

...KLTDSQNFDEYMKALGVDFATRQVGLNLYLVSQEGGKV...

Protein Sequence

Computational methods



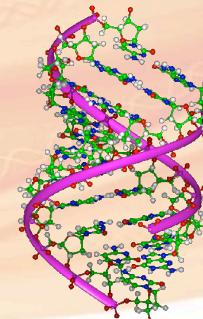
Protein Structure Model

Bioinformatics Computational Methods 1 - BIOL 6308



October 29th 2013

<http://155.33.203.128/cleslin/home/teaching6308F2013.php>

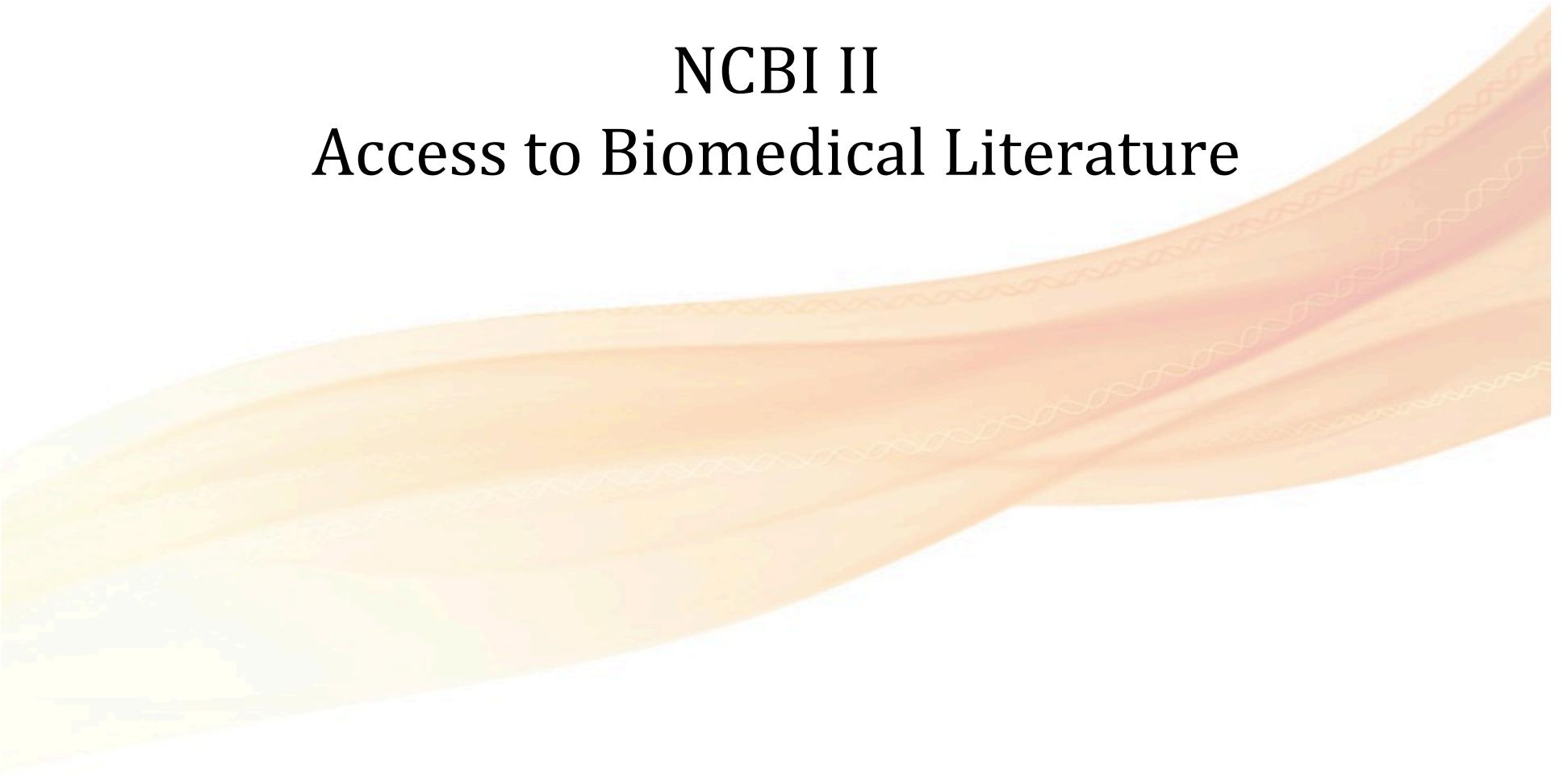


Last Time

- What is an Accession Number
- Examples of RefSeq Accession #'s
- Databases
 - Gene
 - GenPepts
 - Unigene
 - Structure
 - MMDB
 - Vast
 - CDD
- What are SNPs
 - Formal Definition
 - Linked and Causative SNPs
 - Functional Categories
 - Nomenclature
 - dbSNP
- Homework to "discover" the The Discovery Process

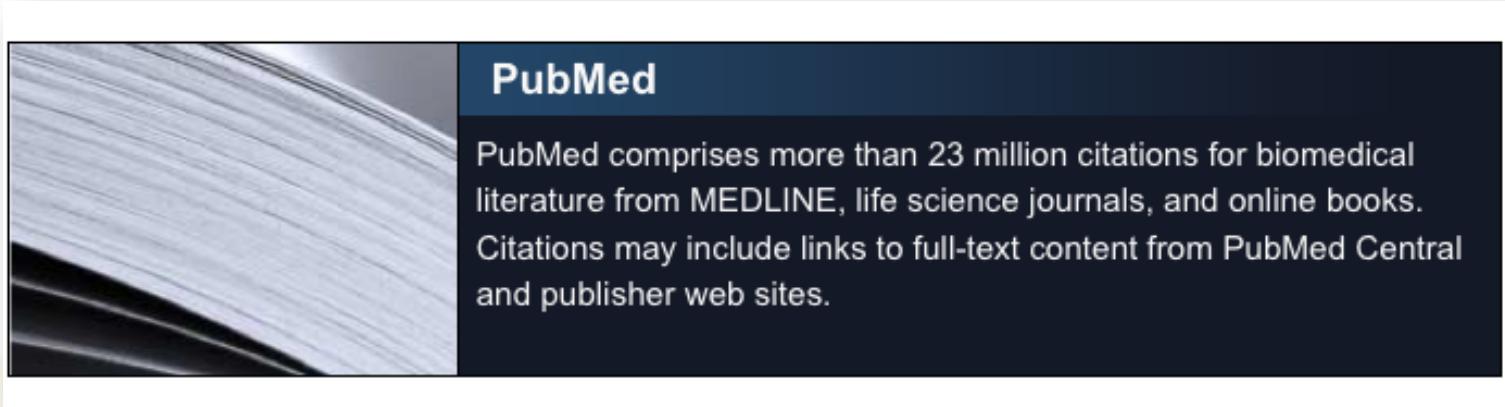
NCBI II

Access to Biomedical Literature



PubMed

- PubMed is the NCBI gateway to MEDLINE
- +23 million citations



- As a bioinformatics **scientist**, you will have to use PubMed
- Watch this video
 - <https://webmeeting.nih.gov/p71971540/?launcher=false&fcsContent=true&pbMode=normal>

Medical Subject Headings

- MeSH : "Medical Subject Headings"
 - National Library of Medicine's (NLM) controlled vocabulary thesaurus for indexing articles for PubMed
 - List of the vocabulary terms used for subject analysis of biomedical literature at NLM
 - Used for indexing journal articles for MEDLINE
 - Imposes uniformity and consistency to the indexing of biomedical literature

The MeSH Tree Structure

Organized by 19 main branches:

- Anatomy
- Organisms
- Diseases
- Chemical and Drugs Analytical, Diagnostic and Therapeutic Techniques and Equipment
- Psychiatry and Psychology
- Biological Sciences Natural Sciences
- Anthropology, Education, Sociology and Social Phenomena
- Technology, Industry, Agriculture
- Humanities
- Information Science
- Named Groups
- Health Care
- Publication Characteristics
- Geographic Location
- **And a few more**

<http://www.ncbi.nlm.nih.gov/mesh/1000048>

You do not have to know these!

MeSH is an Entrez database

- <http://www.ncbi.nlm.nih.gov/mesh>
- MeSH is provided to assist PubMed users locate appropriate terms for MEDLINE searches
- This database provides information about MeSH terms including:
 - Definitions
 - Synonyms for the concept
 - Related terms
 - The position of the term in the MeSH hierarchy

MeSH Example

MeSH MeSH Help

Save search Limits Advanced

Display Settings: Summary Send to:

Results: 5

- Cell Count**
 - 1. The number of CELLS of a specific kind, usually measured per unit volume or area of sample.
Year introduced: 1973(1969)
- Blood Cell Count**
 - 2. The number of LEUKOCYTES and ERYTHROCYTES per unit volume in a sample of venous BLOOD. A complete blood count (CBC) also includes measurement of the HEMOGLOBIN; HEMATOCRIT; and ERYTHROCYTE INDICES.
- CD4 Lymphocyte Count**
 - 3. The number of CD4-POSITIVE T-LYMPHOCYTES per unit volume of BLOOD. Determination requires the use of a fluorescence-activated flow cytometer.
Year introduced: 1995
- Leukocyte Count**
 - 4. The number of WHITE BLOOD CELLS per unit volume in venous BLOOD. A differential leukocyte count measures the relative numbers of the different types of white cells.
- Erythrocyte Count**
 - 5. The number of RED BLOOD CELLS per unit volume in a sample of venous BLOOD.

PubMed search builder

AND

Find related data

Database:

Search details

"cell count"[MeSH Terms]
OR cell count[Text Word]

See more...

Cell Count - MeSH

Cell Count

The number of CELLS of a specific kind, usually measured per unit volume or area of sample.

Year introduced: 1973(1969)

PubMed search builder options

Subheadings:

1

- analysis
- classification
- drug effects
- economics
- ethics
- history
- instrumentation
- methods
- organization and administration
- pharmacology
- physiology
- physiopathology
- radiation effects
- standards
- statistics and numerical data
- supply and distribution
- therapeutic use
- trends
- utilization
- veterinary

Restrict to MeSH Major Topic.

Do not include MeSH terms found below this term in the MeSH hierarchy.

Tree Number(s): E01.370.225.500.195, E05.200.500.195, E05.242.195, G04.170

Entry Terms:

- Cell Counts
- Count, Cell
- Counts, Cell
- Cell Number
- Cell Numbers
- Number, Cell
- Numbers, Cell
- Cell Density
- Cell Densities
- Densities, Cell
- Density, Cell

MeSH term,
definition, and year

Select Subheadings

Major MeSH terms &
Explosion of MeSH terms

“Synonyms” for this term.

2

PubMed search builder

"Cell Count/classification"
[Mesh]

Add to search builder AND Search PubMed

Cell Count – MeSH - Subheadings

- Select specific *subheadings* to describe a particular aspect of a subject (diagnosis, prognosis, treatment etc.)
 - Select as many subheadings as you like to focus your search results
 - The more subheadings you select the broader your search results will be

Darlene Chapman

Cell Count – MeSH - Major MeSH Terms

- Below the subheading list you will see a box where you can “Restrict Search To MeSH Major Topic”
- This will narrow your search results so that the MeSH term you search is the major focus of references retrieved

Darlene Chapman

Cell Count – MeSH - Explosion of MeSH Terms

- Also see a hierarchical arrangement of terms related to the MeSH term you selected
- Display shows how your term relates to other MeSH terms
- PubMed **automatically explodes MeSH terms** to include all narrower terms in the hierarchical list
- Check box beside “*Do Not Include MeSH Term...*”
 - Do not want to include the more specific subject headings indexed below your term in the list

Darlene Chapman

Cell Count - MeSH

Previous Indexing:

- [Cytology \(1966-1968\)](#)

See Also:

- [Blood Cell Count](#)
- [Sperm Count](#)

[All MeSH Categories](#)

[Analytical, Diagnostic and Therapeutic Techniques and Equipment Category](#)

[Diagnosis](#)

[Diagnostic Techniques and Procedures](#)

[Clinical Laboratory Techniques](#)

[Cytological Techniques](#)

Cell Count

[Blood Cell Count](#)

[Erythrocyte Count +](#)

[Leukocyte Count +](#)

[Platelet Count](#)

[Sperm Count](#)

[All MeSH Categories](#)

[Analytical, Diagnostic and Therapeutic Techniques and Equipment Category](#)

[Investigative Techniques](#)

[Clinical Laboratory Techniques](#)

[Cytological Techniques](#)

Cell Count

[Blood Cell Count](#)

[Erythrocyte Count +](#)

[Leukocyte Count +](#)

[Platelet Count](#)

[Sperm Count](#)

Before 1969 ...

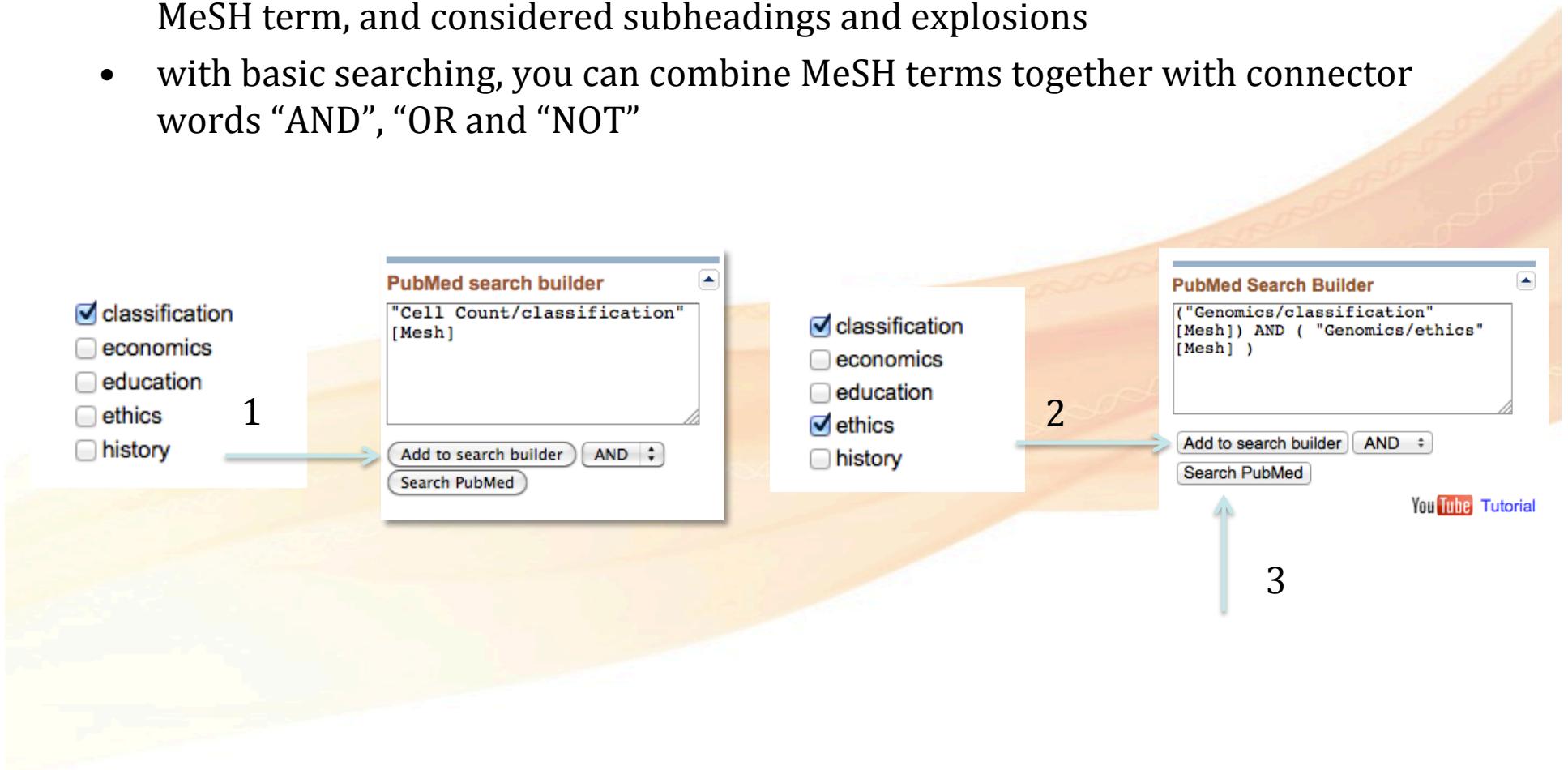
Related Terms of Possible Interest

Where its at in the tree (multiple branches)

You can choose other subject headings from this hierarchical list to broaden or narrow your search results

PubMed Search Builder

- After you've selected MeSH terms, decided if your term should be a Major MeSH term, and considered subheadings and explosions
- with basic searching, you can combine MeSH terms together with connector words "AND", "OR and "NOT"



Check Our MeSH Terms For Bioinformatics

- <http://www.ncbi.nlm.nih.gov/mesh/?term=bioinformatics>



PubMed Search

PubMed search builder

"Cell Count/classification"
[Mesh]

Add to search builder AND Search PubMed

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Article types
more ...

Languages
English
more ...

[Clear all](#)

[Show additional filters](#)

How To | Help | cleslin1 | My NCBI | Sign Out | Search

RSS | Save search | Advanced

Display Settings: Summary, 20 per page, Sorted by Recently Added | Send to: | Filters: Manage Filters

Results: 13

[Asthma phenotypes: consistency of classification using induced sputum.](#)
1. Hancox RJ, Cowan DC, Aldridge RE, Cowan JO, Palmary R, Williamson A, Town GI, Taylor DR.
Respirology. 2012 Apr;17(3):461-6. doi: 10.1111/j.1440-1843.2011.02113.x.
PMID: 22142406 [PubMed - indexed for MEDLINE]
[Related citations](#)

[Reactive peripheral blood plasmacytosis in a patient with acute hepatitis A.](#)
2. Wada T, Maeba H, Ikawa Y, Hashida Y, Okumura A, Shibata F, Tone Y, Inoue M, Koizumi S, Takatori H, Sakai Y, Kaneko S, Yachie A.
Int J Hematol. 2007 Apr;85(3):191-4.
PMID: 17483053 [PubMed - indexed for MEDLINE]
[Related citations](#)

[Contribution of monocytes to viral replication in macaques during acute infection with simian immunodeficiency virus.](#)
3. Kuwata T, Kodama M, Sato A, Suzuki H, Miyazaki Y, Miura T, Hayami M.
AIDS Res Hum Retroviruses. 2007 Mar;23(3):372-80.
PMID: 17411370 [PubMed - indexed for MEDLINE]
[Related citations](#)

[Diagnostic utility of pleural fluid eosinophilia.](#)
4. Nasilowski J, Krenke R, Przybylowski T, Abouchaz B, Dmowska-Sobstyl B, Droszcz W, Chazan R.
Pneumonol Alergol Pol. 2006;74(1):10-5. Polish.
PMID: 17175969 [PubMed - indexed for MEDLINE]
[Related citations](#)

[Lupus erythematosus or human immunodeficiency virus infection? Diagnostic problems in rheumatologic practice--two case reports.](#)
5. [See more...](#)

Titles with your search terms
MBL or CLL: which classification best categorizes the clinical [Leuk Res. 2008]
[How to interpret the red blood cell count--classification of [Infirm Fr. 1973]]
[See more...](#)

2 free full-text articles in PubMed Central
Impairments, activity limitations and participation [Health Qual Life Outcomes. 2004]
Association of medical insurance and other factors [Am J Public Health. 2002]
[See all \(2\)...](#)

Find related data
Database: Select | Find items

Search details
"Cell Count/classification"
[Mesh]

Search | [See more...](#)

NLM MeSH Browser

[Display Settings:](#) Full

[Send to:](#)

Cell Count

The number of CELLS of a specific kind, usually measured per unit volume or area of sample.

Year introduced: 1973(1969)

PubMed search builder options

[Subheadings:](#)

- classification
- drug effects
- economics
- history

- instrumentation
- methods
- radiation effects
- standards

- statistics and numerical data
- trends
- utilization
- veterinary

Restrict to MeSH Major Topic.

Do not include MeSH terms found below this term in the MeSH hierarchy.

Tree Number(s): E01.370.225.500.195, E05.200.500.195, E05.242.195, G04.170

PubMed search builder

"Cell
Count/classification"
[Mesh]

[Add to search builder](#)

AND

[Search PubMed](#)

Related information

[PubMed](#)

[PubMed - Major Topic](#)

[Clinical Queries](#)

[NLM MeSH Browser](#)

NLM MeSH Browser

- Is the tool used by MEDLINE indexers and catalogers
- Great to determine terms

National Library of Medicine - Medical Subject Headings

2013 MeSH

MeSH Descriptor Data

[Return to Entry Page](#)

Standard View. [Go to Concept View](#); [Go to Expanded Concept View](#)

MeSH Heading	Cell Count
Tree Number	E01.370.225.500.195
Tree Number	E05.200.500.195
Tree Number	E05.242.195
Tree Number	G04.170
Annotation	usually NIM; not for micro-organisms
Scope Note	The number of CELLS of a specific kind, usually measured per unit volume or area of sample.
Entry Term	Cell Density
Entry Term	Cell Number
See Also	Blood Cell Count
See Also	Sperm Count
Allowable Qualifiers	CL EC ES HI IS MT SN ST TD UT VE
Previous Indexing	Cytology (1966-1968)
Online Note	use CELL COUNT to search CELL NUMBER 1978-79
History Note	73(69); CELL NUMBER was heading 1978-79
Date of Entry	19990101
Unique ID	D002452

http://www.nlm.nih.gov/mesh/2013/mesh_browser/MBrowser.html

Indexing with MeSH Vocabulary

- Indexers examine articles
 - Assign the most specific MeSH heading(s) appropriate to describe the main concepts discussed
 - When there's no single specific MeSH heading for a concept
 - Indexer uses closest, more general MeSH heading available
 - Assign as many MeSH headings as appropriate to cover the topics of the article (generally 5 to 15)
- MeSH terms that reflect the major points of the article are marked with an asterisk (*) by indexers

Article Title:

Hormone therapy in perimenopausal and postmenopausal women: examining the evidence on cardiovascular disease risks.

Abstract:

Women may live for 30 years or longer after menopause with cardiovascular disease as their highest mortality risk. Menopause may correspond to health alterations for women, yet the use of estrogen during and after this transition has been controversial for the past four decades. The evidence from recent scientific studies does not support the use of hormone therapy for the prevention or treatment of cardiovascular disease, which has resulted in its removal from national guideline recommendations. However, because of concerns related to specific aspects of the research, there are gaps in the evidence. Studies are under way to evaluate alternate methods for hormone delivery, low-dose hormone therapy, and selective estrogen receptor modulators (SERMs) in reducing cardiovascular risks in perimenopausal and postmenopausal women. Implications for clinical nursing practice include education as well as assessment and counseling related to individual risk factors.

Publication Types:

- Review

Mesh Terms:

- Aged
- Cardiovascular Diseases/chemically induced*
- Estrogen Replacement Therapy/adverse effects*
- Evidence-Based Medicine
- Female
- Humans
- Middle Aged
- Perimenopause*
- Postmenopause*
- Risk Factors

Subheadings

- Subheadings further describe a particular aspect of a MeSH heading
- Examples: diagnosis, metabolism, adverse effects



Subheading Groupings

- Related subheadings have been grouped
- Not all subheadings have been placed in these groupings – some do not logically fit

Families of Subheading Explosions	
etiology	physiology
chemically induced	genetics
complications	growth & development
secondary	immunology
congenital	metabolism
embryology	biosynthesis
genetics	blood
immunology	cerebrospinal fluid
microbiology	deficiency
virology	enzymology
parasitology	pharmacokinetics
transmission	urine
	physiopathology

Pharmacologic Action Terms

- Every drug and chemical MeSH heading has been assigned one or more headings that describe known pharmacological actions [PA]
- Beginning in 1996
 - Indexers add the appropriate pharmacological action MeSH heading
 - As well as the specific chemical MeSH heading to a citation when the action of the chemical is being discussed in the article

Example: Pharmacologic Action Terms

The pharmacological actions established for the MeSH Heading, Aspirin:

Pharmacological Action	Anti-Inflammatory Agents, Non-Steroidal
Pharmacological Action	Cyclooxygenase Inhibitors
Pharmacological Action	Fibrinolytic Agents
Pharmacological Action	Platelet Aggregation Inhibitors

- A citation to an article that discusses **aspirin used as an anti-inflammatory agent** will be assigned:
 - Aspirin
Anti-Inflammatory Agents, Non-Steroidal
- A citation to an article that discusses **aspirin used to inhibit blood clotting** will be assigned:
 - Aspirin
Platelet Aggregation Inhibitors

How Was It Indexed

- If you've found the perfect reference in PubMed
 - May want to take a look at how this paper was indexed with MeSH terms to revise your search strategy to find other, similar (and relevant) references

Display Settings: Abstract Send to:

[Science](#). 2013 Aug 30;341(6149):958-9. doi: 10.1126/science.341.6149.958-b.

Call for prudence in whole-genome testing.

[van El CG](#), [Dondorp WJ](#), [de Wert GM](#), [Cornel MC](#).

