelevare i dati contenuti in banche dati nucleotidiche, proteiche, strutturali e via dicendo. Per accedere a tale tipologia di informazione è necessario prima di tutto selezionare, tra le tante disponibili, la classe di dati sulla quale si vuole effettuare la ricerca, selezionando la voce di interesse nel menù a tendina posizionato alla destra del tasto *Search* (in alto a sinistra nella pagina principale di ENTREZ). Si selezioni, per esempio, la voce *Protein* e si inseriscano i termini *serine* e *hydroxymethyltransferase*.

The screenshot shows the NCBI Protein search results page. The search query is 'serine hydroxymethyltransferase'. The results are displayed in a list format, showing the first six items. Each item includes a checkbox, a protein name, and a Gene ID (GI) with a link to the protein page. The results are as follows:

Item	Protein Name	Gene ID (GI)
1.	serine hydroxymethyltransferase	518 aa protein AAA33687.1 GI:169158
2.	serine hydroxymethyltransferase [Mycobacterium sp. MCS]	492 aa protein YP_841292.1 GI:108801095
3.	serine hydroxymethyltransferase [Campylobacter jejuni subsp. jejuni NCTC 11168]	414 aa protein CAL34552.1 GI:112359766
4.	serine hydroxymethyltransferase [Clavibacter michiganensis subsp. michiganensis NCPPB 382]	425 aa protein CAN02604.1 GI:147831636
5.	serine hydroxymethyltransferase [Ehrlichia ruminantium str. Weigelvonden]	421 aa protein CAH58416.1 GI:57161490
6.	serine hydroxymethyltransferase [Streptomyces coelicolor A3(2)]	420 aa protein

Additional features visible in the screenshot include the 'Top Organisms' list (Escherichia coli, Vibrio cholerae, Salmonella enterica, Chlamydia trachomatis, Bacillus cereus, All other taxa) and the 'Recent activity' section showing previous searches.

Notate che la proteina viene identificata da un **codice GI** (*gene identifier*), assegnato da ENTREZ in

maniera univoca a tutte le entry presenti, e da un codice relativo alla banca dati di appartenenza. In alcune pagine del programma possono essere presenti le seguenti opzioni, poste al di sotto dell'area di ricerca:

- *Limits*: consente di limitare i valori di alcuni campi (funzione simile a quella trovata nell'area di ricerca avanzata).
- *Preview/Index*: consente di formare un'interrogazione booleana scegliendo i campi sui quali effettuare la ricerca e combinando i termini di ricerca. Premendo il pulsante *Preview* viene visualizzato il numero di risultati prodotti da quella particolare ricerca.
- *History*: elenca le ricerche già fatte e consente di combinarle fra loro.
- *Clipboard*: consente di salvare temporaneamente i risultati delle ricerche.
- *Details*: mostra la traduzione dell'interrogazione nella sintassi del motore di ricerca.

ESERCITAZIONE

Impostate una ricerca della proteina umana *serine hydroxymethyltransferase*, isoforma 2, lunga 444 residui. (*Suggerimento*: si usino gli strumenti *Limits* e *Preview/Index* dopo aver selezionato la voce *Protein*). Qual è la differenza tra le due isoforme?

Supponiamo ora di voler prelevare la sequenza della proteina cercata in formato **FASTA**. È necessario fare attenzione al formato con il quale le sequenze sono estratte, perché molti programmi di analisi sono in grado di riconoscere solo alcuni. Uno dei formati più semplici è appunto il formato FASTA, nel quale ogni sequenza viene scritta con una riga di intestazione che riporta il nome della entry, preceduta dal simbolo >, seguita nelle righe successive la sequenza stessa. Più sequenze possono essere scritte una sotto l'altra. Altri formati molto utilizzati sono il GCG, PAUP, o PIR.

Per ottenere la sequenza in formato FASTA si deve selezionare la voce *FASTA* nel menù *Format* in alto a sinistra della pagina. Per salvare la sequenza nello stesso formato, invece, bisogna selezionare la voce *Download* in alto a destra.

