COMPUTATIONAL STRUCTURAL BIOLOGY

STRUCTURE, SIMULATION, FUNCTION & PREDICTION

Lecture 6

Michael Levitt
Structural Biology, Stanford

http:/csb.stanford.edu/clas

BIOINFORMATICS I

Data in Biology

Statistics of Comparison.

Data Visualization.

Databases.

Web Resources.

Sequence Comparison.

Data in Biology
Concept 6.1

DATA IN BIOLOGY

Strings (1-D).

Sequence.

Relationships (2-D).

Multiple Sequence Alignments.

Data in 3-D.

Sequence Objects.

Structure Objects.

STRINGS

Proteins sequences:	AVHTIKHERWTQ
DNA sequence:	ATGGCATGACAA
English Text:	A CAT SAT ON
Numbers:	123457980123
Digits of π:	3.1415926535 8979323846 26

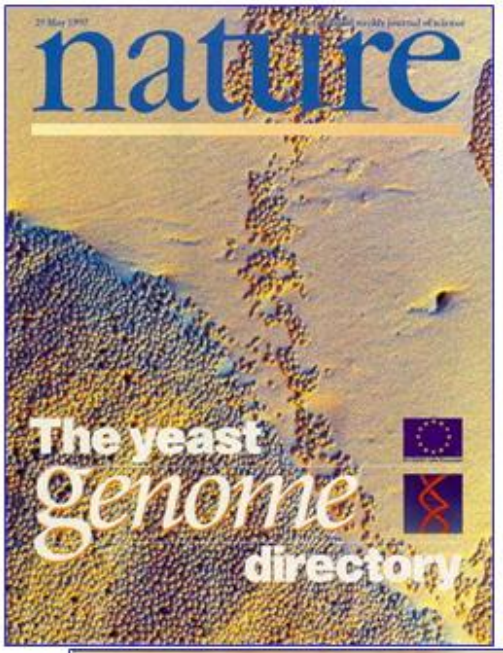
HUMAN GENOME

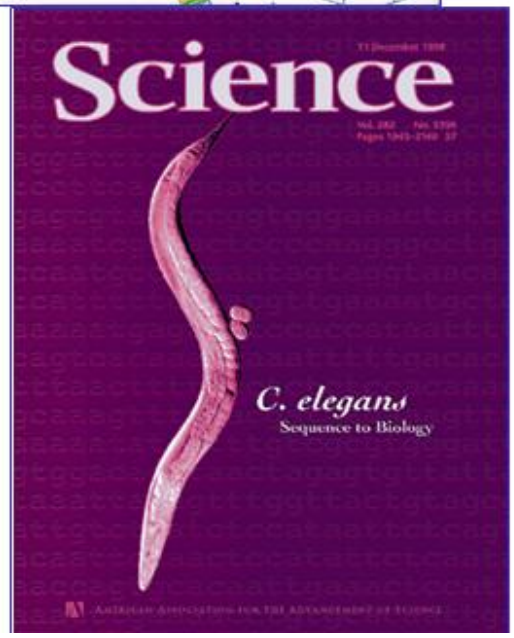
1995
Bacteria,
1.6 Mb,
~1600 genes
[*Science* 269: 496]

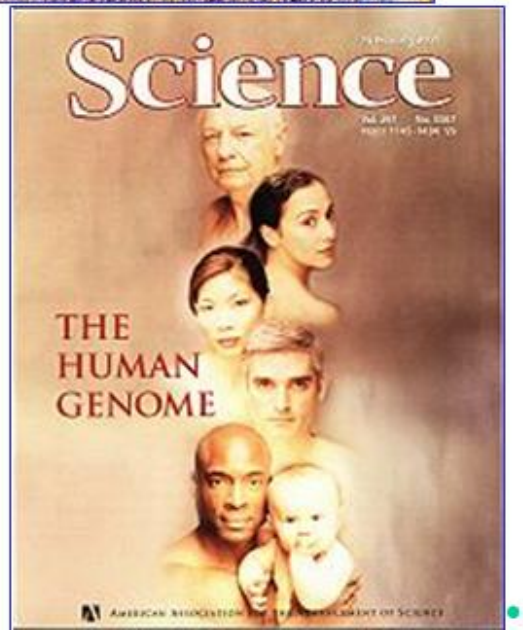
1997
Yeast,
13 Mb,
~6000 genes
[*Nature* 387: 1]

1998
Worm,
100 Mb,
~20,000 genes
[*Science* 282: 1495]

2001
Human,
3,000 Mb,
~30,000 genes
[*Science* 291: 1153]







Science
28 Jul 95

Nature
29 May 97

Science
11 Dec 98

Science
16 Feb 01

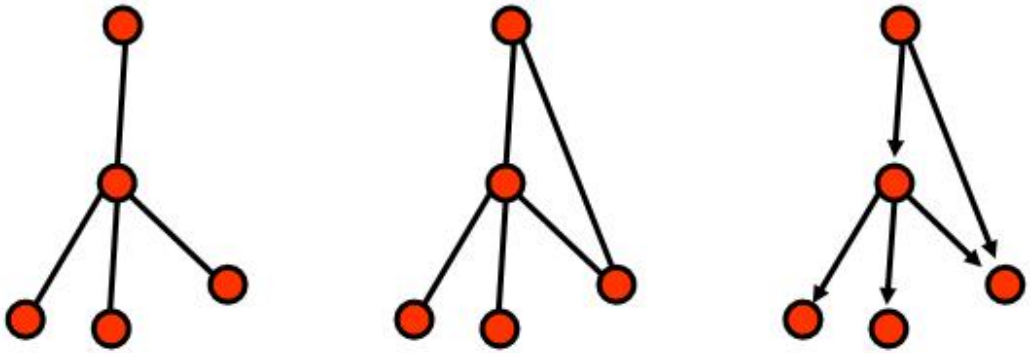
RELATIONSHIPS

Trees.

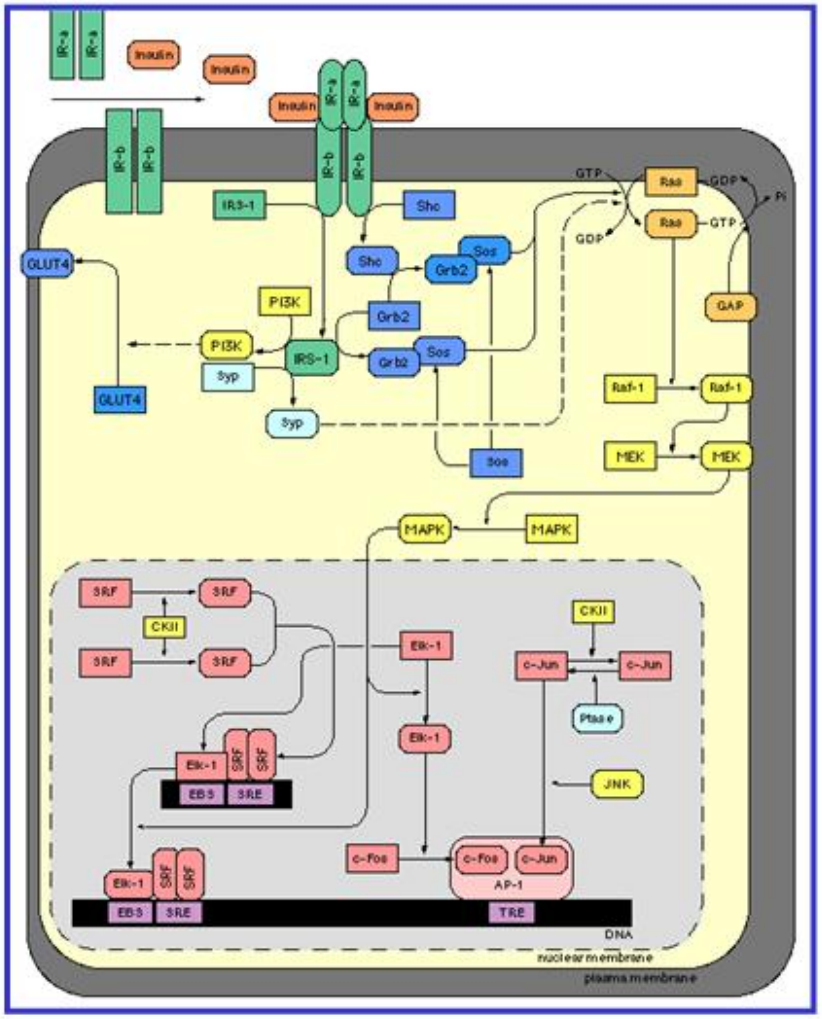
Graphs.

Directed Graphs.

Pathways.



Insulin Pathway



http://www.grt.kyushu-u.ac.jp/spad/pathway/pdgf.html

MULTIPLE SEQUENCE ALIGNMENTS

Initial consensus alignment

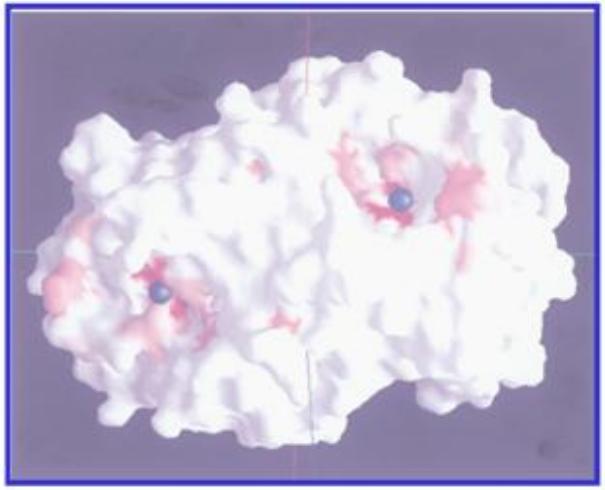
Resolved alignment

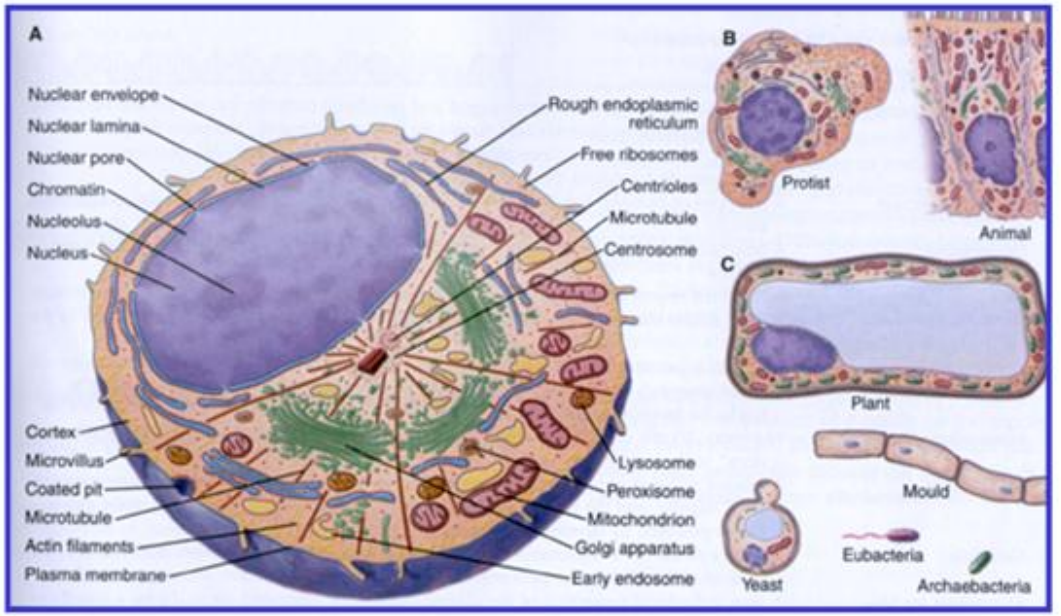
3-D DATA

Protein Structures.

Cells.

Galaxies.





SEQUENCE OBJECTS

- Data:
  - DNA sequences.
  - RNA sequences.
  - Protein sequences.

- Methods (Biological):
  - RNA sequence is derived from DNA sequence: Transcription.
  - Protein sequence is derived from RNA sequence: Translation.

- Methods (Evolutionary):
  - DNA sequence A is similar to DNA B: Homology.
  - RNA sequence A is similar to DNA B: Gene Finding.

These methods are all sequence to sequence.

STRUCTURE OBJECTS

- Data:
  - Organic Molecule.
  - Fibrous Protein.
  - Globular Protein.
  - Membrane Protein.
  - RNA Structure.
- Methods (Biological)
  - Organic Molecule is bound to Membrane Protein: Drugs.
  - Globular Protein is transformed to Fibrous Protein: Mad Cow disease.
  - Globular Protein transforms Organic Molecule: Enzymes.
- Methods (Evolutionary):
  - Protein structure A is similar to protein structure B: Homology.

These methods are structure to structure.

Statistics of Comparison
Concept 6.2

STATISTICS IS IMPORTANT IN BIOINFORMATICS

- All comparisons aim to determine if the observed level of similarity is significant.
- More precisely, what is the probability that the observed level of similarity could have been found between objects that are not-similar?
- In other words: could such a level of similarity be actually observed for non-similar objects?

MEAN

- Roll two die and get total of their values.
- Repeat many times and plot distribution of values.

$$\text{Mean} = \frac{(2\times1+3\times2+4\times3+5\times4+6\times5+7\times6+8\times5+9\times4+10\times3+11\times2+12\times1)}{(1+2+3+4+5+6+5+4+3+2+1)} = 252/36 = 7$$

2=1+1

3=1+2, 2+1

4=2+2, 1+3, 3+1

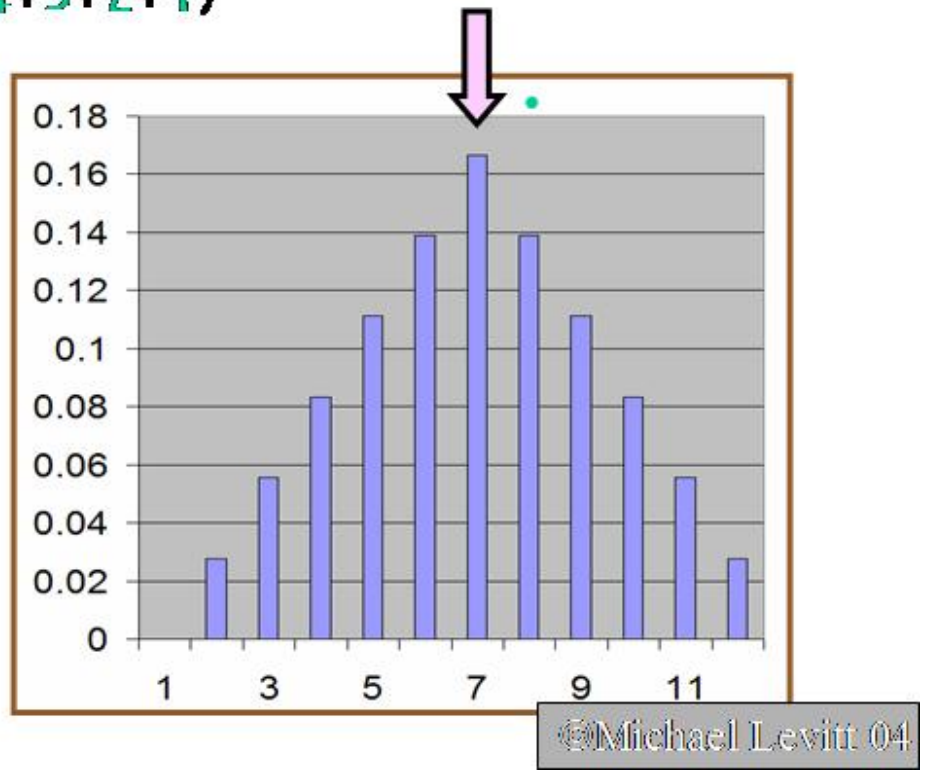
5=2+3, 3+2, 1+4, 4+1

6=3+3, 2+4, 4+2, 1+5, 5+1

7=3+4, 4+3, 2+5, 5+2, 1+6, 6+1

8=4+4, 3+5, 5+3, 2+6, 6+2

9=4+5, 5+4, 3+6, 6+3



STANDARD DEVIATION

Mean Squared

$$= \frac{(2^2 \times 1 + 3^2 \times 2 + 4^2 \times 3 + 5^2 \times 4 + 6^2 \times 5 + 7^2 \times 6 + 8^2 \times 5 + 9^2 \times 4 + 10^2 \times 3 + 11^2 \times 2 + 12^2 \times 1)}{(1+2+3+4+5+6+5+4+3+2+1)}$$

