Bioinformatics

Lecture 2

BCH 550

Arjumand Warsy

Databases

- A **database** is a collection of data for one or more multiple uses. One way of classifying databases involves the type of content, for example: DNA sequence database; protein sequence database; SNP database; bibliographic; image; customer databases; personal database; inventory databases; accounting databases etc etc.
- Databases consist of software-based "containers" that are structured to collect and store information so users can retrieve, add, update or remove such information in an automatic fashion. Database programs are designed for users so that they can add or delete any information needed. The structure of a database is tabular, consisting of rows and columns of information.

Examples of Databases used in biochemistry and molecular biology

- DNA and protein databases
- Pubmed/Medline
- Entrez
- ExPASy Molecular Biology server
- GenBank
- Genome sequence database
- EMBL Nucleotide Sequence Database
- DNA Data Bank of Japan
- DBCAT
- SALSA
- 123 Genomics
- Pfam
- EMALL for DNA sequences or UNIPROT
- UniProtKB
- UniRef
- UniParc
- TIGR
- SWISS-PROT
- Protein and Associated Topics Database
- Protein 3D and Associated Topics Database

Examples of Databases (conti..)

- **Physical Maps, Genetic Maps, Transcript Maps and Integration**
- CHLC Cooperative Human Linkage Center
- An STS-Based Map of the Human Genome
- Généthon Human Genome Research Centre
- CEPH Généthon
- Marshfield
- **SNP databases**
- **Other Sequence Related Databases**
- **Genetic Disease Databases**
- OMIM
- The Human Gene Mutation Database
- **Model Organism Databases**
- Mouse Genome Database
- Drosophila Genome database
- The C. elegans Genome Project
- The Saccharomyces Genome Database
- E. coli Genome Center's Home Page
- Amos's links to organism specific databases

DNA Sites on the Internet

- - Genetics Computer Group (Accelrys) http://www.gcg.com/
- - National Centre for Genome Resources http://www.ncgr.org/
- - National Centre for Biotechnology Information http://www.ncbi.nlm.nih.gov/
- - Baylor College of Medicine Human Genome Centre and Baylor Bioinformatics Unit: http://www.hgsc.bcm.tmc.edu/
- - The Human Genome Database http://www.gdb.org/
- - European Bioinformatics Institute http://www.ebi.ac.uk/
- - The Sanger Centre http://www.sanger.ac.uk/
- - European Molecular Biology Laboratory http://www.embl-heidelberg.de/
- - UK Medical Research Council http://www2.mrc-lmb.cam.ac.uk/
- - SWISS PROJ Databank http://www .isb-sib.ch/
- PERL http://bio.perl.org/
- - Human Genome Organization http://www.gene.ucl.ac.uk/hugo/
- - Human Genome Resources http://www.ncbi.nlm.nih.gov/genome/guide/human/
- - Human Genome Biouser http://genome.ucsc.edu/cgibin/hgGateway

How to Use Bioinformatics- using databases

- **Databases and how to use them**
- Using **PubMed** to become quickly knowledgeable on any biological subject
- Retrieving protein and DNA sequences relevant to our work
- Running our first sequence/database comparison with BLAST

Lets start with a protein we know nothing about: e.g. “dUTPase” by Searching PubMed

1. Using an Internet browser, navigate to www.ncbi.nlm.nih.gov/entrez/ **on the World Wide Web (WWW). The following window opens:**

2. Type in dUTPase in the For window, and click the Go (enter) button.

- For the dUTPase example, we now have more than 400 references at our fingertips, more than enough to start unraveling the mysteries of dUTPase

3. For any entry in the Results list, click the associated author names.

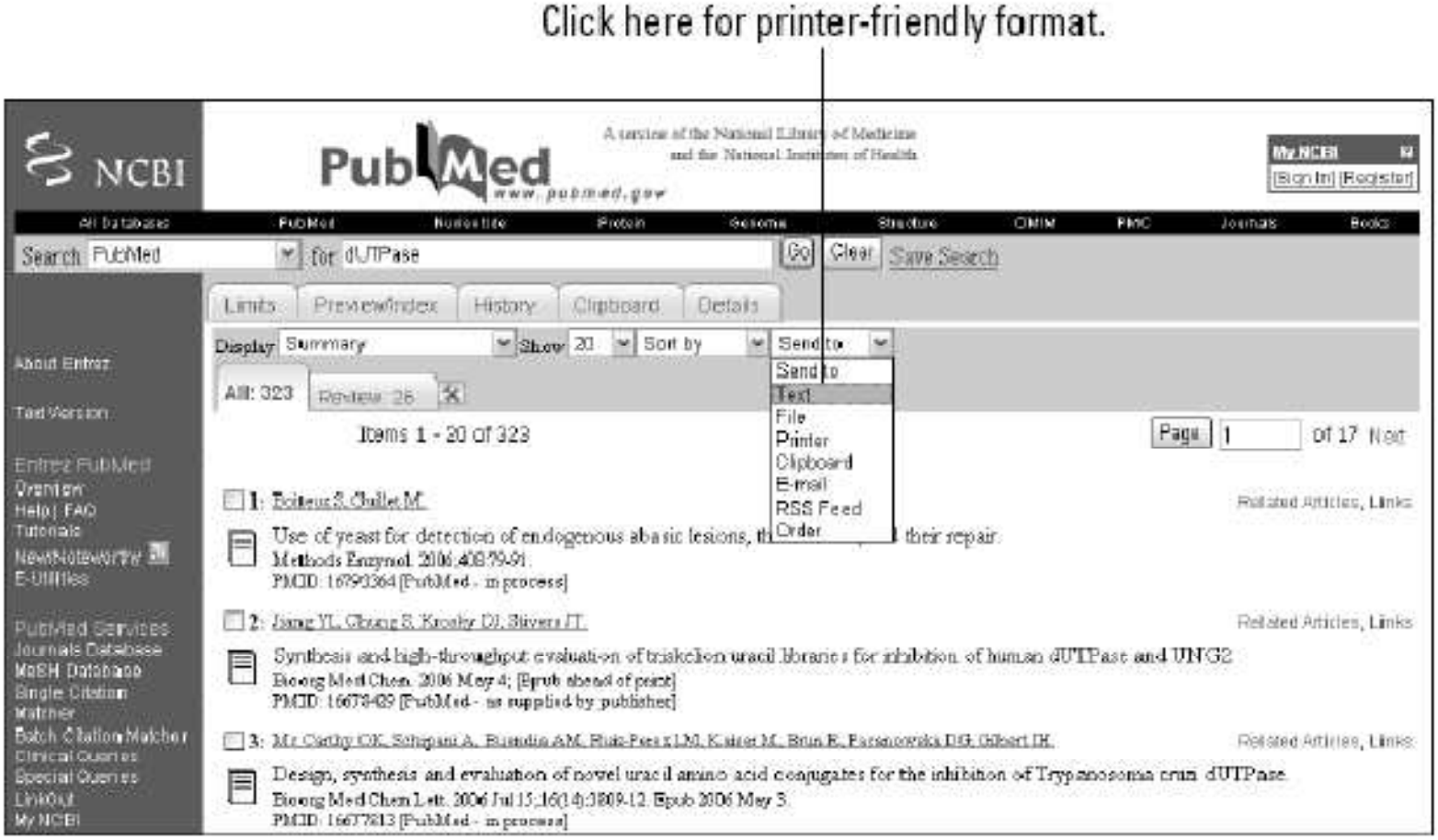
- Titles stand out because they appear in a blue font, underlined —two signs of a clickable hyperlink.