Comment on
[Point-counterpoint. Patient autonomy and incidental findings in clinical genomics. \[Science. 2013\]](#)

PMID: 23990543 [PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms ▲

Publication Types
[Comment](#)
[Letter](#)

MeSH Terms
[Disease/genetics*](#)
[Genetic Predisposition to Disease*](#)
[Genomics/ethics*](#)
[Genomics/standards*](#)
[Humans](#)
[Incidental Findings*](#)
[Practice Guidelines as Topic*](#)

LinkOut - more resources ▼

PubMed - Developing a Search Strategy

- Before you can search for any information, you should first develop a search strategy
- **What is a Search Strategy?**
 - A search strategy is a plan that helps you look for the information you need
- **Search Strategy Tips**
 - Identify the key concepts
 - Determine alternative terms for these concepts, if needed
 - Use the MeSH browser for this
 - Find those terms
 - Understand Boolean Logic (coming slides)
 - Refine your search to dates, study groups, etc., as appropriate
 - Practice helps
 - Strategies and styles will differ according to personal choice and professional discipline

What Happens When Conduct Search from PubMed

- PubMed uses an Automatic Term Mapping feature to search for **unqualified terms**
- When you click Search, PubMed will look for a match in several lists
- It looks first for a match for your phrase as a
 - **Subject** in the
 - MeSH Translation Table. If it doesn't find a match, it looks for your phrase as a
 - **Journal** in the
 - Journals Translation Table, then it searches for
 - **Author and Investigator** names in the
 - Full Author Translation Table in the
 - Author Index, in the
 - Full Investigator Translation Table, and in the
 - Investigator Index

Search from PubMed

- As soon as PubMed finds a match:
 - The mapping stops
 - If a term matches in the MeSH Translation Table
 - PubMed does not continue looking in the next table
- If no match is found
 - PubMed breaks apart the phrase and repeats the process until a match is found

FYI - Automatic Term Mapping

- The process used by PubMed to find a match to unqualified terms that are entered into the query box
- Untagged terms are matched (in this order) against subjects using the MeSH (Medical Subject Headings) translation table, journals using the Journals translation table, and authors and investigators, using the Full Author translation table, Author index, Full Investigator translation table and Investigator index
- If a match is found in any translation table, the mapping stops
- When subject or journal matches are found, the query and individual terms are also searched in All Fields
- If no match is found in any tables, terms are searched in All Fields and AND'd together

FYI - Unqualified Terms

- Terms that are entered into the PubMed query box without a search field tag (e.g., [mh])



FYI - MeSH Translation Table

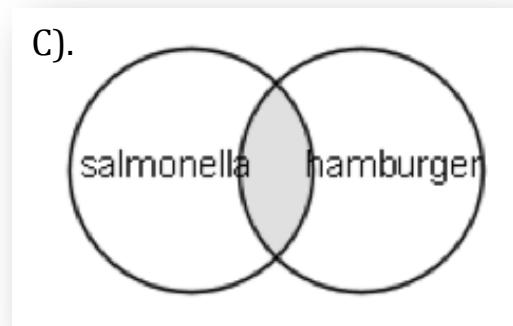
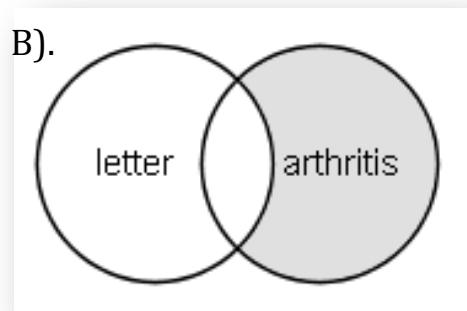
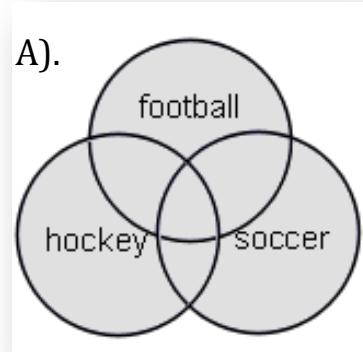
- The PubMed MeSH Translation Table is an alphabetical list of MeSH Terms, Subheadings, See-References for MeSH Terms, Unified Medical Language System mappings, and Names of Substances and synonyms to the Names of Substances

FYI - Journals Translation Table

- The PubMed Journals Translation Table is an alphabetical list of full journal titles, MEDLINE journal title abbreviations, and International Standard Serial Numbers (ISSNs)

FYI - Additional

- **Full Author Translation Table** - A list of full author names, in direct and inverse order, that appear in PubMed citations
- **Author Index** - In PubMed there are several indexes including an All Fields Index as well as indexes for specific search fields including Author Name, Title Word, and Text Word indexes among others
- **Full Investigator Translation Table** - A list of the full names of investigators or collaborators, in direct and inverse order, that appear in PubMed citations
- **Investigator Index** - The list of all values that appear in the investigator name field in PubMed citations



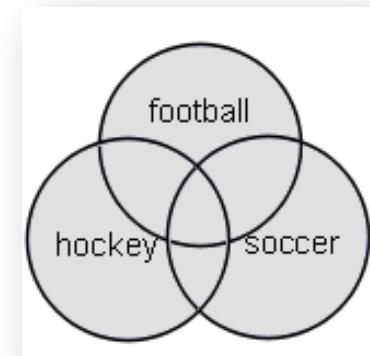
Introduction to Boolean Logic

Introduction to Boolean Logic

- In the context of database searching:
 - Boolean logic refers to the logical relationships among search terms
 - Operators AND, OR, NOT can be used to combine search terms in PubMed
 - In PubMed, Boolean operators **must be entered in uppercase letters**
- **OR:**
 - Used to retrieve a set in which each citation contains *at least one* of the search terms
 - Use OR when you want to pull together articles on similar topics.

Example OR:

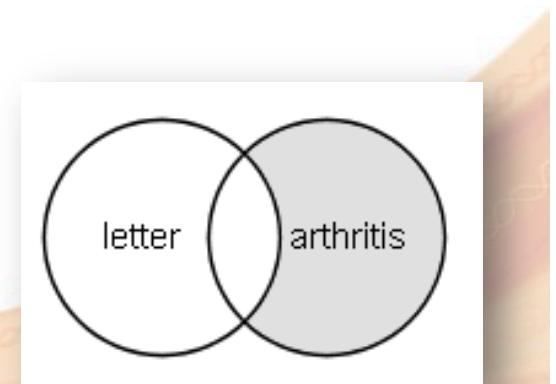
- Search "*football OR hockey OR soccer*"
- Each circle in the diagram to the right represents the retrieval for each term.
- The grey areas represent the retrieval for this example – all records that include any one of these terms



Search terms	Results
football	6325
hockey	2060
soccer	5498
football OR hockey OR soccer	12108

Example NOT:

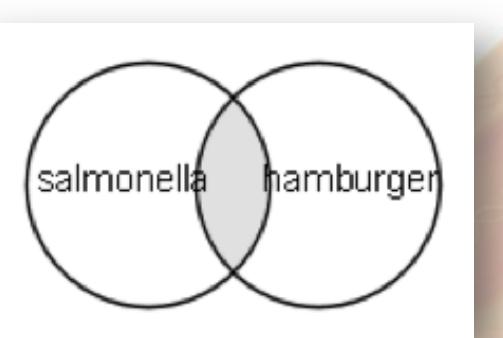
- **NOT:**
 - Retrieves a set from which citations to articles containing specified search terms following the NOT operator are eliminated
 - **Example:** *arthritis NOT letter*
 - the retrieval is a portion of the total retrieval for arthritis – that portion not including the term letter



Search terms	Results
arthritis	243053
letter	859692
arthritis NOT letter	230685

Example AND:

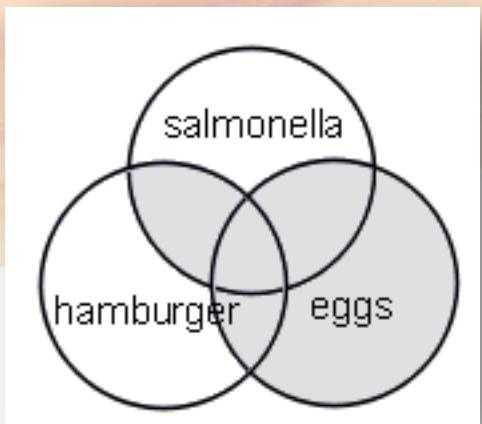
- **AND:**
 - Used to retrieve a set in which each citation contains *all* search terms
 - **Example:** *salmonella AND hamburger*
 - the retrieval is only the overlap of the results for each term – those records in which both terms appear



Search terms	Results
salmonella	74677
hamburger	3142
salmonella AND hamburger	18

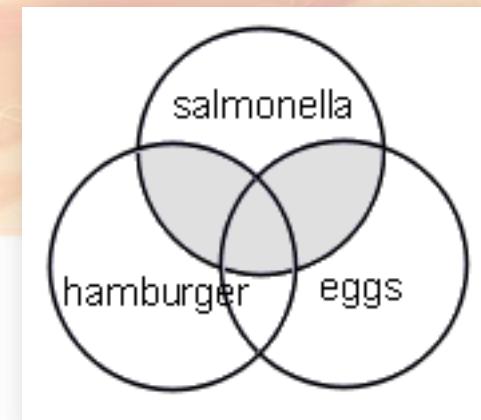
Example Nesting:

- When using multiple Boolean operators in PubMed, they are processed left to right
 - **Example:** *salmonella AND hamburger OR eggs*
 - Retrieve records that include both terms *salmonella* AND *hamburger* as well as all records with the term *eggs*, whether or not they contain the other two terms
- Maybe what you wanted? If not...
 - Need Nesting



Example Nesting:

- To change the order in which terms are processed, enclose the terms(s) in parentheses
- The terms inside the set of parentheses will be processed as a unit and then incorporated into the overall strategy
- **This is called nesting**
- **Example:** *salmonella AND (hamburger OR eggs)*
- Retrieve records that contain the term *salmonella*, as well as one or both of the terms *hamburger* OR *eggs*



("Cell Count/classification"[Mesh]) AND ("Child, Preschool"[Mesh])

Boolean Rules and Syntax

- Boolean operators: AND, OR, and NOT:
 - Must be entered in UPPERCASE
 - AND is the **default operator** used in
 - If you do not include Boolean operators in your search
 - PubMed will automatically use AND between terms
- PubMed processes Boolean connectors in a left-to-right sequence
- You can change order in which PubMed processes a search statement by nesting an **individual concept in parentheses**
 - Terms inside parentheses will be **processed as a unit** and then **incorporated** into the overall strategy

Searching PubMed

NCBI Resources How To cleslin1 My NCBI Sign Out

PubMed estradiol [NM] Search Help

US National Library of Medicine National Institutes of Health

Show additional filters

Text availability

Abstract available

Free full text available

Full text available

Publication dates

5 years

10 years

Custom range...

Species

Humans

Other Animals

Article types

Clinical Trial

Display Settings: Summary, 20 per page, Sorted by Recently Added Send to: makes for a long weekend!

Results: 1 to 20 of 70460 << First < Prev Page 1 of 3523 Next > Last >>

[Stroke and myocardial infarction with hormonal contraception.](#)

1. Vitale J.
N Engl J Med. 2012 Sep 27;367(13):1264; author reply 1264-5. No abstract available.
PMID: 23013086 [PubMed - indexed for MEDLINE]
[Related citations](#)

[Stroke and myocardial infarction with hormonal contraception.](#)

2. Siegel AJ.
N Engl J Med. 2012 Sep 27;367(13):1264; author reply 1264-5. No abstract available.
PMID: 23013085 [PubMed - indexed for MEDLINE]
[Related citations](#)

[\[Estrogen-like effects of Menoprogen on female ovariectomized rats\].](#)

3. Li X, Ma H, Lv Y, Hattori M, Mi HC.
Zhongguo Zhong Yao Za Zhi. 2012 Jun;37(11):1646-50. Chinese.
PMID: 22994000 [PubMed - indexed for MEDLINE]

Filters: Manage Filters

Results by year

Titles with your search terms

The bioidentical hormone debate: are bioidentical horr [Postgrad Med. 2009]

What's new in hormone replacement ther

Search details

estradiol[NM]

Search See more...

Searching PubMed

NCBI Resources How To cleslin1 My NCBI Sign Out

PubMed estradiol [NM] Search Help

Show additional filters Clear all

Text availability Abstract available Free full text available Full text available

Publication dates 5 years 10 years Custom range... Species clear Humans Other Animals

Display Settings: Summary, 20 per page, Sorted by Recently Added Send to: Still a long weekend! Results: 1 to 20 of 30981

Filters activated: Humans

Stroke and myocardial infarction with hormonal contraception.

- Vitale J. N Engl J Med. 2012 Sep 27;367(13):1264; author reply 1264-5. No abstract available. PMID: 23013086 [PubMed - indexed for MEDLINE]
[Related citations](#)
- Siegel AJ. N Engl J Med. 2012 Sep 27;367(13):1264; author reply 1264-5. No abstract available. PMID: 23013085 [PubMed - indexed for MEDLINE]
[Related citations](#)
- Estrogen therapy in Turner syndrome: does the type, dose and mode of delivery matter?

Results by year

Titles with your search terms

The bioidentical hormone debate: are bioidentical hormones [Postgrad Med. 2009]

What's new in hormone replacement therapy

Search details

estradiol [NM] AND "humans" [MeSH Terms]

Search See more...

Searching PubMed

NCBI Resources ▾ How To ▾ cleslin1 My NCBI Sign Out

PubMed estradiol [NM] Search Help

US National Library of Medicine National Institutes of Health

RSS Save search Advanced

Show additional filters Clear all

Text availability Abstract available Free full text available Full text available

Publication dates 5 years 10 years Custom range...

Species Humans Other Animals

Additional filters

- Text availability
- Publication dates
- Species
- Article types
- Languages
- Sex
- Subjects
- Journal categories
- Ages
- Search fields

Show

20 per page, Sorted by Recently Added Send to: ▾ Filters: Manage Filters

<< First < Prev Page 1 of 1550 Next > Last >>

ar all

arction with hormonal contraception.

7(13):1264; author reply 1264-5. No abstract available.
Indexed for MEDLINE]

arction with hormonal contraception.

7(13):1264; author reply 1264-5. No abstract available.
Indexed for MEDLINE]

Results by year

Titles with your search terms

The bioidentical hormone debate: are bioidentical horr [Postgrad Med. 2009]

What's new in hormone replacement therapy: focus on ti [Climacteric. 2012]

Regional differences in estradiol effects on numbers [Brain Res. 2010]

See more...

Searching PubMed

The screenshot shows the PubMed search results for the query "estradiol [NM]". The results are sorted by Recently Added, with 20 items per page, currently on page 1 of 2. A yellow box highlights the search terms "estradiol [NM]" in the search bar and the title "A great weekend!" above the results. The results list includes several studies, such as one on leuprolide acetate depot treatment for central precocious puberty and another on neurodevelopmental follow-up of extremely low birth weight infants after postnatal replacement of 17 β -estradiol and progesterone.

Display Settings: Summary, 20 per page, Sorted by Recently Added
Send to: Filters: Manage Filters

A great weekend!

Results: 1 to 20 of 21

Filters activated: published in the last 5 years, Humans, Clinical Trial, Male, Core clinical journals [Clear all](#)

1. Efficacy and safety of leuprolide acetate 3-month depot 11.25 milligrams or 30 milligrams for the treatment of central precocious puberty.
Lee PA, Klein K, Mauras N, Neely EK, Bloch CA, Larsen L, Mattia-Goldberg C, Chwalisz K.
J Clin Endocrinol Metab. 2012 May;97(5):1572-80. Epub 2012 Feb 16.
PMID: 22344198 [PubMed - indexed for MEDLINE]
[Related citations](#)

2. Neurodevelopmental follow-up at five years corrected age of extremely low birth weight infants after postnatal replacement of 17 β -estradiol and progesterone.
Trotter A, Steinmacher J, Kron M, Pohlundt F.
J Clin Endocrinol Metab. 2012 Mar;97(3):1041-7. Epub 2012 Jan 18.
PMID: 22259065 [PubMed - indexed for MEDLINE]
[Related citations](#)

3. The TLR-mediated response of plasmacytoid dendritic cells is positively regulated by estradiol in vivo through cell-intrinsic estrogen receptor α signaling.
Seillet C, Laffont S, Trémollières F, Rouquié N, Ribot C, Arnal JF, Douin-Echinard V, Gourdy P, Guéry JC.
Blood. 2012 Jan 12;119(2):454-64. Epub 2011 Nov 16.
PMID: 22096248 [PubMed - indexed for MEDLINE]
[Related citations](#)

4. Higher testosterone levels are associated with less loss of lean body mass in older men.
LeBlanc ES, Wang PY, Lee CG, Barrett-Connor E, Cauley JA, Hoffman AR, Laughlin GA, Marshall LM, Orwoll ES.
J Clin Endocrinol Metab. 2011 Dec;96(12):3855-63. Epub 2011 Oct 5.
PMID: 21976718 [PubMed - indexed for MEDLINE]
[Related citations](#)

Gender, sex-steroid, and secretagogue-selective recovery from growth hormone-

Titles with your search terms

The bioidentical hormone debate: are bioidentical hormon [Postgrad Med. 2009]
What's new in hormone replacement therapy: focus on trans [Climacteric. 2012]
Regional differences in estradiol effects on numbers of HSD2-cc [Brain Res. 2010]

See more...

1795 free full-text articles in PubMed Central

High salt intake down-regulates colonic mineralocorticoid recep [PLoS One. 2012]
Genome-wide association study of circulating estradiol, te [PLoS One. 2012]
Plasma kisspeptin levels in girls with prematu [J Clin Res Pediatr Endocrinol....]

See all (1795)...

Find related data

Database: Select

Find items

Search details

```
estradiol[NM] AND ("2007/10/30"[PDate] : "2012/10/27"[PDate] AND "humans"[MeSH Terms] AND Clinical Trial[ptyp] AND "male"[MeSH Terms] AND jsubsetain[text])
```

Search See more...

Now have query
for an
automated
search!

Wait... How?

Search Field Descriptions and Tags

Search Field Descriptions and Tags

Affiliation [AD]
Article Identifier [AID]
All Fields [ALL]
Author [AU]
Author Identifier [AUID]
Book [book]
Comment Corrections
Corporate Author [CN]
Create Date [CRDT]
Completion Date [DCOM]
EC/RN Number [RN]
Editor [ED]
Entrez Date [EDAT]
Filter [FILTER]
First Author Name [1AU]
Full Author Name [FAU]
Full Investigator Name [FIR]
Grant Number [GR]
Investigator [IR]
ISBN [ISBN]
Issue [IP]
Journal [TA]
Language [LA]
Last Author [LASTAU]
Location ID [LID]
MeSH Date [MHDA]
MeSH Major Topic [MAJR]

MeSH Subheadings [SH]
MeSH Terms [MH]
Modification Date [LR]
NLM Unique ID [JID]
Other Term [OT]
Owner
Pagination [PG]
Personal Name as Subject [PS]
Pharmacological Action [PA]
Place of Publication [PL]
PMID [PMID]
Publisher [PUBN]
Publication Date [DP]
Publication Type [PT]
Secondary Source ID [SI]
Subset [SB]
Supplementary Concept[NM]
Text Words [TW]
Title [TI]
Title/Abstract [TIAB]
Transliterated Title [TT]
UID [PMID]
Version
Volume [VI]

<http://www.ncbi.nlm.nih.gov/books/NBK3827/>

NCBI Entrez

- We can do a lot using a web-browser
 - Look up a specific record for PubMed
 - Search for matches to a gene or disease name
 - Find the search query needed for an automated process
- **Sometimes we need to automate the process**
 - Use Entrez:
 - to select and return the items of interest
 - rather than download, parse, and select

Automagically!

- http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_example.pl
- Download and Run from a terminal window:
- perl eutils_example.pl Accept the defaults and inspect the output
 - Database [Pubmed] :
 - Query [zanzibar] :
 - Report [abstract] :
- **Different outputs!**
- **Try them with the Perl program**

→

Display Settings: Summary, 20 per page, Sorted by Recently Added

Format	Items per page	Sort by
<input checked="" type="radio"/> Summary	<input type="radio"/> 5	<input checked="" type="radio"/> Recently Added
<input type="radio"/> Summary (text)	<input type="radio"/> 10	<input type="radio"/> Pub Date
<input type="radio"/> Abstract	<input checked="" type="radio"/> 20	<input type="radio"/> First Author
<input type="radio"/> Abstract (text)	<input type="radio"/> 50	<input type="radio"/> Last Author
<input type="radio"/> MEDLINE	<input type="radio"/> 100	<input type="radio"/> Journal
<input type="radio"/> XML	<input type="radio"/> 200	<input type="radio"/> Title
<input type="radio"/> PMID List		

[Stroke and myocardial infarction with hormonal contraception.](#)

1. [Siegel AJ.](#)
N Engl J Med. 2012 Sep 27;367(13):1264; author reply 1264-5. No abstract available.
PMID: 23013085 [PubMed - Indexed for MEDLINE]
[Related citations](#)

Output

MacBook-Pro-SSD:~ cleslin\$ perl eutils_example.pl

Database [Pubmed] :

Query [zanzibar]: estradiol[NM] AND ("2007/10/30"[PDat] : "2013/10/27"[PDat] AND "humans"[MeSH Terms] AND Clinical Trial[ptyp] AND "male"[MeSH Terms] AND jsubsetaim[text])

Report [abstract] :

1. J Clin Endocrinol Metab. 2012 May; 97(5):1572-80. Epub 2012 Feb 16.

Efficacy and safety of leuprolide acetate 3-month depot 11.25 milligrams or 30 milligrams for the treatment of central precocious puberty.

Lee PA, Klein K, Mauras N, Neely EK, Bloch CA, Larsen L, Mattia-Goldberg C, Chwalisz K.

Penn State College of Medicine, The Milton S. Hershey Medical Center, P.O. Box 850, Hershey, Pennsylvania 17033-0850, USA. plee@psu.edu

CONTEXT: GnRH agonist (GnRHa) monthly injections are frequently used in the treatment of central precocious puberty (CPP). The 3-month leuprolide depot 11.25- and 30-mg formulations are newly approved treatment options.

OBJECTIVE: The aim of the study was to investigate the safety and efficacy of leuprolide acetate 3-month depot formulations for the treatment of CPP in children.

DESIGN: This was a phase III, randomized, open-label, dose-ranging 6-month study.

SETTING: Twenty-two U.S. medical centers (including Puerto Rico) participated.

PATIENTS: Children diagnosed with CPP ($n = 84$), who were either treatment naive or previously treated with GnRHa, were recruited. Chronological age at onset of pubertal signs was less than 8 yr in girls and less than 9 yr in boys, and bone age was advanced over chronological age at least 1 yr.

INTERVENTION: Leuprolide acetate depot (11.25 or 30 mg) was administered im every 3 months.

Entrez



NCBI Entrez

- Powerful web-portal for NCBI's online databases
 - Nucleotide
 - Protein
 - PubMed
 - Gene
 - Structure
 - Taxonomy
 - OMIM
 - etc...

GQuery
NCBI Global Cross-database Search

Search NCBI databases

LIME1 Help

Literature

(none) PubMed: scientific & medical abstracts/citations
27 PubMed Central: full-text journal articles
(none) NLM Catalog: books, journals and more in the NLM Collections
(none) MeSH: ontology used for PubMed indexing
(none) Books: books and reports
1 Site Search: NCBI web and FTP site index

Health

(none) PubMed Health: clinical effectiveness, disease and drug reports
(none) MedGen: medical genetics literature and links
(none) GTR: genetic testing registry
(none) dbGaP: genotype/phenotype interaction studies
6 ClinVar: human variations of clinical significance
1 OMIM: online mendelian inheritance in man
(none) OMIA: online mendelian inheritance in animals

Organisms

(none) Taxonomy: taxonomic classification and nomenclature catalog

Nucleotide Sequences

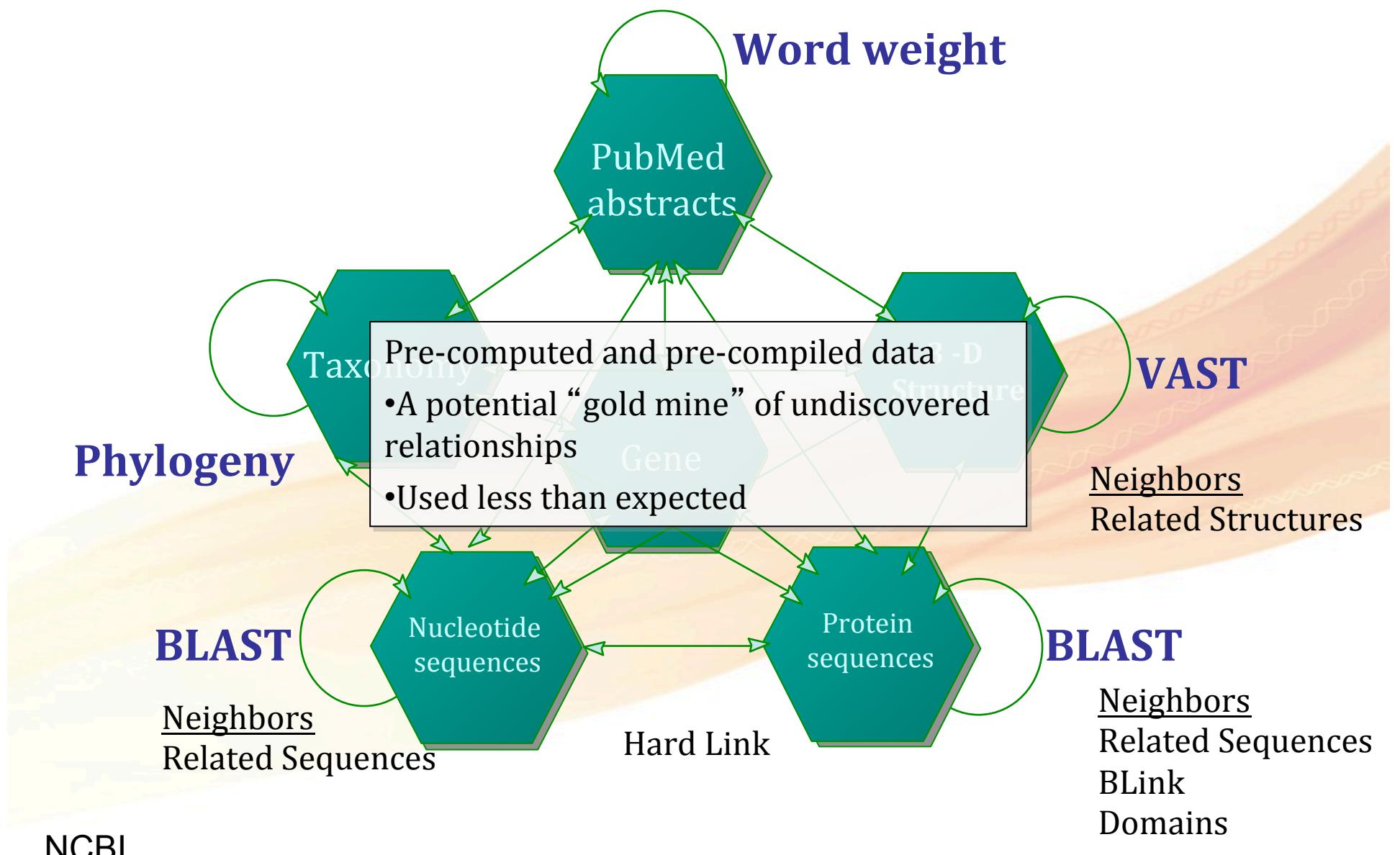
129 Nucleotide: DNA and RNA sequences
(none) GSS: genome survey sequences
7 EST: expressed sequence tag sequences
(none) SRA: high-throughput DNA and RNA sequence read archive
1 PopSet: sequence sets from phylogenetic and population studies
94 Probe: sequence-based probes and primers

Genomes

3 Genome: genome sequencing projects by organism
(none) Assembly: genomic assembly information
(none) Epigenomics: epigenomic studies and display tools
6 UniSTS: sequence-tagged sites for genome mapping
274 SNP: short genetic variations
162 dbVar: genome structural variation studies
(none) BioProject: biological projects providing data to NCBI
(none) BioSample: descriptions of biological source materials
294 Clone: genomic and cDNA clones

Edwards

Entrez: A Discovery System



The Discovery Initiative

- Make the most relevant and interesting results more obvious and readily accessible
 - Through easier to use interfaces
 - Promote higher quality resources
 - Gene – A great place to start!
 - RefSeqs
- Expose the power of pre-computed similarities and pre-compiled links
- There are three main categories of discovery components :
 - Sensors
 - Database Ads
 - Analysis Tools

<http://www.ncbi.nlm.nih.gov/books/NBK7039/>

Sensors

- Detects certain types of search terms & provides access to **potentially** more relevant results
- For ex. PubMed:
 - Citation & Gene Sensor activated when someone searches with a literature citation
- For ex. Nucleotide Search
 - Accession Sensor that provides a direct link to the sequence databases when someone searches with an NCBI sequence identifier

PubMed Search PB2

[See 22 articles about PB2 gene function](#)

See also: [PB2 PB2 protein in the Gene database](#)

[pb2 in Influenza A virus A/Puerto Rico/8/1934H1N1](#) | [Influenza A virus A/Goose/Guangdong/1/196H5N1](#) | [Influenza A virus A/New York/392/2004H3N2](#) | [All 9 Gene records](#)

Citation Sensor

Gene Sensor

Incorrect DB Search in Nucleotide

 The following term was not found in PubMed: NM_000795.