$= 1974/36 = 54.83$

Mean Squared

$(\text{Mean})^2$

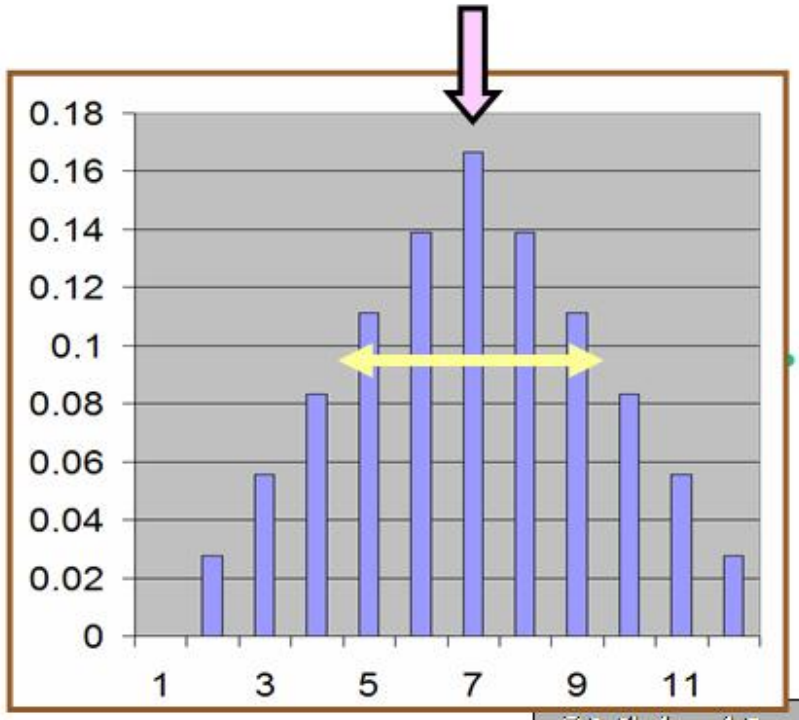
$$\text{Variance} = (54.83 - 7^2)$$

$$= 5.83$$

$$SD = \text{sqrt}(\text{Variance})$$

$$= \text{sqrt}(5.83)$$

$$= 2.415$$



UNIFORM DISTRIBUTION

$P(x) = 1/\max(x)$ for $0 < x < \max(x)$

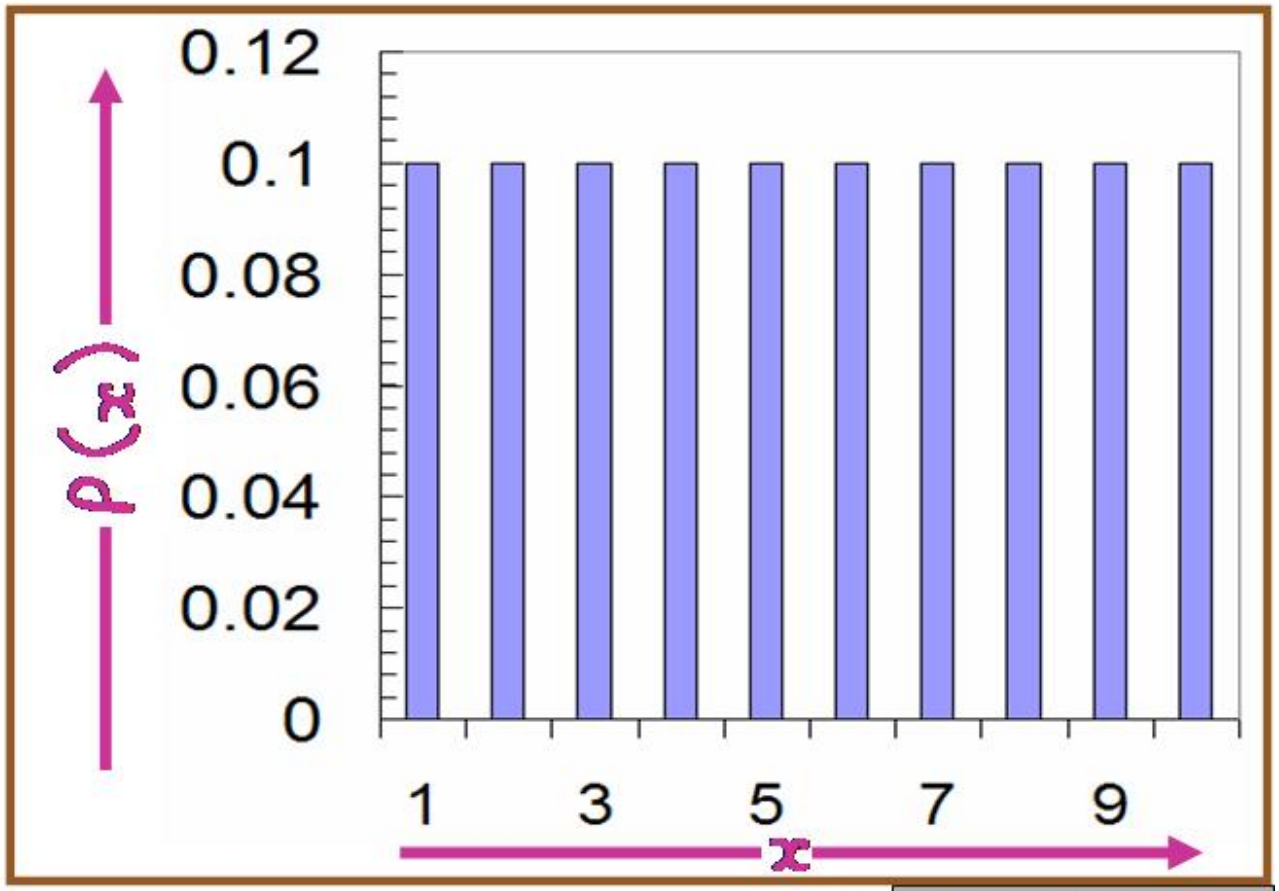
$= 0$ otherwise

Mean $= \min+\max/2$

$= 5.5$

$SD^2 = \max-\min/6$

$= 1.5$



BINOMIAL DISTRIBUTION

$$P(n) = \frac{N!}{n!(N-n)!}\, p^n (1-p)^{N-n}$$

$$\text{Mean} = Np$$

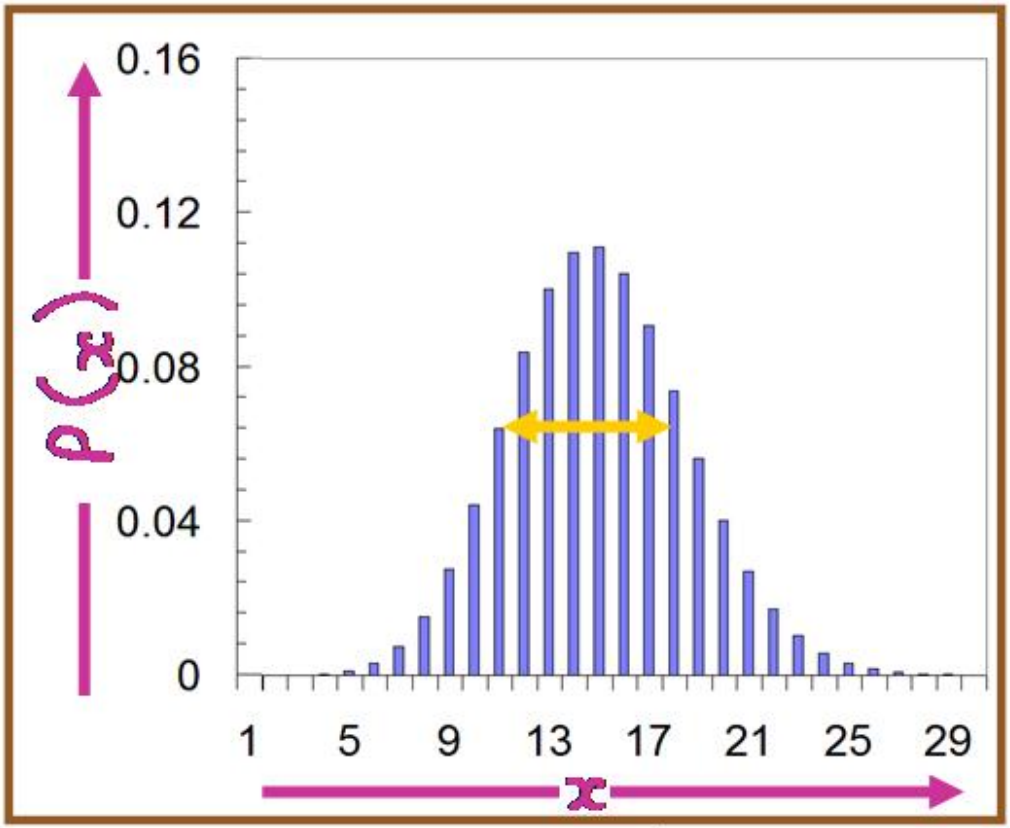
$$SD^2 = Np(1-p) = \text{Mean} \times (1-p)$$

RANDOM PATTERN (p=0.5)											#1's
0	1	1	1	0	0	0	0	0	0	0	3
0	1	1	1	1	0	0	0	0	1	0	1	6
1	1	0	0	1	0	1	1	0	0	0	0	4
1	1	1	0	0	0	0	1	0	0	0	1	5
1	1	1	1	1	1	1	0	1	0	1	1	8
1	1	1	0	1	1	0	1	0	0	0	0	5
0	0	0	1	0	0	0	1	1	1	1	1	5
1	0	1	0	0	1	1	1	1	0	0	1	7
1	0	1	0	1	0	0	0	0	1	0	1	5
0	0	1	0	1	0	0	0	1	1	0	0	4

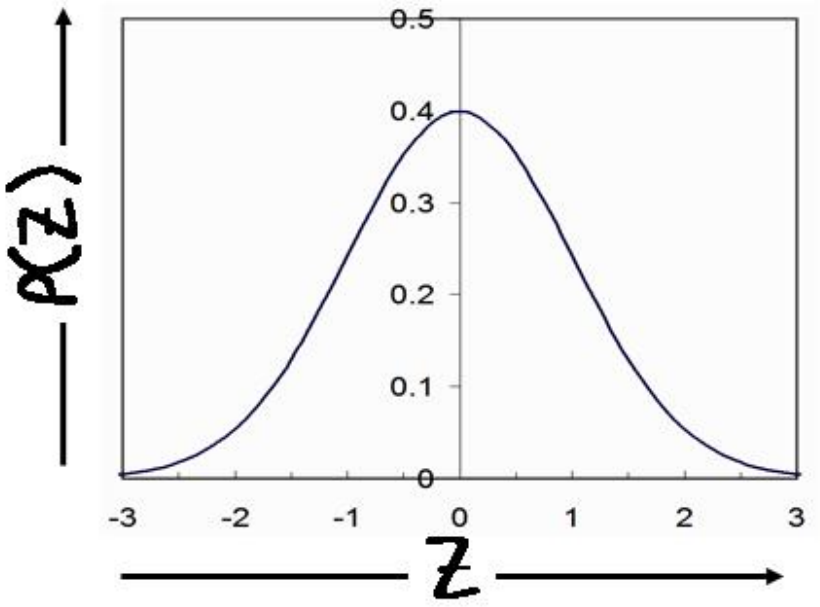
For N=30, p=0.5,
Mean = 15, SD^2 = 7.5

For N=100, p = 0.15,
Mean = 15, SD^2 = 15×0.85

If p small SD = $(\text{Count})^{1/2}$



NORMAL DISTRIBUTION

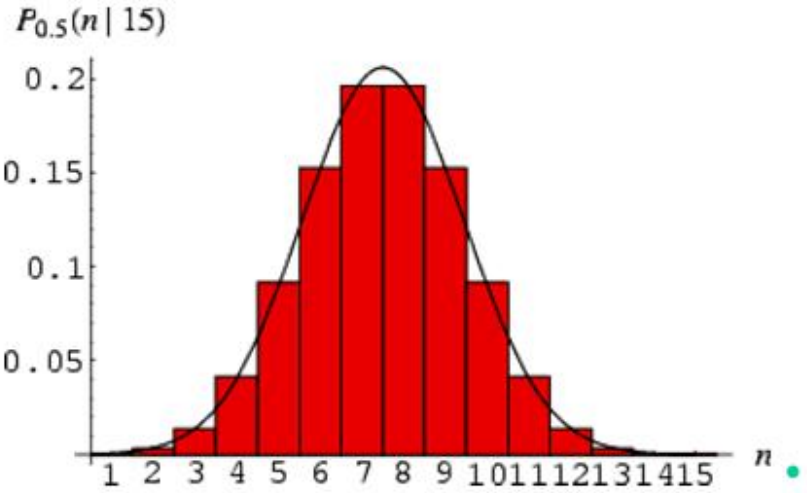


Gaussian:

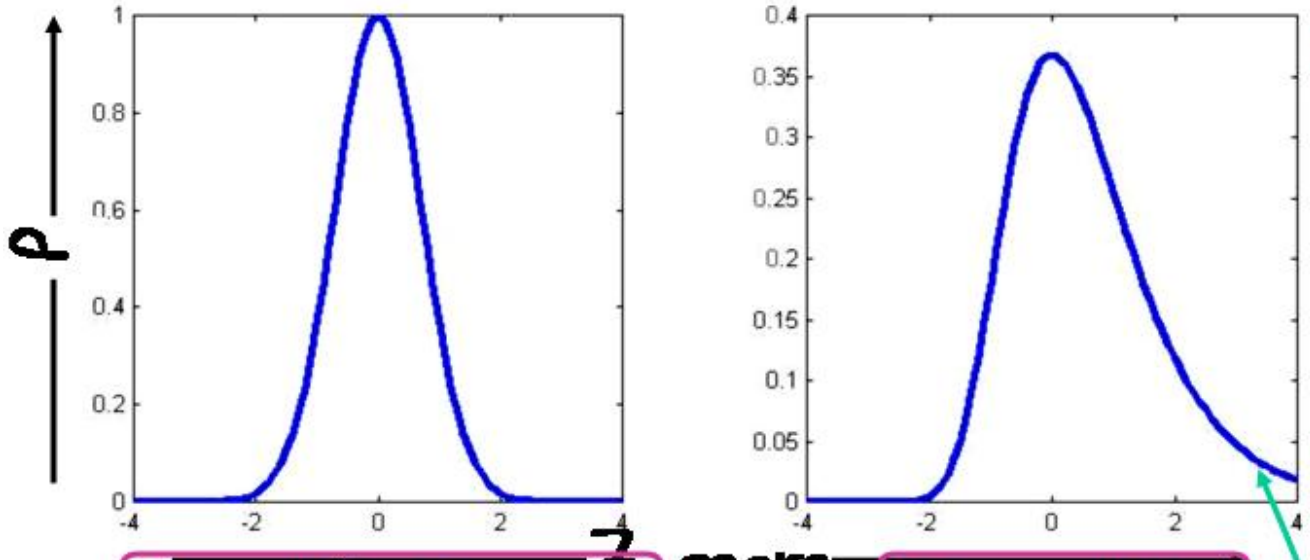
$$P(Z) = 1/(2\pi)^{1/2}\exp(-Z^2/2)$$

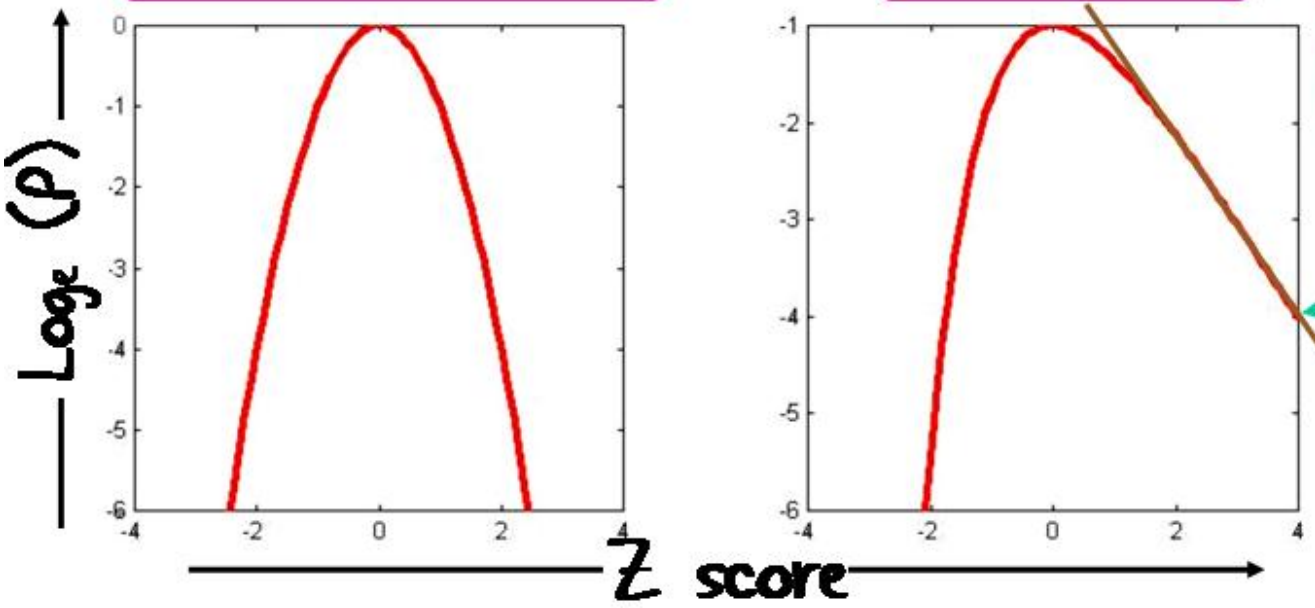
Mean = 0
SD = 1

The Binomial distribution Becomes like a Normal Distribution as the sample size increases as the sample Size increases.