36.
Active site of mycobacterial dUTPase: structural characteristics and a built-in sensor.
Varga B, Barabás O, Takács E, Nagy N, Nagy P, Vértessy BG.
Biochem Biophys Res Commun. 2008 Aug 15;373(1):8-13. Epub 2008 Jun 2.PMID: 18519027 [PubMed - indexed for MEDLINE]**Related articles**
37.
Sequence analysis of a non-classified, non-occluded DNA virus that causes salivary gland hypertrophy of Musca domestica, MdSGHV.
Garcia-Maruniak A, Maruniak JE, Farmerie W, Boucias DG.
Virology. 2008 Jul 20;377(1):184-96. Epub 2008 May 21.PMID: 18495197 [PubMed - indexed for MEDLINE]**Related articlesFree article**
38.
ORF018R, a highly abundant virion protein from Singapore grouper iridovirus, is involved in serine/threonine phosphorylation and virion assembly.
Wang F, Bi X, Chen LM, Hew CL.
J Gen Virol. 2008 May;89(Pt 5):1169-78.PMID: 18420794 [PubMed - indexed for MEDLINE]**Related articlesFree article**
39.
Vaccinia virus lacking the deoxyuridine triphosphatase gene (F2L) replicates well in vitro and in vivo, but is hypersensitive to the antiviral drug (N)-methanocarbathymidine.
Prichard MN, Kern ER, Quenelle DC, Keith KA, Moyer RW, Turner PC.
Virol J. 2008 Mar 5;5:39.PMID: 18321387 [PubMed - indexed for MEDLINE]**Related articlesFree article**
40.
Mechanism of dTTP inhibition of the bifunctional dCTP deaminase:dUTPase encoded by Mycobacterium tuberculosis.
Helt SS, Thymark M, Harris P, Aagaard C, Dietrich J, Larsen S, Willemoes M.
J Mol Biol. 2008 Feb 15;376(2):554-69. Epub 2007 Dec 5.PMID: 18164314 [PubMed - indexed for MEDLINE]**Related articles**

4- Clicking the link calls up a page containing a rather detailed summary of the paper we chose.

- Kinetic mechanism of human dUTPase, an essential nucleotide pyrophosphatase enzyme. Tóth J, Varga B, Kovács M, Málnási-Csizmadia A, Vértessy BG. J Biol Chem. 2007 Nov 16;282(46):33572-82. Epub 2007 Sep 11.PMID: 17848562 [PubMed - indexed for MEDLINE]**Related articlesFree article**

4. Save what you like to your hard drive by choosing your browser's File➪Save As option.



You can select many references and get their abstracts together.

e.g. The abstract of the paper we selected.

- J Biol Chem. 2007 Nov 16;282(46):33572-82. Epub 2007 Sep 11.
- **Kinetic mechanism of human dUTPase, an essential nucleotide pyrophosphatase enzyme.**
- Tóth J, Varga B, Kovács M, Málnási-Csizmadia A, Vértessy BG.
- Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Karolina út 29, 1113 Budapest, Hungary. tothj@enzim.hu
- Human dUTPase is essential in controlling relative cellular levels of dTTP/dUTP, both of which can be incorporated into DNA. The nuclear isoform of the enzyme has been proposed as a promising novel target for anticancer chemotherapeutic strategies. The recently determined three-dimensional structure of this protein in complex with an isosteric substrate analogue allowed in-depth structural characterization of the active site. However, fundamental steps of the dUTPase enzymatic cycle have not yet been revealed. This knowledge is indispensable for a functional understanding of the molecular mechanism and can also contribute to the design of potential antagonists. Here we present detailed pre-steady-state and steady-state kinetic investigations using a single tryptophan fluorophore engineered into the active site of human dUTPase. This sensor allowed distinction of the apoenzyme, enzyme-substrate, and enzyme-product complexes. We show that the dUTP hydrolysis cycle consists of at least four distinct enzymatic steps: (i) fast substrate binding, (ii) isomerization of the enzyme-substrate complex into the catalytically competent conformation, (iii) a hydrolysis (chemical) step, and (iv) rapid, nonordered release of the products. Independent quenched-flow experiments indicate that the chemical step is the rate-limiting step of the enzymatic cycle. To follow the reaction in the quenched-flow, we devised a novel method to synthesize gamma-(32)P-labeled dUTP. We also determined by indicator-based rapid kinetic assays that proton release is concomitant with the rate-limiting hydrolysis step. Our results led to a quantitative kinetic model of the human dUTPase catalytic cycle and to direct assessment of relative flexibilities of the C-terminal arm, critical for enzyme activity, in the enzyme-ligand complexes along the reaction pathway.
- PMID: 17848562 [PubMed - indexed for MEDLINE]
- Publication Types, MeSH Terms, Substances, Grant Support
- **Publication Types:**

- When the search yields many references, the best move is to start scanning a few pages — and select the most promising papers for future use by checking; the corresponding boxes will be different.