 See the search [details](#).

 No items found.

Accession Sensor

[See Nucleotide sequence data for NM_000795](#)

Homo sapiens dopamine receptor D2 (DRD2), transcript variant 1, mRNA [Homo sapiens]

- Promotes related information in other databases that may be more useful or may provide unexpected connections

Database Ads

[Display Settings:](#) Abstract

See 1 citation found by title matching your search:

[Cell](#), 2009 Jan 23;136(2):352-63. doi: 10.1016/j.cell.2008.11.038.

Large-scale structural analysis of the classical human protein tyrosine phosphatome.

Barr AJ, Ugochukwu E, Lee WH, King ON, Filippakopoulos P, Alfano I, Savitsky P, Burgess-Brown NA, Müller S, Knapp S. University of Oxford, Structural Genomics Consortium, Old Road Campus Research Building, Roosevelt Drive, Headington, Oxford, OX3 7DQ, UK. alastair.barr@sgc.ox.ac.uk

Abstract

Protein tyrosine phosphatases (PTPs) play a critical role in regulating cellular functions by selectively dephosphorylating their substrates. Here we present 22 human PTP crystal structures that, together with prior structural knowledge, enable a comprehensive analysis of the classical PTP family. Despite their largely conserved fold, surface properties of PTPs are strikingly diverse. A potential secondary substrate-binding pocket is frequently found in phosphatases, and this has implications for both substrate recognition and development of selective inhibitors. Structural comparison identified four diverse catalytic loop (WPD) conformations and suggested a mechanism for loop closure. Enzymatic assays revealed vast differences in PTP catalytic activity and identified PTPD1, PTPD2, and HDPTP as catalytically inert protein phosphatases. We propose a "head-to-toe" dimerization model for RPTPgamma/zeta that is distinct from the "inhibitory wedge" model and that provides a molecular basis for inhibitory regulation. This phosphatome resource gives an expanded insight into intrafamily PTP diversity, catalytic activity, substrate recognition, and autoregulatory self-association.

Comment in

[The PTP family photo album](#). [Cell. 2009]

PMID: 19167335 [PubMed - indexed for MEDLINE] PMCID: PMC2638020 [Free PMC Article](#)

Images from this publication. See all images (7) [Free text](#)

Publication Types, MeSH Terms, Substances, Secondary Source ID, Grant Support

LinkOut - more resources

Structure

[Send to:](#) [Free in PMC](#) full-text archive

Save items

Full Text

Related citations in PubMed

Crystal structures and inhibitor identification for PTPN [Biochem J. 2006]
Crystal structure of PTP-SL/PTPBR7 catalytic domain: implic [J Mol Biol. 2001]
Crystal structure of the catalytic domain of protein-tyrosine pl [J Biol Chem. 1998]
Review Mechanistic studies on protein ty [Prog Nucleic Acid Res Mol Biol. 2003]
Review An overview of the protein tyrosine pho [Curr Top Med Chem. 2003]

[See reviews...](#)

PubMed

[See all...](#)

Cited by 57 PubMed Central articles

Cadmium is a potent inhibitor of PPM phosphatases and target [Sci Rep. 2013]
Targeting density-enhanced phosphatase [Cell Commun Signal. 2013]
A loss-of-function screen for phosphatases that reg [PLoS One. 2013]

Full Text

[See all...](#)

Structures reported by this article

Crystal Structure Of The C472s Mutant Of Human Protein PDB: 2CJZ Source: Homo sapiens Method: X-Ray

Analysis Tools

- Provide on-the-fly analysis
- Important components of the discovery initiative
- Ex. Sequence records include a direct link:
 - BLAST search with the sequence as well as a link to run a conserved domain search for protein records

The screenshot shows the NCBI Protein page for the apolipoprotein E precursor from Homo sapiens. The page includes:

- Display Settings:** GenPept
- Send to:** Change region shown
- Customize view**
- Analyze this sequence** (highlighted in yellow):
 - Run BLAST
 - Identify Conserved Domains
 - Highlight Sequence Features
 - Find in this Sequence

Basic sequence information displayed:

LOCUS	NP_000032	317 aa	linear	PRI 28-OCT-2012
DEFINITION	apolipoprotein E precursor [Homo sapiens].			
ACCESSION	NP_000032			
VERSION	NP_000032.1 GI:4557325			
DBSOURCE	REFSEQ: accession NM_000041.2			
KEYWORDS	.			

What is Entrez

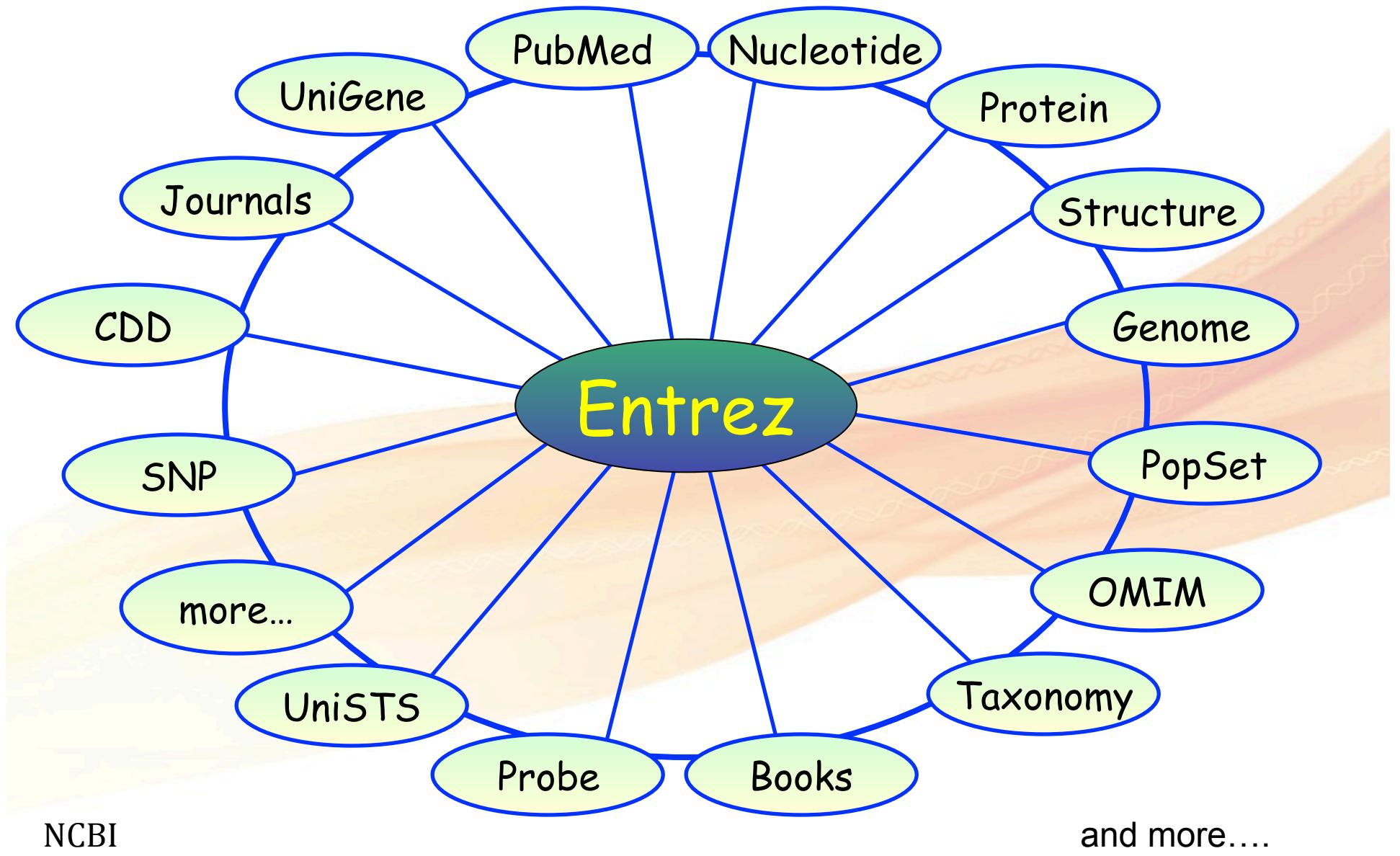


- “The life sciences search engine”
- Text-based search and retrieval system at NCBI
 - Single query string
 - Supports boolean operators
 - Search term tags to limit parts of the search statement of particular fields
- Covers all major databases at National Center for Biotechnology Information (NCBI)
 - [PubMed](#); [Nucleotide](#); [Protein](#); [Structure](#); [Genome](#); [Taxonomy](#)
- More than one way to do things
- Go through the following:
 - <http://www.ncbi.nlm.nih.gov/books/NBK3837/>

Entrez Databases

- All Molecular Database entries are organized by organism (*Taxonomy Database*)
- Each record is assigned a UID
 - A “unique integer identifier” for internal tracking
- Each record is indexed by data fields
 - [author], [title], [organism], and many others
- Each record is given a Document Summary
 - A summary of the record’s content (DocSum)
- Each record is manually or computationally assigned links to biologically related UIDs in and across databases

The (ever) Expanding Entrez System



Accessing Entrez

NCBI Resources How To

All Databases Search

cleslin1 My NCBI Sign Out

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [NCBI News](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

Genomic Structural Variation

dbVar archives large scale genomic variation data and associates defined variants with phenotypic information [GO](#)

II 1 2 3 4 5 6 7 8

Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI Announcements

New NCBI Insights Blog Post: Joining PubMed Commons - A step-by-step guide Oct 23, 2013

GenBank Release 198.0 is Available Oct 22, 2013

The new release for GenBank is now available via FTP, as well as in the [Entrez](#), [BLAST](#), and [BLAST+](#)

PubMed Commons is now live! Oct 22, 2013

NCBI has released PubMed Commons, currently in pilot phase, which is a new

GQuery
NCBI Global Cross-database Search

Search NCBI databases

alzheimer's disease

Literature

- [94387 PubMed](#): scientific & medical abstracts/citations
- [63926 PubMed Central](#): full-text journal articles
- [19115 NLM Catalog](#): books, journals and more in the NLM Collections

- [4 MeSH](#): ontology used for PubMed indexing
- [2660 Books](#): books and reports
- [73 Site Search](#): NCBI web and FTP site index

Health

- [526 PubMed Health](#): clinical effectiveness, disease and drug reports
- [26 MedGen](#): medical genetics literature and links
- [15 GTR](#): genetic testing registry
- [2031 dbGaP](#): genotype/phenotype interaction studies

- [17 ClinVar](#): human variations of clinical significance
- [161 OMIM](#): online mendelian inheritance in man
- [185 OMIA](#): online mendelian inheritance in animals

Organisms

- (none) Taxonomy: taxonomic classification and nomenclature catalog

Nucleotide Sequences

- [92760 Nucleotide](#): DNA and RNA sequences
- (none) GSS: genome survey sequences
- (none) EST: expressed sequence tag sequences

- [485 SRA](#): high-throughput DNA and RNA sequence read archive
- [2 PopSet](#): sequence sets from phylogenetic and population studies

Genomes

- [1 Genome](#): genome sequencing projects by organism
- (none) Assembly: genomic assembly information
- [11 Epigenomics](#): epigenomic studies and display tools
- (none) UniSTS: sequence-tagged sites for genome mapping
- (none) SNP: short genetic variations

- [6579 dbVar](#): genome structural variation studies
- [105 BioProject](#): biological projects providing data to NCBI
- [13466 BioSample](#): descriptions of biological source materials
- (none) Clone: genomic and cDNA clones

Genes

- [4130 Gene](#): collected information about gene loci
- (none) HomoloGene: homologous gene sets for selected organisms
- [9 UniGene](#): clusters of expressed transcripts

- [399510 GEO Profiles](#): gene expression and molecular abundance profiles
- [1729 GEO DataSets](#): functional genomics studies

Proteins

- [7592 Protein](#): protein sequences
- [45 Conserved Domains](#): conserved protein domains

- (none) Protein Clusters: sequence similarity-based protein clusters
- [744 Structure](#): experimentally-determined biomolecular structures

Chemicals

- [4 PubChem Compound](#): chemical information with structures, information and links
- [46 PubChem Substance](#): deposited substance and chemical information
- [8673 PubChem BioAssay](#): bioactivity screening studies

Pathways

- [543 BioSystems](#): molecular pathways with links to genes, proteins and chemicals

Global Query: All NCBI Databases



The Entrez system: 42 (and counting) integrated databases

Current Data in Entrez

Search across databases Help

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

20269920 PubMed: biomedical literature citations and abstracts	299648 Books: online books
2065122 PubMed Central: free, full text journal articles	21226 OMIM: online Mendelian Inheritance in Man
18830 Site Search: NCBI web and FTP sites	2664 OMIA: online Mendelian Inheritance in Animals
42100588 Nucleotide: Core subset of nucleotide sequence records	109476 dbGaP: genotype and phenotype
67040108 EST: Expressed Sequence Tag records	4361652 UniGene: gene-oriented clusters of transcript sequences
29103596 GSS: Genome Survey Sequence records	40774 CDD: conserved protein domain database
35939669 Protein: sequence database	313714 3D Domains: domains from Entrez Structure
12572 Genome: whole genome sequences	529104 UniSTS: markers and mapping data
68700 Structure: three-dimensional macromolecular structures	120464 PopSet: population study data sets
665610 Taxonomy: organisms in GenBank	63811486 GEO Profiles: expression and molecular abundance profiles
83559953 SNP: single nucleotide polymorphism	29922 GEO DataSets: experimental sets of GEO data
510291 dbVar: Genomic structural variation	524 Epigenomics: Epigenetic maps and data sets
7707243 Gene: gene-centered information	142056 Cancer Chromosomes: cytogenetic databases
29115 SRA: Sequence Read Archive	488155 PubChem BioAssay: bioactivity screens of chemical substances
137417 BioSystems: Pathways and systems of interacting molecules	28862823 PubChem Compound: unique small molecule chemical structures
123767 HomoloGene: eukaryotic homology groups	72246317 PubChem Substance: deposited chemical substance records
99026 GENSAT: gene expression atlas of mouse central nervous system	507133 Protein Clusters: a collection of related protein sequences
10249154 Probe: sequence-specific reagents	322 Peptidome: MS/MS proteomic experiments
5958 Genome Project: genome project information	
26190 Journals: detailed information about the journals indexed in PubMed and other Entrez databases	220543 MeSH: detailed information about NLM's controlled vocabulary
1422450 NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections	

2010 - Oct

GQuery
NCBI Global Cross-database Search

Search NCBI databases

Literature

[23200640](#) PubMed: scientific & medical abstracts/citations
[2878555](#) PubMed Central: full-text journal articles
[1487441](#) NLM Catalog: books, journals and more in the NLM Collections

[243770](#) MeSH: ontology used for PubMed indexing
[225947](#) Books: books and reports
[21929](#) Site Search: NCBI web and FTP site index

Health

[42748](#) PubMed Health: clinical effectiveness, disease and drug reports
[169668](#) MedGen: medical genetics literature and links
[29200](#) GTR: genetic testing registry
[155457](#) dbGaP: genotype/phenotype interaction studies

[48153](#) ClinVar: human variations of clinical significance
[23163](#) OMIM: online mendelian inheritance in man
[2844](#) OMIA: online mendelian inheritance in animals

Organisms

[1167918](#) Taxonomy: taxonomic classification and nomenclature catalog

Nucleotide Sequences

[105344133](#) Nucleotide: DNA and RNA sequences
[36983541](#) GSS: genome survey sequences
[74938973](#) EST: expressed sequence tag sequences

[516394](#) SRA: high-throughput DNA and RNA sequence read archive
[177805](#) PopSet: sequence sets from phylogenetic and population studies
[31394098](#) Probe: sequence-based probes and primers

Genomes

[11004](#) Genome: genome sequencing projects by organism
[19183](#) Assembly: genomic assembly information
[6634](#) Epigenomics: epigenomic studies and display tools
[545913](#) UniSTS: sequence-tagged sites for genome mapping
[309795501](#) SNP: short genetic variations

[3594158](#) dbVar: genome structural variation studies
[104528](#) BioProject: biological projects providing data to NCBI
[2137395](#) BioSample: descriptions of biological source materials
[33135862](#) Clone: genomic and cDNA clones

Genes

[14728937](#) Gene: collected information about gene loci
[133548](#) HomoloGene: homologous gene sets for selected organisms
[6473284](#) UniGene: clusters of expressed transcripts

[91392791](#) GEO Profiles: gene expression and molecular abundance profiles
[1073972](#) GEO DataSets: functional genomics studies

Proteins

[98311196](#) Protein: protein sequences
[48034](#) Conserved Domains: conserved protein domains

[382691](#) Protein Clusters: sequence similarity-based protein clusters
[94271](#) Structure: experimentally-determined biomolecular structures

Chemicals

[47646830](#) PubChem Compound: chemical information with structures, information and links
[119989988](#) PubChem Substance: deposited substance and chemical information
[717519](#) PubChem BioAssay: bioactivity screening studies

Pathways

[534816](#) BioSystems: molecular pathways with links to genes, proteins and chemicals

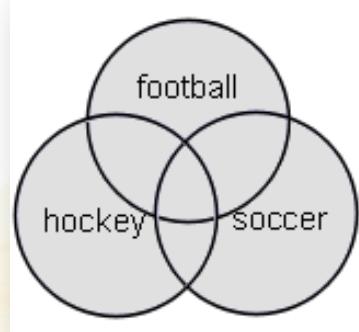
2013 - Oct

Searching NCBI Databases with Entrez

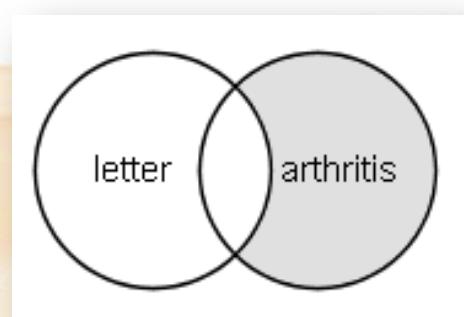
- Use words and phrases to construct precise searches of NCBI databases, either individually or searching across multiple databases
 - **Many features can be used in many or all databases**
 - **Others are specialized for specific databases**
- **Global Query**
 - Search across all Entrez databases by following the “All Databases” link from the NCBI homepage
 - Quick way to search NCBI content
 - But the search interface lacks many of the features we will discuss
- **Limits**
 - This tab is different for each database and provides an easy way to focus your search
 - Click the tab to see the options for the database you are using
 - e.g. limit to a particular species/taxonomy in Gene, limit to mRNA in Nucleotide, or limit to a particular chromosome in OMIM

Entrez - Boolean Operators

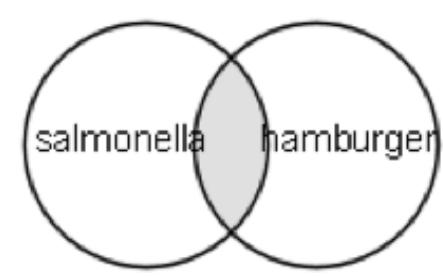
- As with PubMed we can use Booleans
- Use the upper case Boolean operators AND, OR and NOT in conjunction with parentheses to narrow or expand your search.
 - e.g. breast cancer AND (brca1 OR brca2)



OR



NOT



AND

Search Terms

- Databases consist of “records” and “fields.”
- Records represent genes, proteins, journal articles or other objects depending on the database
- Each record is composed of fields that describe the object
- Fields can include Title, Accession Number, Organism, Author and more, depending on the record
- Using search “terms” to specify fields allows you to increase search precision, for example **mouse[ORGN] AND polya signal[FKEY]**
 - **Do nucleotide search for:**
 - **mouse AND polya signal**
 - **mouse[ORGN] AND polya signal[FKEY]**

<http://dl.dropboxusercontent.com/u/1304874/entrez/index.html>

Gene Search

NCBI Resources How To

Nucleotide Nucleotide Save search Limits Advanced Help

Display Settings: Summary, 20 per page, Sorted by Default order Send to: Filter your results:

Results: 1 to 20 of 30 << First < Prev Page 1 of 2 Next > Last >>

[Oryza sativa](#) Indica Group cultivar IRGC 16339 **Xa21** gene, promoter region and 5' UTR
1. 2,108 bp linear DNA
Accession: JN564633.1 GI: 379139219
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Oryza sativa](#) Indica Group cultivar IRGC. 27045 **Xa21** gene, promoter region and 5' UTR
2. 2,008 bp linear DNA
Accession: JN564632.1 GI: 379139218
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Oryza sativa](#) receptor kinase-like protein (**Xa21**) gene, complete cds
3. 3,921 bp linear DNA
Accession: U37133.1 GI: 1122442
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Oryza sativa](#) (indica cultivar-group) clone Xa923 receptor kinase-like protein (**Xa21**) gene, partial cds
4. 1,020 bp linear DNA
Accession: DQ374737.1 GI: 86990757

All (30)
Bacteria (0)
[INSDC \(GenBank\) \(30\)](#)
mRNA (0)
RefSeq (0)
[Manage Filters](#)

Top Organisms [\[Tree\]](#)
[Oryza sativa](#) Indica Group (19)
[Oryza sativa](#) Japonica Group (11)

Find related data
Database:

Search details

A Better Gene Search

NCBI Resources ▾ How To ▾ cleslin1 My NCBI Sign Out

Nucleotide Nucleotide ↴ xa21[Gene Name] AND Oryza sativa[Organism] AND complete CDS[TITLE] × Search Save search Limits Advanced Help

Display Settings: Summary, Sorted by Default order Send to: Filter your results: All (3)

Results: 3

- [Oryza sativa receptor kinase-like protein \(Xa21\) gene, complete cds](#)
 - 1. 3,921 bp linear DNA
Accession: U37133.1 GI: 1122442
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Oryza sativa Indica Group Xa21 gene for receptor kinase-like protein, complete cds, cultivar:zheda8220](#)
 - 2. 4,623 bp linear DNA
Accession: AB212798.1 GI: 94481120
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Oryza sativa Indica Group Xa21 gene for receptor kinase-like protein, complete cds, cultivar:il you 8220](#)
 - 3. 4,623 bp linear DNA
Accession: AB212799.1 GI: 94481122
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

Display Settings: Summary, Sorted by Default order Send to: Filter your results: All (3)

Filter your results:

- All (3)
- Bacteria (0)
- [INSDC \(GenBank\) \(3\)](#)
- mRNA (0)
- RefSeq (0)

[Manage Filters](#)

Analyze these sequences Run BLAST

Find related data Database: Select Find items

Search details

Writing Advanced Search Statements

FIELD	Entrez Search Tips
LOCUS GenBank division	gbdiv_est [PROP], gbdiv_pln [PROP], etc.
Molecule type	biomol_genomic [PROP], biomol_mRNA [PROP], etc.
DEFINITION	complete CDS [TITL]
ACCESSION	AP001111 [ACCN]
SOURCE Organism	Oryza sativa [ORGN]
FEATURES	TATA_signal [FKEY], CDS [FKEY], D-loop [FKEY], etc.