EXTREME VALUE DISTRIBUTIONS





- **Normal:**

$$P(Z) = \exp(-Z^2)$$

$$\log_e(P(Z)) = -Z^2$$

- **Extreme Value:**

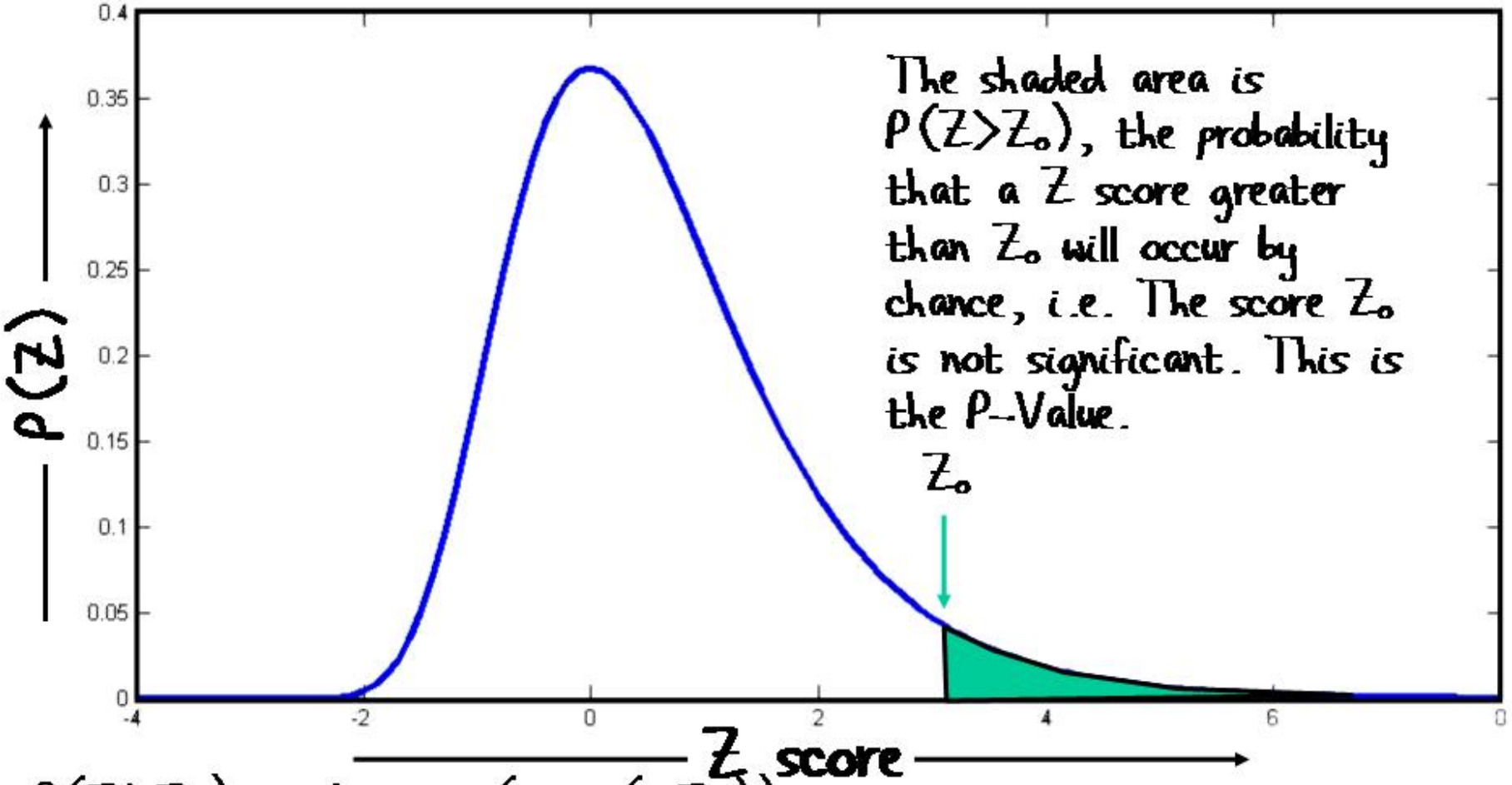
$$P(Z) = \exp(-Z - \exp(-Z))$$

$$\log_e(P(Z)) = -Z - \exp(-Z)$$

where $Z = (\text{score} - \text{mean})/SD$

The Extreme Value distribution has a long tail.

EXPECTATION VALUES



$$P(Z>Z_0) = 1 - \exp(-\exp(-Z_0))$$

The expectation value, $E(Z>Z_0)$ is the expected number of errors.

It is $E(Z>Z_0) = N_{db}\, P(Z>Z_0)$, where N_{db} is number of queries.

EXPECTATION VALUES

- Database searches use the expectation value, $E(Z>Z_o)$, to indicate whether the score, S_o, is significant.

  Note, $Z_o = (S_o - \text{Mean})/SD$, where SD is the standard deviation.

- Typically one requires that the expectation value be less than 10^{-20} for a sequence search and less than 10^{-4} for a structural match.

- The expectation value depends of the size of the database: A score of 100 might be best when there are 1000 comparisons but would be much less good when there are 1,000,000 comparisons.

Data Visualization Concept 6.3

DATA VISUALIZATION

Hierarchical Clustering.

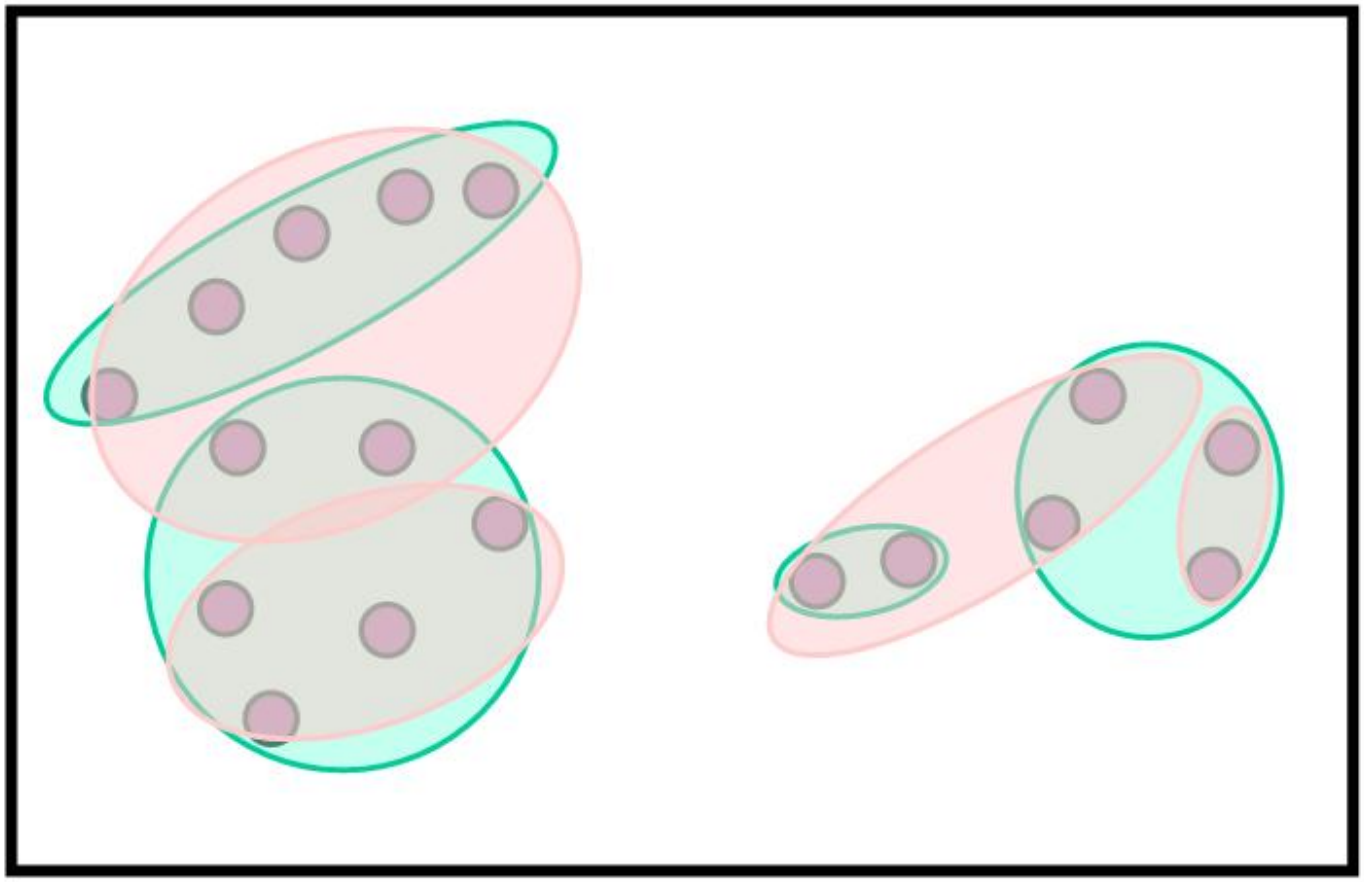
K-Means Clustering.

ROC Curves.

Views of Structure Space.

Multidimensional Scaling.

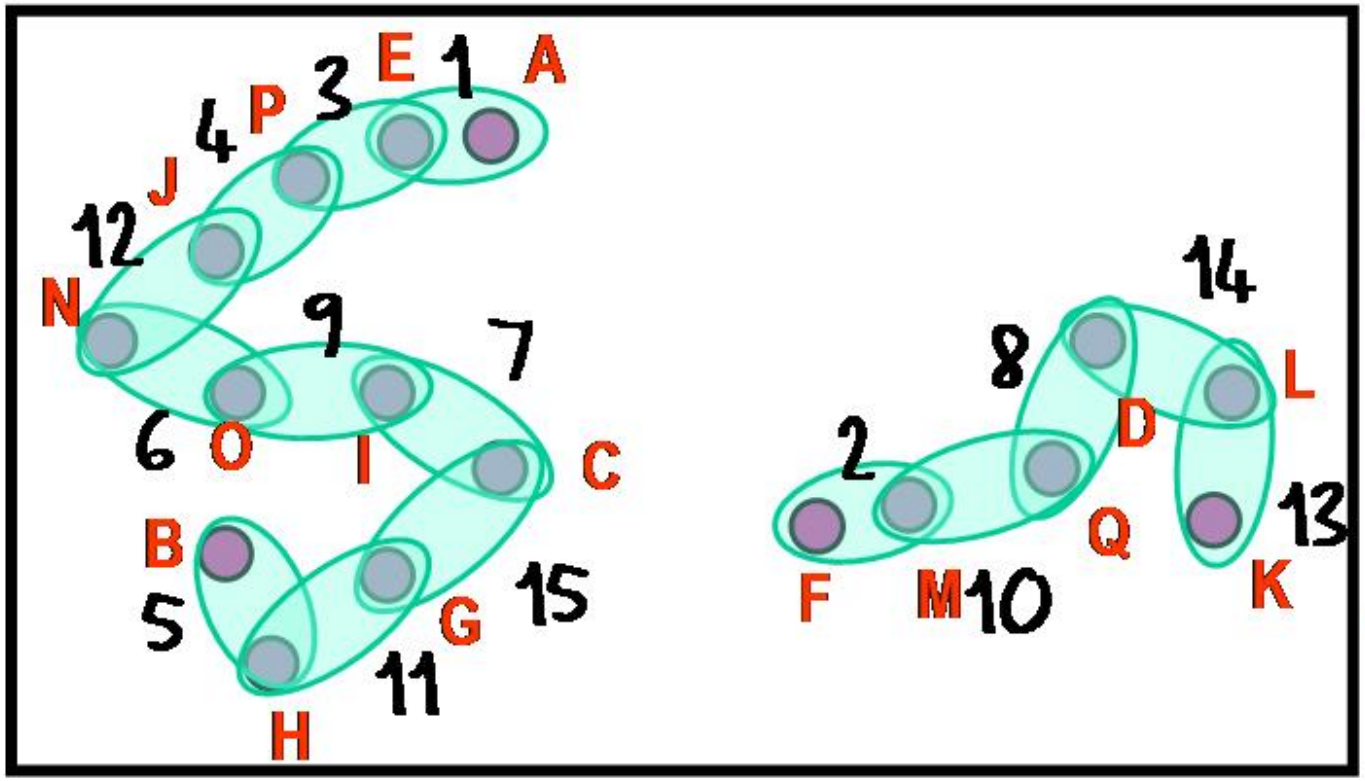
INTUITIVE CLUSTERING



- Many possibilities.
- Not so easy.
- What is best clustering?

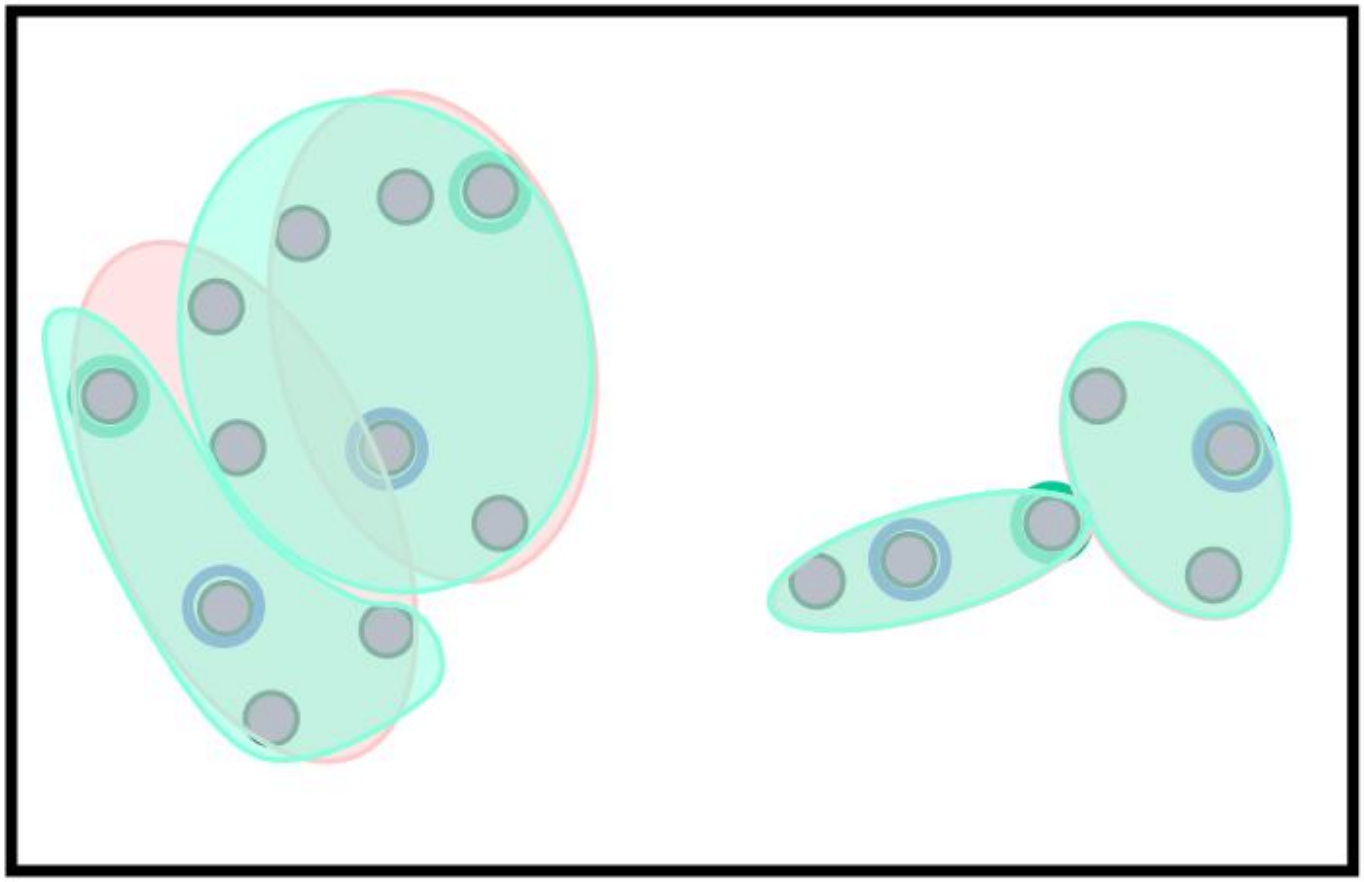
- Clustering seems easy and intuitive but it is actually very hard. Is there a solution?