The screenshot shows the NCBI Protein database search results for the protein *serine hydroxymethyltransferase 1 (soluble) isoform 2 [Homo sapiens]*. The search bar contains the protein name, and the results page displays the FASTA sequence. The sequence is as follows:

```
>gi|22547189|ref|NP_683718.1| serine hydroxymethyltransferase 1 (soluble) isoform 2 [Homo sapiens]
MTMPVNGAHKADLHSSHDKMLAQLKDSDEVVYNIKKESNRQRVGLLELIASENFASRAVLEALGSLN
NKYSQYPOQRYYGOTEFIDELETLCQRALQAKLDPQCHGVNVVQYSGSPAMFAVYTTALVEPGRIMG
LDLFDGQHLHGFMDKKKISATGIFPESMYPVNVDPQYINVDLEEARLFPKLLIAGTCYSRML
YARLRKIDANGAYLMDMAHISGLVAAGVVPSPFBCHVVTTHKTLRCCRAGMIFYRKGVAVALKQA
MTLEFKVYHQVAVNCRALSEALTELCYKIVTQGSNDHLLVDLRSKOTDGGRAEKVLEACSACNKNTC
PGDRSALRPSGLRLGTPALTSRGLLEKDFQKAVHFIRGIELTLQIQSDTGVRAITLKEFKERLAGDKYA
AVQALREEVESFASLFLPLGLPDF
```

The page also includes a search bar, navigation tabs (Limits, Preview/Index, History, Clipboard, Details), and a format selection menu (GenPept, FASTA, Graphics, More Formats). On the right side, there are sections for 'Change Region Shown', 'Sequence Analysis Tools' (BLAST Sequence, Conserved Domains), 'Articles about the SHMT1 gene', 'Identical Proteins for NP_683718.1', 'RefSeq mRNA', 'RefSeq Protein Isoforms', and 'More about the SHMT1 gene'.

Banche di dati genomiche

Concentriamoci ora sulle informazioni che è possibile acquisire dal genoma umano. Effettueremo passo dopo passo una ricerca di questo tipo.

1. Cliccate su *Links* sulla destra della vostra sequenza e selezionate *gene*.

1: SHMT1 serine hydroxymethyltransferase 1 (soluble) [Homo sapiens]
GeneID: 6470 updated 8-Nov-2009

Official Symbol SHMT1 provided by HGNC

Official Full Name serine hydroxymethyltransferase 1 (soluble) provided by HGNC

Primary Source HGNC:10850

See related Ensembl:ENSG00000176974; HPRD:01643; MIM:182144

Gene type protein coding

RefSeq status REVIEWED

Organism *Homo sapiens*

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as SHMT; CSHMT; MGC15229; MGC24556; SHMT1

Summary This gene encodes the cellular form of serine hydroxymethyltransferase, a pyridoxal phosphate-containing enzyme that catalyzes the reversible conversion of serine and tetrahydrofolate to glycine and 5,10-methylene tetrahydrofolate. This reaction provides one carbon units for synthesis of methionine, thymidylate, and purines in the cytoplasm. This gene is located within the Smith-Magenis syndrome region on chromosome 17. Alternative splicing of this gene results in 2 transcript variants encoding 2 different isoforms. Additional transcript variants have been described, but their biological validity has not been determined. [provided by RefSeq]

Genomic regions, transcripts, and products

(minus strand) Go to [reference sequence details](#) Try our new [Sequence Viewer](#)

NC_000017.10

18266856 18291187

NP_149918.1 NP_682718.1 isoform 2 CCDS11197.1
NP_151625.2 NP_151168.2 isoform 1 CCDS11398.1

■ = coding region ■ = untranslated region

2. Selezionate poi *See SHMT1 in MapViewer*:

Homo sapiens (human) Build 37.1 (Current)
Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 [17] 18 19 20 21 22 X Y MT

Query: 6470[[gene_id](#)] [clear](#)

Master Map: Genes On Sequence

Region Displayed: 18,227K-18,271K bp

Model [Hs Uni6](#) [ensGenes](#) [RefSeq](#) [RNO](#) [Genes_seq](#) [Symbol](#)

SHMT1 + OMIM HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 17p11.2 serine hydroxymethyltransferase 1 (soluble)

Nella pagina che apparirà troverete informazioni sulla struttura del gene della serina idrossimetiltrasferasi (SHMT) e dei geni adiacenti nel cromosoma 17, oltre a diversi collegamenti a molte altre pagine. Vi invitiamo a esplorare il contenuto di questa pagina:

- Disegnate schematicamente la struttura del gene dell'SHMT. Come sono indicati, secondo voi, gli introni, gli esoni e la regione 3' non tradotta?
- Disegnate schematicamente la posizione, sul cromosoma 17, in corrispondenza della quale si trova l'SHMT.
- Aiutandovi con gli strumenti di zoom disponibili, tentate di individuare i due geni adiacenti all'SHMT. Di quali geni si tratta?
- Qual è l'intervallo di nucleotidi in cui si trova, approssimativamente, il gene dell'SHMT?
- Individuate il collegamento a OMIM (*Online Mendelian Inheritance in Man*) e fornite una brevissima descrizione della malattia genetica che potrebbe derivare dalla delezione della porzione genica nella quale mappa l'SHMT.
- Nella pagina che avete visualizzato sono presenti solo alcune delle mappe genomiche disponibili. Aiutandovi con lo strumento *Maps and Options* (in alto a destra), potete modificare la visualizzazione di tali mappe, e aggiungere anche mappe di altri organismi.