Writing Advanced Search Statements

```
→ LOCUS OSU37133 3921 bp DNA linear PLN 14-DEC-1995
→ DEFINITION Oryza sativa receptor kinase-like protein (Xa21) gene, complete
               cds.
→ ACCESSION U37133
→ VERSION U37133.1 GI:1122442
→ KEYWORDS .
→ SOURCE Oryza sativa Indica Group
→ ORGANISM Oryza sativa Indica Group
               Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
               Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BEP
               clade; Ehrhartoideae; Oryzeae; Oryza.
→ REFERENCE 1 (bases 1 to 3921)
→ AUTHORS Song,W.-Y., Wang,G.-L., Chen,L.-L., Kim,H.-S., Pi,L.-Y.,
           Holsten,T., Wang,G.B., Zhai,W.-X., Zhu,L.-H., Fauquet,C. and
           Ronald,P.
→ TITLE A receptor kinase-like protein encoded by the rice disease
           resistance gene, Xa21
→ JOURNAL Science 270 (5243), 1804-1806 (1995)
→ PUBMED 8525370
→ REFERENCE 2 (bases 1 to 3921)
→ AUTHORS Ronald,P., Song,W.-Y., Wang,G.-L., Kim,H.-S. and Pi,L.-Y.
→ TITLE Direct Submission
→ JOURNAL Submitted (26-SEP-1995) Pamela Ronald, Plant Pathology, UC Davis,
           Hutchison Hall, Davis, CA 95616, USA
→ FEATURES Location/Qualifiers
→ source 1..3921
           /organism="Oryza sativa Indica Group"
           /mol_type="genomic DNA"
           /strain="IRBB21"
           /db_xref="taxon:39946"
           /chromosome="11"
           /map="11q, RG103"
→ gene 1..3921
           /gene="Xa21"
→ CDS join(1..2677,3521..3921)
           /gene="Xa21"
           /note="Xa21 disease resistance gene"
           /codon_start=1
```

Some Common Search Fields

Some Common Search Fields

Field Name	Field Tag	Database Examples	Features and Examples
Accession number	[ACCN]	Gene, Nucleotide, Protein, Structure, SNP, Probe	Searches sequence accession numbers for nucleic acids and proteins. Use a colon to search a range of accession numbers.
Organism	[ORGN]	Gene, Nucleotide, Protein, Structure, EST, GSS, Probe	Use either common names or scientific names.
Creation Date	[CDAT]	All	Date the record was created. Use a colon to search for a range of dates, e.g. 2003/15/01:2004/15/01[CDAT] . Month and day are optional.
Gene Name	[GENE NAME]	Gene, Nucleotide, Protein, OMIM, Probe, UniGene	Retrieves your search term as the exact gene name. Use a * to truncate your search; e.g. Irh*[GENE NAME] retrieves LRH1, LRHA, LRHR, etc.
Disease Name	[DIS]	Gene, OMIM, Probe, UniGene	Searches for the name of a disease or phenotype associated with genetic mutations.
Gene Ontology	[GO]	Gene, SNP, UniGene, Probe	Searches for the term in Gene Ontology annotations.
Molecular Weight	[MOLWT]	Protein	Use a colon to indicate a range of molecular weights; e.g. 1500:1550[MOLWT] .
Sequence Length	[SLEN]	Gene, Nucleotide, Protein	Use a colon to indicate a range of sequence lengths; e.g. 900:1000[SLEN] .

Entrez Search Terms

ALL, All Fields, All terms from all searchable fields
UID, Unique number assigned to each sequence
FILT, Filter, Limits the records
WORD, Text Word, Free text associated with record
TITL, Title, Words in definition line
KYWD, Keyword, Nonstandardized terms provided by submitter
AUTH, Author, Author(s) of publication
JOUR, Journal, Journal abbreviation of publication
VOL, Volume, Volume number of publication
ISS, Issue, Issue number of publication
PAGE, Page Number, Page number(s) of publication
ORGN, Organism, Scientific and common names of organism, and all higher levels of taxonomy
ACCN, Accession, Accession number of sequence
PACC, Primary Accession, Does not include retired secondary accessions
GENE, Gene Name, Name of gene associated with sequence
PROT, Protein Name, Name of protein associated with sequence
ECNO, EC/RN Number, EC number for enzyme or CAS registry number
PDAT, Publication Date, Date sequence added to GenBank
MDAT, Modification Date, Date of last update
SUBS, Substance Name, CAS chemical name or MEDLINE Substance Name
PROP, Properties, Classification by source qualifiers and molecule type
SQID, SeqID String, String identifier for sequence
GPRJ, Genome Project, Genome Project
SLEN, Sequence Length, Length of sequence
FKEY, Feature key, Feature annotated on sequence
RTYP, Replicon type, Replicon type
RNAM, Replicon name, Replicon name
ORGL, Organelle, Organelle

http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html

http://155.33.203.128/teaching/BIOL6308-Fall2013/local/Literature/entrez_fields.pdf

Using Truncation to Expand Field Searches

- Incorporating Entrez's truncation symbol *
 - An asterisk
 - e.g. pseudopod* @ Taxonomy
- Will retrieve the following:
 - Pseudopoda namkhan
 - Pseudopodoces humilis
 - Pseudopodoces
 - Pseudopoda
 - Pseudopodisma nagyi
 - Pseudopodisma

Taxonomy Taxonomy pseudopod* Search Help

Save search Limits Advanced

Display Settings: Summary, 20 per page Send to: Filters: [Manage Filters](#)

Results: 6

1. [Pseudopoda namkhan](#)
species, spiders
[Nucleotide](#) [Protein](#)

2. [Pseudopodoces humilis](#)
species, birds
[Nucleotide](#) [Protein](#)

3. [Pseudopodoces](#)
genus, birds
[Nucleotide](#) [Protein](#)

4. [Pseudopoda](#)
genus, spiders
[Nucleotide](#) [Protein](#)

5. [Pseudopodisma nagyi](#)
species, grasshoppers
[Nucleotide](#) [Protein](#)

6. [Pseudopodisma](#)
genus, grasshoppers
[Nucleotide](#) [Protein](#)

Find related data
Database: [Select](#) Find items

Search details
pseudopoda[All Names] OR pseudopoda namkhan[All Names] OR pseudopoda namkhan jager, pathoumthong velei, 2006[All Names] OR pseudopodisma[All Names] OR pseudopodisma

Search See more...

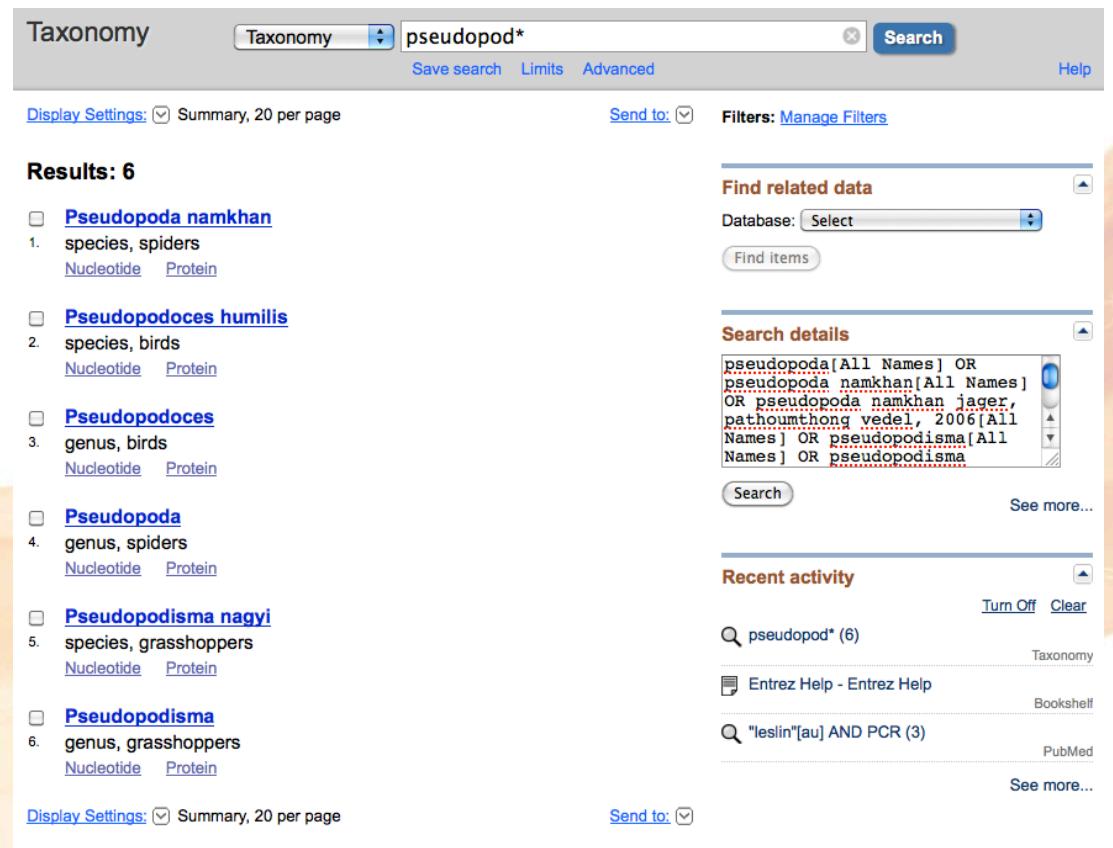
Recent activity
Turn Off Clear

pseudopod* (6) Taxonomy

Entrez Help - Entrez Help Bookshelf

"leslin"[au] AND PCR (3) PubMed

Display Settings: Summary, 20 per page Send to:



Properties and Filters

- In addition to search fields like [ACCN] and [ORGN]
 - Entrez databases use properties and filters
 - These search features limit searches in many different ways depending on the database
- Some examples are searching in Gene to retrieve only records that have a link to Protein, or a searching in Nucleotide to retrieve only tRNA molecule records

Some Examples of Properties and Filters

Search	Databases	Features and Examples
gene nucleotide[FILTER]	Gene	Gene records with links to Nucleotide records
genotype protein coding[PROPERTIES]	Gene	Gene records for protein coding genes
biomol trna[PROPERTIES]	Nucleotide	Nucleotide records for tRNA molecules
uncultured[FILTER]	Nucleotide	Nucleotide records for sequences from uncultured microorganisms
gene in plasmid[PROPERTIES]	Protein	Protein records in which the coding gene is on a plasmid
protein homologene[FILTER]	Protein	Protein records with links to the Homologene database

human[Organism]) AND RNR1[Gene Name]

human[Organism]) AND RNR1[Gene Name] AND gene nucleotide[FILTER]

Details & History

- Some databases use “Automatic Term Mapping” for your searches
 - “Search details” allows you to see how your term has been mapped and make any adjustments necessary
 - search for “human” in the Gene database
 - “*Homo sapiens*” [Organism] OR human[All Fields]
- Search History – Found in Advanced tab
 - Displays all of your recent searches in chronological order
 - Use the History tab to rerun, combine or modify previous searches
 - Searches are lost after 8 hours of inactivity on NCBI unless they are saved to a myNCBI account

Search details

```
pseudopoda[All Names] OR
pseudopoda namkhan[All Names]
OR pseudopoda namkhan jager,
pathoumthong vedel, 2006[All
Names] OR pseudopodisma[All
Names] OR pseudopodisma
```

Search

[See more...](#)

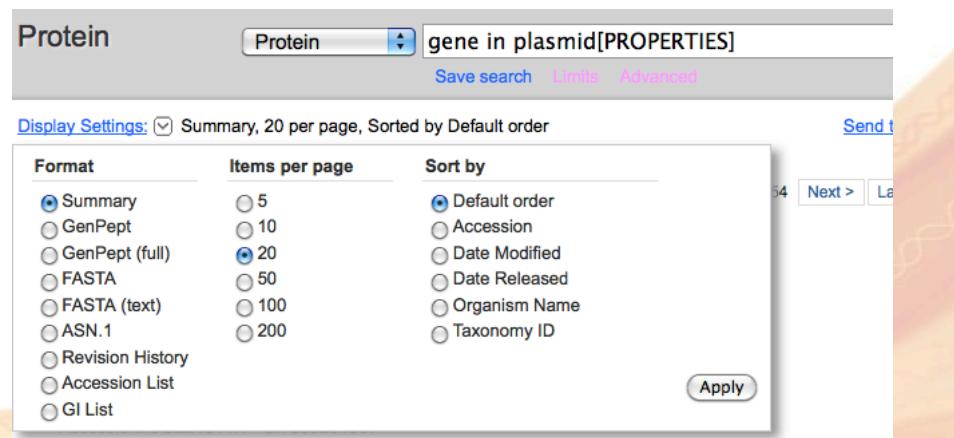
“ – not needed

MyNCBI

- MyNCBI allows you to:
 - Save results and search strategies indefinitely
 - As well as set up email alerts and change displays to your preferences
- After registering:
 - You can access your results and preferences at any computer you use to access NCBI database

Display Setting

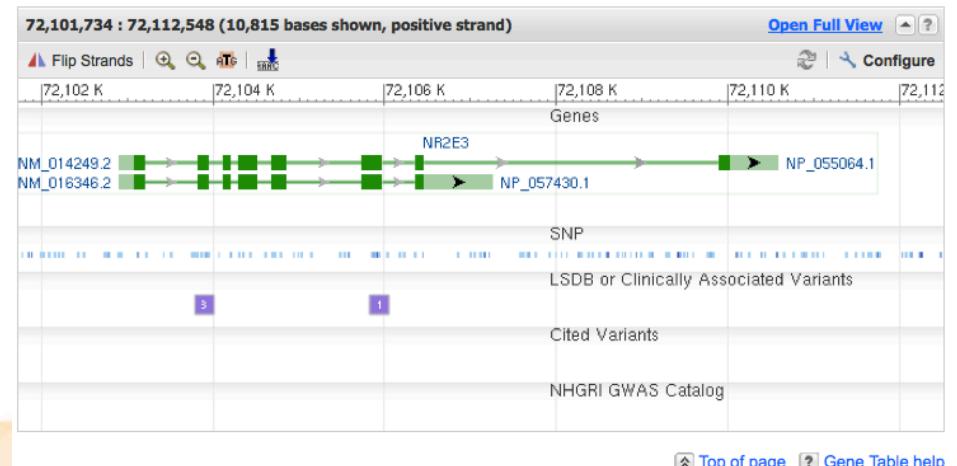
- The Display drop-down menu
 - View your search results in different formats depending on the database
 - Typically, displays are available that show:
 - Record summaries,
 - Links to other databases
 - Sequence data



Display Setting

- Some databases have specialized displays;
 - Such as the Gene Table display in Entrez Gene
 - Provides a graphical overview of the gene and includes intron and exon information
 - Ex: search for **NR2E3** in Gene

The screenshot shows the Entrez Gene search results for 'Gene'. A dropdown menu is open under 'Display Settings' with the 'Gene Table' option selected. An arrow points from this menu to the right towards the detailed gene visualization.



Exon table for mRNA NM_014249.2 and protein NP_055064.1				
Exon	Coding	Length (bp)		
		Exon	Coding	Intron
72102894-72103201	72103084-72103201	308	118	621
72103823-72103949	72103823-72103949	127	127	156
72104106-72104209	72104106-72104209	104	104	85
72104295-72104516	72104295-72104516	222	222	159
72104676-72104851	72104676-72104851	176	176	877
72105729-72105976	72105729-72105976	248	248	376
72106353-72106458	72106353-72106458	106	106	3434
72109893-72110600	72109893-72110025	708	133	

Send To

- The Send to drop-down menu allows you to send your results (in your chosen Display format) to a file, clipboard or collection

A screenshot of a protein search interface. At the top, there is a search bar with the query "NR2E3". Below the search bar, there are buttons for "Save search", "Limits", and "Advanced". The main area displays search results for "NR2E3" across various species. Each result includes a checkbox, the protein name, its length, accession number, GI number, and links to "GenPept", "FASTA", "Graphics", and "Related Sequences". A "Send to:" dropdown menu is open, showing options: "File" (selected), "Clipboard", and "Collections". It also shows a count of 55 items available for download in FASTA format. A "Create File" button is present. To the right of the dropdown, a sidebar lists taxonomic categories with their counts: "Homo sapiens (11)", "Mus musculus (9)", "Pan troglodytes (4)", "Bos taurus (4)", "Danio rerio (4)", "All other taxa (23)", and a "More..." link.

Protein

Protein NR2E3

Save search Limits Advanced

Display Settings: Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 55

<< First < Prev > Next >> Last

[NR2E3 \[Pan troglodytes\]](#)
1. 63 aa protein
Accession: ABM92077.1 GI: 124111341
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[NR2E3 \[Pan troglodytes\]](#)
2. 34 aa protein
Accession: ABM92076.1 GI: 124111340
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[NR2E3 \[Pongo pygmaeus\]](#)
3. 42 aa protein
Accession: ABM89412.1 GI: 124054378
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

Send to: Filter your results:

Choose Destination

File Clipboard Collections (55)

Download 55 items.

Format

FASTA

Create File

Homo sapiens (11)
Mus musculus (9)
Pan troglodytes (4)
Bos taurus (4)
Danio rerio (4)
All other taxa (23)
More...

Output

```
MacBook-Pro-SSD:~ cleslin$ perl eutils_example.pl
```

```
Database [Pubmed]: protein
```

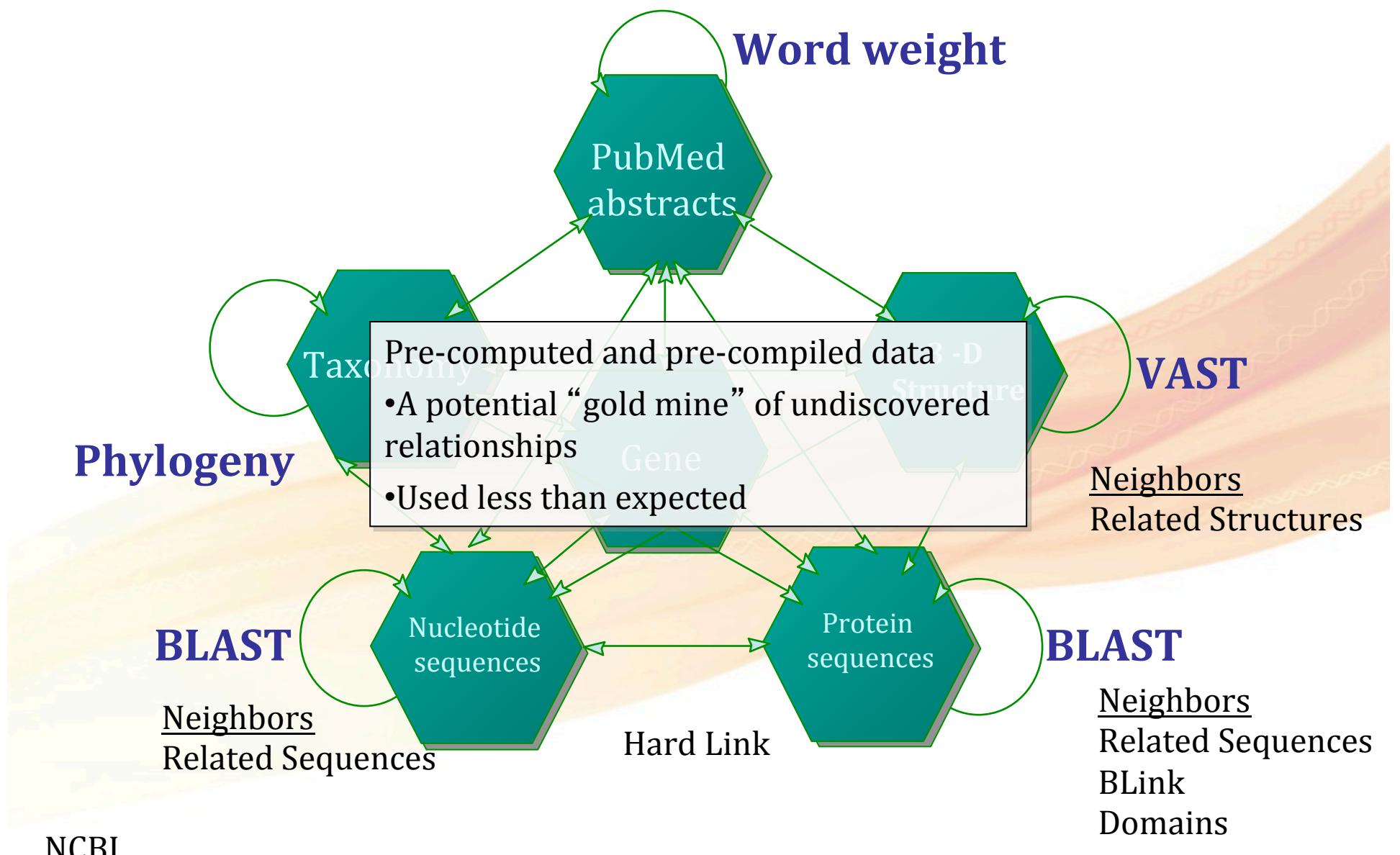
```
Query [zanzibar]: NR2E3
```

```
Report [abstract]: fasta
```

```
>gi|7706515|ref|NP_057430.1| photoreceptor-specific nuclear receptor isoform  
a [Homo sapiens]  
METRPTALMSSTVAAAAPAAAGAASRKESPGRWGLGEDPTGVSPSLQCRVCGDSSSGKHYGIYACNGCSEF  
FKRSVRRLIYRCQVGAGMCPVDKAHRNQCACRLKKCLQAGMNQDAVQNERQPRSTAQVHLDMSMESNTE  
SRPESLVAPPAPAGRSPRGPTPMASAARALGHFMASLITAETCAKLEPEDADENIDVTSNDPEFPSSPYS  
SSSPCGLDISETSARLLFMAVKWAKNLPVFSSLFPRDQVILLEAWSELFLGAIQWSLPLDSCPLLAP  
PEASAAGGAQGRLTLASMETRVLQETISRFRALAVDPTEFACMKALVLFKPETRGLKDPEHVEALQDQSQ  
VMLSQHSKAHHPSQPVR  
  
>gi|122132423|sp|Q08E02.1|NR1D1_BOVIN RecName: Full=Nuclear receptor subfamily 1 group D member 1; AltName: Full=Rev-erbA-alpha; AltName: Full=V-erbA-related protein 1; Short=EAR-1  
MTTLDNNNTGGVITYIGSSGSSPNRTSPESLYSDSSNGSFQS LTQGCPTYFPPSPTGS LTQDPARSFGS  
IPPSLGDDGSPSSSSSSSSSSSFYNGSPPGGLQVALEDNSRVSPSKSTSNIKLNGMVLLCKVCGDVA  
SGFHYGVHACEGCKGFFRRSIQQNIQYKRCLKNENCSIVRINRNRCQQCRFKKCLSVGMSRDAVRGRI  
KREKQRMLAEMQSAMNLANNQLSSQCPLETPPTQHPTPGPMGPSPPPAPAPSPLVGF SQFPQQLTPPRSP  
SPEPTVEDVISQVARAHREIFTYAHDKLGTSPGNFNAHASGNRPATTPHRWESQGCPPANDNIMAAQRH  
NEALNSLRQASSSYFPPWPPGAHHHSCHQPNNSNGHRLCPTHVYPAPEGEAPVNSPRGNSKNILLACPMN  
MYPHGRSGRTVQEIWEDFSMSFTPRAVREVVEFAKHIPGFRDLSQHDQVTLLKAGTFEVLMVRFASLFNVK  
DQTVMFLSRTTYSLQELGAMGMGDLNAMFDFSEKLNSLATEEELGLFTAVVILVSADRSGMENSASVEQ  
LQETLLRALRALVLKNRPSSETSRFTKLLLKPDLRTLNNMHSEKLLSFRVDAQ
```

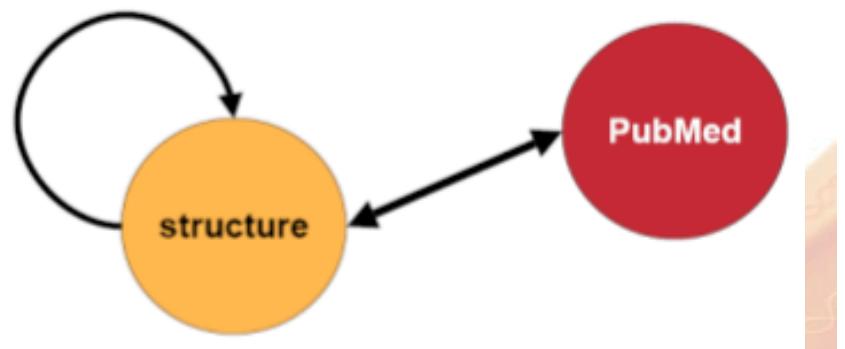
http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_example.pl

Entrez: A Discovery System



Following Links

**Follow links to related data
in the same database
or in others!**



**“Hard” Links: *Curated links based on biology*
for example:**

nucleotide → Taxonomy (based on organism identifier)

protein → Domain relatives (based on domain assignment)

domains → PubMed (based on supporting literature)

**“Soft” Links: *Pre-computed analyses*
for example:**

nucleotide → Related sequences (BLASTn neighbors)

protein → Conserved domains (RPS-BLAST search)

gene → Map viewer (map position of annotated gene)

Precomputed Services

Take a look at protein [CAA18341](#)

Some Entrez Links:

A List of Similar Sequences:

[Related Sequences](#)

Structures with Similar Sequences:

[Related Structures](#)

The Multifunctional Blast Link:

[Blink](#)

Gene Records

[Gene](#)

CDD (*Functional Domains*)

[Conserved Domains](#)

Taxonomic information

[Taxonomy](#)

Domain Relatives

[Domain Relatives](#)

What is BLink

- BLink stands for "BLAST Link"
- Foundation is protein sequence similarity information derived from protein BLAST search
- It is a link available for existing protein record
- It displays the proteins that most similar to that entry
- In addition, it also contains a collection of precomputed results/links to external data pertaining to that given protein sequence

How Does NCBI Generate the BLINK Data

- NCBI performs:
 - Scheduled BLAST searches against other existing proteins
 - for **every protein** record released into Entrez Protein database
 - Results are saved for "Related Sequences" Entrez links
- Additional information also saved to form the complete dataset for Blink
 - Taxonomic distribution
 - 3D structure matches
 - Conserved domain matches

Searching for Sequences: Hard and Soft Links

- Go to Entrez
 - Do a search for “U41100.1”
 - Click on nucleotide database
 - View the GenBank FlatFile
 - Investigate the **hard links** available from this nucleotide sequence record
 - » Full Text in PMC
 - » PubMed
 - » Protein
 - » Taxonomy
 - Investigate the **soft links**
 - » PubMed (Weighted) - identified using a word weight algorithm- similar words in their title, abstracts and Medical Subject Headings
 - » Related Sequences

An Easier Way: Entrez Protein Record

Display Settings: GenBank

Homo sapiens carcinoembryonic antigen-related cell adhesion molecule 3 (CEACAM3), mRNA

NCBI Reference Sequence: NM_001815.2

FASTA Graphics

Go to:

LOCUS NM_001815 1151 bp mRNA linear PRI 13-AUG-2011

DEFINITION Homo sapiens carcinoembryonic antigen-related cell adhesion molecule 3 (CEACAM3), mRNA.

ACCESSION NM_001815

VERSION NM_001815.2 GI:121114299

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

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

REFERENCE 1 (bases 1 to 1151)

AUTHORS Buntru,A., Kopp,K., Voges,M., Frank,R., Bachmann,V. and Hauck,C.R.