HIERARCHICAL CLUSTERING



- Link the closest pairs. Keep going until no more close pairs.
- Single linkage clustering. Bad as can have distant members in same cluster.

K-MEANS CLUSTERING

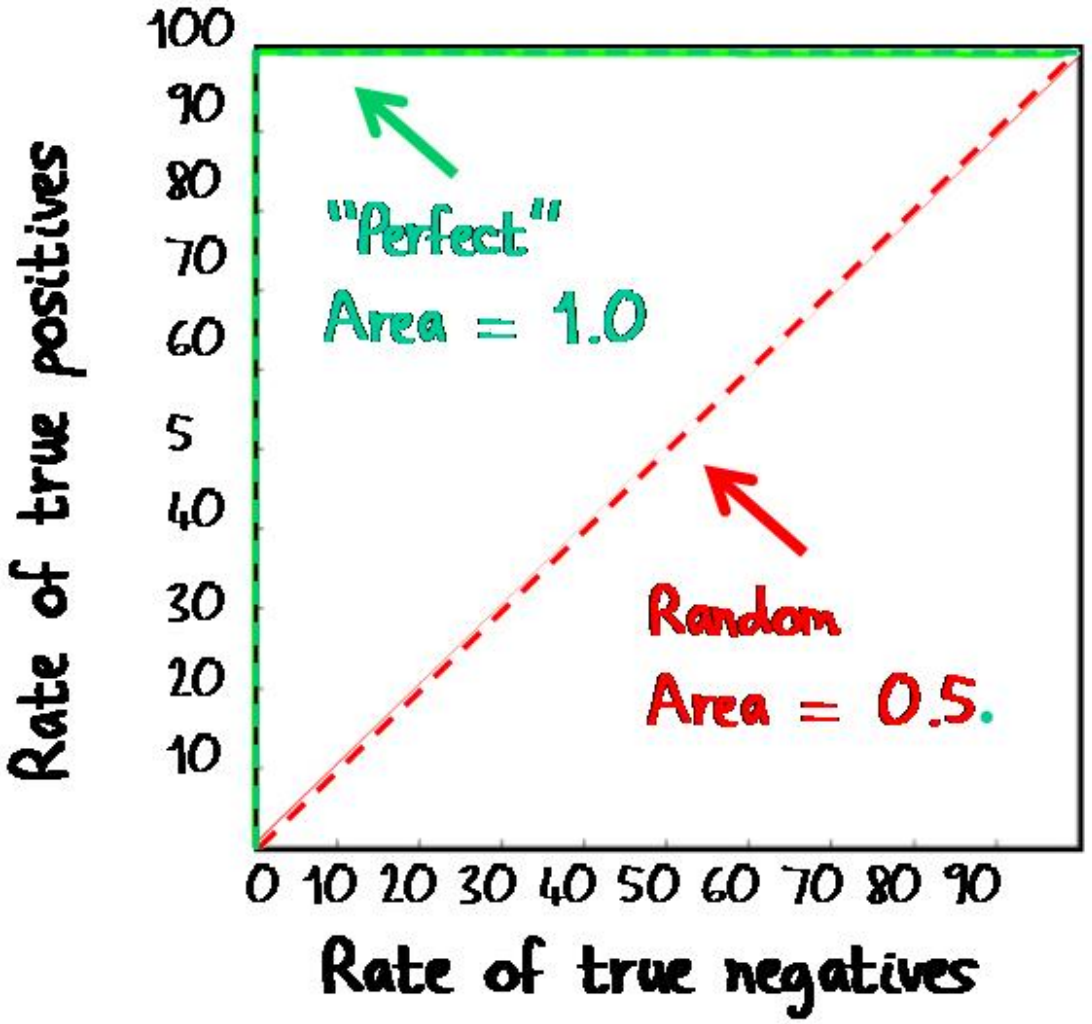


- Select K points at random.
- Associate all points with K point nearest it.
- Calculate a new mid point (K)
- Repeat till no change.

- This can fail badly if some regions are very dense.

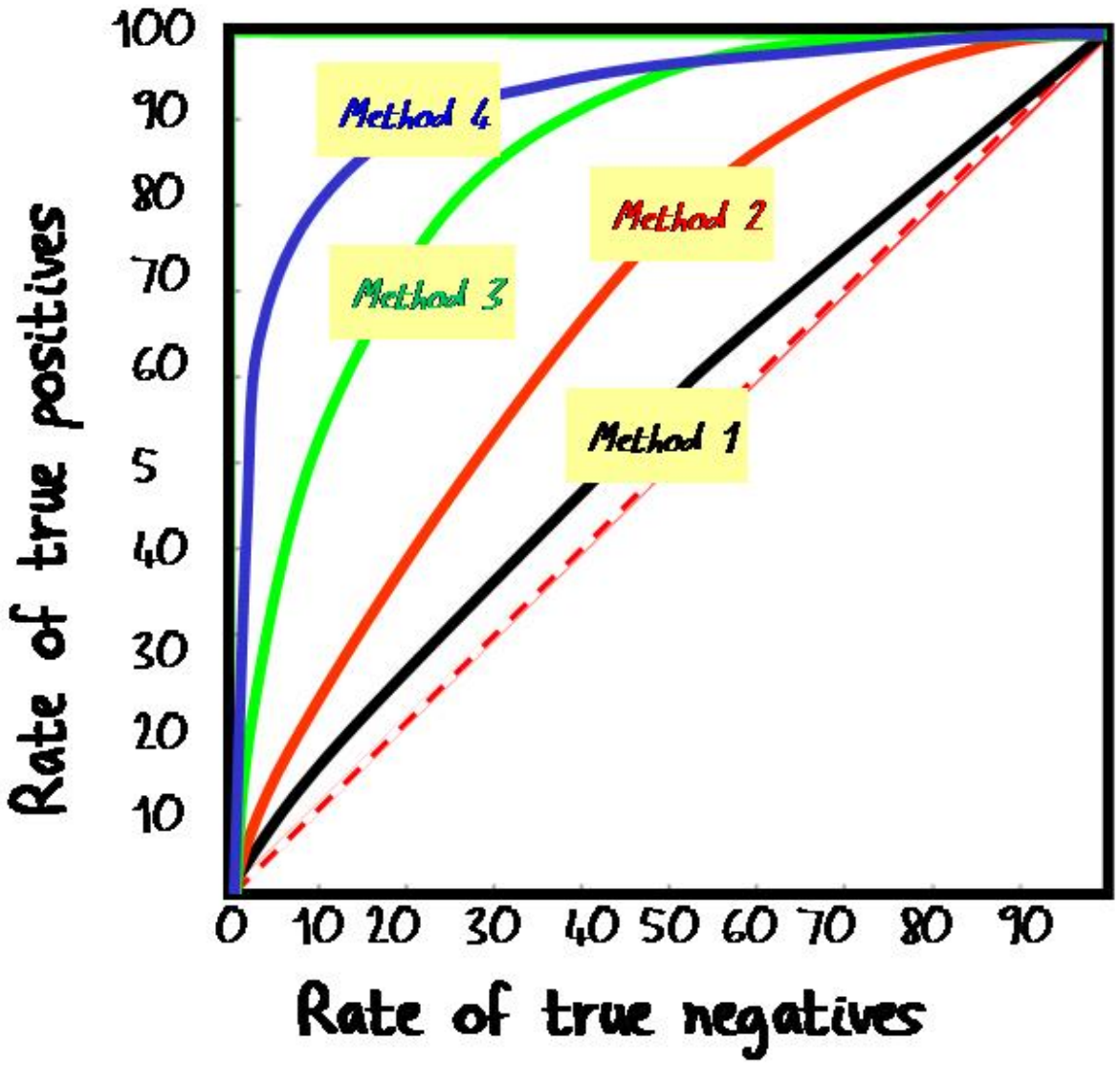
ROC ANALYSIS

(Receiver Operating



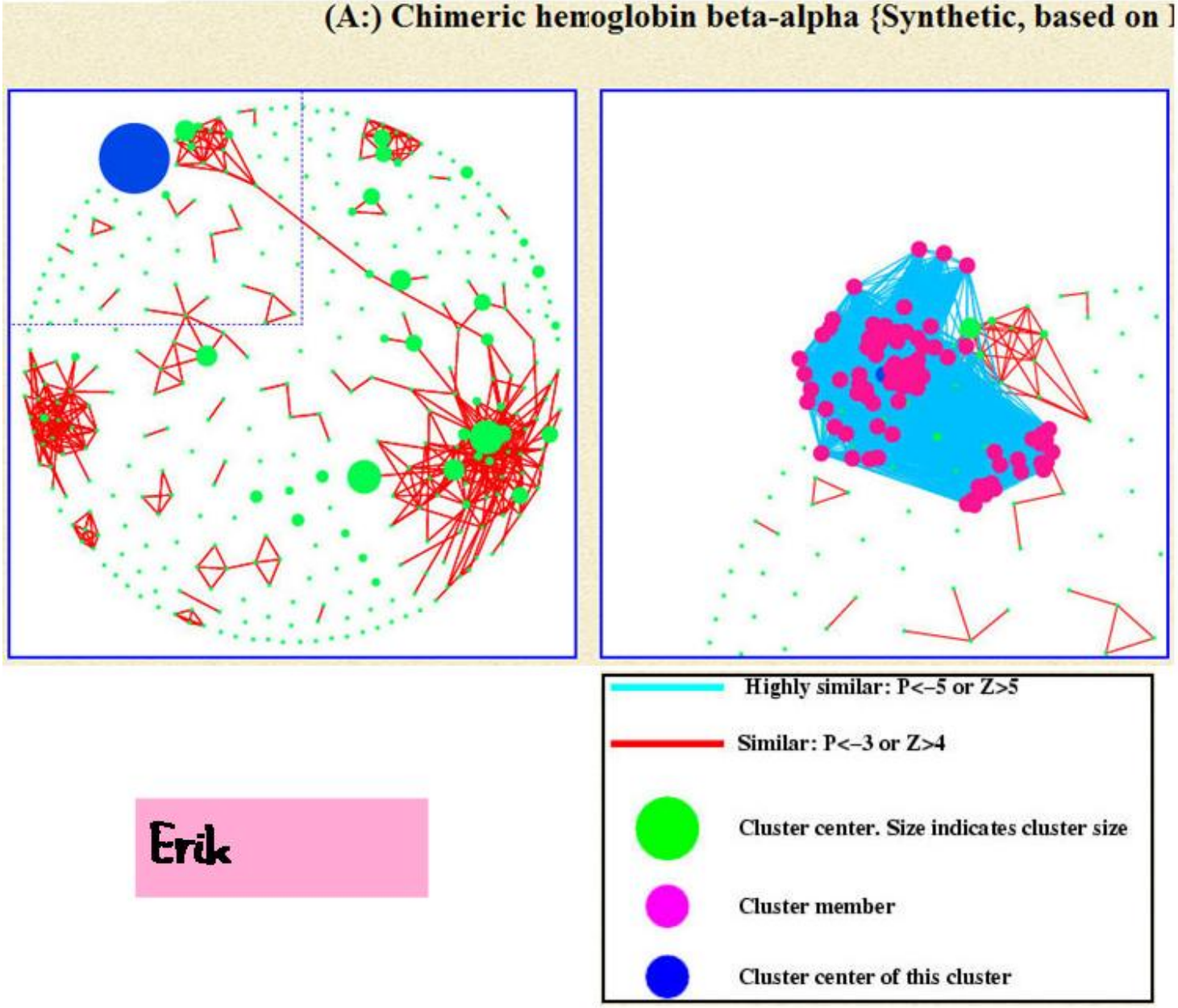
- Can a method sort items into correct and incorrect (true and false)?
- Compare the results with a "Gold Standard".

ROC CURVE EXAMPLES



- The best classification has the largest area under the curve.
- Too sensitive to errors in the "gold standard" classification.

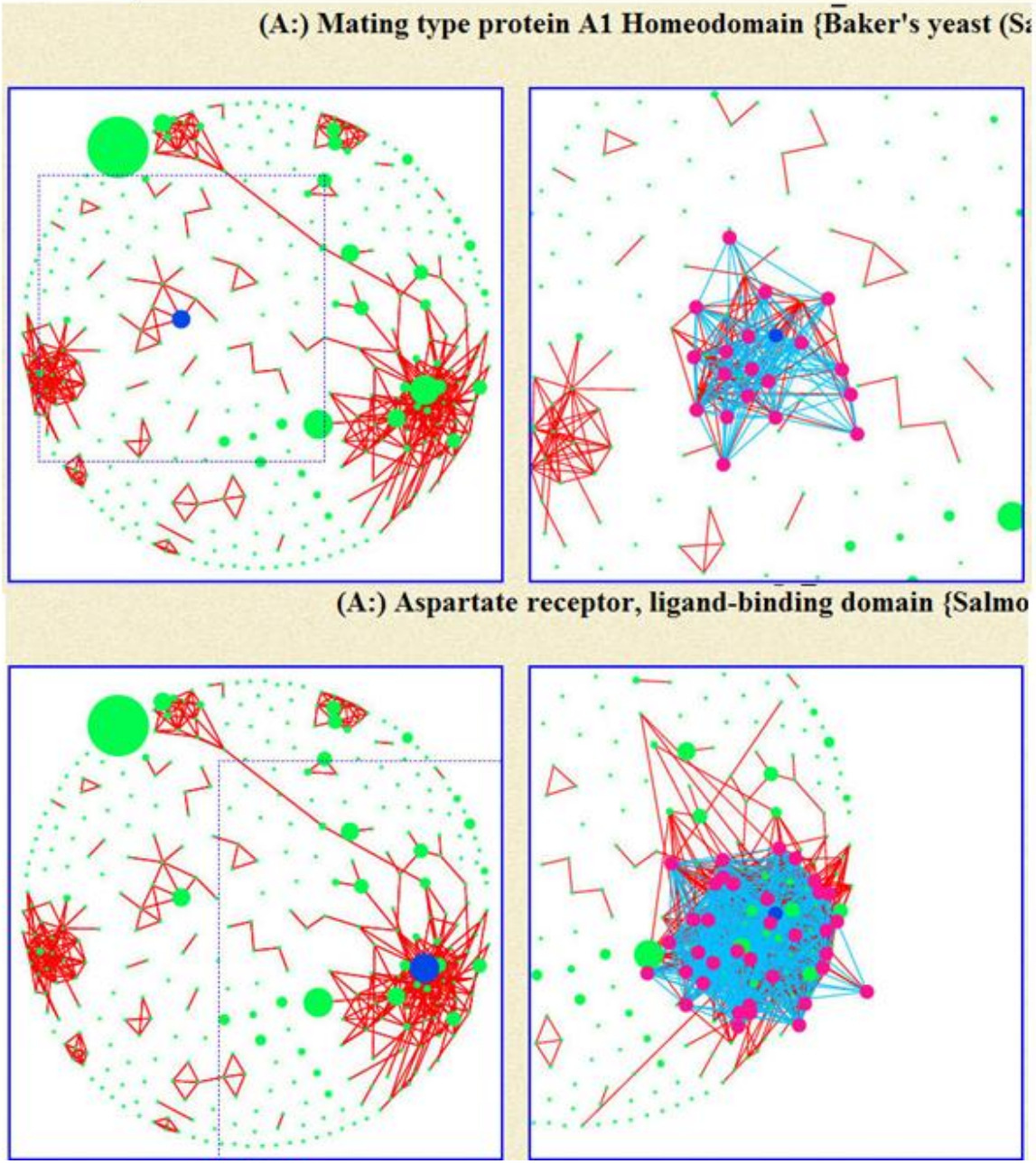
VIEWS OF STRUCTURE SPACE

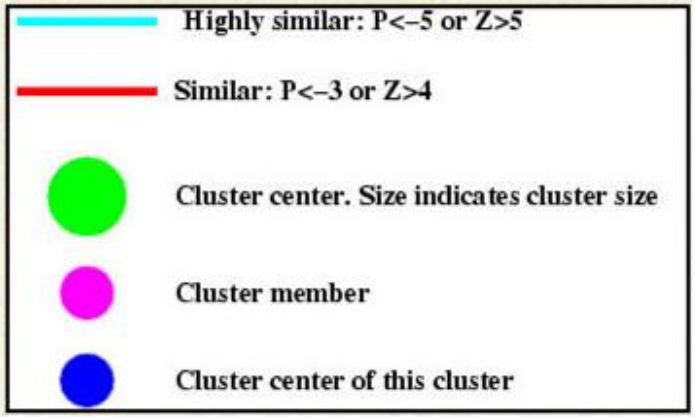


How can one see 5000 structure space?