Banche di dati strutturali

Il punto di partenza per qualsiasi ricerca di tipo strutturale è certamente la Protein Data Bank (PDB), accessibile dal sito www.pdb.org (è possibile accedere a PDB anche attraverso ENTREZ, selezionando la voce *STRUCTURE* dal menù a tendina posizionato a destra del tasto *Search*.). Collegatevi a PDB e inserite nel campo di ricerca la stringa *serine hydroxymethyltransferase*; cliccando su *Search* dovrete ottenere una pagina simile alla seguente:

The screenshot shows the PDB website interface. At the top, there's a navigation bar with 'MyPDB Login' and 'A MEMBER OF THE PDB'. Below that, a search bar contains the query 'SERINE HYDROXYMETHYLTRANSFERASE'. The results page displays two entries:

Structure ID	Title	Release Date	Exp. Method	Resolution
3G8M	Serine Hydroxymethyltransferase Y55F Mutant	10-Nov-2009	X-RAY DIFFRACTION	3.30 Å
3H7F	Crystal structure of serine hydroxymethyltransferase from <i>Mycobacterium tuberculosis</i>	05-May-2009	X-RAY DIFFRACTION	1.50 Å

Each entry includes a 3D ribbon diagram of the protein structure and detailed classification information such as EC number, polymer type, and length.

- Cercate la struttura tridimensionale che ha codice PDB 1BJ4. Vi invitiamo a esplorare il contenuto di questa pagina. A quale valore di risoluzione è stata risolta questa struttura? Come viene classificata questa struttura in **SCOP** e in **CATH**? La struttura è stata risolta

con qualche ligando? Se sì, quale? In quali processi biologici è coinvolta? Tramite quale collegamento è possibile visualizzare, in formato «testo», le coordinate tridimensionali della proteina, come sono mostrate di seguito?

```

.....
ATOM   104  OE1  GLN  A   24      75.541  32.811 208.061  1.00 88.43      O
ATOM   105  NE2  GLN  A   24      76.753  31.070 207.312  1.00 87.47      N
ATOM   106   N   PRO  A   25      78.636  35.443 204.218  1.00 74.17      N
ATOM   107  CA   PRO  A   25      78.291  35.681 202.811  1.00 71.67      C
ATOM   108   C   PRO  A   25      76.909  35.136 202.473  1.00 70.44      C
ATOM   109   O   PRO  A   25      76.453  34.158 203.072  1.00 71.15      O
ATOM   110  CB   PRO  A   25      79.371  34.904 202.055  1.00 72.23      C
ATOM   111  CG   PRO  A   25      80.521  34.894 203.005  1.00 75.28      C
ATOM   112  CD   PRO  A   25      79.848  34.614 204.319  1.00 73.02      C
ATOM   113   N   LEU  A   26      76.261  35.768 201.498  1.00 68.47      N
.....

```

- Tentate di dare una breve descrizione dei campi presenti in un file .pdb.
- Utilizzate il collegamento *View in Jmol* per visualizzare la struttura.
- Cercate l'articolo originale che si riferisce alla determinazione della struttura tridimensionale della proteina.
- Esplorate i collegamenti ipertestuali esterni a PDB.

Per maggiori informazioni sull'utilizzo di PDB vi invitiamo a seguire il tutorial relativo, disponibile all'indirizzo: <http://www.rcsb.org/pdbstatic/tutorials/tutorial.html>

ESERCITAZIONE

Dopo aver selezionato una proteina umana dall'elenco in basso, attraverso l'utilizzo delle basi di dati trattate nell'esercitazione, determinatene:

1. La funzione
2. La sequenza in formato FASTA
3. L'organizzazione in domini
4. La classificazione in SCOP e CATH
5. La struttura del gene
6. Eventuali malattie associate

- A) *Acetylcholinesterase*
- B) *Hemoglobin*
- C) *BDNF*
- D) *Trypsin*