TITLE Phosphatidylinositol 3'-kinase activity is critical for initiating the oxidative burst and bacterial destruction during CEACAM3-mediated phagocytosis

JOURNAL J. Biol. Chem. 286 (11), 9555-9566 (2011)

PUBMED 21216968

REMARK GeneRIF: the ability of CEACAM3 to coordinate signaling events that not only mediate bacterial uptake, but also trigger the killing of internalized pathogens.

REFERENCE 2 (bases 1 to 1151)

AUTHORS Rose,J.E., Behm,F.M., Drgon,T., Johnson,C. and Uhl,G.R.

TITLE Personalized smoking cessation: interactions between nicotine dose, dependence &

JOURNAL Mol. Med. 16 (1), 1-10 (2010)

PUBMED 20379614

REMARK GeneRIF: Clustering of smoking cessation genes based on gene-environment interaction. (HUGO Navigator)

REFERENCE 3 (bases 1 to 1151)

AUTHORS Stanke,F., Becker,T., Hedtfeld,S., Tamm,S., Wienker,T.F. and Tummler,B.

TITLE Hierarchical fine mapping of the cystic fibrosis modifier locus on 19q13 identifies an association with two elements near the genes CEACAM3 and CEACAM6

JOURNAL Hum. Genet. 127 (4), 383-394 (2010)

PUBMED 20047061

REMARK GeneRIF: CEACAM6 and a regulatory element near the 3' end of CEACAM3 are associated with cystic fibrosis disease severity and intrapair discordance

REFERENCE 4 (bases 1 to 1151)

AUTHORS Tsavaris,N., Kosmas,C., Papadoniou,N., Kopteridis,P., Tsigritis,K., Dokou,A., Sarantinis,J., Skopelitis,H., Tzivras,M., Gennatas,K., Polyzos,A., Papastratis,G., Karatzas,G. and Papalambros,A.

TITLE CEA and CA-19.9 serum tumor markers as prognostic factors in patients with locally advanced (unresectable) or metastatic pancreatic adenocarcinoma: a retrospective analysis

J. Clin. Oncol. 27 (6), 673-680 (2009)

PUBMED 20071292

REMARK GeneRIF: CA-19.9 serum tumor marker levels retain independent prognostic value for poor survival in patients with advanced pancreatic cancer.

REFERENCE 5 (bases 1 to 1151)

AUTHORS Ali,C.W., Kaye,T.F., Adamson,D.J., Tait,I.S., Polignano,F.M. and Highley,M.S.

TITLE CA 19-9 and survival in advanced and unresectable pancreatic adenocarcinoma and cholangiocarcinoma

J. Gastrointest. Cancer 38 (2-4), 108-114 (2007)

PUBMED 19089662

REMARK GeneRIF: Increased CA 19-9 level is associated with pancreatic adenocarcinoma and cholangiocarcinoma.

Discovery Column

Send: Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Find in this Sequence

Articles about the CEACAM3 gene

Phosphatidylinositol 3'-kinase activity is critical for initiating the oxidative burst and bacterial destruction [J Biol Chem. 2011]

Personalized smoking cessation: interactions between nicotine dose, dependence [Mol Med. 2010]

CEA and CA-19.9 serum tumor markers as prognostic factors in patients w [J Chemother. 2009]

See all...

Reference sequence information

RefSeq protein product

See the reference protein sequence for carcinoembryonic antigen-related cell adhesion molecule 3 precursor (NP_001806.2).

More about the CEACAM3 gene

This gene encodes a member of the family of carcinoembryonic antigen-related cell adhesion molecules (CEACAMs), which are used by several ba...

Also Known As: CD66D, CEA, CGM1, MGC11...

LinkOut to external resources

Ensembl [Ensembl]

CEACAM3(NM_001815) ORF [OriGene Technologies]

Order full-length cDNA clone [GeneCopoeia Inc.]

Gene Expression Assays [TaqMan® probe and primer sets...]

imaGenes [imaGenes]

Silencer® Select siRNAs [siRNAs from Ambion, Inc. An A...]

UCSC Genome Browser [UCSC Genome Browser]

protein and peptide [ExactAntigen/Labome]

others [ExactAntigen/Labome]

siRNA and shRNA [ExactAntigen/Labome]

cDNA clone [ExactAntigen/Labome]

antibody [ExactAntigen/Labome]

Analysis Tools

PubMed Citations

RefSeq

Additional Information

External Records

Taxonomy



Taxonomy Database

- Browser for the major divisions of living organisms
 - archaea, bacteria, eukaryota, and viruses
- Taxonomy information such as genetic codes

The screenshot shows the NCBI Taxonomy Browser interface. At the top, there's a navigation bar with links for Entrez, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. Below the navigation bar is a search bar with the text "Search for 4530" and dropdown options "as complete name" and "lock". A "Go" button and a "Clear" button are also present. To the right of the search bar, there's a "Display" dropdown set to "3" and a "levels using filter: none" dropdown. Below these are several checkboxes for different database components, with "BLAST" checked. The main content area displays the taxonomic lineage for the ID 4530, which is "Oryza sativa". The lineage is listed as follows:

Lineage (full): root; cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; Liliopsida; commelinids; Poales; Poaceae; BEP clade; Ehrhartoideae; Oryzeae; **Oryza**

Below this, there's a section for "Oryza sativa" (rice) with links for MapView, LinkOut, and BLAST page. It also includes a note: "Click on organism name to get more information." and two sub-links: "Oryza sativa Indica Group" (Indian rice) and "Oryza sativa Japonica Group" (Japanese rice), each with its own MapView, LinkOut, and BLAST page links.

Compiled Data at Taxonomy DB

- Subtree links
 - Entries linked to any taxa in this subtree of the taxonomy database
- Direct links
 - Entries linked directly to this node in the taxonomy database

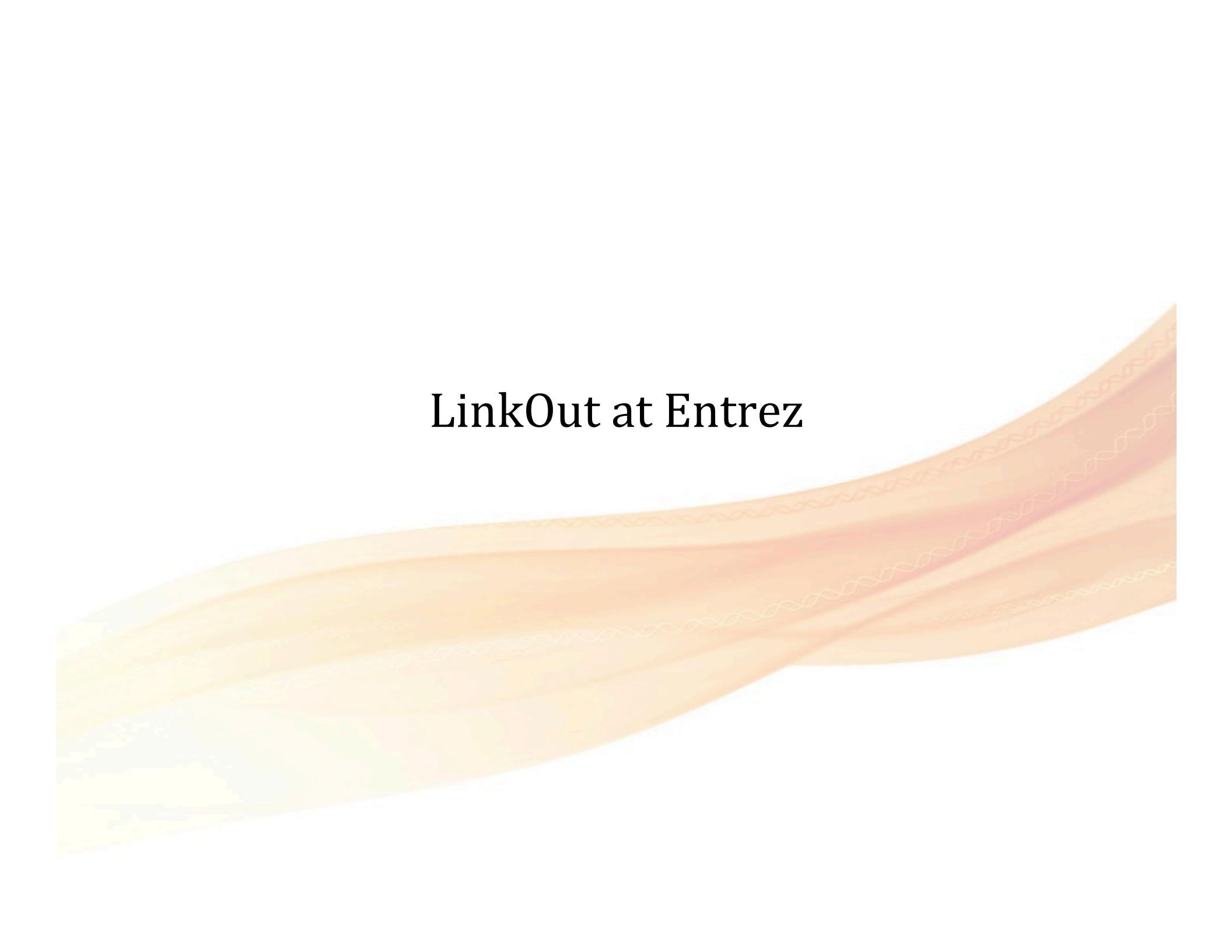
Oryza sativa

Taxonomy ID: 4530
Genbank common name: rice
Inherited blast name: monocots
Rank: species
Genetic code: [Translation table 1 \(Standard\)](#)
Mitochondrial genetic code: [Translation table 1 \(Standard\)](#)

Other names:
common name: red rice
authority: *Oryza sativa* L.

[Lineage \(full\)](#)
[cellular organisms](#); [Eukaryota](#);
[Viriplantae](#); [Streptophytina](#);
[Streptophytina](#); [Embryophytina](#);
[Tracheophytina](#); [Euphyllophyta](#);
[Spermatophytina](#); [Magnoliophytina](#);
[Liliopsida](#); [commelinids](#); [Poales](#);
[Poaceae](#); [BEP clade](#); [Ehrhartoideae](#);
[Oryzeae](#); [Oryza](#)

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	247,110	99,901
Nucleotide EST	1,253,557	62,231
Nucleotide GSS	500,493	3,769
Protein	281,241	17,717
Structure	108	34
Genome	1	1
Popset	854	754
SNP	11,023,172	11,023,172
Domains	19	11
GEO Datasets	7,300	4,928
UniGene	44,118	44,118
UniSTS	1,073	1,073
PubMed Central	9,517	9,517
Gene	84,179	1
HomoloGene	10,431	10,431
SRA Experiments	2,126	1,692
Probe	182,866	89,901
Assembly	6	-
Bio Project	452	298
Bio Sample	32,386	3,296
Bio Systems	1,130	661
GEO Profiles	670,939	670,939
PubChem BioAssay	106	70
Protein Clusters	12,990	-
Taxonomy	5	1

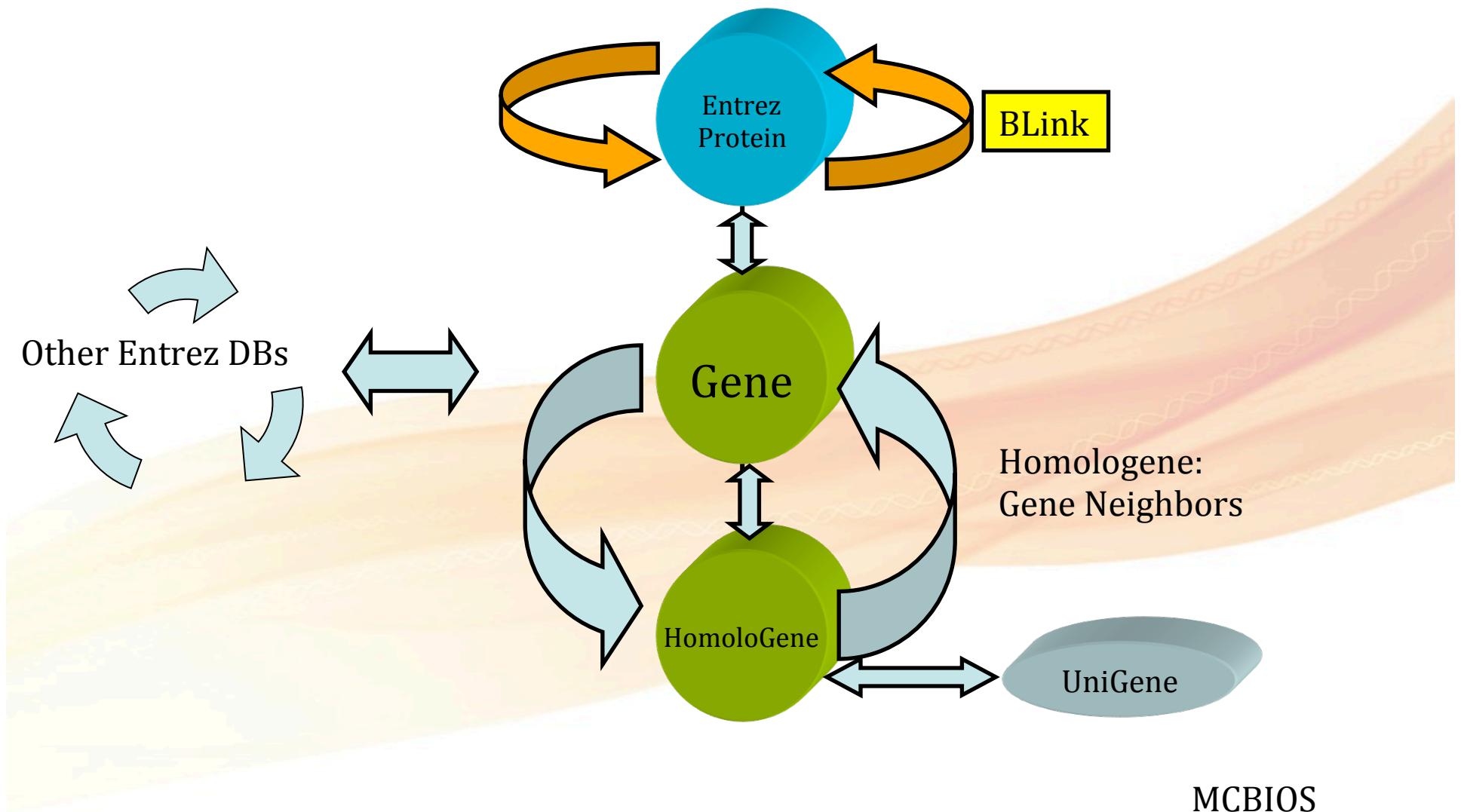


LinkOut at Entrez

LinkOut

- LinkOut is a service of Entrez:
 - Link directly from PubMed and other Entrez databases to a wide range of information and services beyond the Entrez system
 - Aims to facilitate access to relevant online resources in order to extend, clarify, and supplement information found in the Entrez databases
- Examples of LinkOut Resources include:
 - Full-text publications
 - Biological databases
 - Consumer health information
 - Research tools
 - and more

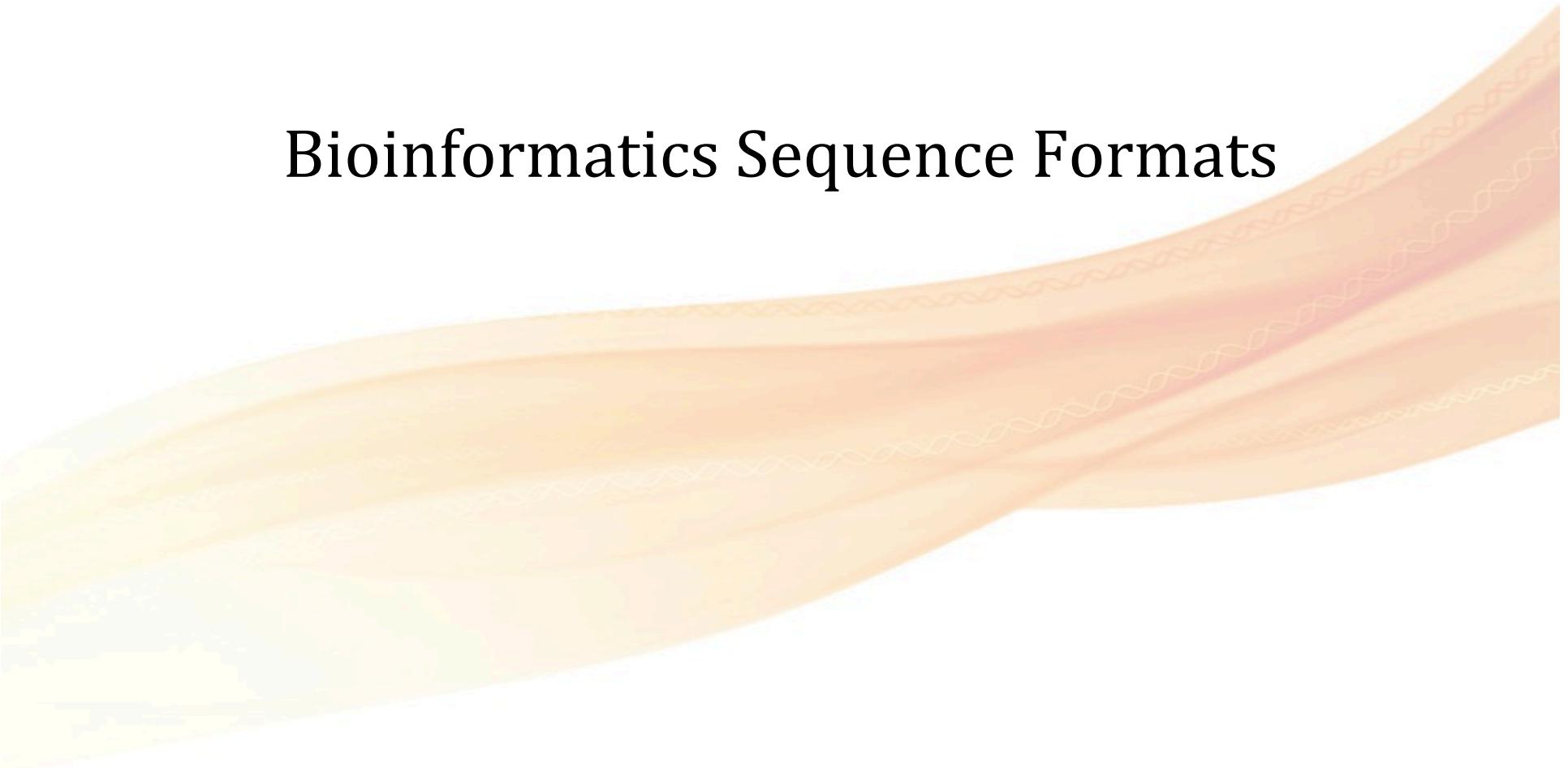
FYI - Entrez Tip: Start Searches in Gene



Where Else Can I get Help?

- Service Addresses
 - General Help info@ncbi.nlm.nih.gov
 - BLAST blast-help@ncbi.nlm.nih.gov
 - Telephone support: 301- 496- 2475

Bioinformatics Sequence Formats



Sequence Formats

- Microsoft WORD format is not a sequence format!!!
- Sequences can be read and written in a variety of formats
 - These can be very confusing for users
 - Ex. converting from one sequence editor to another
 - Now have your existing sequences in a format that is specific for that package
 - If you don't hold your sequence in a recognized standard format, you will not be able to analyze your sequence easily
- As a bioinformaticist you should be able to identify and use different formats

What a Sequence Format is NOT

- NOT any sort of program-specific format like a word processor format or text formatting language
 - Note: 'WORD', 'WORDPAD', 'PostScript', 'PDF', 'RTF', 'TeX', 'HTML'
- If you have somehow managed to type a sequence into a word-processor, you should:
 - # Save the sequence to a file as ASCII text (try selecting: File, Save As, Text)
 - # Stop using word-processors to write sequences
 - Investigate using simple text editor
 - like notepad Windows
 - Terminal or TextWrangler OSX

What a Sequence Format IS

- Sequence formats are ASCII TEXT
- They are the required arrangement of:
 - characters, symbols and keywords that specify:
 - What things such as the sequence, ID name, comments, etc. look like in the sequence entry
 - Where in the entry the program should look to find them
- There are generally no hidden, unprintable 'control' characters in any sequence format
- All standard sequence formats can be printed out or viewed simply by displaying their file (less or more)

Why So Many Formats?

- Dozens common sequence
- Some are much more common than others
- Formats were designed to be able to hold the sequence data and other information about the sequence
- Nearly every sequence analysis package has its own format
- Nearly every collection of sequences that dares call itself a database has stored its data in its own format

<http://emboss.sourceforge.net/docs/themes/SequenceFormats.html>

A Traditional GenBank Record

```
LOCUS      AF124527          2540 bp   mRNA    linear   PLN 29-JAN-2004
DEFINITION Prunus persica ethylene receptor (ETR1) mRNA, complete cds.
ACCESSION  AF124527
VERSION    AF124527.1  GI:6841074
KEYWORDS   .
SOURCE     Prunus persica (peach)
ORGANISM   Prunus persica
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
Spermatophyta; Magnoliophyta; eudicots; core eudicots;
rosids; eurosids I; Rosales; Rosaceae; Amygdaloideae; Prunus.
REFERENCE  1 (bases 1 to 2540)
AUTHORS   Bassett,C.L., Artilip,T.S. and Callahan,A.M.
TITLE     Characterization of the peach homologue of the ethylene receptor,
          PpETR1, reveals some unusual features regarding transcript
          processing
JOURNAL   Planta 215 (4), 679-688 (2002)
PUBMED    12172852
REFERENCE  2 (bases 1 to 2540)
AUTHORS   Bassett,C.B., Artilip,T.S. and Nickerson,M.L.
TITLE     Direct Submission
JOURNAL   Submitted (29-JAN-1999) Appalachian Fruit Research Station,
          USDA-ARS, 45 Wiltshire Road, Kearneysville, WV 25430, USA
FEATURES
  source
    Location/Qualifiers
    source
      1..2540
      /organism="Prunus persica"
      /mol_type="mRNA"
      /cultivar="Loring"
      /db_xref="taxon:3760"
      /dev_stage="III B/C fruit"
  gene
    1..2540
    /gene="ETR1"
  CDS
    269..2485
    /gene="ETR1"
    /codon_start=1
    /product="ethylene receptor"
    /protein_id="AAF28893.1"
    /db_xref="GI:6841075"
    /translation="MEACNCIEPQNPADELLMKYQYISDFFIALAYFSIPLLEIYFVK
KSAVFPYRVLVQFGAFITVLCGATHLINLWTFSMHRSRTVAIVMTTAKVLTAVVSCATA
ILMLVHIIPDLLSVKTRELFLXNKAAELDREMGHLIRTQEETGRHVRLMLTHEIRSTLDRH
TILKTTLVELGRTLALEZCALNMPTRTGLELQLSYTLRQQNPVGYYTVPIHLPVINQVF
SSNRALKISPNSPVARMERPLAGKHMPGEVAVRVPLLHLNSFQINDWPELSLKRYALM
VLMLFPSDSARQWHVHEZLVEVVADQVAVALSHAAILEESMRARDLIMEQNIALDLAR
REAETAIRARNDFLAVMNHEMPTPMHAIIALSSLQETELTPQRIMVETILKSSHLL
ATLINDVLDLSRLEDGSLQLEIATFNLHSVFREVHNLIKFWASVKKLSVSINLJAADLP
VQAVGDEKRILQGIVLNVVGNAVKFSKEGSISITAFVAKSESRLDFRAPEFFPAQSDNH
FYLRVQVVDGSGGINFQDIPKLFTKFAQTSLATRNSGGSGGLGLAICKRFVNLMEGHI
WIESEGPGKCTAIFIVKLGFQFAERSNESKLPLFLTKVQANHVVQTNFPLKVLVMDONGS
VTKGLLVHLGCDVTTVSSIDEFLHVVISQEHKVVFMDVCMEGIDGYELAVRIHEKPTKR
HERPVLVALTGNIDKMTKENCMRVGMDGVILKFPSVDKMRSVLSELLEHRVLFEAM"
ORIGIN
  1  gcacgagggc tcaccggagcg agctatgtct tcaggagtc aaggcttctgg gtgaggggaa
  61  gaagaagaag cttctttgtat gtgtgggtt gccaatctaa agaggaaagaa gaaggccct
  121  aatgtattga ggtcggtct ctggggctc gatctgtgtt gaatggatag tttggtagag
  181  atgcttcacaa gacatagggtt ggctgaaaaaag ggtttgaaga aagtggaaagg gaaaccacaa
  ...
  2401 tataactgaaa cctgtctcag ttgataaaaat gaggaggtt ttatcagaac tgttggaca
  2461  tcgagggtta tttagggctta tgtaagatat aggaaaatgtt ttctatgttggaa gggaaatgtt
  2521  aaatggaaaa aaaaaaaaaaa
//
```

Header

The Flatfile Format

Feature Table

Sequence

FASTA format (a.k.a. Pearson Format)

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAI PYIGTNLV
EWIWGGF SVDKATLNRRFAHFILPFTMVALGVHLTFLHETGSNNPLGLTSDDSKIPFHPYYTIKDFLG
LLILLLLLLALLSPDMLGDPDNHMPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTWIGSQVEPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

```
>sp_ac|P02769_WOSIGO \ID=ALBU_BOVIN \DE="Serum albumin precursor
RGVFRRDTHKSEIAHRFKDLGEHFKGVLIAFSQYLQQCPFDEHVVKLVNELTEFAKTCV
ADESHAGCEKSLHTLFGDELCKVASLRETYGDMADCCEKQEPERNECFLSHKDDSPDLPK
LKPDPTLCDEFKADEKKFWGKYLYEIARRHPFYAPELLYYANKYNGFQECQCQAEDKG
ACLLPKIETMREKVLAASSARQLRCASIQKFGERALKAWSVARLSQKFPKAEFVEVTKLV
TDLTKVHKCCCHGDLLECADDRADLAKYICDNQDTISSLKECCDKPLLEKSHCIAEVEK
DAIPENLPLTADFAEDKDVKNYQEAKDAFLGSFLYEYSRRHPEYAVSVLLRLAKEYEAT
TLEECCA KDDPHACYSTVFDKLKHLVDEPQNLIKQNCDQFEKLGEYGFQNALIVRYTRKV
PQVSTPTLVEVSRSRSLGKVGTCCRCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTK
CCTESLVNRRPCFSALTPDETYVPKA FDEKLFTFHADICLPLDTEKQIKKQTALVELLKH
KPKATEEEQLKTVMENFVAFVDKCCAADDKEACFAVEGPKLVVSTQ TALA
```

A sequence file in FASTA format can contain several sequences. With the **Header line**, beginning with '>', gives a name and/or unique identifier for the sequence, and other information as well

CLUSTAL W Format

CLUSTAL W (1.82) multiple sequence alignment

FOSB_MOUSE	MFOQAFPGDYDGSRCSSSPSAESQYLVSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA	60
FOSB_HUMAN	MFOQAFPGDYDGSRCSSSPSAESQYLVSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA	60

FOSB_MOUSE	ITTSQDLQWLQVQPTLISSMAQSQQPLASQPAPVDPYDMPGTSYSTPGLSAYSTGGASGS	120
FOSB_HUMAN	ITTSQDLQWLQVQPTLISSMAQSQQPLASQPVVDPYDMPGTSYSTPGMSGYSSGGASGS	120

FOSB_MOUSE	GGPSTTTSGPV SARPARARPRPREETLTPEEEKRRVRERNKLAAKCRNRRRELT	180
FOSB_HUMAN	GGPSTSGT TSGPGPARPARPRPREETLTPEEEKRRVRERNKLAAKCRNRRRELT	180

FOSB_MOUSE	DRLQAETDQLEEEKAELAEIAELQKEKERLEFVLVAHKPGCKI PYEEGP GPGLAEVRD	240
FOSB_HUMAN	DRLQAETDQLEEEKAELAEIAELQKEKERLEFVLVAHKPGCKI PYEEGP GPGLAEVRD	240

FOSB_MOUSE	LPGSTS AKE DGF GWLL PPPPPPPLPFQSSRDAPPNL TASL FTHSEVQ VLGDPFPV VSPSY	300
FOSB_HUMAN	LPGSAPAKEDGFSWLL PPPPPPPLPFQTSQDAPPNL TASL FTHSEVQ VLGDPFPV VNP SY	300

FOSB_MOUSE	TSSFVLT CPEVSAFAGAQR TGSEQPSDPLNSPSLLAL	338
FOSB_HUMAN	TSSFVLT CPEVSAFAGAQR TGSDQPSDPLNSPSLLAL	338

1. First line in the file must start with the words "CLUSTAL W". Other information in the first line is discarded
2. One or more empty lines
3. One or more blocks of sequence data. Each block consists of:
 - * One line for each sequence in the alignment. Each line consists of:
 1. the sequence name
 2. white space
 3. up to 60 sequence symbols
 4. optional - white space followed by a cumulative count of residues for the sequences
 - * A line showing the degree of conservation for the columns of the alignment in this block
 - * One or more empty lines.