Erik

VIEWS OF STRUCTURE SPACE





Make distorted 2-D view that changes as you move over it.

Databases
Concept 6.4

DATABASES

Protein Data-Base

RNA Database.

Membrane Database.

Small Molecule Databases.

Pathway Databases.

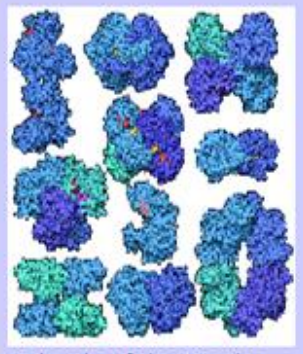
PROTEIN DATABASE RCSB

http://www.rcsb.org/pdb

DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
BETA XML files

Current Holdings

24248 Structures
Last Update: 10-Feb-2004
PDB Statistics



Molecule of the Month:
The Glycolytic Enzymes

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB).

The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences

RCSB PDB
PROTEIN DATA BANK

RCSB Home | wwPDB Home | Contact Us | Help

Did you find what you wanted?

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

ABOUT PDB | NEW FEATURES | USER GUIDES | FILE FORMATS | DATA UNIFORMITY | STRUCTURAL GENOMICS | SOFTWARE | PUBLICATIONS | EDUCATION

Search the Archive

Enter a PDB ID or keyword — Query Tutorial

Search

PDB ID Authors Full Text Search
match exact word remove similar sequences

SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

News — Complete News Newsletter — pdb-l Archive Subscribe

10-Feb-2004
"The PDB: A case study in management of community data" published in *Current Proteomics* A paper describing the development of the PDB and the systems in place for deposition and distribution, has been published in the inaugural issue of *Current Proteomics* (http://www.bentham.org/cp)... [MORE...]

PDB Mirrors

Please bookmark a mirror site

San Diego Supercomputer Center, UCSD*
Rutgers University*
Center for Advanced Research in Biotechnology, NIST*
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Universidade Federal de Minas Gerais, Brazil
Max Delbrück Center for Molecular Medicine, Germany

OTHER SITES

In citing the PDB please refer to:

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp.

6,400 Backward

PROTEIN DATABASE PDBLITE

www.pdblite.org

PDB Lite for Searching and Downloading Macromolecules

PDB Lite is designed for nonspecialists who search for atomic coordinate ("PDB") files at the Protein Data Bank on an occasional basis. It is especially targeted towards students and educators. See PDB Lite: What and Why? See also Nature of 3D Structural Data.

These sites were last tested, and these lists updated on 08/12/2003.

PDB Lite is available, including direct links to **Protein Explorer 2 Beta**, updated weekly with new entries from RCSB, from:

- Australia (The Walter and Eliza Hall Institute of Medical Research, Melbourne)
- India (Bioinformatics Centre, University of Pune)
- Israel (Bioinformatics, Weizmann Institute of Science, Rehovot)
- Israel (Tel Aviv University)
- Poland (Interdisciplinary Centre for Modeling, Warsaw University)
- United Kingdom (Cambridge Crystallographic Data Centre)
- United Kingdom (EMBL Outstation, European Bioinformatics Institute, Hinxton)
- USA (BioMolecular Engineering Research Center, Boston U)

For advanced searches, see Jaime Prilusky's OCA as an alternative to SearchFields at RCSB. OCA can find some things that SearchFields cannot. For example, OCA has query fields for Kingdom, Gene, Disease, and Function. On the other hand, SearchFields can find some things more easily than OCA. For example it can limit searches to entries that contain coordinates for RNA but neither protein nor DNA, and it can find "phospholipase C" while OCA ignores the "C".

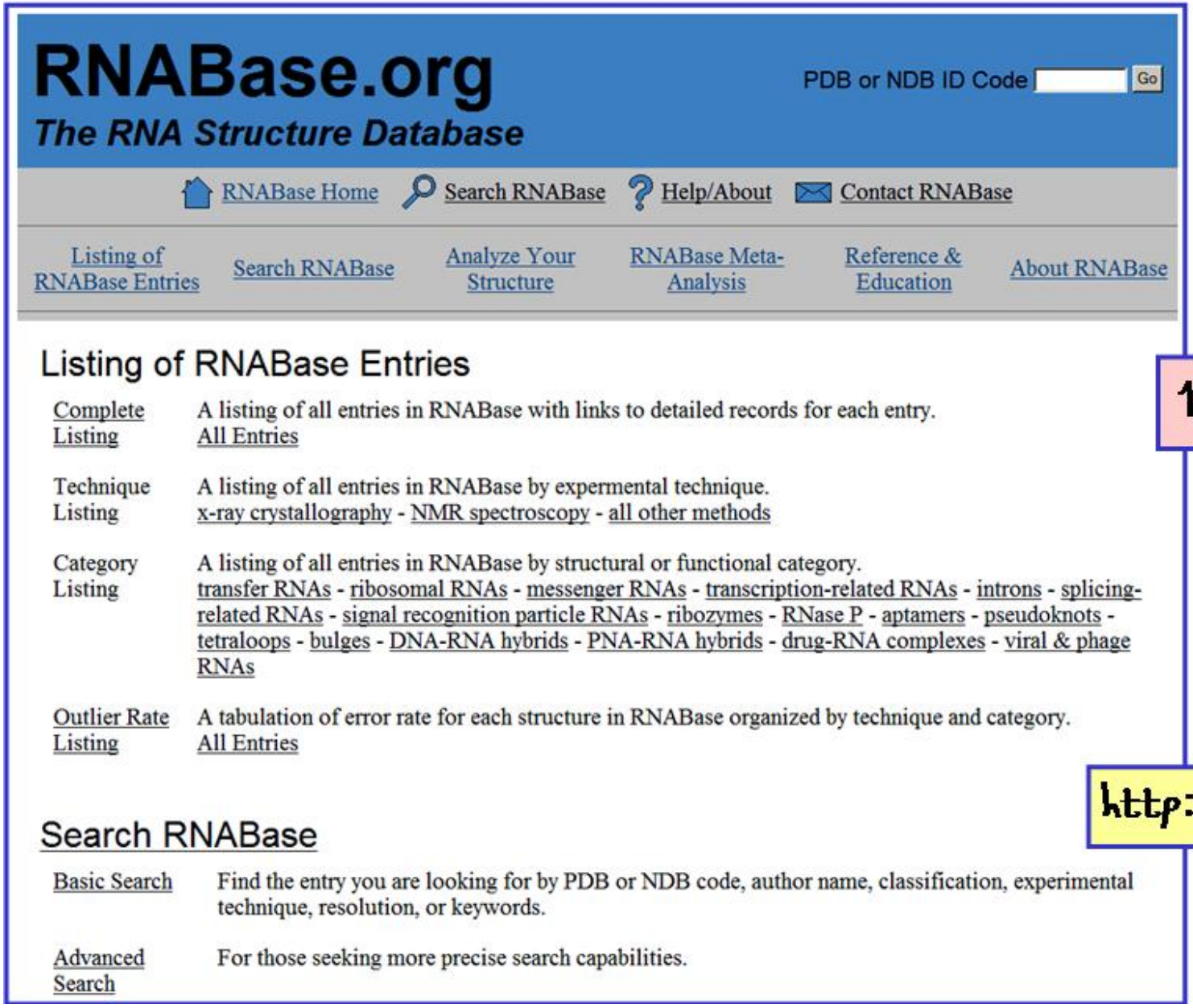
The above PDB Lite sites were keeping their databases up to date with new releases when checked on 08/12/2003. Other former mirror sites of the former Protein Data Bank (PDB) at Brookhaven National Laboratory (now closed) or OCA mirror sites were out of date, or were no longer offering PDB Lite (at least in fully functional form).

Best place to find the PDB file you are looking for.

83 Backward

http://www.umass.edu/microbio/rasmol/pdblite.ht

RNA STRUCTURE DATABASE



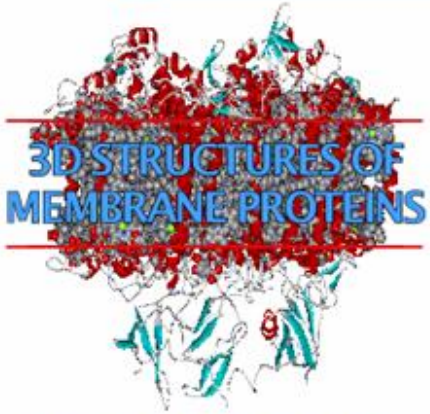
183 Backward

http://www.rnabase.org

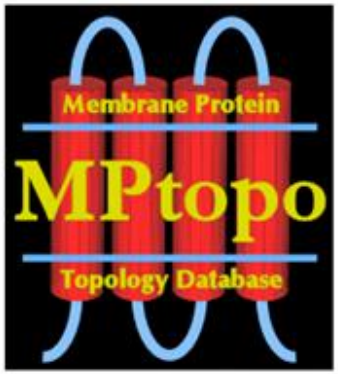
MEMBRANE PROTEIN DATA BASE

http://blanco.biomol.uci.edu/MemPro_resources.htm

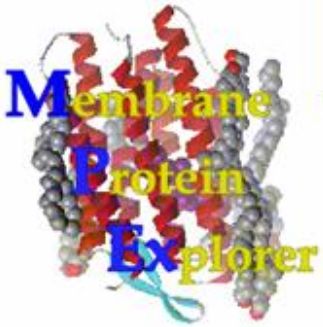
MEMBRANE PROTEIN RESOURCES



[Membrane Protein Structures]



[Membrane Protein Topology Database]

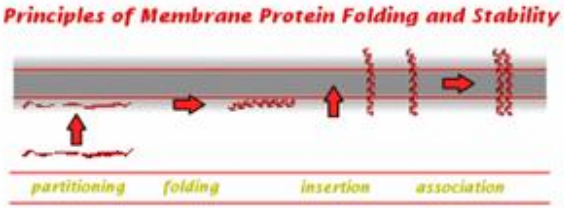


[Membrane Protein Explorer]

5 Backward

Additional Resources

Energetics of Protein-Bilayer Interactions
Experiment-Based Hydrophobicity Scales
Structure of Fluid Lipid Bilayers
Transport Protein Database from UCSD



[Membrane Protein Folding and Stability]

Additional Resources

MSB Bibliographic Database
Structural Biophysics Web Resources
Meetings of Interest to MP biophysicists

from the Stephen White Laboratory at UC Irvine

Author: Stephen White

ISI Current Web Contents ™

Page last updated: 7 Nov 2003

ORGANIC MOLECULE STRUCTURES

Welcome To the MathMol Library

Menu

- Water and Ice
- Carbon
- Hydrocarbons
- Amino Acids
- Nucleotides
- Lipids
- Sugars
- Photosynthesis
- Drugs

Other Databases

Visit The MathMol Hypermedia Textbook

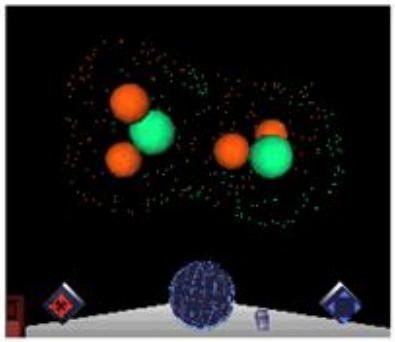
Back to

52 Backward

These pages are being designed at the New York University Scientific Visualization Center

We welcome feedback and comments at mr23@nyu.edu

Library of 3-D Molecular Structures



If you are using CosmoPlayer click on the above image of a water dime

[About the Database]

To enter the library, click on the appropiate buttons below

Water and Ice

Carbons

Hydrocarbons

Molecules of Life

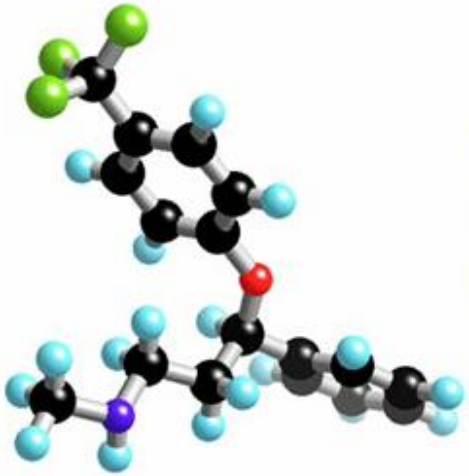
Drugs

http://www.nyu.edu/pages/mathmol/librar

MOLECULES OF THE MONTH

Mar 2002 - Astemizole

212 Backward



this site extensively uses two types of multimedia plug-ins for world wide web browsers; a plug-in to display 3D coordinate chemical structures, such as the: Chemscape Chime plug-in, or CS Chem3D plug-in or even external chemical viewers such as MSI's WebLab Pro or RasMol. Also required is a VRML plug-in to display the extended polyhedra arrays.

these page have won two HMS Beagle's three star awards as indispensable web pages (July 1998) and these pages were their "Web Pick of the Day". Additionally, they were highlighted in Science Magazine's Netwatch and also have been featured in the New Scientist and by the Royal Society of Chemistry.

you can find more chemistry based multimedia at the University of Oxford's Chemistry IT Centre and the Virtual Laboratory.

other "Molecules of the Month" are at the University of Bristol and at Imperial College of Science, Technology and Medicine, London

http://www.chem.ox.ac.uk/mo

SMALL MOLECULE DATABASE

NIST Chemistry WebBook

NIST Standard Reference Database Number 69 - March, 2003 Release

View: Search Options, Models and Tools, Documentation, Notes

Show Credits

NIST reserves the right to charge for access to this database in the future.

Search Options top

General Searches

- Formula
- Name
- CAS registry number
- Reaction
- Author
- Structure

Physical Property Based Searches

- Ion energetics properties
- Vibrational and electronic energies
- Molecular weight

Models and Tools top

- Thermophysical Properties of Fluid Systems High accuracy data for a select group of fluids.
- Group Additivity Based Estimates Estimates of gas phase thermodynamic properties based on a submitted structure.

Documentation top

- Frequently asked questions

1,210 Backward

http://webbook.nist.gov/chemistr

PATHWAY DATABASES

87 Backward

Signaling Pathway Database

The Signaling PAthway Database (SPAD) is an integrated database for genetic information and signal transduction systems.

There are multiple signal transduction pathways: cascade of information from plasma membrane to nucleus in response to an extracellular stimulus in living organisms. Extracellular signal molecule binds specific intracellular receptor, and initiates the signaling pathway. Now, there is a large amount of information about the signaling pathway which controls the gene expression and cellular proliferation. We have developed an integrated database SPAD to understrand the overview of signaling transduction. SPAD is divided to four categories based on extracellular signal molecules (Growth factor, Cytokine, and Hormone) and stress, that initiate the intracellular signaling pathway. SPAD is compiled in order to descrive information on interaction between protein and protein, protein and DNA as well as information on sequences of DNA and proteins.

There are two methods for retrieving this database. Please select one of the two items.