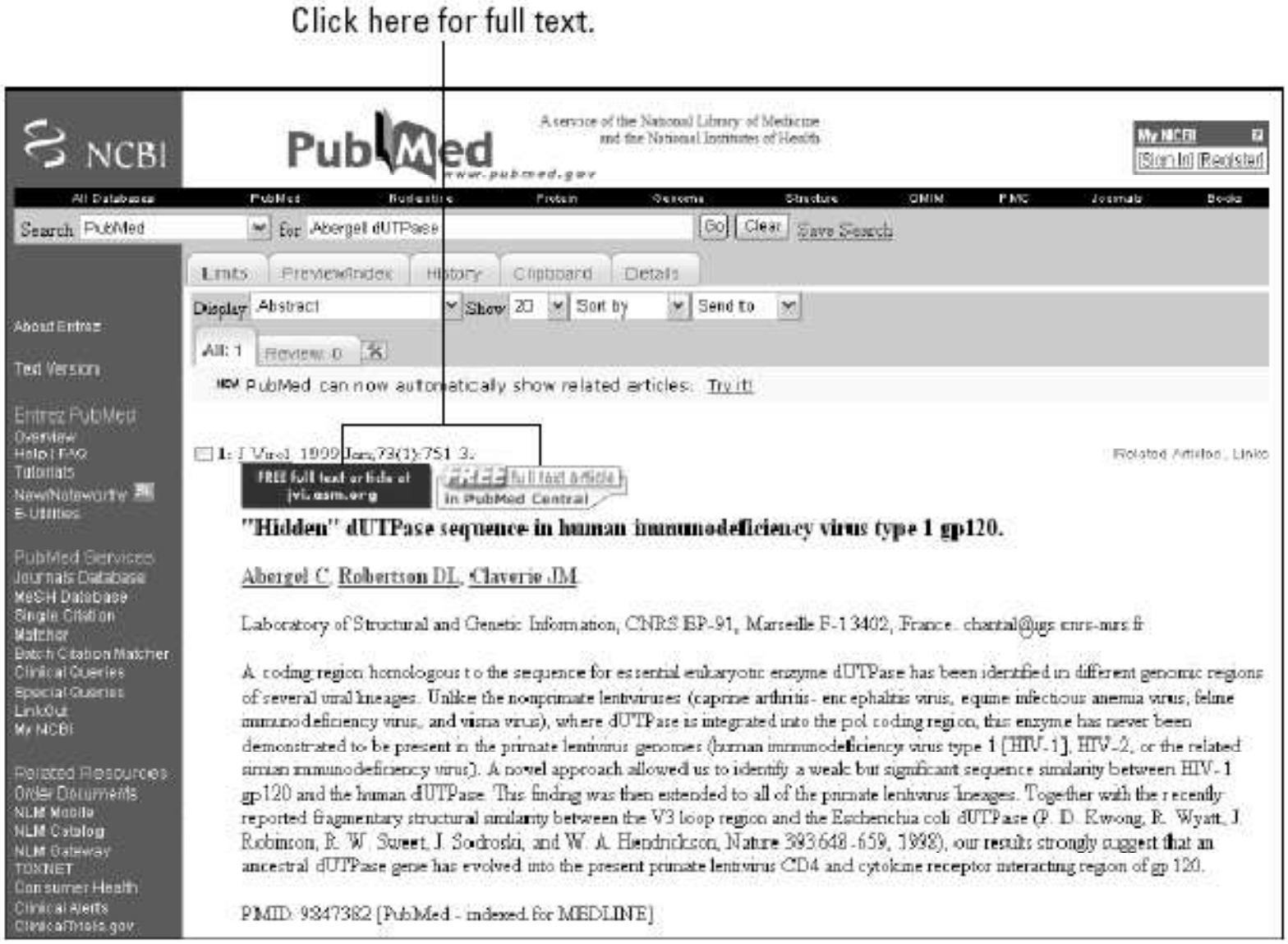
When you think you have selected enough references, just do the following:

1. **Choose Abstract from the Display drop-down menu.**
   The abstracts of all the papers you selected in the various pages appear for you to read.
2. **If you want to print this display as is, choose File⇨Print from your browser menu.** Click here for printer-friendly format.
3. **If you want to print this display in a more printer-friendly format, first: select Text from the scroll-down menu on the far right of the menu bar; to display this information in a non-HTML format, and then choose File⇨Print from your browser menu.**
4. **If you'd like to save the file in the format of your choice, choose File⇨Save As from your browser menu, enter a new filename for the file, and then choose the file format you want to save to.**