Phylip Format

```
4 131
IXI_234 TSPASIRPPA GPSSRPAMVS SRRTRPSPPG PRRPTGRPCC SAAPRRPQAT
IXI_235 TSPASIRPPA GPSSR----- RPSPPG PRRPTGRPCC SAAPRRPQAT
IXI_236 TSPASIRPPA GPSSRPAMVS SR--RPSPPP PRRPPGRPCC SAAPPRPQAT
IXI_237 TSPASLRPPA GPSSRPAMVS SRR-RPSPPG PRRPT----C SAAPRRPQAT

GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRK MRAACSRSA G
GGWKTCSGTC TTSTSTRHRG RSGW----- RASRK MRAACSRSA G
GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRK MRAACSR--G
GGYKTCGTC TTSTSTRHRG RSGYSARTTT AACLRASRK MRAACSR--G

SRPNRFAPTL MSSCITSTTG PPAWAGDRSH E
SRPNRFAPTL MSSCITSTTG PPAWAGDRSH E
SRPPRFAPPL MSSCITSTTG PPPPAGDRSH E
SRPNRFAPTL MSSCLTSTTG PPAYAGDRSH E
```

The PHYLIP format came from Joe Felsenstein's phylogeny inference package and is now used by several phylogenetics programs. PHYLIP file names often have have a .phy or .ph extension.

PIR Format

```
>P1; Cow
Cow
MAYPMQLGFQDATSPIEELLHFHDHTLMIVFLISSLVLYIISLMLTTKLHTSTMADAQEVE
TIWTLPAIILILI
ALPSLRILYMMDEINNPSLTVKTMGHQWYWSYEYTDYEDLSFDSYMIPTSELKPGELRLLEV
DNRVVLPMEMTIRM
LVSSEDVLHSWA VPSLGLKTD AIPGRLNQ TLMSSRPGLYYGQCSEICGSNHSFMPIVLE
VPLKYFEKWSASML-
-----*
>P1; Carp
Carp
MAHPTQLGFKD AAMPVMEELLHFHDHALMIVLLISTLVLYIITAMVSTKLTNKYILD
SQEIEIWWTILPAVILVLI
ALPSLRILYLMDEINDPHLTIKAMGHQWYWSYEYTDYENLGFD SYMVPTQDLAPGQFR
LLET DHRMVPVMESPVRV
LVSAEDVLHSWA VPSLGVKMDA VPGRLNQ AAFIASRPGVFY GQCSEICGANHSFMPIV
VEAVPLEHFENWSSLME
DASLGS*
>P1; Chicken
Chicken
MANHSQ LGFQDASSPIEELVEFHDHALMVALAICSLVLYLLTLM MEKLS-SNTVDAQEVELI
WTLPAIVLVL
ALPSLQI LYMMDEIDE PDLTLKAIGHQWYWTY EYTD FKDL SFDSYMTT DPLGHFR
LLEV DHRIVIPMESPIRV
I ITADDVLHSWA VPALGVKTDA I PGRLNQ TSFITTRPGVFY GQCSEICGANHSYMP
IVVESTPLKHF
EAWSL
---LSS*
```

Protein Information Resource (PIR) is an annotated, non-redundant and cross-referenced database of protein sequences at the NBRF. PIR file names often have a .pir extension.

The header line of a nucleotide sequence file in pir-format begins with a greater than sign ">" followed by DL.

The header line of protein sequence file in pir-format begins with a greater than sign ">" followed by P1.

EMBL Format

```
ID  AB000263 standard; RNA; PRI; 368 BP.
XX
AC AB000263;
XX
DE Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ Sequence 368 BP;
acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg      60
ctgccttgc cctggagggt ggccccaccc gccgagacag cgagcatatg caggaagcgg      120
caggaataag gaaaagcage ctcctgactt tcctcgctt gtggtttgag tggacctccc      180
aggccagtgc cggccccctc ataggagagg aagctcgga ggtggccagg cggcaggaag      240
gcgcacccccc ccagaatcc gcgcgcggg acagaatgcc ctgcaggaac ttcttctgga      300
agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca      360
gacctgaa
//
```

A sequence file in EMBL format can contain several sequences.

One sequence entry starts with an identifier line ("ID"), followed by further annotation lines. The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").

The EMBL flat format is a rich format for storing sequences and their associated meta-information, feature coordinates, and annotations. It shares details with the GenBank sequence format.

```

#NEXUS
[TITLE: NoName]

begin data;
dimensions ntax=3 nchar=384;
format interleave datatype=protein    gap=- symbols="FSTNKEYVQMCLAWPHDRIG";

matrix
CYS1_DICDI      ----MKVIL LFVLAVFTVF VSS----- -----RG IPPEEQ-----
ALEU_HORVU      MAHARVLLA LAVLATAAVA VASSSSFADS NPIRPVTDRA ASTLESAVLG
CATH_HUMAN      -----MWAT LPLLCAGAWL LGV----- -PVCGAAELS VNSLEK-----

CYS1_DICDI      -----SQ FLEFQDKFNK KY-SHEEYLE RFEIFKSNLG KIEELNLIAI
ALEU_HORVU      ALGRTRHALR FARFAVRYGK SYESAEEVRR RFRIFSESLE EVRSTN-----
CATH_HUMAN      -----FH FKSWMSKHRK TY-STEEYHH RLQTFASNWR KINAHN-----

CYS1_DICDI      NHKADTKFGV NKFDLSSDE FKNYYLNNKE AIFTDDLPVA DYLDDEFINS
ALEU_HORVU      RKGLPYRLGI NRFSDMSWEF QOATRL-GAA QTCSATLAGN HLMRDA--AA
CATH_HUMAN      NGNHTFKMAL NQFSDMSFAE IKHKYLWSEP QNCSAT--KS NYLRGT--GP

CYS1_DICDI      IPTAFDWTR G-AVTPVKNQ GOCGSCWSFS TTGNVEGQHF ISQNKLVSLS
ALEU_HORVU      LPETKDWRED G-IVSPVKNQ AHCGSCWTFS TTGALEAAYT QATGKNISLS
CATH_HUMAN      YPPSVDWKK GNFVSPVKNQ GACGSCWTFS TTGALESIA IATGKMLSLA

CYS1_DICDI      EQNLVDCDHE CMEYEGEREAC DEGCNGGLQP NAYNYIICKNG GIQTESSYPY
ALEU_HORVU      EQQLVDCAGG FNNF----- --GCNGGLPS QAFEYIKYNG GIDTEESYPY
CATH_HUMAN      EQQLVDCAQD FNNY----- --GCQGGLPS QAFEYILYNK GIMGEDTYPY

CYS1_DICDI      TAETGTQCNF NSANIGAKIS NFTMIP-KNE TVMAGYIVST GPLAIAAADAV
ALEU_HORVU      KGVNNG-CHY KAENAABQVL DSVNITLNAE DELKNAVGL RPVSVAFQVI
CATH_HUMAN      QGKDGY-CKF QPGKAIGFVK DVANITIYDE EAMVEAVALY NPVSFAFEVT

CYS1_DICDI      E-WQFYIGGV F-DIPCN--P NSLDHGILIV GYSAKNTIFR KNMPYWIVKN
ALEU_HORVU      DGFRQYKSGV YTSDHCGTPP DDVNHAVLAV GYGVENGV-- ---PYWLKLN
CATH_HUMAN      QDFMMYRTGI YSSTSCHKTP DKVNHAVLAV GYGEKNGI-- ---PYWIVKN

CYS1_DICDI      SWGADWGEQG YIYLRRGKNT CGVSNFVSTS II--
ALEU_HORVU      SWGADWDNG YFKMEMGKNM CAIATCASYP VVAA
CATH_HUMAN      SWGPQWGMNG YFLIERGKNM CGLAACASYP IPLV

```

Nexus Format

This is the file format used by many popular programs like GARLI, GDA, MacClade, Mesquite, ModelTest, MrBayes and PAUP*. Nexus file names often have a .nxs or .nex extension

GFF Format

- Latest version GFF3
- It's a tabular plain-text format for genome or sequence annotation
- Can contain also the sequences, alignments, and dependencies between features
- Is extensible
- One of the most recommended & “standard” format

```
##gff-version 3
ctg123 . operon      1300 15000 . + . ID=operon001;Name=superOperon
ctg123 . mRNA        1300 9000  . + . ID=mrna0001;Parent=operon001;Name=sonichedgedehog
ctg123 . exon         1300 1500  . + . Parent=mrna0001
ctg123 . exon         1050 1500  . + . Parent=mrna0001
ctg123 . exon         3000 3902  . + . Parent=mrna0001
ctg123 . exon         5000 5500  . + . Parent=mrna0001
ctg123 . exon         7000 9000  . + . Parent=mrna0001
ctg123 . mRNA        10000 15000 . + . ID=mrna0002;Parent=operon001;Name=subsonicsquirrel
ctg123 . exon         10000 12000 . + . Parent=mrna0002
ctg123 . exon         14000 15000 . + . Parent=mrna0002
```

FASTQ

- FASTQ – FASTA “with an attitude” (embedded quality scores)
- Originally developed at the Sanger to couple (Phred) quality data with sequence
- Now common to specify raw read output data from NGS machines in this format
- Various flavors:
 - fastq-sanger
 - fastq-illumina
 - fastq-solexa
- “...the Sanger version of the FASTQ format has found the broadest acceptance, supported by many assembly and read mapping tools ...Therefore, most users will do this conversion very early in their workflows...”



```
@NCYC361-11a03.q1k bases 1 to 1576  
GCGTGCCCGAAAAAATGCTTTGGAGCCGCGCGTGAAT...  
+NCYC361-11a03.q1k bases 1 to 1576  
! ))))))****(( (**%(( (((*(((+, **(( (+**+, -...
```

http://en.wikipedia.org/wiki/FASTQ_format

<http://maq.sourceforge.net/fastq.shtml>

<http://nar.oxfordjournals.org/content/early/2009/12/16/nar.gkp1137.full>

VCF (Variant Call Format) Files

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:2
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7
20 1234567 microsat1 GTCT G,GTACT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4
```

FYI - Sequence Formats

ASN.1
DNAStrider
EMBL
Fitch
GCG
GenBank/GB
IG/Stanford
MSF
NBRF
Olsen
PAUP/NEXUS
Pearson/**Fasta**
Phylip
PIR/CODATA
Plain/Raw
Pretty
Zuker

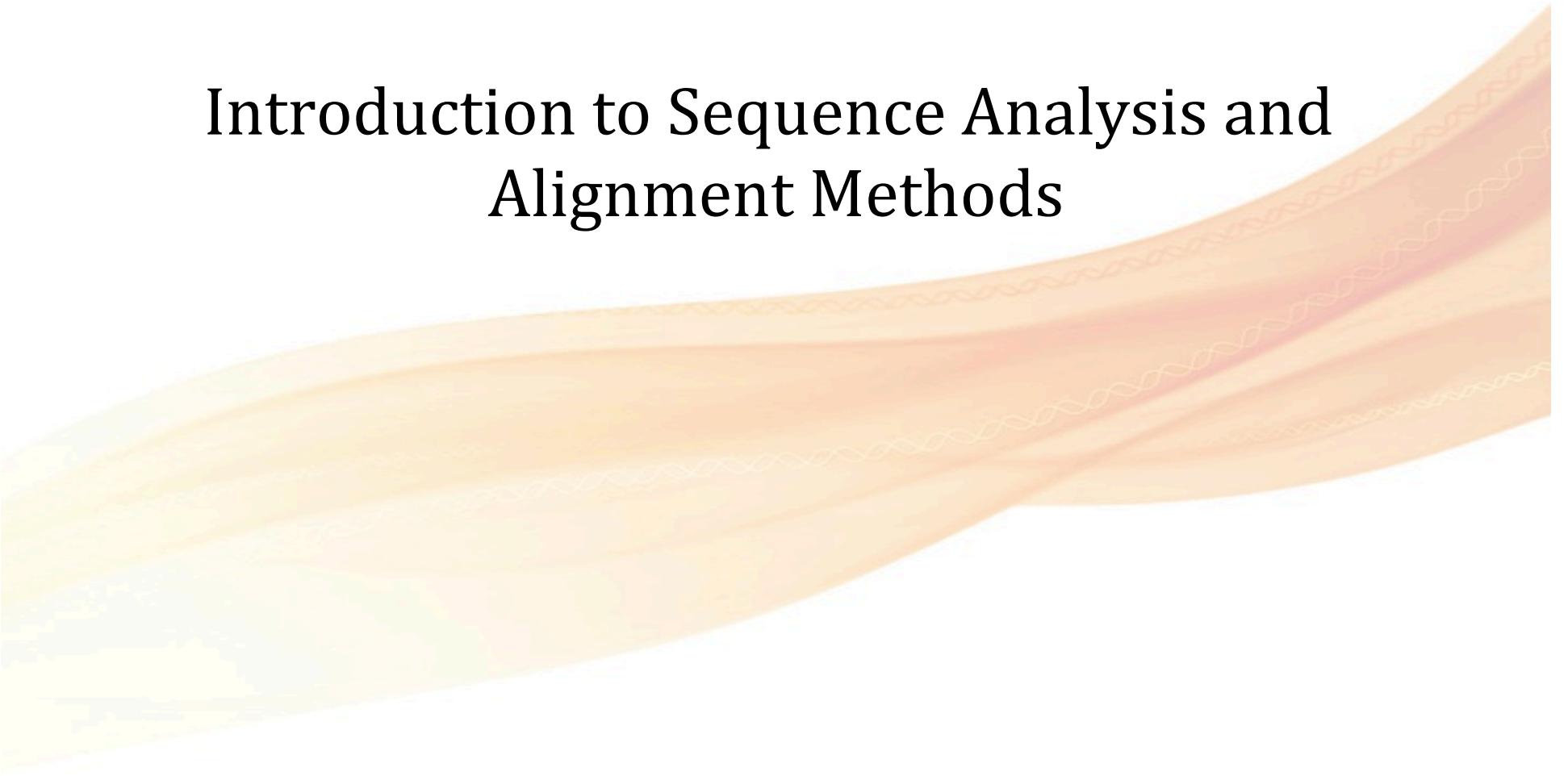
- Convert sequence formats
 - ReadSeq Web based
 - <http://www.ebi.ac.uk/cgi-bin/readseq.cgi>
 - SeqVerter (stand alone app Windows)
 - <http://www.genestudio.com/seqverter.htm>
- NOTE:
 - FASTA is a popular sequence format
 - It is also a sequence similarity tool used by EMBL-EBI and NCBI

Now we know where to get sequence data,
and the formats, so what's next?

5 minutes



Introduction to Sequence Analysis and Alignment Methods



Pairwise Sequence Alignment Methods

- Primary methods of producing pairwise alignments are:
 - Manual & Dot plot
 - Automated Methods
 - Dynamic programming
 - Word Methods
 - FASTA
 - BLAST

Why Align Sequences?

- Many sequences:
 - Have no known ancestry, structure, or function
 - Some of them do
- If they align (with meaning), they are similar
 - If they are similar:
 - They might have the same:
 - ancestry
 - similar structure
 - function
 - **If one of them has known ancestry, structure, or function, then alignment to the others yields insight about them**

Sequence Alignment

- In bioinformatics
 - A Sequence alignment - way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity
 - Similarities may be a consequence of:
 - functional
 - structural
 - or evolutionary relationships between sequence
 - molecular evolution

AAB24882	TYHMCQFHCRYVNNHSGEKL <ins>KLYECNERSKAFSCP</ins> SHLQCHKRRQIGEKTHEHNQCGKAFPT	60
AAB24881	----- <ins>YE</ins> CNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK	40

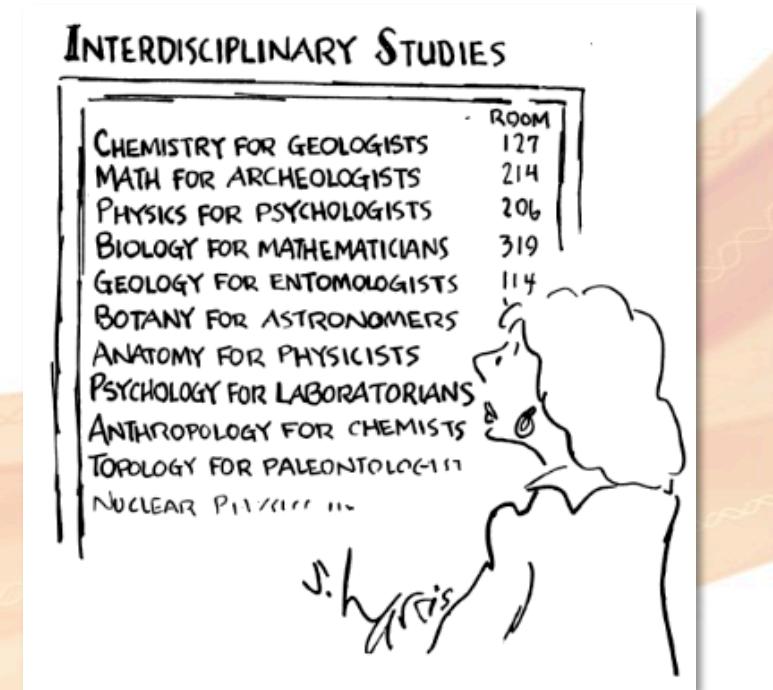
*****: . ***: * * :** * :****. :* **** * ..

AAB24882	PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ-	116
AAB24881	HSHLQCHKR <ins>THTGEKPYECNQCGKAFSQH</ins> LLQRHKRTHTGEKPYMNVINMVKPLHNS	98

***** * :*****:*****:***:***. : . *****:***** : * . : :

Molecular Evolution is an Interdisciplinary Endeavor

- **The discipline of Molecular Evolution deals only with events occurring after the emergence of biological systems that possess replicable genetic material**
- **Deals with two subjects**
 - The evolution of:
 - molecular entities, e.g., genes, genomes, proteins, introns, chromosomal arrangements
 - organisms and biological complexes as deduced from molecular data
- How is this done?
- Must perform Sequence Analysis



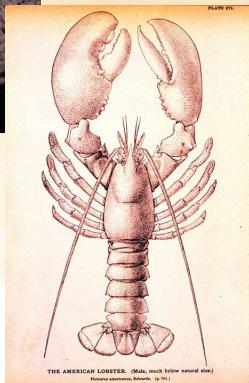
Adopted from Dan Graur, Ph.D.

Sequence Alignment

- One of the most common bioinformatics tasks
 - We will comprehensively learn about:
 - The different methods
 - Why we use BLAST and not dynamic programming for large datasets
 - How a simple dot plot provides information
 - Other things as well
- Remains **large area of research**
 - One of the major topics in bioinformatics
 - Present in almost every R&D activity:
 - Across the many industries in the area of life sciences
 - » Academia, biotech, services, software, pharma, and hospitals

End Results

- Analysis of **organismal** diversity/similarity



**ARMADILLIDIUM
CARCINUS
HOMARUS**

RIFDTSCKGFYDRGLFAQLDRCEDCYNLYRKPH--
 QIYDTSCKGVYDRALF-DLEHVCDDCYNLYRTSYVA
 QVFDQACKGVYDRNLFKLDRVCEDCYNLYRKPFVA
 * *** *** ** * *** *** ***

**ARMADILLIDIUM
CARCINUS
HOMARUS**

AECRRDCYTTEVFESCLKDLMMHDFINEYKEMALMV
SACRSNCYSNLVFRQCMDDLLMMDEFDQYARKVQMV
TTCRENCYSNWVFRQCLDDLLL-NVIDEYVSNVQM-
 ** ** ** * ** * * * * *

images from Dan Graur, Ph.D.

Sequence Similarity Searching

- Availability of increasingly expanding databases poses a major challenge to bioinformatics experts:
 - Must develop effective programs/Web servers to extract maximum information from these databases
 - Possibly the fastest, cheapest, and most powerful experiment a biologist can conduct
- As databases become more *complete*?
 - Searches more likely to reveal database sequences with statistically significant similarity
 - Thus inferred **homology** to a query sequence

Homology: General Definition

- Homology designates a qualitative relationship of common descent between entities
- Two genes are either **homologous** or they are **not!**
 - it doesn't make sense to say:
 - “two genes are 43% homologous”
 - “Linda is 43% pregnant”

Homology Terms

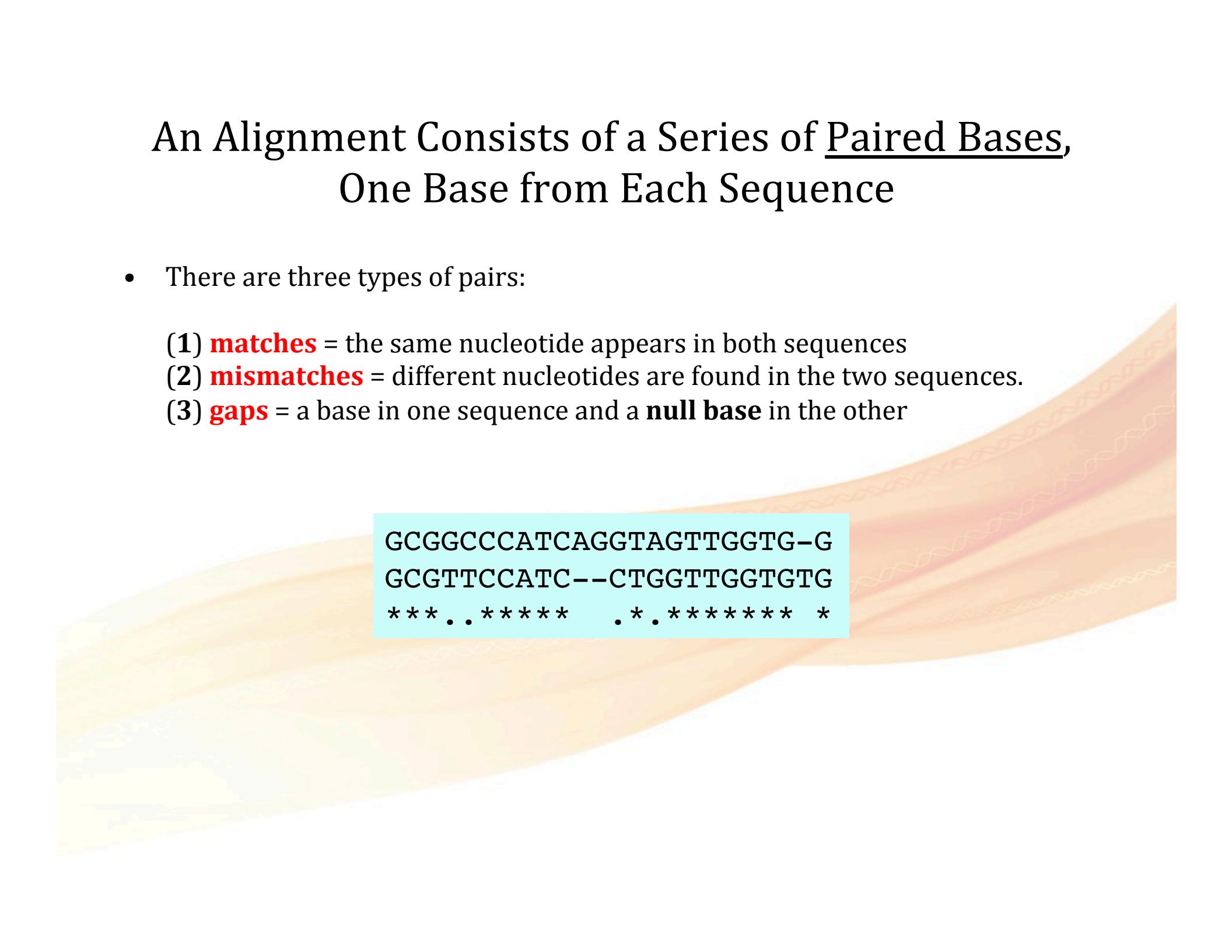
- Two genes are **orthologs**
 - Originated from a single ancestral gene in the most recent common ancestor of their respective genomes
- Two genes are **paralogs**
 - Related by gene duplication
 - Two genes are **ohnologs** if they are related by gene duplication due to whole-genome duplication (WGD)
- **Xenologs**
 - Homologs resulting from horizontal gene transfer between two organisms are termed
- When comparing sequences, we are interested in positional homology,
 - We identify positional homology through?
 - sequence alignment

Interpretation

- If two sequences share a common ancestor:
 - Mismatches can be interpreted as point mutations
 - Gaps as indels
 - These have been introduced in one or both lineages since they diverged from one another
- In proteins:
 - The degree of similarity between a.a. occupying a particular position in the sequence can be interpreted as:
 - A rough measure of how conserved a particular region or sequence motif is among lineages
 - Absence of substitutions (or conserved substitutions) suggests this region has structural or functional importance

An Alignment Consists of a Series of Paired Bases, One Base from Each Sequence

- There are three types of pairs:
 - (1) **matches** = the same nucleotide appears in both sequences
 - (2) **mismatches** = different nucleotides are found in the two sequences.
 - (3) **gaps** = a base in one sequence and a **null base** in the other



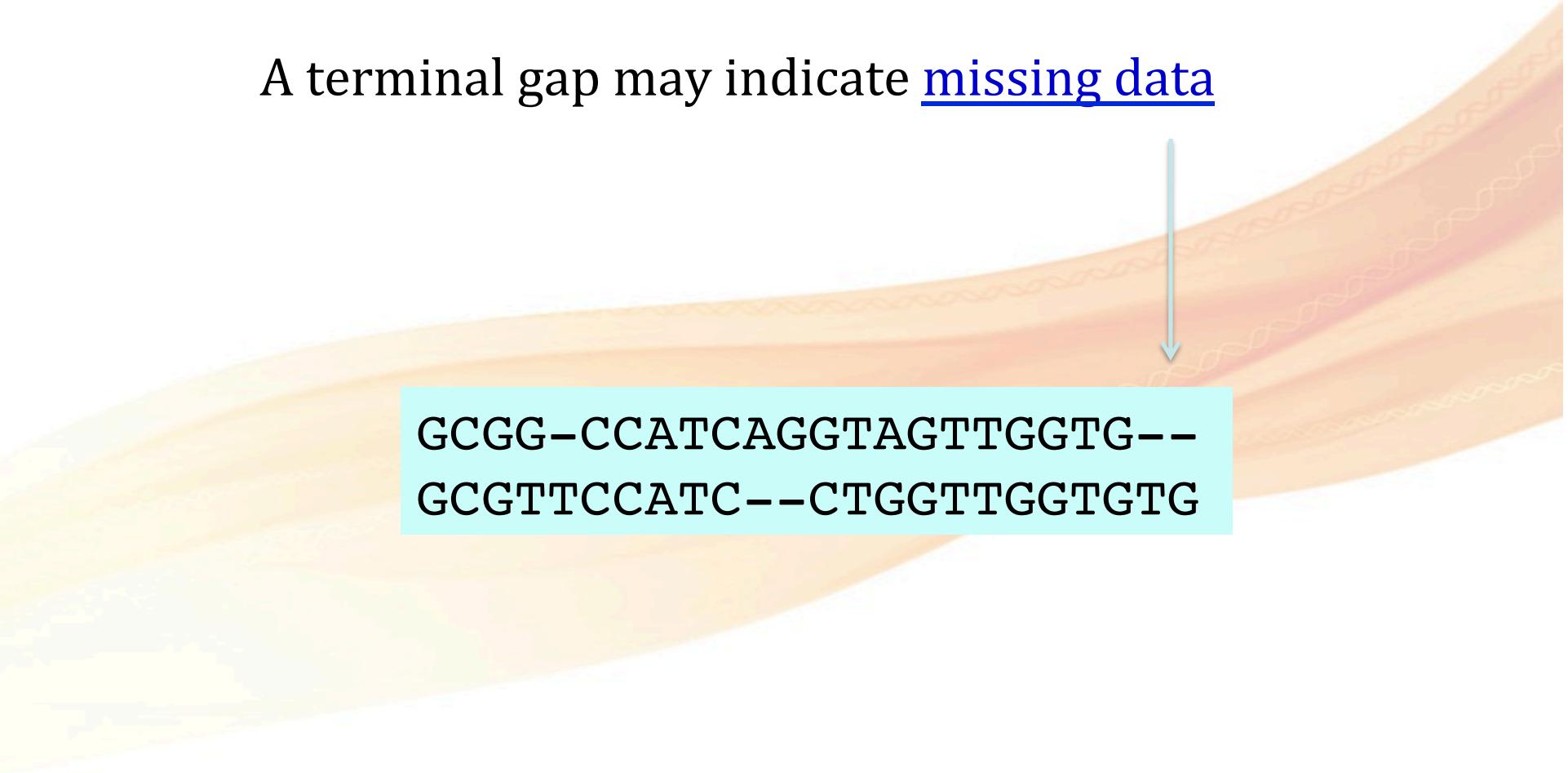
```
GCGGCCCATCAGGTAGTTGGTG-G  
GCGTTCCATC--CTGGTTGGTGTG  
*** .. ***** . * . ***** * *
```

There are **internal** and **terminal** gaps