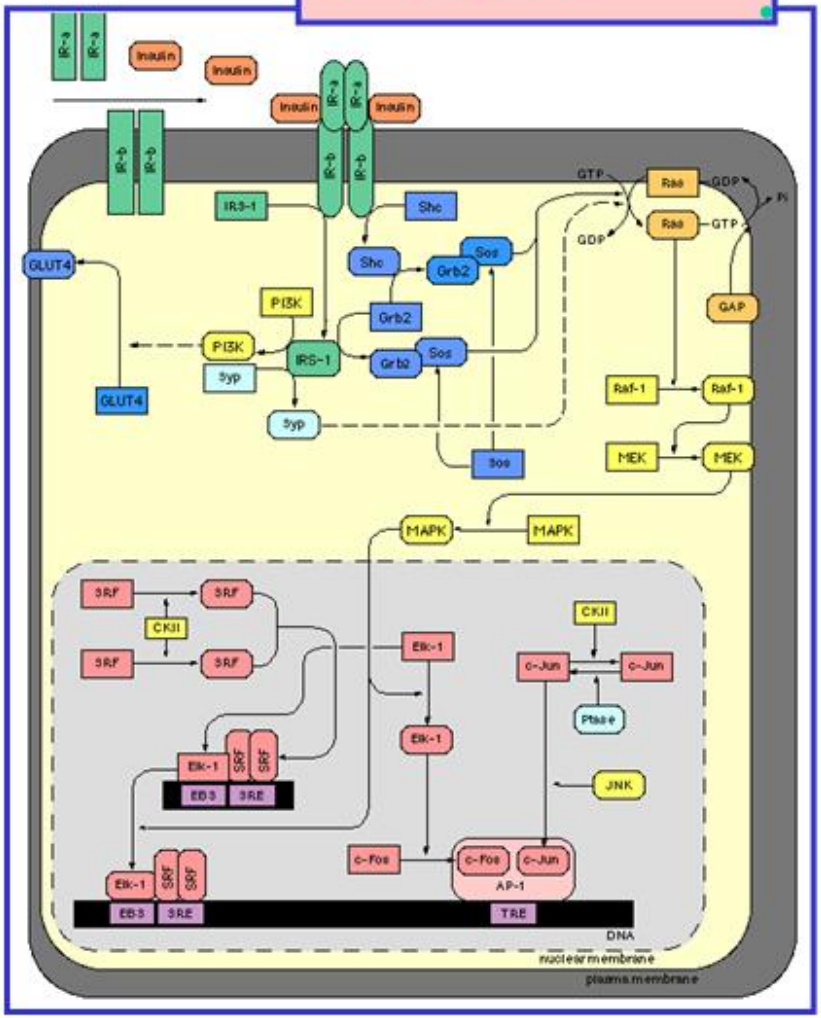
- **Extracellular Signal Molecules**

This WWW service "SPAD"is still under development.

mail to: sachivo@

Molecular Gene Technics
Hakozaki Higashi-ku,
Fukuoka, 812-8581, Japan
Graduate School of Genetic Resources Technology
Kyushu University

Last Update Oct 13, 1998



http://www.grt.kyushu-u.ac.jp/spad/pathway/pdgf.html

PATHWAY DATABASES

KEGG - Table of Contents

KEGG2 PATHWAY GENES LIGAND EXPRESSION BRITE XML API DBGET

1. KEGG Databases

Category	Database		Search & Compute	DBGET Search	
Pathway information	KEGG PATHWAY Database	XML	Search objects in KEGG pathways Color objects in KEGG pathways Generate possible reaction paths	PATHWAY	
Genomic information	KEGG GENES Database	KO	Search similar GENES sequences Search similar GENOME sequences	KO GENES GENOME	
Chemical information	KEGG LIGAND Database	RC	Search similar compound structures Search similar glycan structures Search similar reactions	COMPOUND GLYCAN REACTION ENZYME	LIGAND

2. KEGG Gene Catalogs

2.1 Genomes in KEGG

Category	Genome	DBGET Search
Organism	Complete genomes in KEGG	GENES
	Complete genomes (taxonomy)	DGENES
Virus	Complete viral genomes	VGENES
Organelle	Complete mitochondrial genomes Complete plastid genomes Complete nucleomorph genomes	OGENES

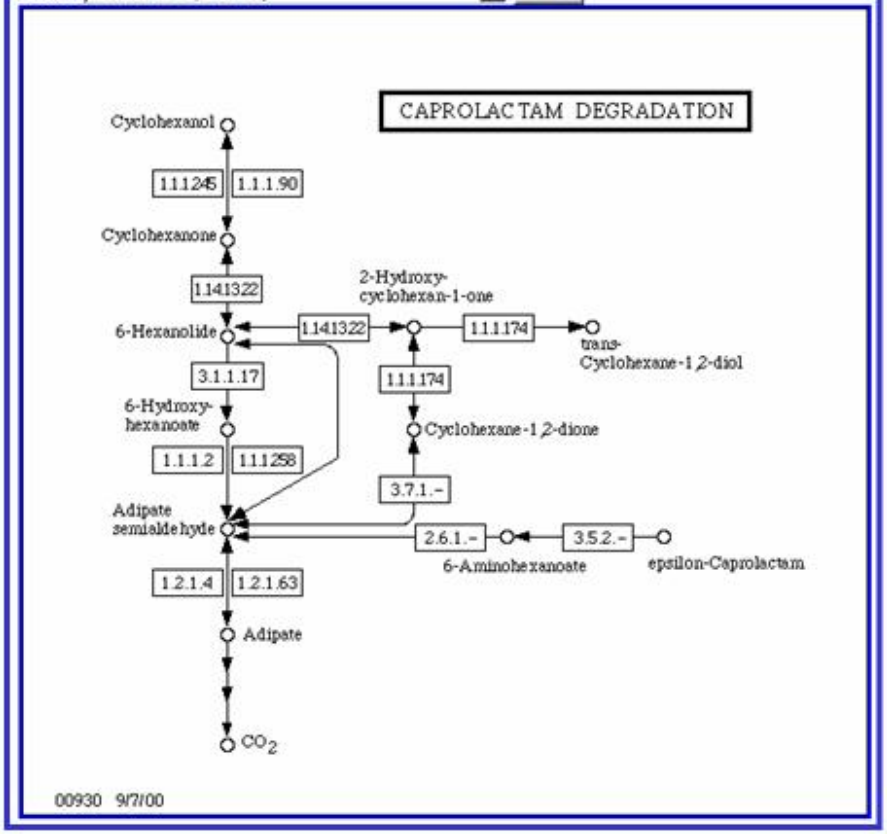
http://www.genome.ad.jp/kegg/kegg2.html

238 Backward

Caprolactam degradation - Reference pathway

Go to: [LinkDB search]

Go to: Reference pathway Exec



Web Resources
Concept 6.5

WEB RESOURCES

EBI (European Bioinformatics Institute).

NCBI
(National Center for Biotechnology

NCBI for Sequences.

EBI for Tools.

ENSEMBL EBI SANGER GENOME VIEWER

http://www.ensembl.or

project Ensembl — The Wellcome Trust Sanger Institute — EBI

Ensembl Genome Browser

Search Ensembl

Search all species for: Anything with: Lookup

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on metazoan genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on the species buttons to the right to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints [details].

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Species - Ensembl v19

Human	34a	NCBI 34	15 Dec 2003
Chimp (pre!)	1	BROAD 1	TBA
Mouse (pre!)	30	NCBI m30	06 May 2003
Rat	3a	RGSC 3.1	15 Dec 2003
Zebrafish (pre!)	2	WTSI Zv2	02 Jul 2003
Fugu	2	Fugu v2.0	03 Mar 2003
Mosquito	2a	MOZ 2	01 Oct 2003
Fruitfly	3a	BGDP 3.1	02 Jul 2003
C. elegans	102	WS 102	02 Jul 2003
C. briggsae	25	cb25.agp8	02 Jul 2003

Sequence Similarity searches — BLAST/SSAHA
Batch data/sequence retrieval — EnsMart
Vertebrate Genome Annotation (VEGA) — Vega
Access to whole genome shotgun data (includes additional species) — Trace Server
Download Ensembl data via FTP — Download

Help and documentation

- Take the Ensembl tour, go through a step-by-step worked example, or read these short papers (Jan 2002, Jan 2003) in Nucleic Acids Research.
- For help on any web page click: Help
- There is also an index of help pages, and a set of guided How do I...? trails.

Recent Ensembl news — News
Display your own data in Ensembl — DAS
Apollo genome browser — Apollo
Questions or suggestions? Try the Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site) — Help Desk / Documentation

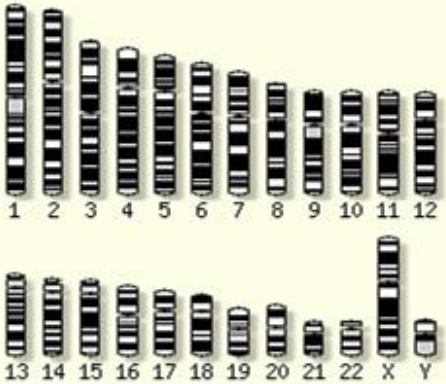
Have you tried ....?

Ensembl Chimp Preview Browser

A preliminary Chimp assembly is now available at http://pre.ensembl.org

Click for more information

Browse a Chromosome

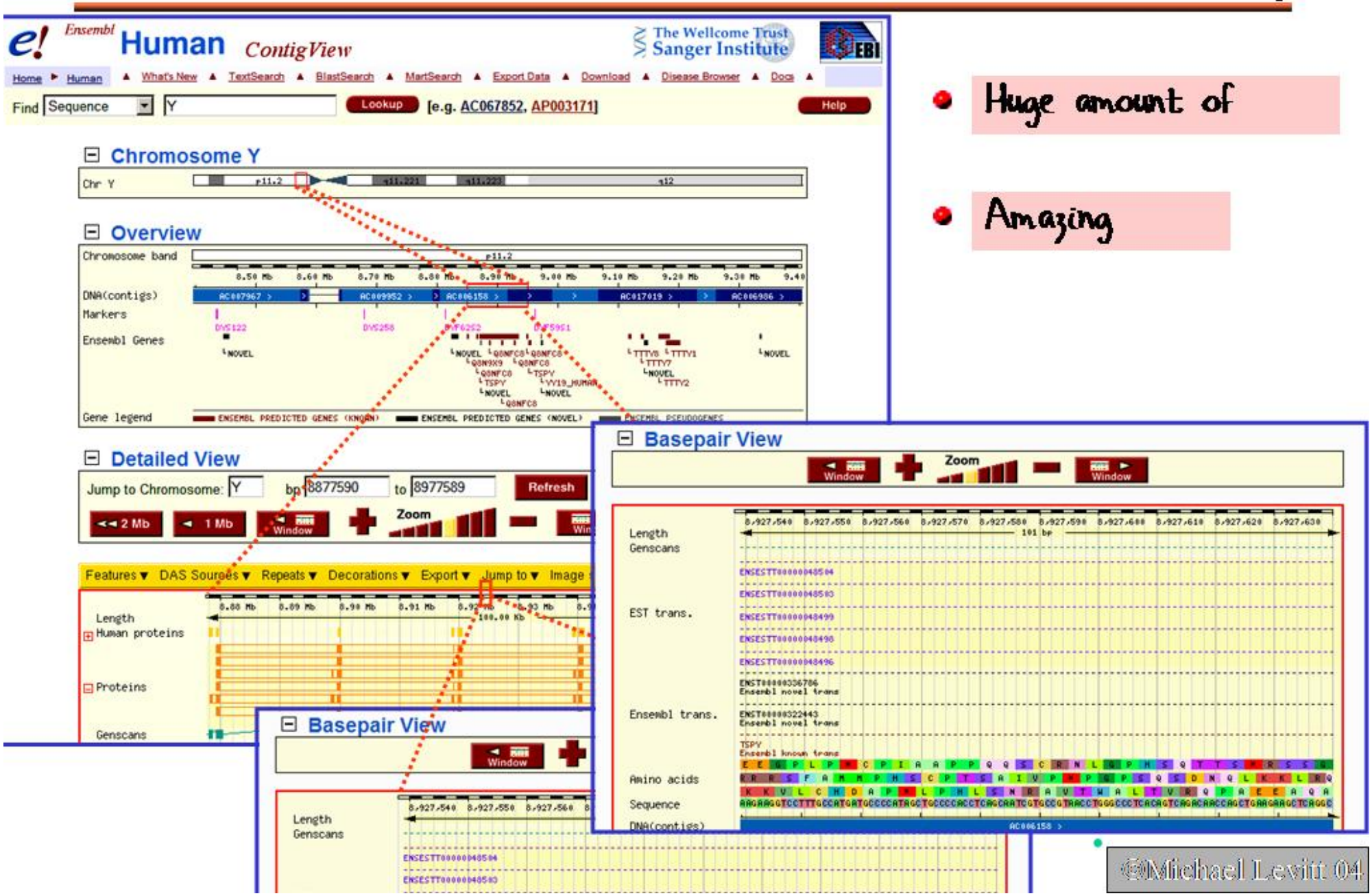


Current Release 19.34a.1

This release is based on the NCBI 34 assembly of the human genome.
View the status history of the human assemblies.
Last Update: 08-01-2004

Ensembl gene predictions (incl. 1744 pseudogenes):	23531
Genscan gene predictions:	65010
Ensembl gene exons:	225897
Ensembl gene transcripts:	31609
Contigs:	26614
Clones:	26614
Base Pairs:	3201762515
Golden Path Length:	2841366484