Searching PubMed using author's names

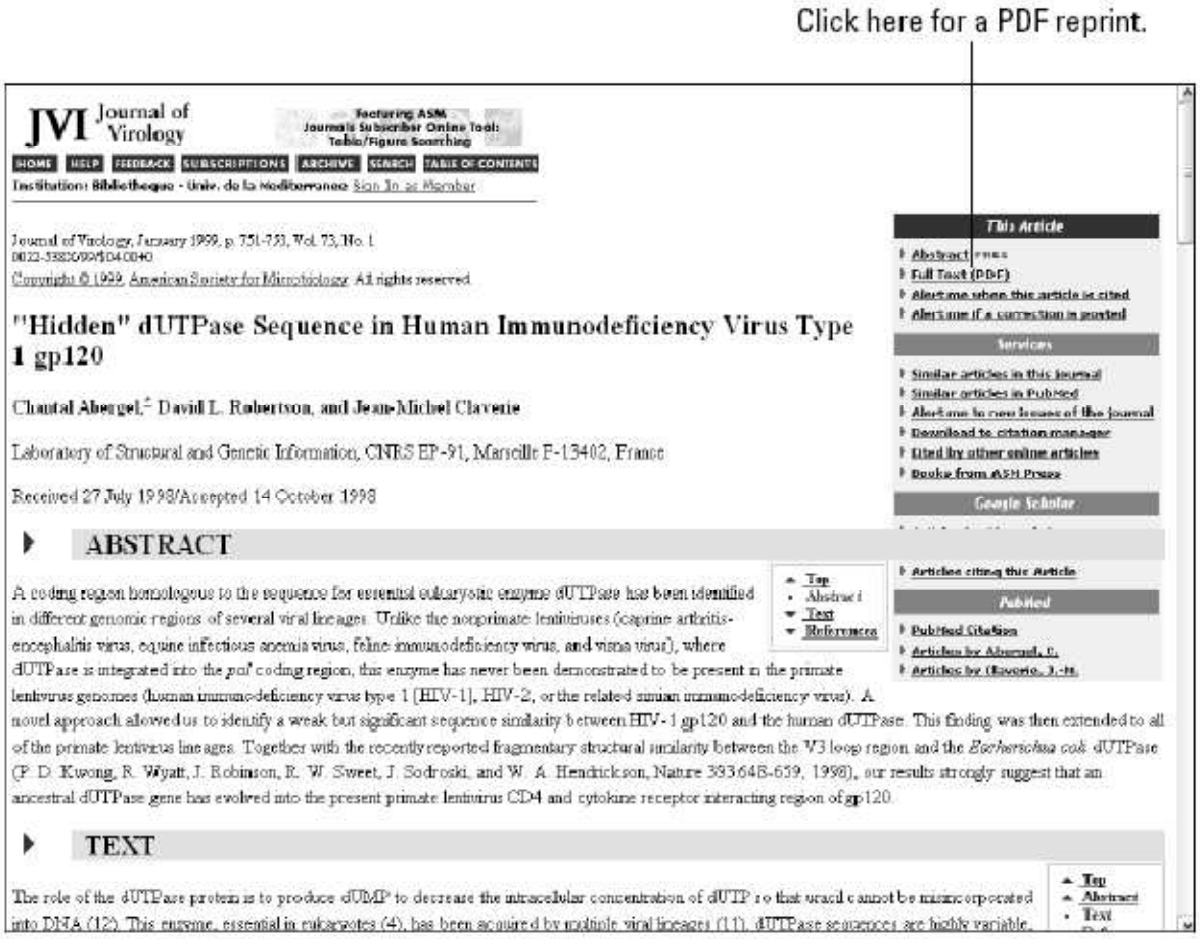
1. **Point your browser to** www.ncbi.nlm.nih.gov/entrez/**.**
2. **Type** Abergel **— the name of a prospective author — in the For window, and then click the Go button.**
3. **Results appear as follows:**

The result of a standard combined search for Abergel and dUTPase.



Some journals that offers free access to the full text (as indicated by the large colored rectangles that appear above the title).

4. Click the blue rectangle on the left.



Searching PubMed using fields

- PubMed offers many more ways to narrow down bibliographic searches to more specific topics. To get a handle on these different ways, however, we need to know more about the internal structure of Medline entries.

1. **Click the small arrow to the right of the Display drop-down menu.**

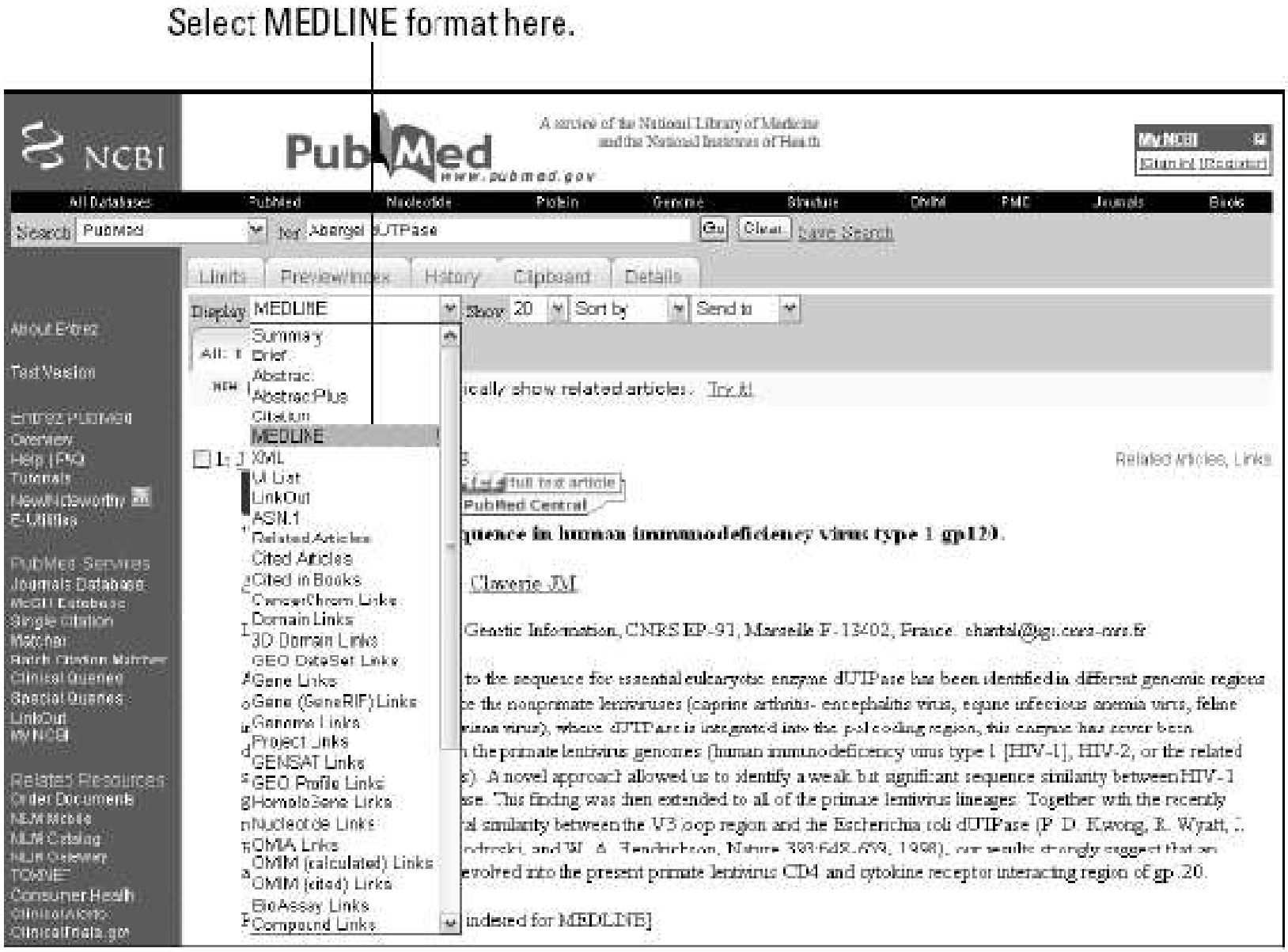
   The contents of the drop-down menu appear: These options are different ways to display information related to the current article, or are links to related information.

