

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
897932384626**

HUMAN GENOME

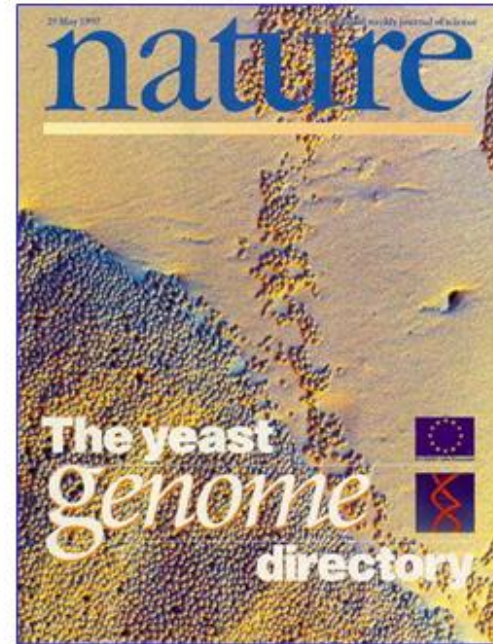
Science
28 Jul 95

Nature
29 May 97

Science
11 Dec 98

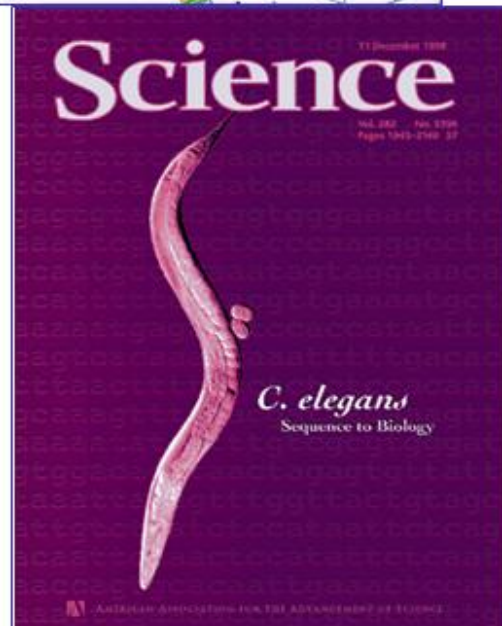
Science
16 Feb 01

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]

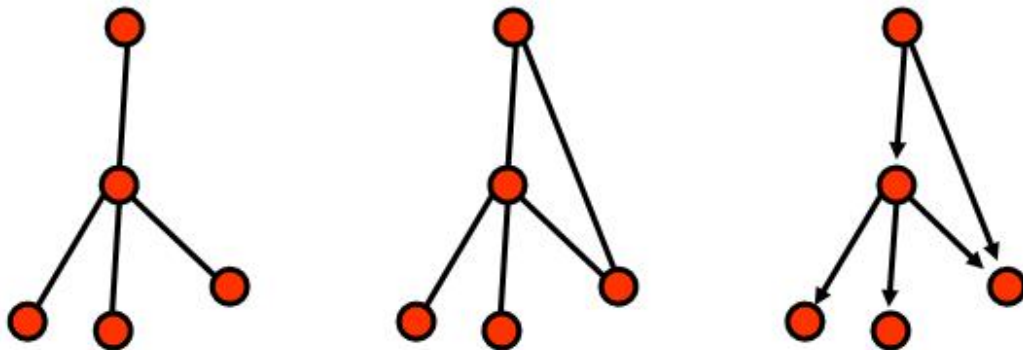
RELATIONSHIPS

Trees.