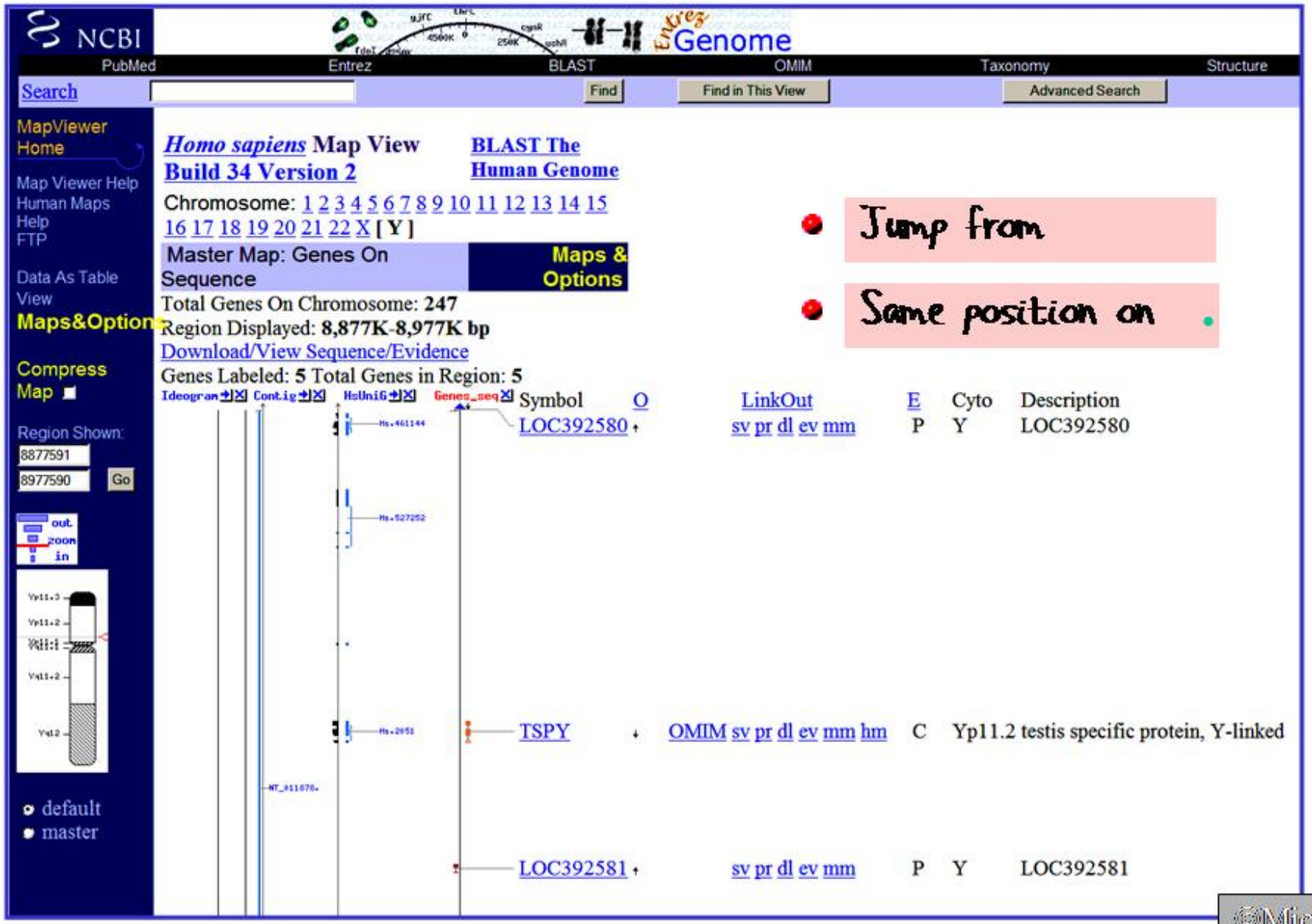
ENSEMBL EBI SANGER GENOME VIEWER



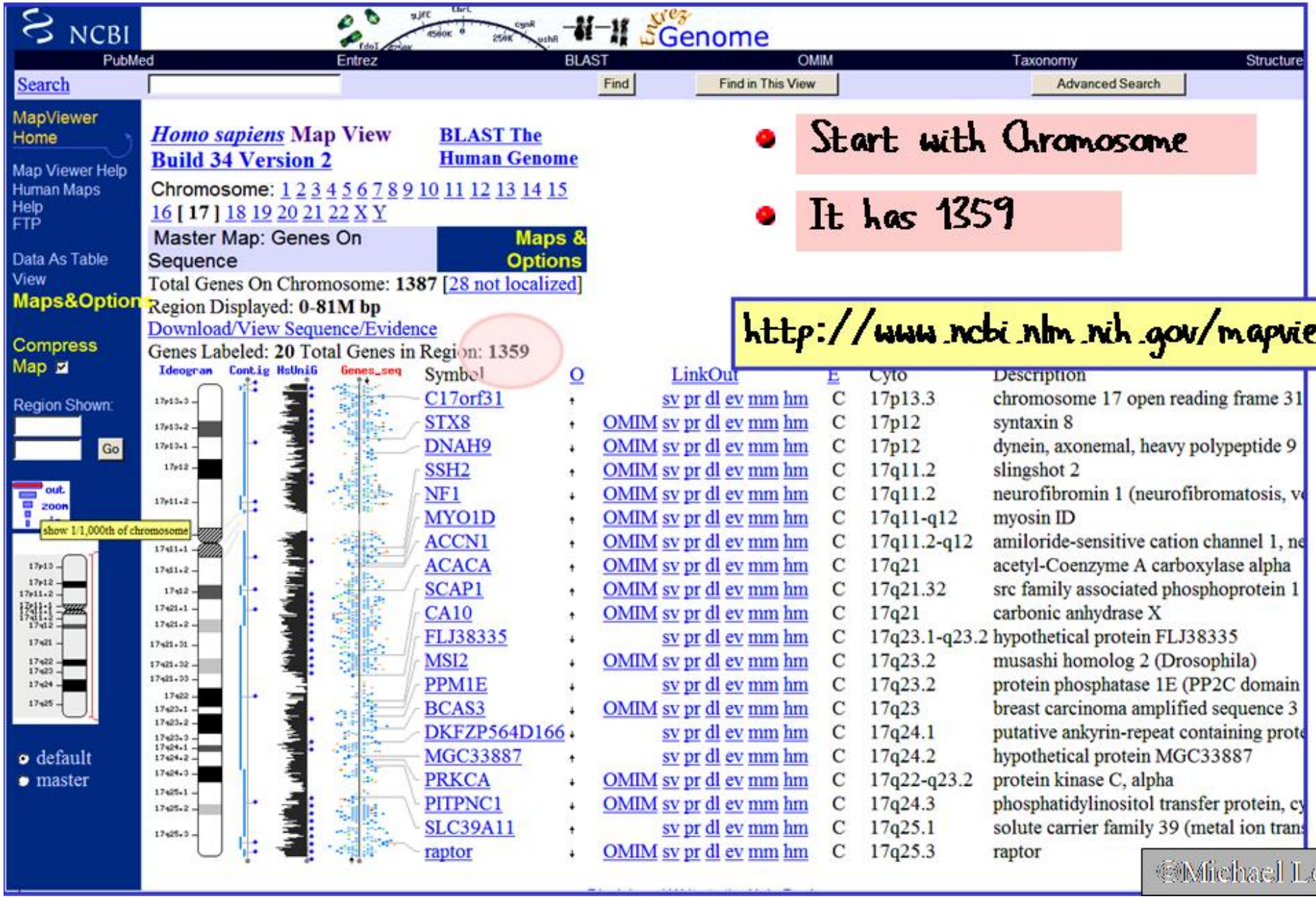
- Huge amount of
- Amazing

©Michael Levitt 04

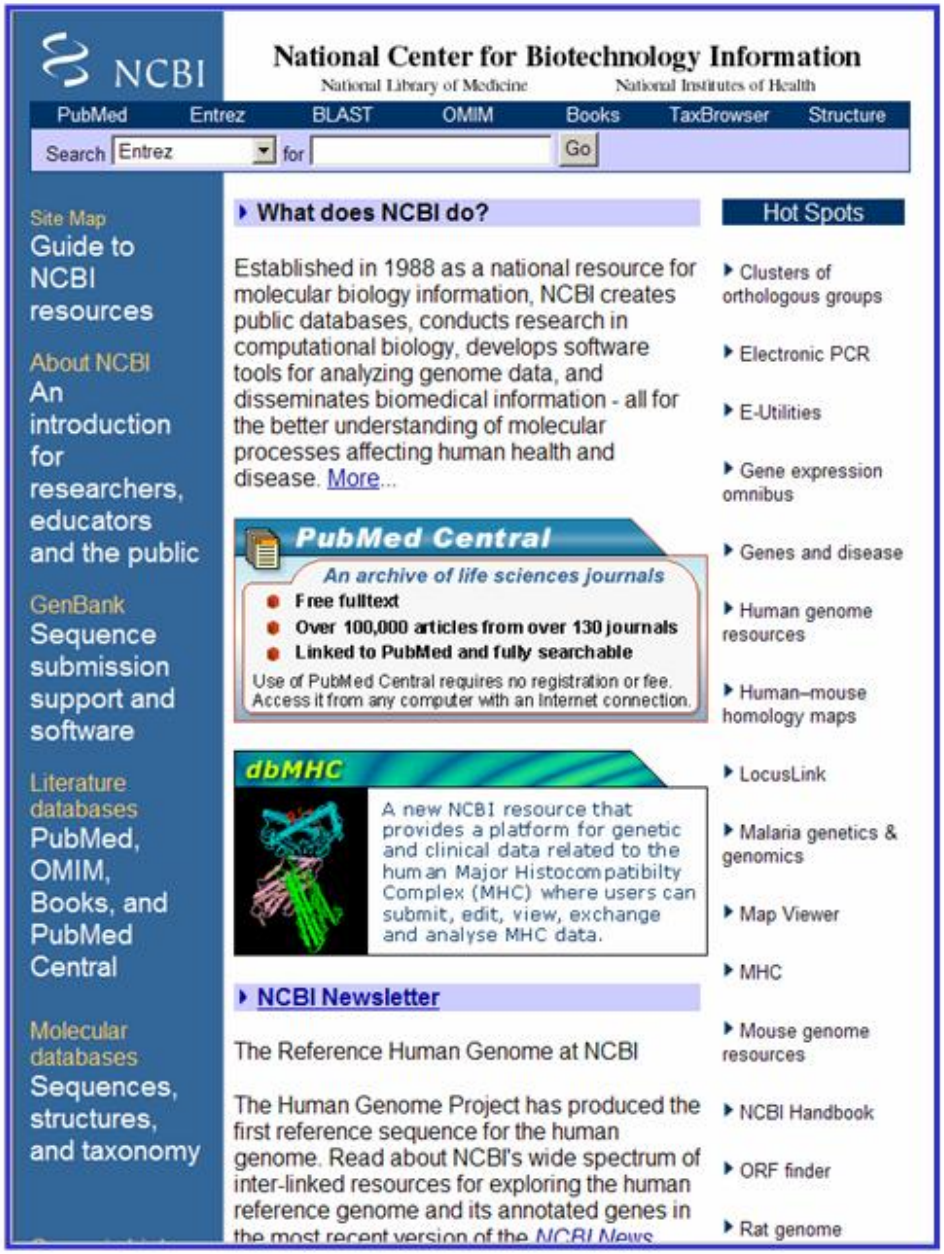
NCBI GENOME VIEWER



NCBI GENOME VIEWER



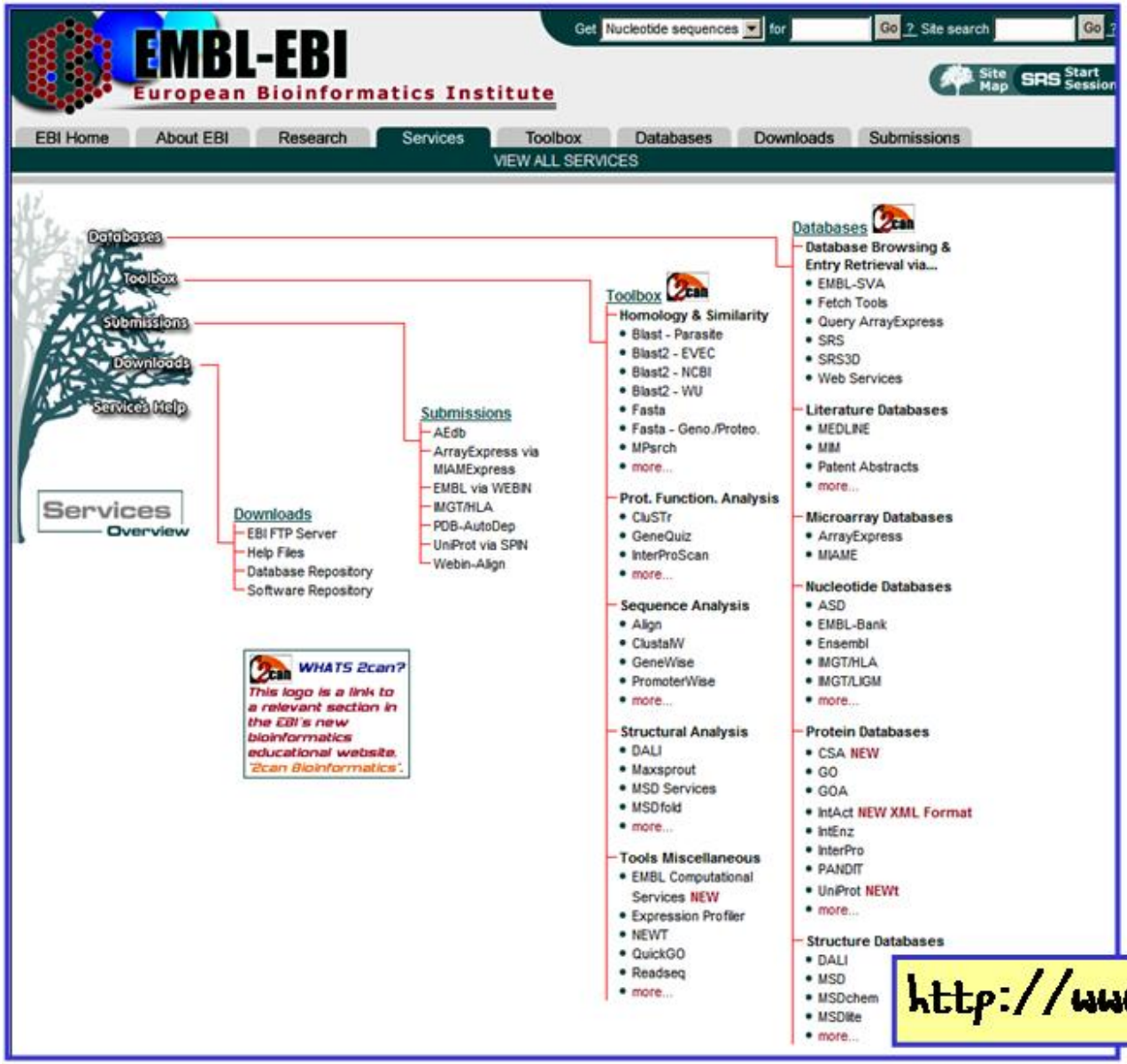
NCBI IS THE PLACE FOR SEQUENCE



- PubMed Central (book
- Entre
- Map
- 17,300 Backward

http://www.ncbi.nlm.nih.go

EBI IS THE PLACE FOR TOOLS



- FAST
- Clustal
- Expression
- DAL
- 7,200 Backward

http://www.ebi.ac.uk/services/index.html

Sequence Comparison Concept 6.6

SEQUENCE COMPARISON

Ungapped

Gapped Comparison.

Scores and Penalties.

Advanced Methods.

IDENTICAL COMPARISON

Sequence A: Y G T P W R S A A Q

Sequence B: Y G T P W R S A A Q

Identical

	Y	G	T	P	W	R	S	A	A	Q
Y	1									
G		1								
T			1							
P				1						
W					1					
R						1				
S							1			
A								1	1	
A								1	1	
Q										1

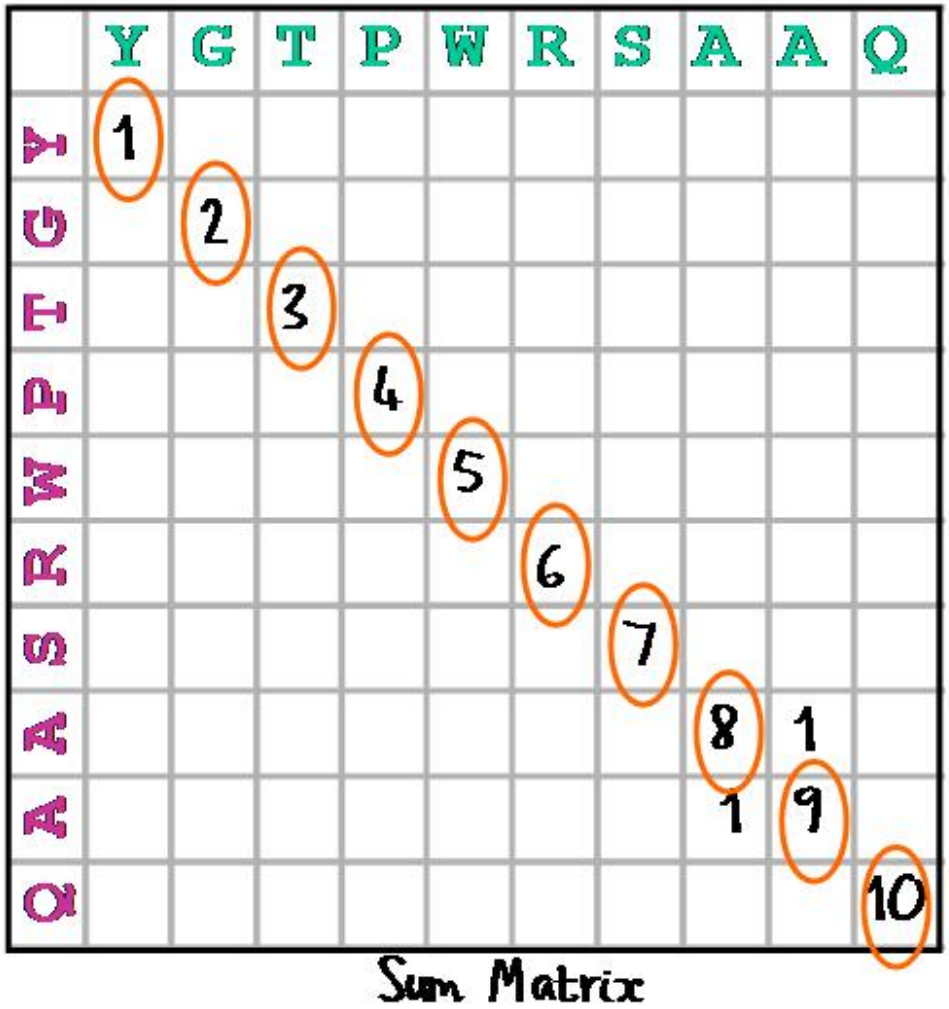
- Assume that score is 1 for a match and 0 otherwise.
- Mark all the matches.

IDENTICAL TRACE

Sequence A: Y G T P W R S A A Q

Sequence B: Y G T P W R S A A Q

Identical



Sum Matrix

- Start at the top left and move down the diagonal from high-scoring match to high-scoring match.
- Add the scores along the path.
- Find the path that goes from top left to bottom right that collects the highest score.

Total score is 10

MISMATCH COMPARISON

Sequence A: Y G T P W R S A A Q

Sequence B: Y G P T W R S A Q A

Mismatches

	Y	G	P	T	W	R	S	A	Q	A
Y	1									
G		1								
T				1						
P			1							
W					1					
R						1				
S							1			
A								1		1
A								1		1
Q									1	

Similarity Matrix

- The sequences are the same length.
- There are four differences or mismatches.

MISMATCH TRACE

Sequence A: Y G T P W R S A A Q

Sequence B: Y G P T W R S A Q A

Mismatches

	Y	G	P	T	W	R	S	A	Q	A
Y	1									
G		2								
T				1						
P			1							
W					2					
R						3				
S							4			
A								5		1
A								1		1
Q									1	

Sum Matrix

- As the gap is two positions wide, it costs more than a single width gap.
- Assume the gap cost is 1.
- Include it in the count after skipping the gap.