2. **Select the MEDLINE option** .

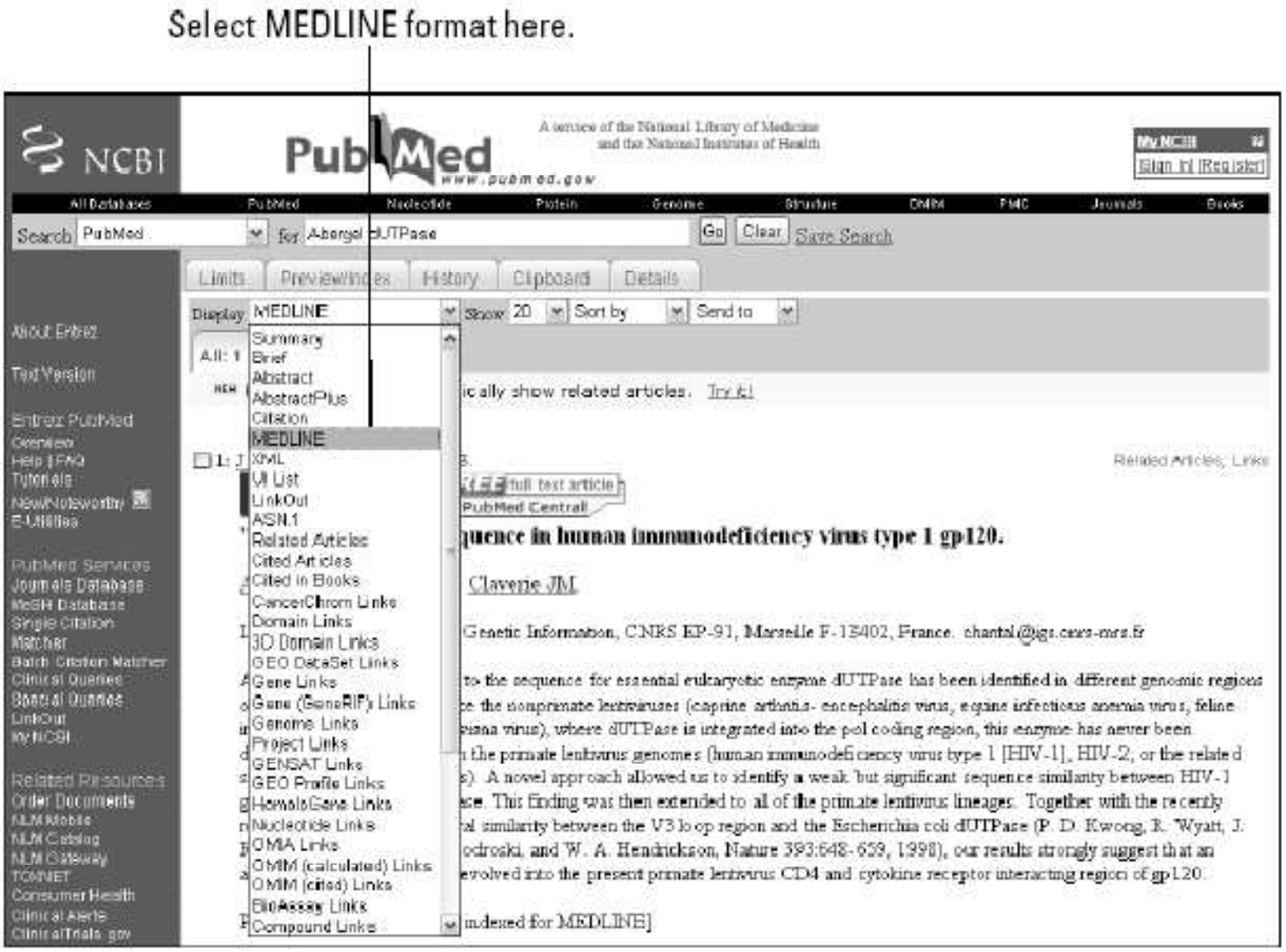
   The Medline page appears, as shown in Figure

Frequently one can get flooded with an overwhelming number of *hits* (the standard term for search results) if we formulate queries that contain common names (such as *Smith* or *Cohen*) or use search terms (such as *Down*) that can occur in different contexts. To alleviate this problem, one can query PubMed by restricting the search for each word of the query to a given field.