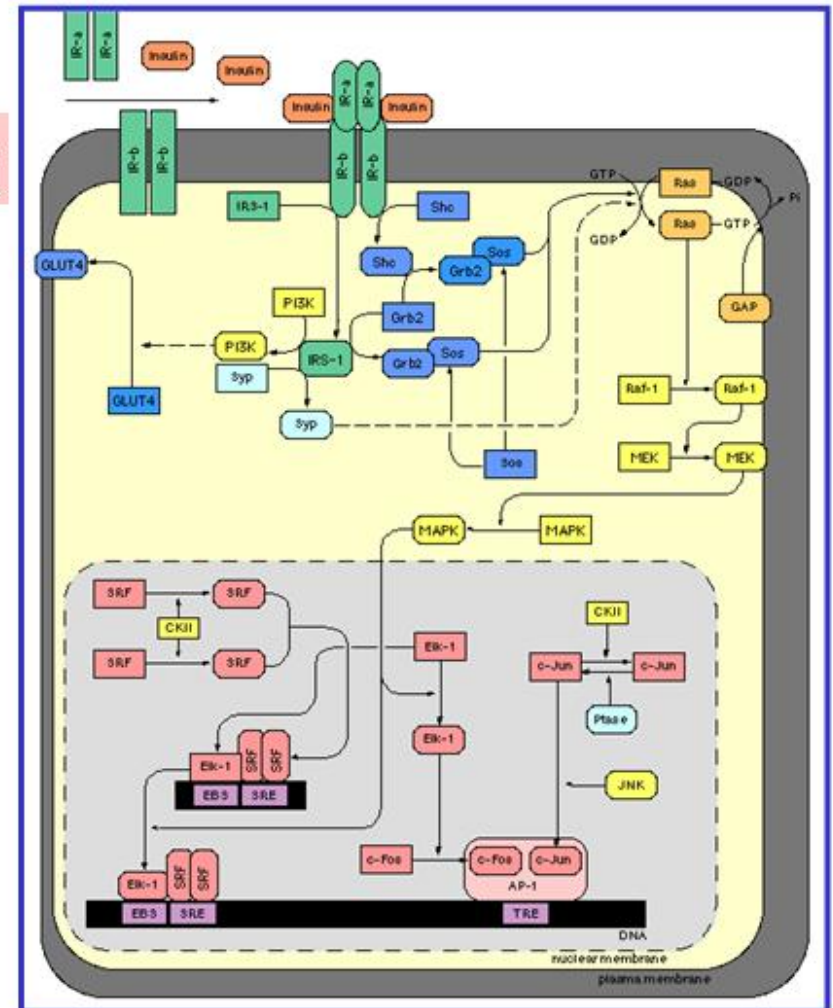
Graphs.

Directed Graphs.

Pathways.



Insulin Pathway



<http://www.grt.kyushu-u.ac.jp/spad/pathway/pdf.html>

MULTIPLE SEQUENCE ALIGNMENTS

Initial consensus alignment

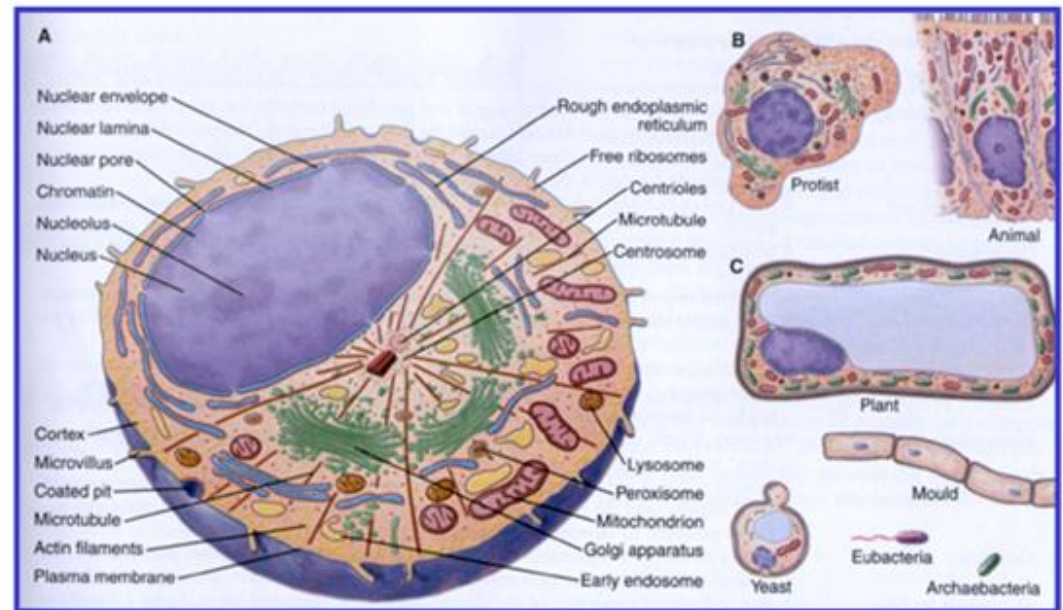
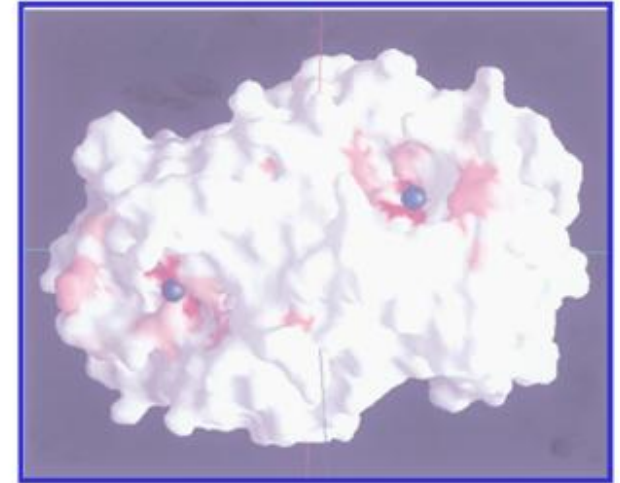
```
1CEL > >SACTLQSETHPPLTWQKCSSGGTCTQQTGSVVIDANWRWTHATNSSTNCYDGNTWSSTLCPDNETCAK---NCCLL
1EG1 > >QPGTSTPEVHPKLT TYKCTKSGGCVAQDTSVVL DWN YRWMH-DANYNSCTVNGGVNTTLC P---DEATCGKNC FIE
2A39 > >KPGET-KEVHPQLTTFRCKRGGCKPATNFIVLDSL SHPIHRAEGLPGGGCGDWGNPPP KDVC PDV ESCAKNCIME
2OVW > >TPDKA-KEQHPKLETYRCKASGCKKQTNYIVADAGIHGIRQKNG---AGCGDWGQKPNATACPDEASCAKNCILS
```