```
GCGG-CCATCAGGTAGTTGGTG--  
GCGTTCCATC--CTGGTTGGTGTG
```

A terminal gap may indicate missing data



GCGG-CCATCAGGTAGTTGGTG--
GCGTTCCATC--CTGGTTGGTGTG

An internal gap indicates that a **deletion** or an **insertion** has occurred in one of the two lineages



```
GC GG-CCATCAGGTAGTTGGTG--  
GCGTTCCATC--CTGGTTGGTGTG
```

A Word of Caution

- The alignment is the first step in many evolutionary and functional studies
- Errors in alignments tend to be amplified in later computational stages
 - Analyze the output from a method
 - Garbage in -> Garbage out
 - Does output seem correct?
 - Save time by doing this!

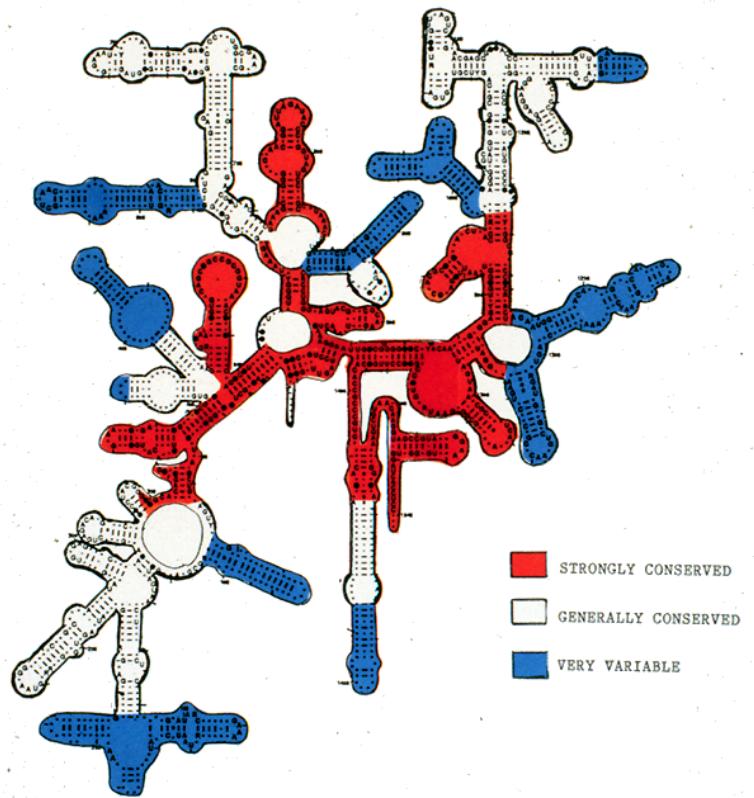
Manual Alignment

- When there are few gaps
- When two sequences are not too different from each other
- A reasonable alignment can be obtained by visual inspection