Changing the default abstract format to MEDLINE.



Selecting a field



- In Figure below, the internal structure of a Medline database record is shown. The information is spread out over separate sections, called *fields,* each one preceded by a specific abbreviation — TI for title fields, AB for abstracts, AD for the laboratory address, AU for the authors, SO for the journal abbreviation, and so forth. This structure applies to all Medline records.

The internal structure of a Medline record.

PMID - 9847382
OWN - NLM
STAT - MEDLINE
DA - 19990128
DCOM - 19990128
LR - 20041117
PUBM - Print
IS - 0022-538X (Print)
VI - 73
IP - 1
DP - 1999 Jan
TI - "Hidden" dUTPase sequence in human immunodeficiency virus type 1 gp120.
PG - 751-3
AB - A coding region homologous to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions of several viral lineages. ----------, our results strongly suggest that an ancestral dUTPase gene has evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

Conti…

The internal structure of a Medline record (conti...)

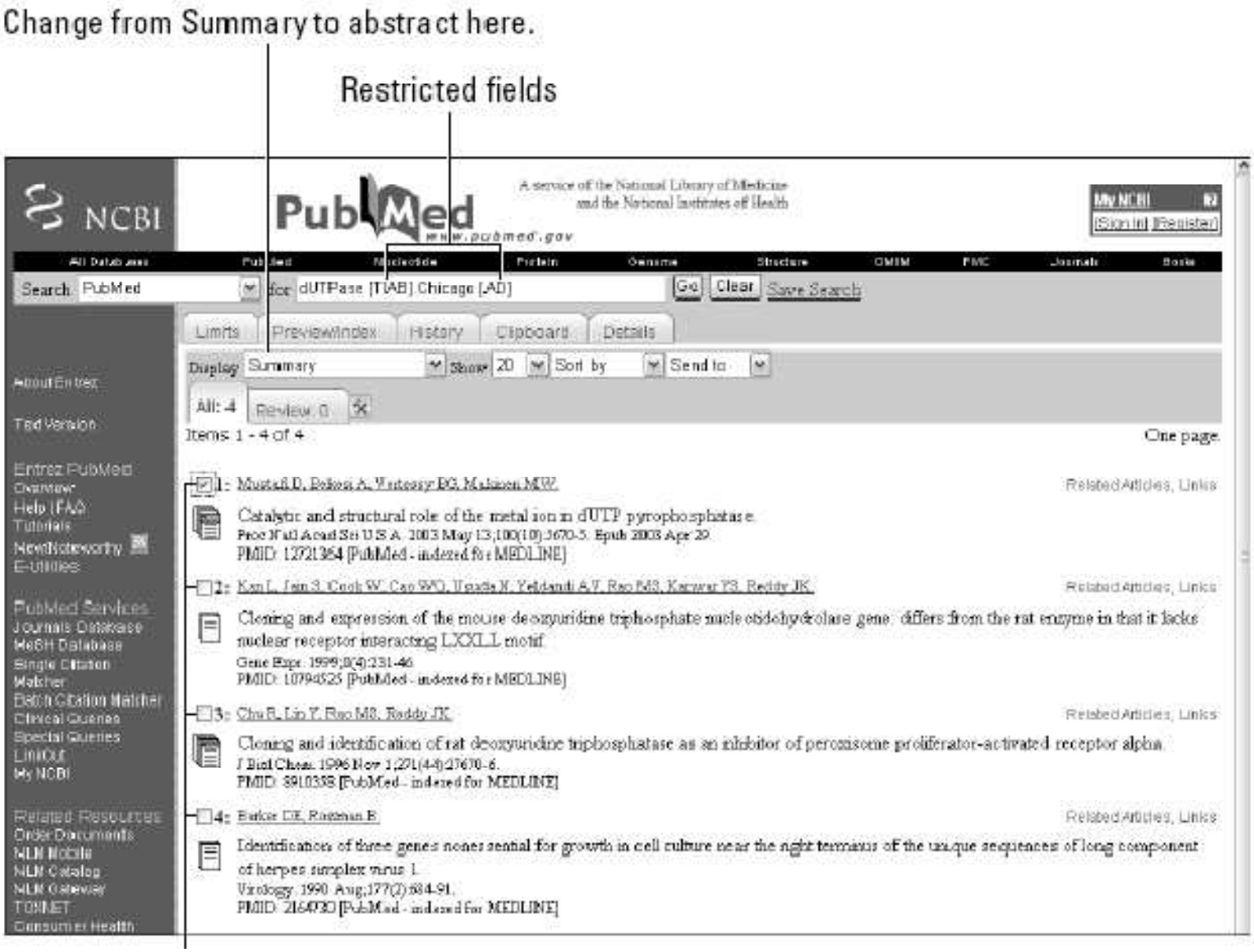
AD	- Laboratory of Structural and Genetic Information, CNRS EP-91, Mars F-12302, France. chantal@igs.cnrs-mrs.fr
FAU	- Abergel, C
AU	- Abergel C
FAU	- Robertson, D L
AU	- Robertson DL
FAU	- Claverie, J M
AU	- Claverie JM
LA	- eng
PT	- Journal Article
PL	- UNITED STATES
TA	- J Virol
JT	- Journal of virology.
JID	- 0113724
RN	- 0 (HIV Envelope Protein gp120)
RN	- EC 3.6.1. - (Pyrophosphatases)
RN	- EC 3.6.1.23 (dUTP pyrophosphatase)
SB	- IM
SB	- X
MH	- Amino Acid Sequence
MH	- HIV Envelope Protein gp120/*chemistry
MH	- HIV-1/* chemistry
MH	- Humans
MH	- Molecular Sequence Data
MH	- Pyrophosphatases/*chemistry/genetics
MH	- Research Support, Non-U.S. Gov't
EDAT	- 1998/12/16
MHDA	- 1998/12/16 00:01
PST	- ppublish
SO	- J Virol. 1999 Jan;73(1):751-3.