Total score is 5

DELETION COMPARISON

Sequence A: Y G T P W R S A A Q

Sequence B: Y G T W R S A A Q

Deletion in B

	Y	G	T	W	R	S	A	A	Q	
Y	1									
G		1								
T			1							
P										
W				1						
R					1					
S						1				
A							1			
A								1	1	
Q								1	1	

Similarity Matrix

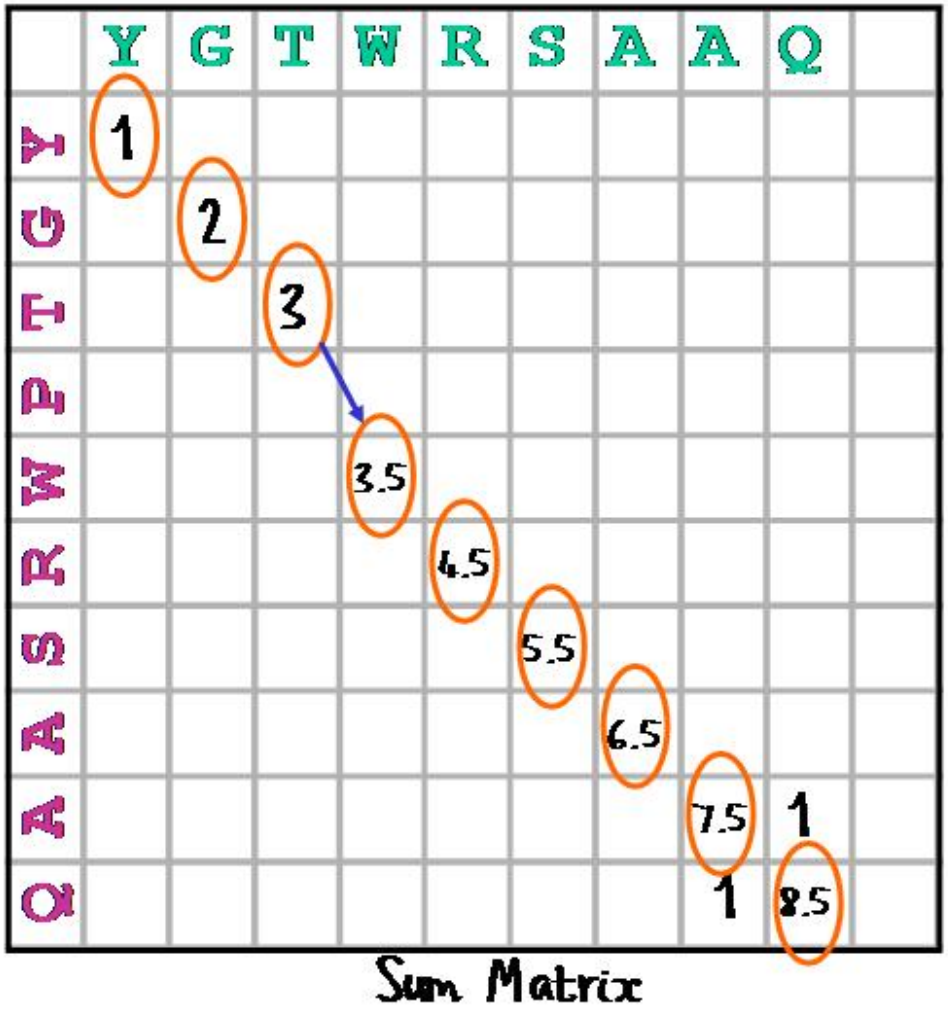
- Sequence B is now shorter than sequence A.

DELETION TRACE

Sequence A: Y G T P W R S A A Q

Sequence B: Y G T W R S A A Q

Deletion in B



Sum Matrix

- Assume that the gap costs 0.5.
- Include this cost in the total score after the crossing the gap.

Total score is 8.5

DELETION/INSERTION COMPARISON

Sequence A: Y G T P W R S A A Q

Sequence B: Y G W R S Y G A A Q

	Y	G	W	R	S	Y	G	A	A	Q
Y	1					1				
G		1					1			
T										
P										
W			1							
R				1						
S					1					
A								1	1	
A								1	1	
Q										1

Similarity Matrix

Deletion and insertion in B

- The two sequences are the same length.
- There appears to be a long mis-match but it is really a deletion followed by an insertion.

DELETION/INSERTION TRACE

Sequence A: Y G T P W R S – – A A Q

Sequence B: Y G – – W R S Y G A A Q

	Y	G	W	R	S	Y	G	A	A	Q
Y	1					1				
G		2					1			
T										
P										
W			2							
R				3						
S					4					
A								4	1	
A								1	5	
Q										6

Sum Matrix

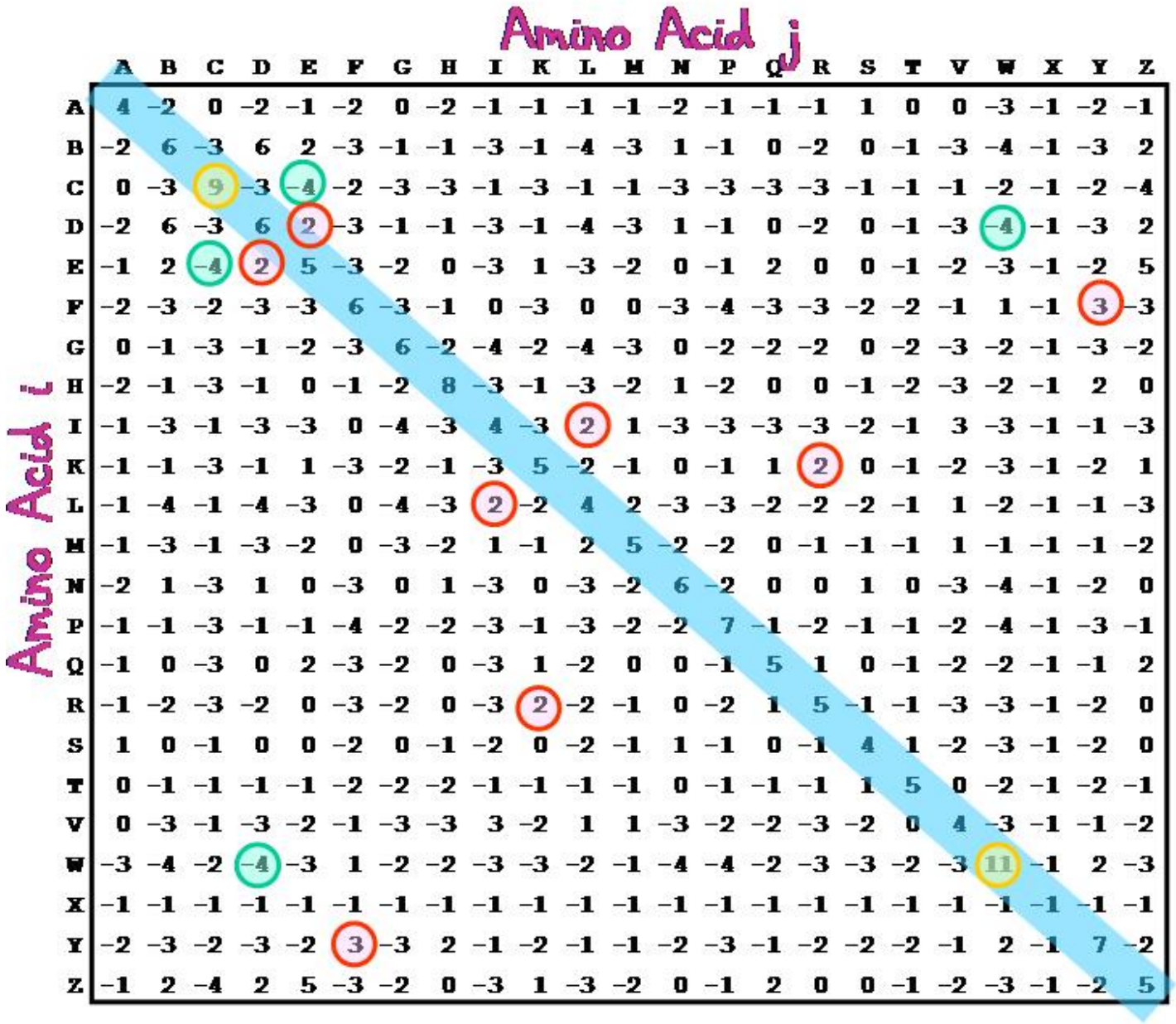
Deletion and insertion in B

- Each gap skips over two positions and is assumed to cost 1.
- Add this cost into the score after crossing each gap.

Total score is 6

Blosum62 score matrix (Henrikoff)

SCORING MATRIX



	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	4	-2	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-1	-2	-1
B	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
C	0	-3	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-1	-2	-4
D	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
E	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5
F	-2	-3	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	-1	3	-3
G	0	-1	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-1	-3	-2
H	-2	-1	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	-1	2	0
I	-1	-3	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1	-1	-3
K	-1	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-1	-2	1
L	-1	-4	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1	-1	-3
M	-1	-3	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1	-1	-2
N	-2	1	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-1	-2	0
P	-1	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-1	-3	-1
Q	-1	0	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1	-1	2
R	-1	-2	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-1	-2	0
S	1	0	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-1	-2	0
T	0	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-1	-2	-1
V	0	-3	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1	-1	-2
W	-3	-4	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	-1	2	-3
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-2	-3	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	-1	7	-2
Z	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5

- Matrix entry is the score of amino acid i matched to amino j.
- Symmetric.
- Strongest matches:
  C..C = 9
  W..W = 11
- Weakest matches:
  C..E = -4
  W..D = -4
- Similar matches:
  F..Y = 3
  K..R = 2
  D..E = 2
  I..L = 2

GAP PENALTIES

- The gap penalty is usually proportional to the number of positions skipped as follows:

GAP	PENALTY
0	0
1	a
2	a + b
3	a + 2b
n	a + (n−1)b

- The optimum gap value depends on the Scoring Matrix.
- For the Blosum62 matrix, the best values are:

  a = −10, b = −2

Non-linear gaps are possible but make calculation much slower.

DYNAMIC PROGRAMMING

Best To I,J = Score at I,J + Best in block + Cost of gaps.

	Y	G	W	R	S	Y	G	A	A	Q
Y	1					1				
G		1					1			
T										
P										
W			1							
R				1						
S					1					
A								1	1	
A								1	1	
Q										1

Similarity Matrix

DYNAMIC PROGRAMMING

In local alignment, score can never be less than 0.

	Y	G	W	R	S	Y	G	A	A	Q
Y	1					1				
G		1					1			
T										
P										
W			1							
R				1						
S					1					
A								1	1	
A								1	1	
Q										1

Similarity Matrix

	Y	G	W	R	S	Y	G	A	A	Q
Y	1	0	0	0	0					
G	0	2	0.5	0	0					
T	0	0.5	2	1.5	1					
P	0	0	1.5	2	1.5					
W	0	0	2	1.5	2					
R										
S										
A										
A										
Q										

Sum Matrix

DYNAMIC PROGRAMMING

BestTo[I,J] = Score[I,J] + max { BestTo[i,j] + gap_IJ_to_ij

where gap_IJ_to_ij = max{I-i-1, J-j-1}/2, say.

	Y	G	W	R	S	Y	G	A	A	Q
Y	1					1				
G		1					1			
T										
P										
W			1							
R				1						
S					1					
A								1	1	
A								1	1	
Q										1

Similarity

	Y	G	W	R	S	Y	G	A	A	Q
Y	1	0	0	0	0	0				
G	0	2	0.5	0	0	0				
T	0	0.5	2	1.5	1	0.5				
P	0	0	1.5	2	1.5	1				
W	0	0	2	1.5	2	1.5				
R	0	0	0.5	3	1.5	2				
S										
A										
A										
Q										

Sum

ADVANCED METHODS

- Dynamic programming is slow as one needs to calculate a score for every cell of the similarity and score matrix. This is $O(n^2)$ at best.

- It can be speeded up by only looking in regions of the similarity matrix where there are high scores.

FASTA

Look for identities of single amino acids or pairs.

Mark every single identity.

	Y	G	W	R	S	Y	G	A	A	Q
Y	1					1				
G		1					1			
T										
P										
W			1							
R				1						
S					1					
A								1	1	
								1	1	
Q										1

Pearson & Lipman, PNAS, 85, 2444 (1988).

Similarity Matrix

Sum scores along diagonals with identities.

	Y	G	W	R	S	Y	G	A	A	Q
Y	1					1				
G		2					2			
T										
P										
W			1							
R				2						
S					3					
A								1		
A									2	
Q										3

Sum Matrix

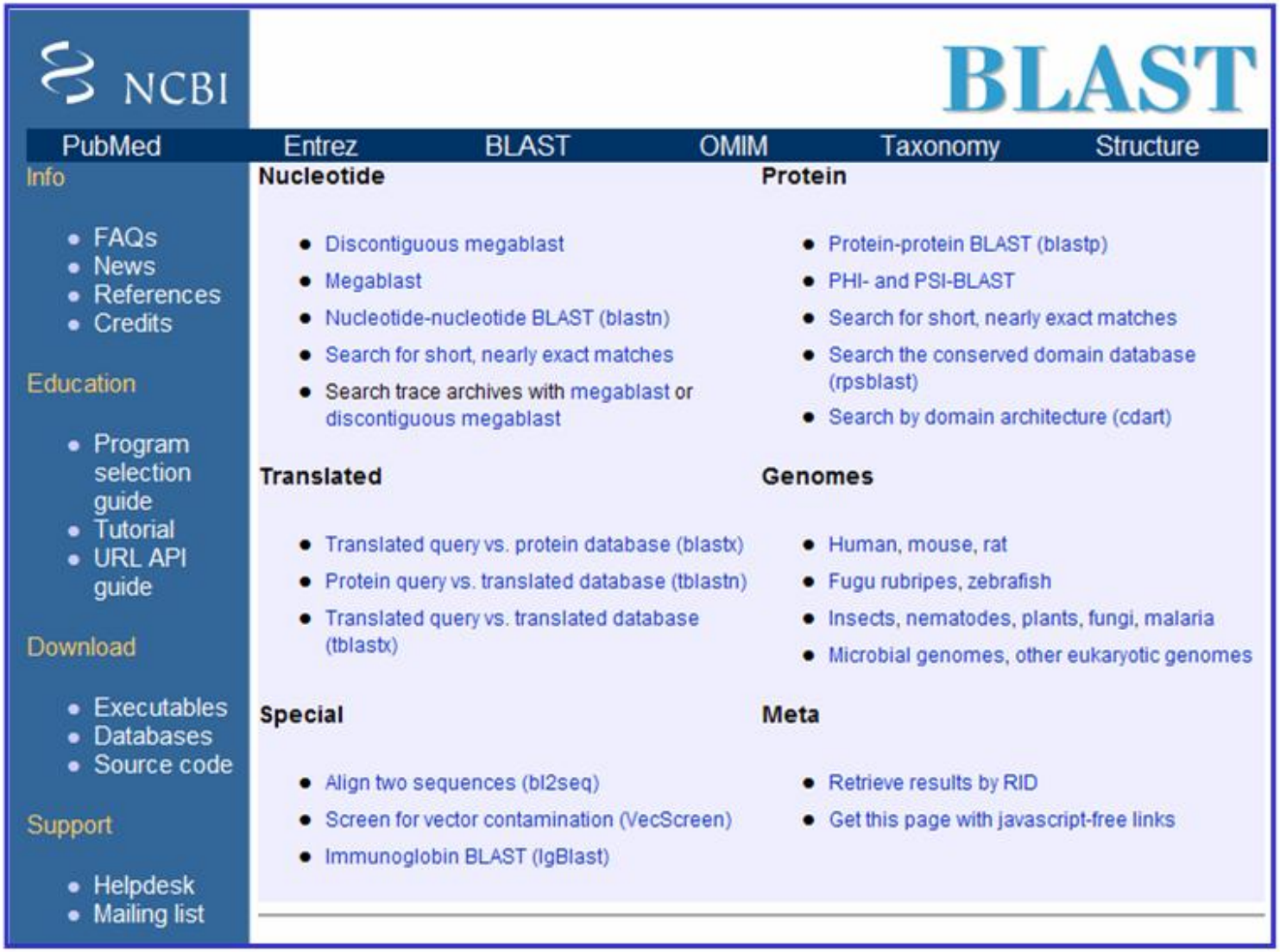
BLAST AND PSI-BLAST

- Look for triplets that have high match scores. For example with Blosum62: **Y K D** is a good match to **F R E** with a score of 7.
- Mark these on the Similarity Matrix.
- Extend these diagonals in the Sum Matrix.
- Merge separate fragments.

Altschul et al. J Mol. Biol. 215: 403 (1990)

PSI-BLAST works with a profile rather than a sequence (more later). It is very, very clever

BLAST: COMPARE SEQUENCES



Blast is a amazing resource.

Play with it.

This is the only way to learn.

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

THE END

of Lecture 6