```
GCG-TCCATCAGGTAGTTGGTGTG  
GC GTTCCATCAGGTGGTTGGTGTG  
*** * * * * * * * . * * * * * *
```

Advantages of Manual Alignments

- Using a powerful and trainable tool, your brain!
- Ability to integrate additional data:
 - **Domain structures**
 - **Biological function**
 - **Your knowledge**
- But....
 - Too much to ask the brain to compare large, diverged sequences
 - The method is subjective and unscalable!

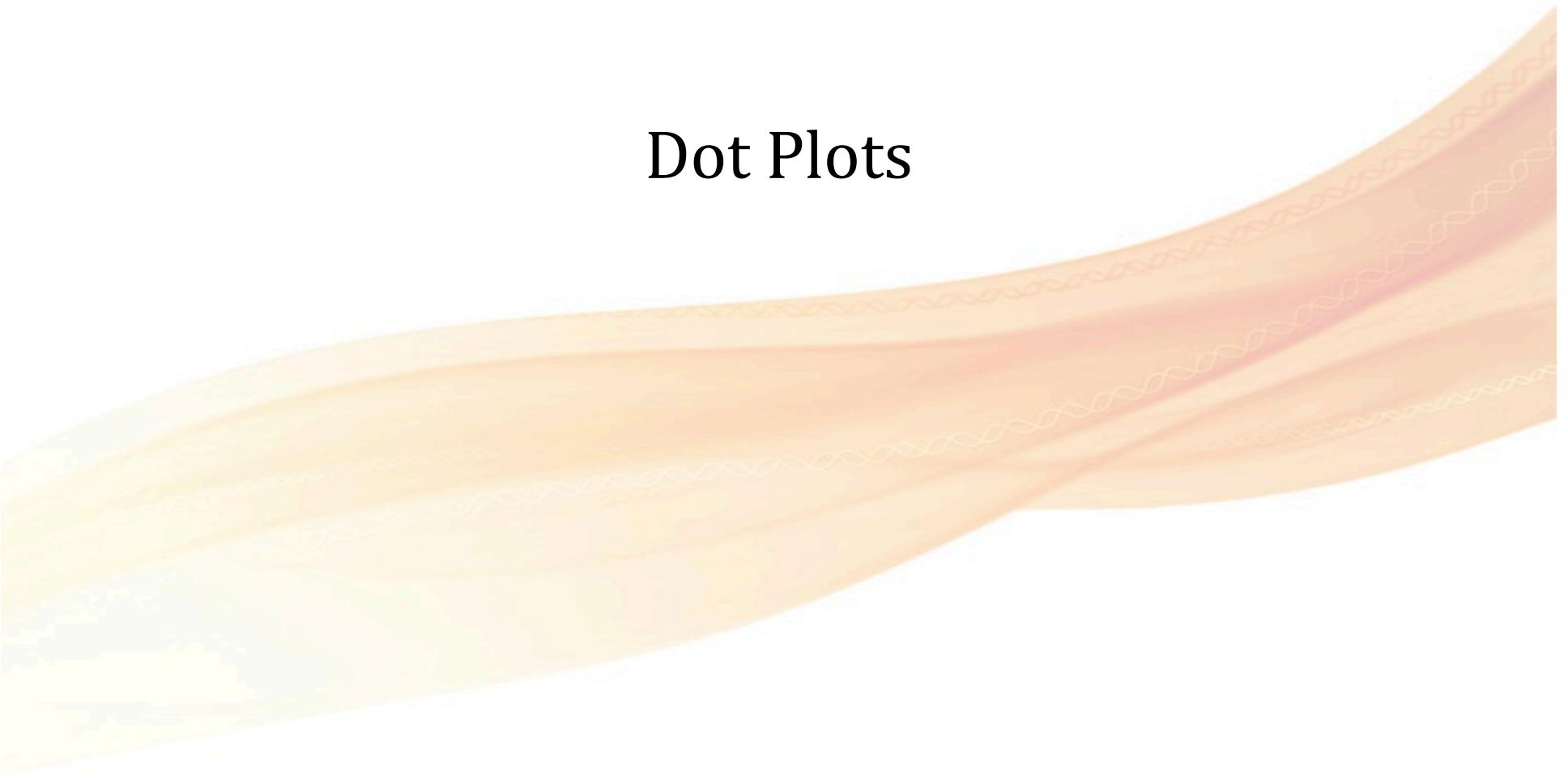


- Manually made ribosomal dna alignment – 2nd structure

- We can also visually inspect (manually) a Dot Plot

images from Dan Graur, Ph.D.

Dot Plots

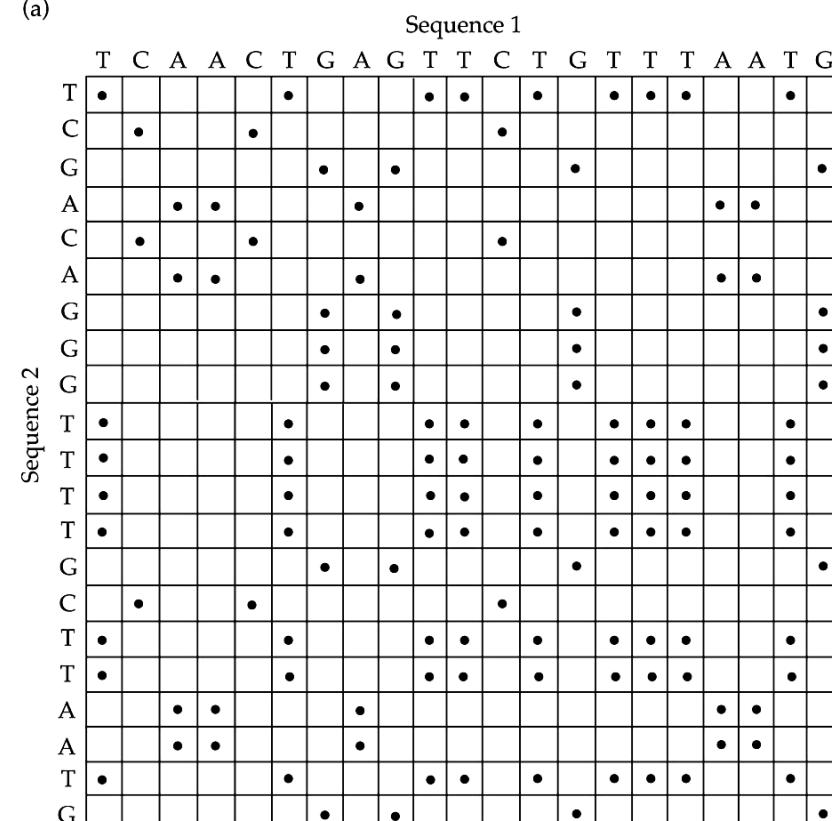


Dot Plot

- Also known as a Dot Matrix
- Contact plot or Residue contact map
- Dot plots are a simple way of seeing alignments
 - We really like to see good visual demonstrations, not just tables of numbers
 - It's a grid: put one sequence along the top and the other down the side, and put a dot wherever they match
- Locates regions of similarity between two DNA or protein sequences
- Implicitly produces a family of alignments for individual regions
 - Its both qualitative and simple
 - Visualize recombinatorial events
 - But time consuming to analyze on a large scale

Dot Plot

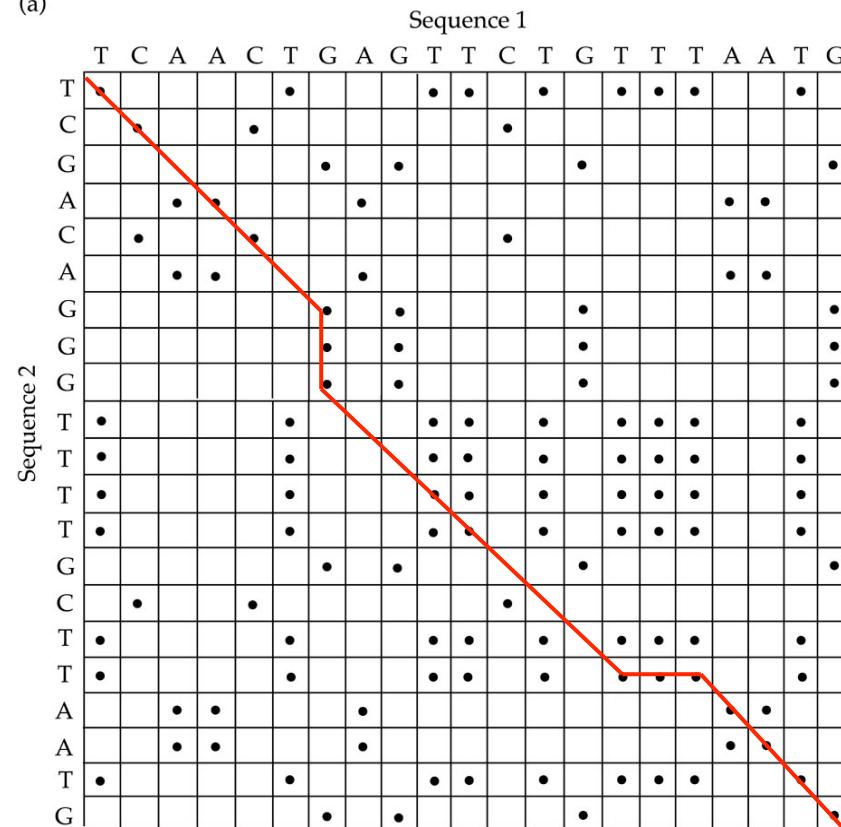
(a)



Gibbs and McIntyre, 1970

Dot Plot

(a)

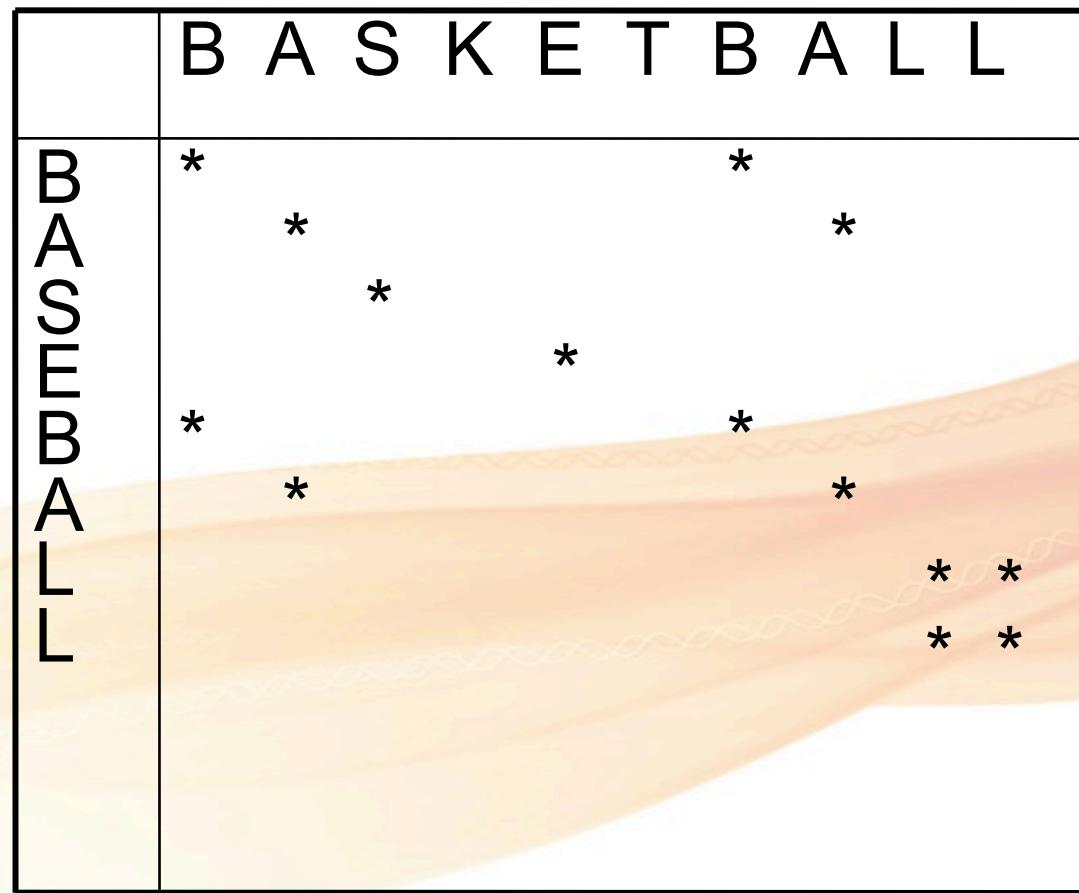


Gibbs and McIntyre, 1970

How to Generate a Dot Plot

- A Dot Plot is created by:
 - Designating one sequence to be the subject
 - Place it on the horizontal axis
 - Designate the second sequence to be the query
 - Place on the vertical axis of the matrix
- Each position within the matrix
 - A point is plotted if the horizontal and vertical elements are identical
 - Diagonal lines within the resulting matrix indicate regions of similarity

A Rather Simple Dot Plot



Dot Plot Visualization

- Using a Dot Plot, plot it's easy to visualize:
 - Insertions
 - Deletions
 - Repeats or Inverted Repeats
 - Dot plot of very closely related sequences will appear as a single line along the matrix's main diagonal
- Dot plots can also be used to assess repetitiveness in a single sequence?
 - Sequence plotted against itself
 - Regions sharing significant similarity will appears as lines off the main diagonal
 - Occur in proteins which consists of multiple similar structural domains

Dot Plot with Noise Reduction

- A certain percentage of the matches between sequence elements can be expected to be random?
 - Random matches are considered “noise”
 - Need to be filtered out to enhance the diagonal lines “signal”
- How to reduce noise in a Dot Plot:
 - Center a substring of elements of the query sequence over each element in the subject sequence
 - Determining the number of corresponding elements within this “window”
 - If the number of corresponding elements \geq a specified (“stringency”) threshold then a point is plotted for the center element

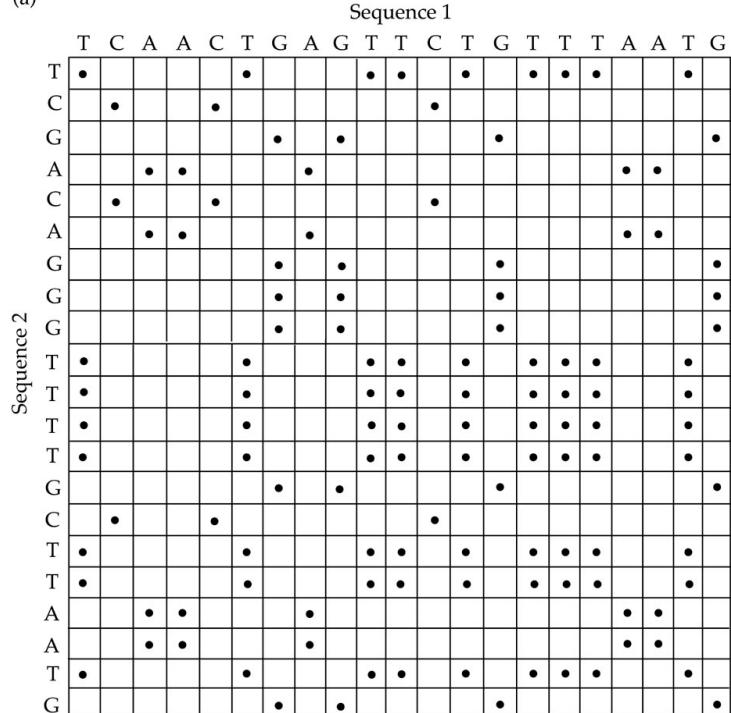
Dot Plot w/ Simple Example Filtering



window size = 5
stringency = 3

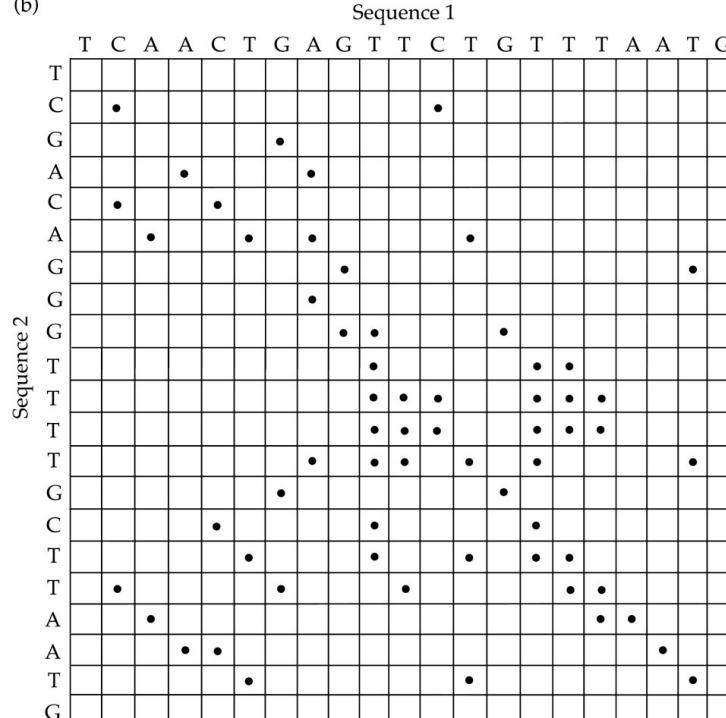
Dot Plot w/ Actual Filtering

(a)



window size = 1
stringency = 1

(b)



window size = 3
stringency = 2

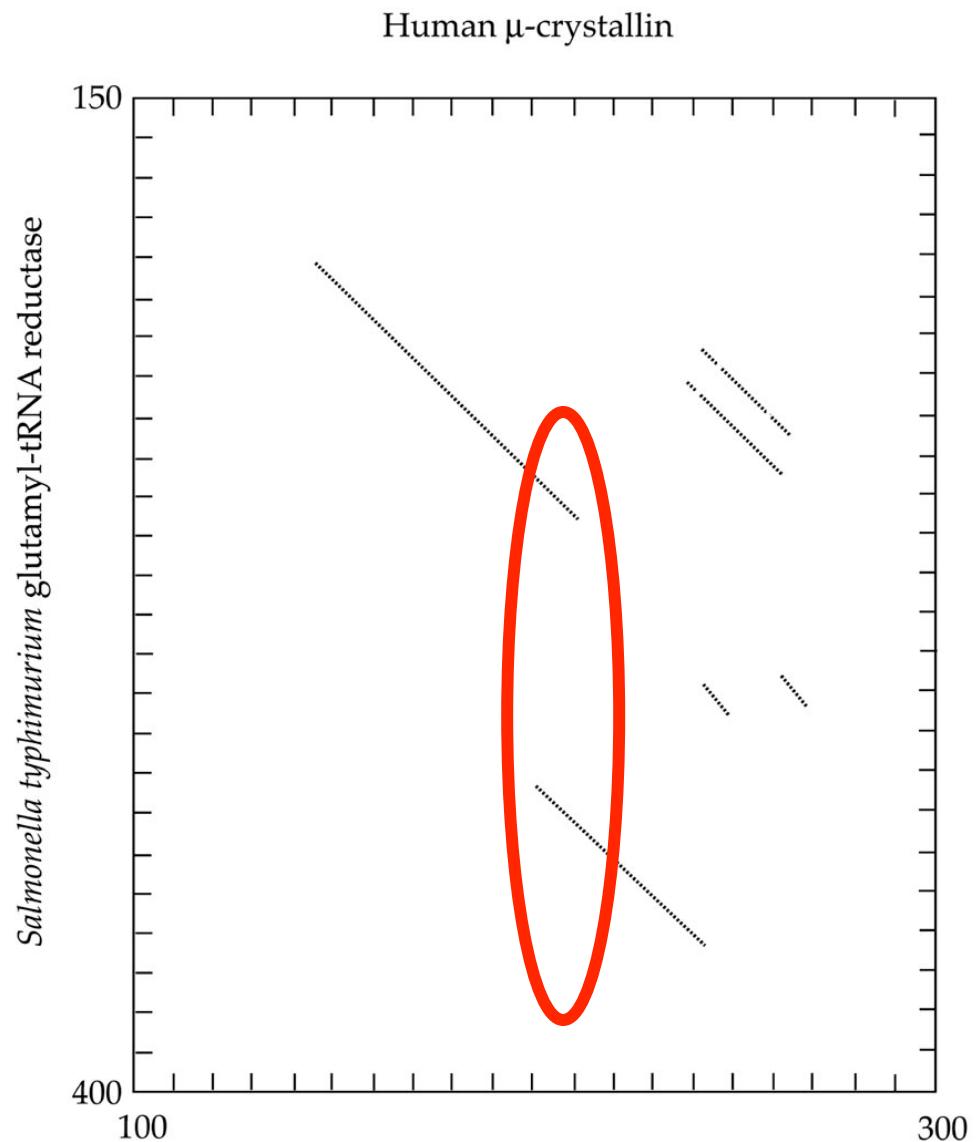
Dot Plot Advantages

- Although simple to understand, and quite primitive in complexity
 - Dot Plot is extremely powerful
 - Visualization aspect may unravel information regarding the evolution of sequences
 - Highlighting this can be of great importance to a molecular biologist
 - Remember, the core of Bioinformatics is all about visualization!
 - Still used in current literature
- Useful for complete genome comparison b/t microbes
 - Since all possible matches of residues b/t two sequences are found
 - The investigator left with choice of identifying the most significant ones by examinations of long diagonals
- Logic led to FASTA algorithm (as we'll see)

Advantages: Highlighting Information

The vertical gap indicates that a coding region corresponding to ~75 amino acids has either been deleted from the human gene or inserted into the bacterial gene, which one is it?

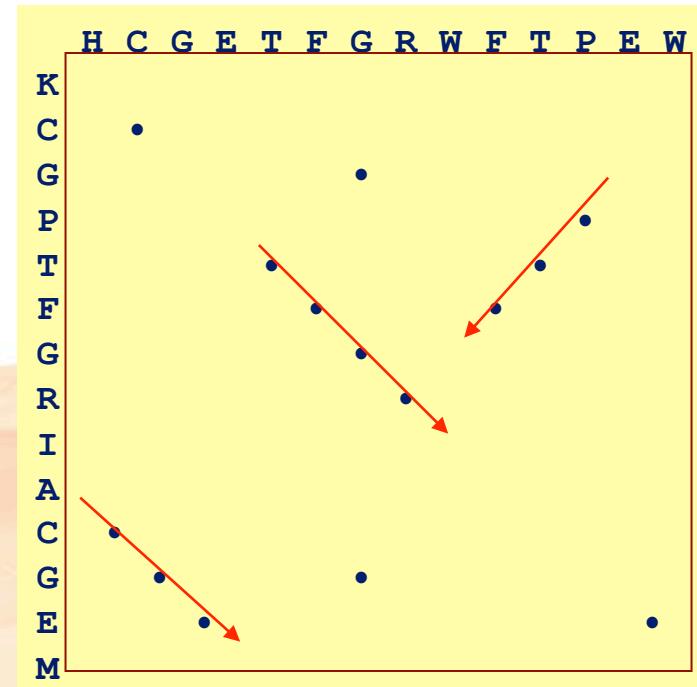
From a Dot Plot you really can't answer this question, ok why?



Window size = 60 amino acids; Stringency = 24 matches

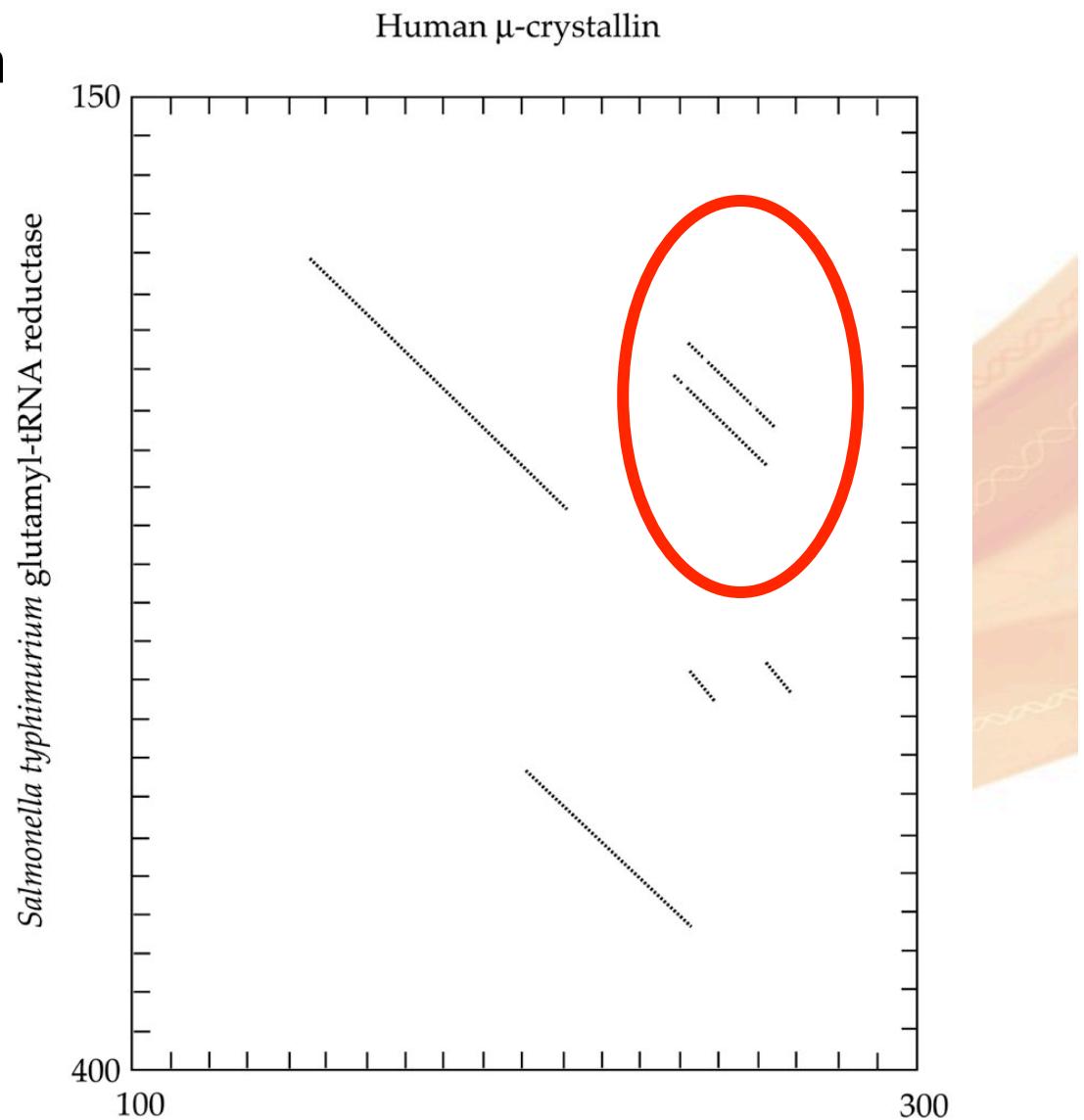
Dot Plot Display

- Diagonal rows (↘) of dots reveal sequence similarity or repeats
- Anti-diagonal rows (↙) of dots represent inverted repeats
- Isolated dots represent random similarity



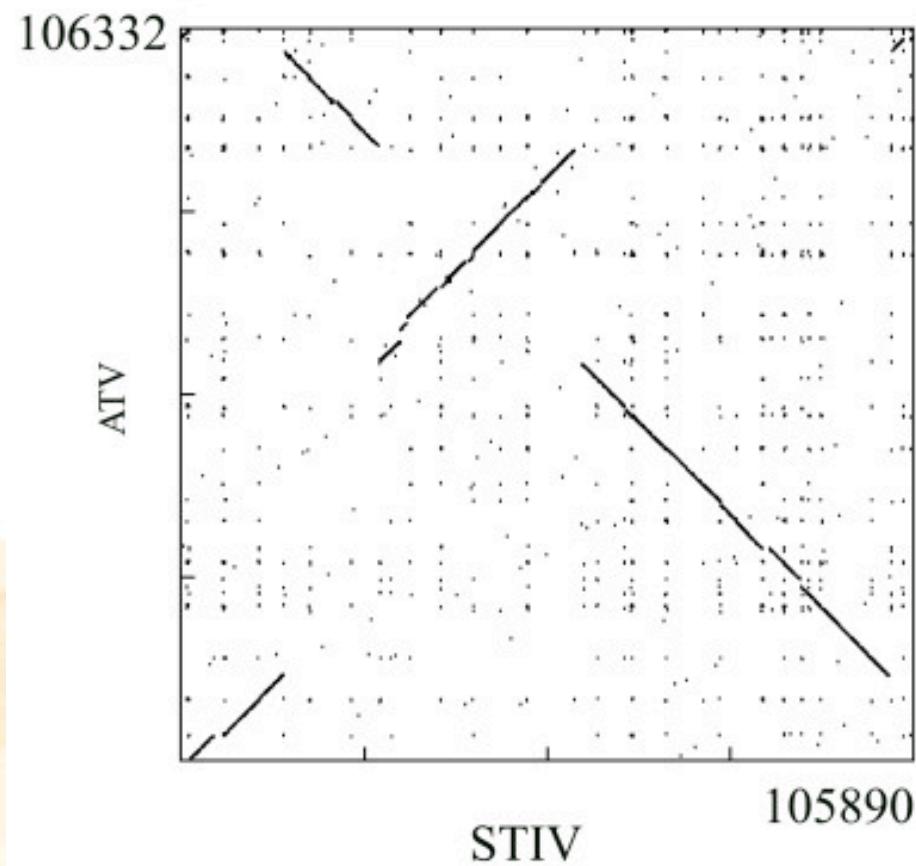
Advantages: Highlighting Information

The two diagonally oriented parallel lines most probably indicate that a small internal duplication has occurred in the bacterial gene



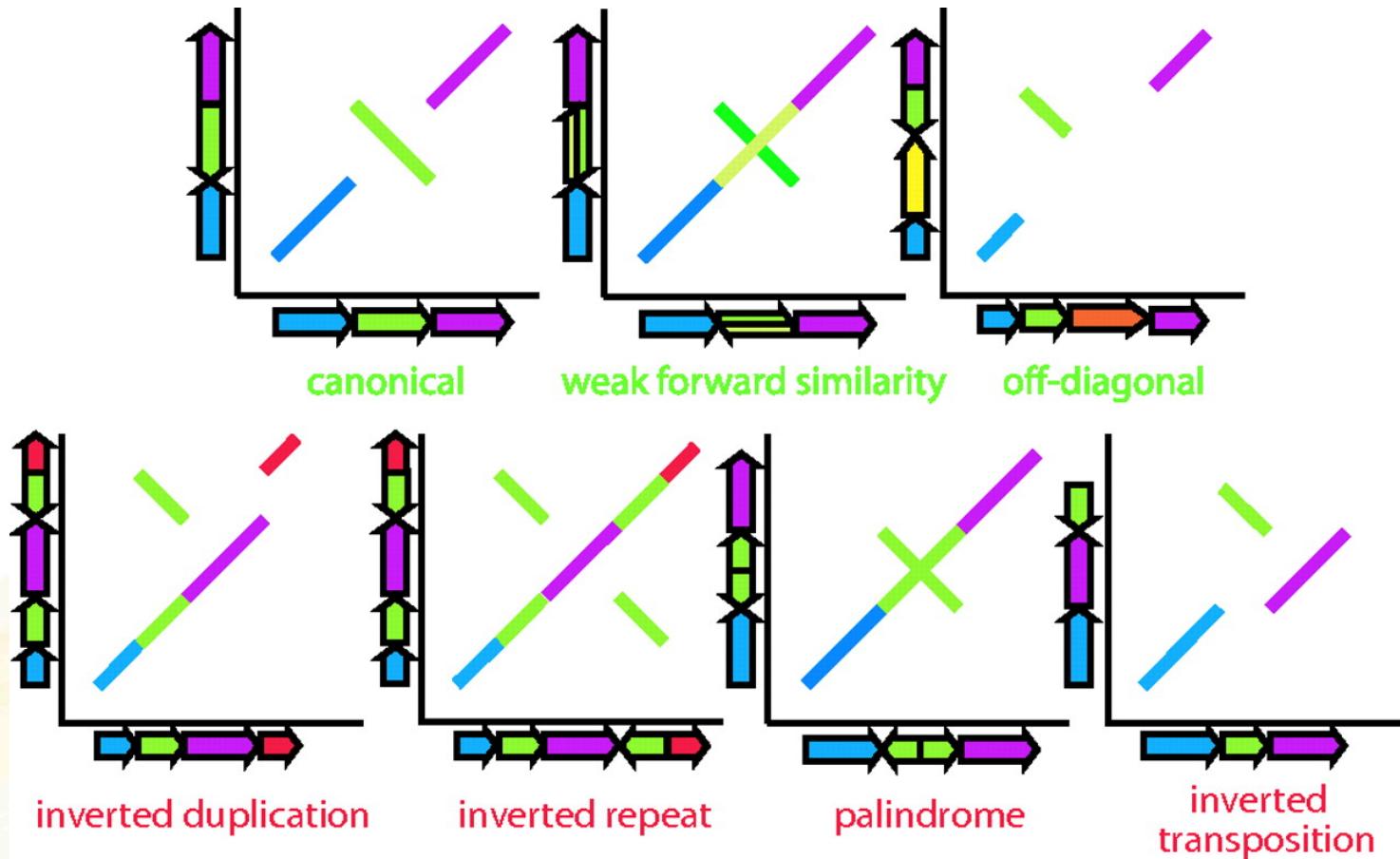
Window size = 60 amino acids; Stringency = 24 matches

Dot Plot with Inversions and Rearrangements



Huang et al. BMC Genomics 2009 10:224

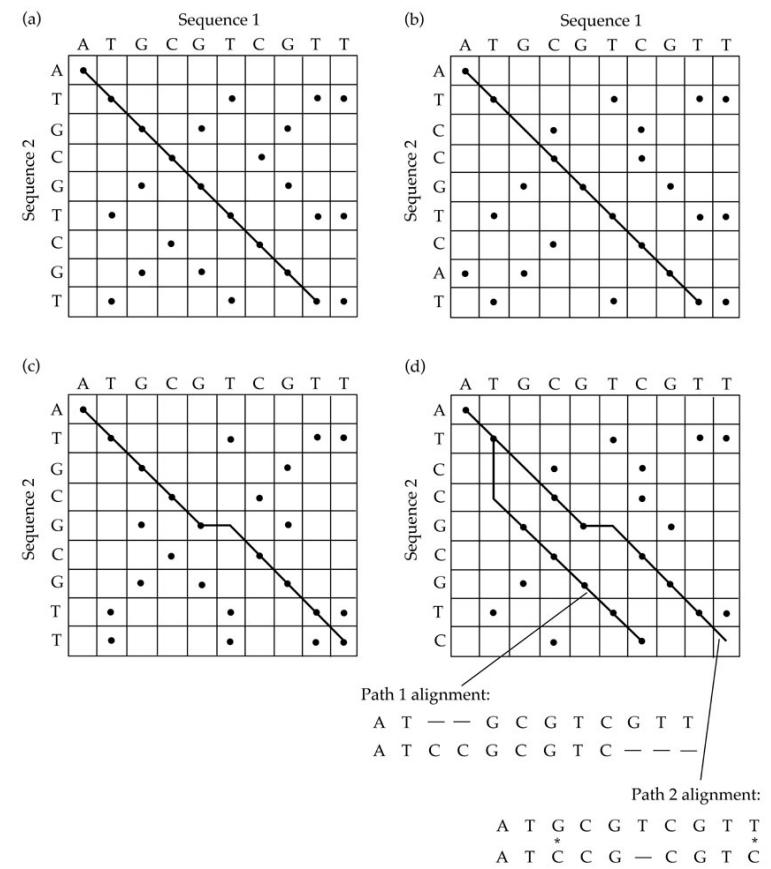
FYI - More Complex Situations in Dot Plot



Chaisson M J et al. PNAS 2006;103:19824-19829

Dot Plot Disadvantage:

- Relies on visual analysis
- Difficult to find optimal alignments
- Difficult to estimate significance of alignments
- Insensitive to conserved substitutions (e.g. L \leftrightarrow I or S \leftrightarrow T) if no **substitution matrix** applied
- Compares only two sequences (vs. multiple alignment)
- Time consuming
 - 1,000 bp vs. 1,000 bp = 10^6 operations
 - 1,000,000 vs. 1,000,000 bp = 10^{12} operations



Advantages of Manual Alignments

- Using a powerful and trainable tool, your brain!
- Ability to integrate additional data:
 - Domain structures
 - Biological function
 - Your knowledge
- **But.....**
 - **Too much to ask the brain to compare large, diverged sequences**
 - **The method is subjective, irreproducible, and unscalable!**

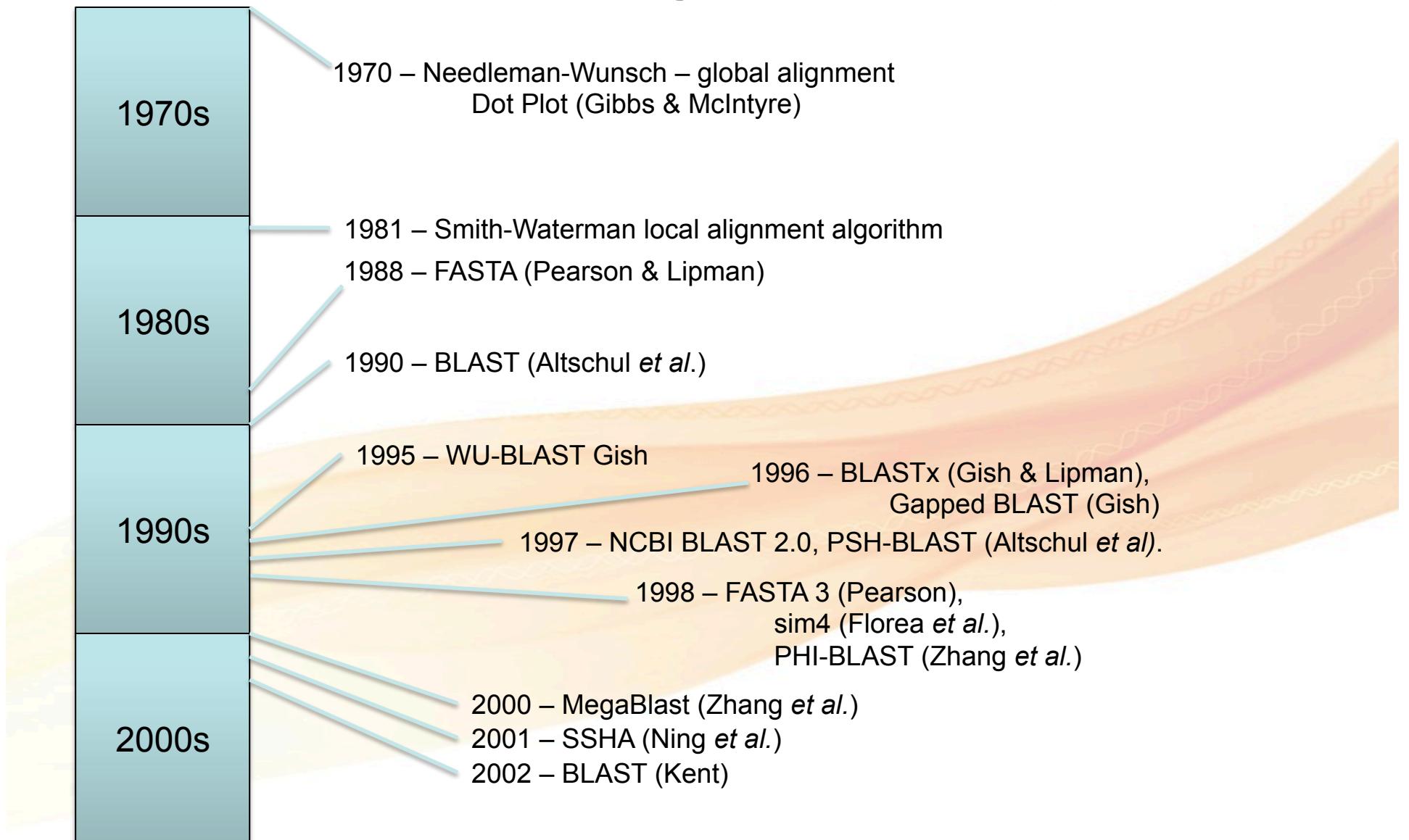
Alignment Methods - Computational Complexity

- Dynamic programming and Word Methods , as we'll see through coming lectures
- Can be divided into two categories:
 - For reasons of **computational complexity**
 - **Pairwise alignment** (i.e., the alignment of two sequences)
 - **Multiple-sequence alignment** (i.e., the alignment of three or more sequences)
- **Pairwise alignment** problems have **exact** solutions
- **Multiple-sequence alignment** problems only have approximate (**heuristic**) solutions
- We'll Discuss Pairwise first

Why Do We Use Alignment Methods?

- Short sequences can be aligned by hand manually
 - Many sequences or long sequence require the use of computer algorithms
- Algorithmic details usually fall into two categories:
 - Global alignments
 - Form of global optimization that “forces” the alignment to span the entire length of all query sequences
 - Needleman-Wunsch algorithm
 - Local alignments
 - Identify regions of similarity with long sequences that are often widely divergent overall
 - Smith-Waterman algorithm

Sequence Alignment History



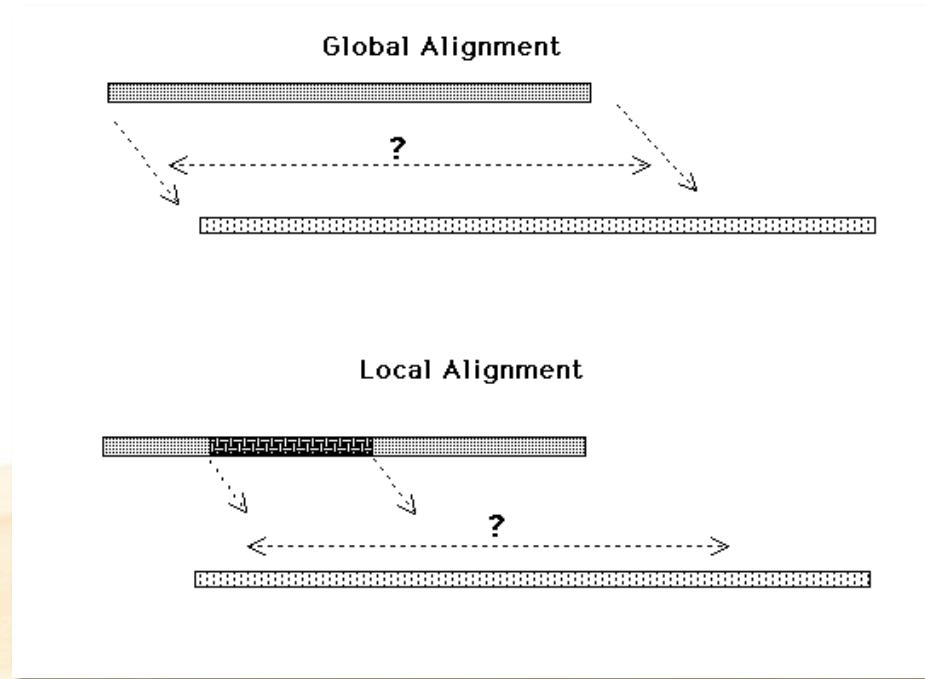
Global vs Local Similarity

- Global similarity uses completely aligned sequences - total % matches
 - Dot plots and Needleman & Wunch algorithm
- Local similarity looks for best internal matching region between 2 sequences
 - Dot plots
 - Smith-Waterman algorithm
 - BLAST and FASTA

Global vs. Local Similarity

- When should I use a certain method?
- Should result of alignment include all amino acids or just those that match?
 - If yes, a global alignment is desired
 - If no, a local alignment is desired

Global vs. Local Alignments



You Might Ask?

- Why do we need local alignments?
 - To compare a:
 - short sequence to a large one
 - single sequence to an entire database
 - partial sequence to the whole
 - Identify newly determined sequences
 - Compare new genes to known ones
 - Annotate functions for entire genomes full of ORFs of unknown function

For Thursday

- Go over tonight's lecture
- Understand code here:
 - http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_example.pl
 - Test it out!
- Get http://155.33.203.128/teaching/BIOL6308-Fall2013/local/Literature/entrez_fields.pdf
- Read
 - <http://www.ncbi.nlm.nih.gov/books/NBK7039/>