- One can follow each term with the code (in brackets) that identifies the field in which to find the search term. Changing the field can totally modify the result of the searches. For instance, by entering three different queries — Down [AU], Down [TI], and Down [AD] into the For text box at the PubMed site.
- For example by running these queries, 275, 16,318, and 1,213, respectively, totally unrelated references were obtained. This narrowed the search a bit.

Using fields to find experts.

1. **Point your browser to the PubMed site** (www.ncbi.nlm.nih.gov/entrez/).
2. **In the For window, enter** dUTPase [TIAB] Chicago [AD]**, and click Go.** By specifying the [TIAB] field, you'll be scanning the titles and abstracts of potential articles; the [AD] field specifies the address of the main laboratory associated with these articles. A couple of papers show up.
3. **Click the list of authors (in blue, underlined).** You get the abstract of the corresponding article, together with the main laboratory address. Alternatively, to get all the abstracts at once, follow Steps 4 and 5.
4. **Go back to the previous URL (after Step 2) using the Go Back button of your browser.**
5. **In the Display drop-down menu, change the display option from *Summary* to *Abstract.*** A list of abstracts appears. At the top of each abstract, the name of the relevant laboratory in Chicago is displaced. With the information contained in the Medline records, names, street addresses, and sometimes e-mail addresses can be obtained. If necessary, use a telephone book (or a Web search engine) to supplement this information and find out how to contact these experts.

Searching for experts on dUTPase in Chicago.

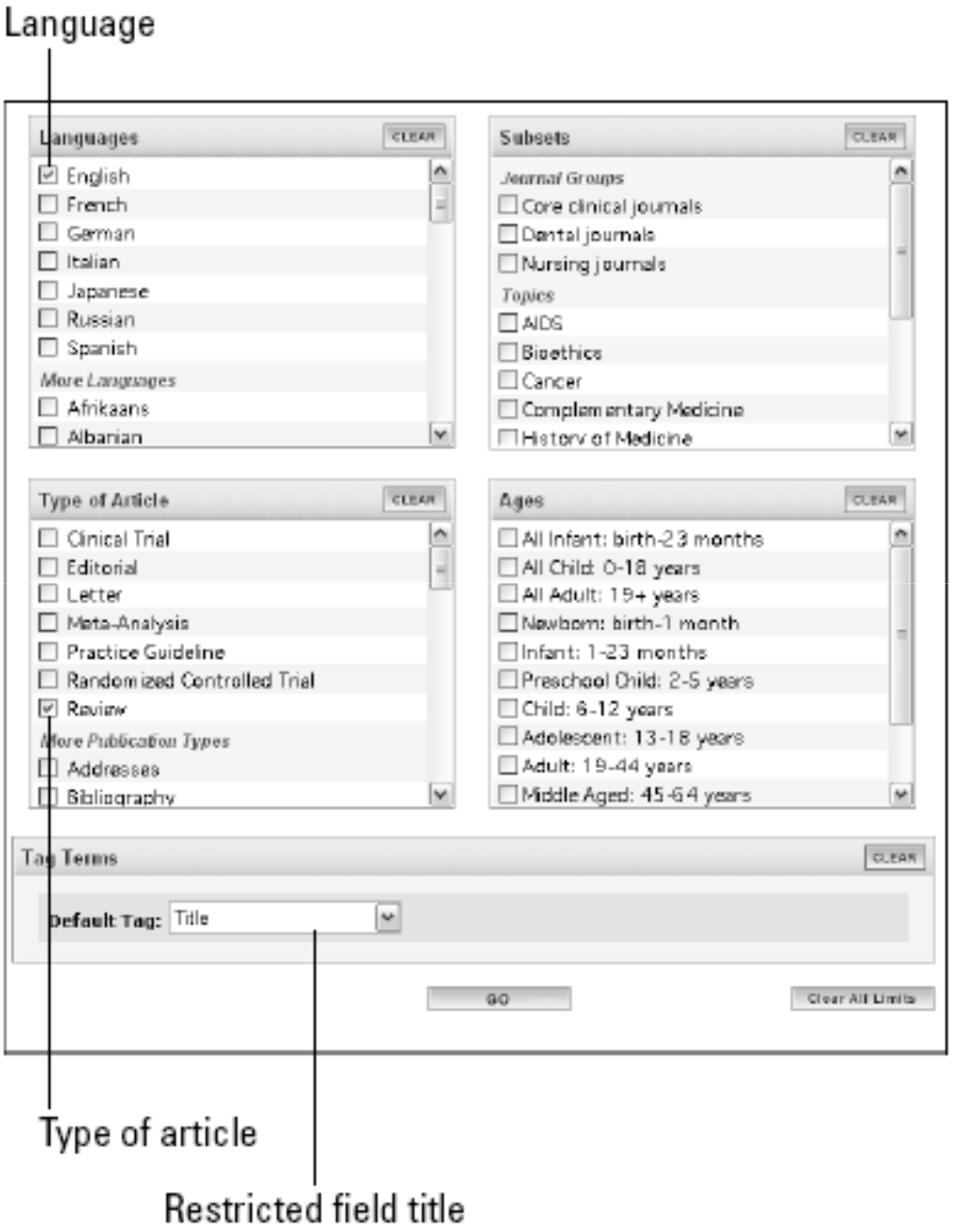


Searching PubMed using limits

- PubMed offers yet another way of fine-tuning your queries: one can pre-define ranges for different attributes in different fields before running the search.

1. Point your browser to the PubMed site at www.ncbi.nlm.nih.gov/entrez/
2. Type dUTPase **in the For text box.** If you clicked the Go button at this point, you'd get a list of more than 320 articles — a bit too much reading for a nonspecialist! Restricting this list to the most recent review articles written in English would be practical.
3. Click the Limits tab, located just beneath the little arrow for the pulldown menu of the Search window. The Limits screen appears, as shown in Figure. Here, below the Limited To line, you now have plenty of fields and attributes you can use for setting limits. You can go back later to this page and explore these various options.
4. Check the English box in the Language section.
5. Check the Review box in the Type of Article section.
6. Choose Title from the Default Tag drop-down menu.
Choosing Title restricts the search to articles for which the topic of dUTPase is central.
7. Finally, click the Go button.
Your (new and more concise) Results page appears.

Limiting a search for dUTPase to the titles of review articles in English.



A few more tips about PubMed

How to get the most out of the query:

- Quoted queries (for example, “down syndrome”) behave as a single word, and are a great way to improve the relevance of your search.
- Impress your colleagues by starting using logical connectors (AND, OR, NOT) in your queries, as in dUTPase[TI] OR pyrophosphatase[TI] NOT Smith[AU]
- Adding initials to proper names (for example, “Abergel C”) can greatly reduce the number of hits.
- Write down the PubMed Identifier (the number in the PMID field) of that interesting paper you just found. It can be very useful in any subsequent searches for related items, such as associated gene and protein sequences.
- etc

Retrieving Protein Sequences

- **Using other sites it is possible to retrieve relevant protein sequences from the Web to find out more about the subject at the molecular level.**
- **Using: ExPASy: A prime Internet site for protein information.**
- **The ExPASy server, is managed by Prof. Amos Bairoch, and is a world-leading resource for protein information.**

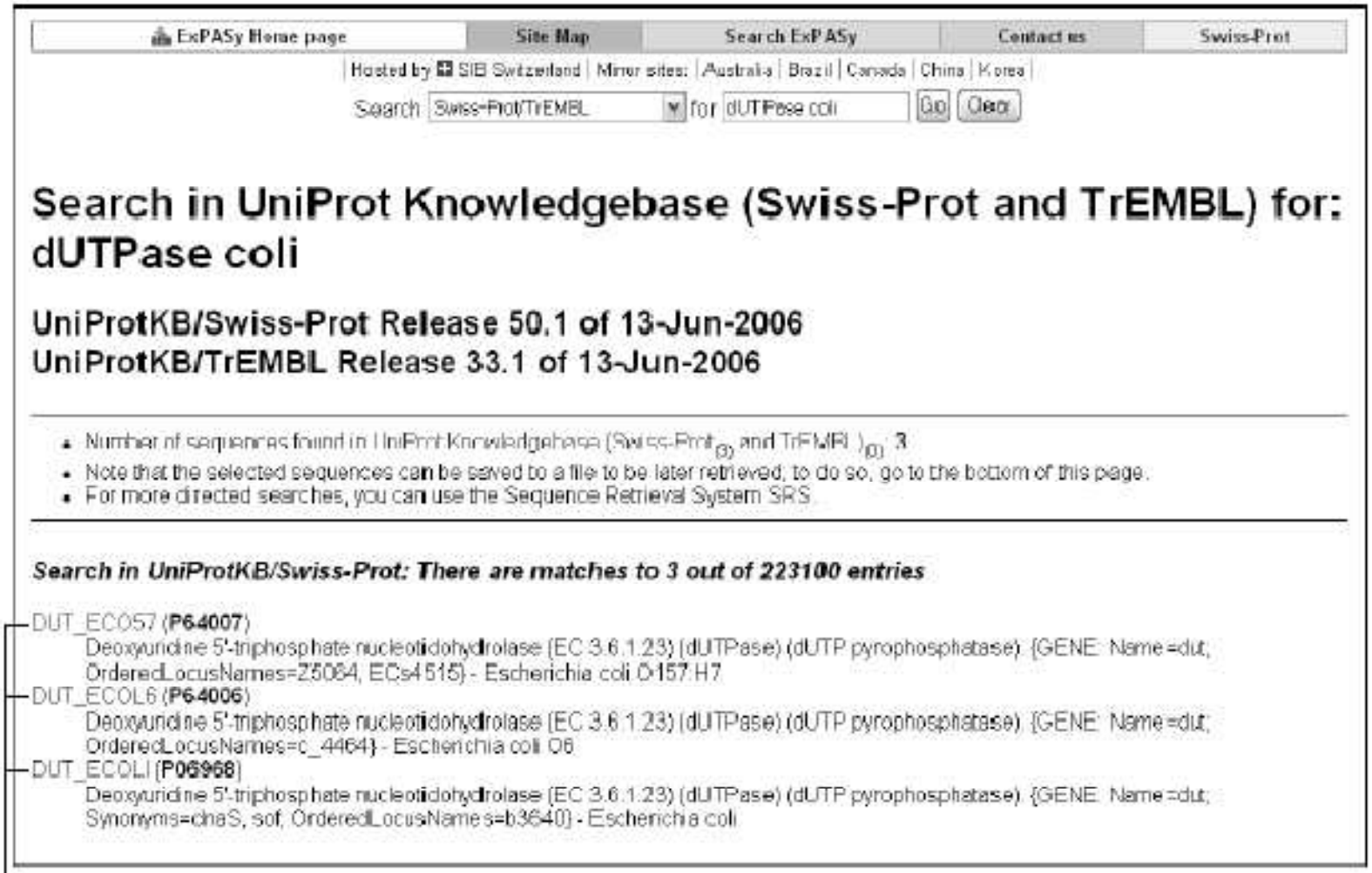
Using ExPASy

1. **Point your browser to** www.expasy.org/sprot/**, the Swiss-Prot database home page.**
2. **Type dUTPase coli in the Search window, and then click the Search button.**

- A list of three relevant protein sequences appears.
- Now, when you click the last DUT_ECOLI (P06968) link, a full page of information about this dUTPase protein *of E. coli* appears on-screen, as shown in Figures 2-12 and 2-13. (No single browser window can hold such a wealth of information, so we've broken up the Results page into two figures.)

The Search window at the top of the Swiss-Prot home page.

Three dUTPase sequences from the Escherichia coli bacterium.



3 relevant dUTPase entries found.