Resolved alignment

```
1CEL > >SACTLQSETHPPLTWQKCSSGGTCTQQTGSVVIDANWRWTHA-TNSSTNCYDGN---TWSSTLCPDNETCAKNCCLDG
1EG1 > >QPGTSTPEVHPKLT TYKCTKSGGCVAQDTSVVL DWN YRWMHD--ANYNSCTVNG---GVNTTLC PD EATCGKNC FIEG
2A39 > >KPGE-TKEVHPQLTTFRCKRGGCKPATNFIVLDSL SHPIHRA-EGLPGGGCGDWGNPPP KDVC PDV ESCAKNCIMEG
2OVW > >TPDK-AKEQHPKLETYRCKASGCKKQTNYIVADAGIHGIRQ----KNGAGCGDWGQKPNATACPDEASCAKNCILSG
```

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.

$$\text{Mean} = \frac{(2 \times 1 + 3 \times 2 + 4 \times 3 + 5 \times 4 + 6 \times 5 + 7 \times 6 + 8 \times 5 + 9 \times 4 + 10 \times 3 + 11 \times 2 + 12 \times 1)}{(1 + 2 + 3 + 4 + 5 + 6 + 5 + 4 + 3 + 2 + 1)} = \frac{252}{36} = 7$$

$$2 = 1 + 1$$

$$3 = 1 + 2, 2 + 1$$

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

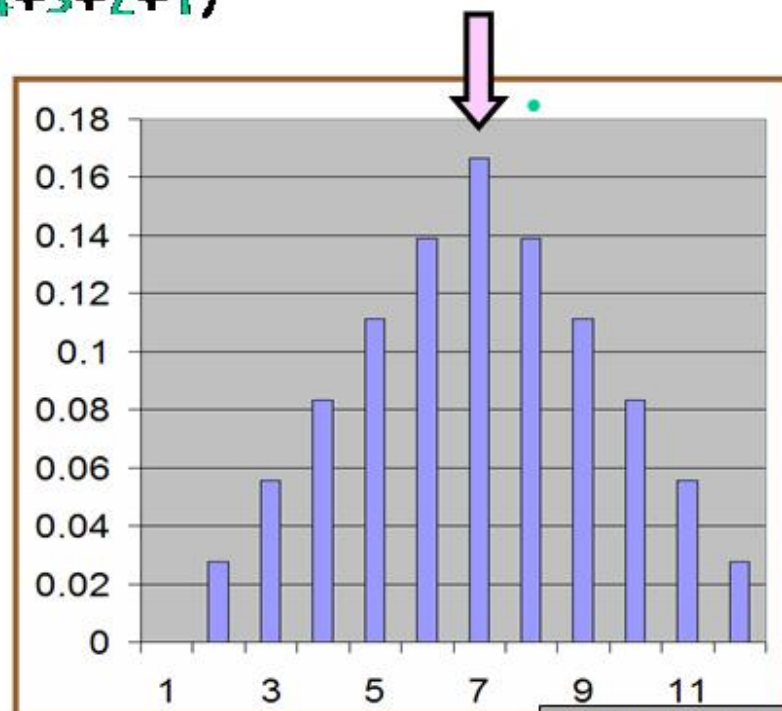
$$5 = 2 + 3, 3 + 2, 1 + 4, 4 + 1$$

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

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

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

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



©Michael Levitt 04

STANDARD DEVIATION

Mean Squared

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

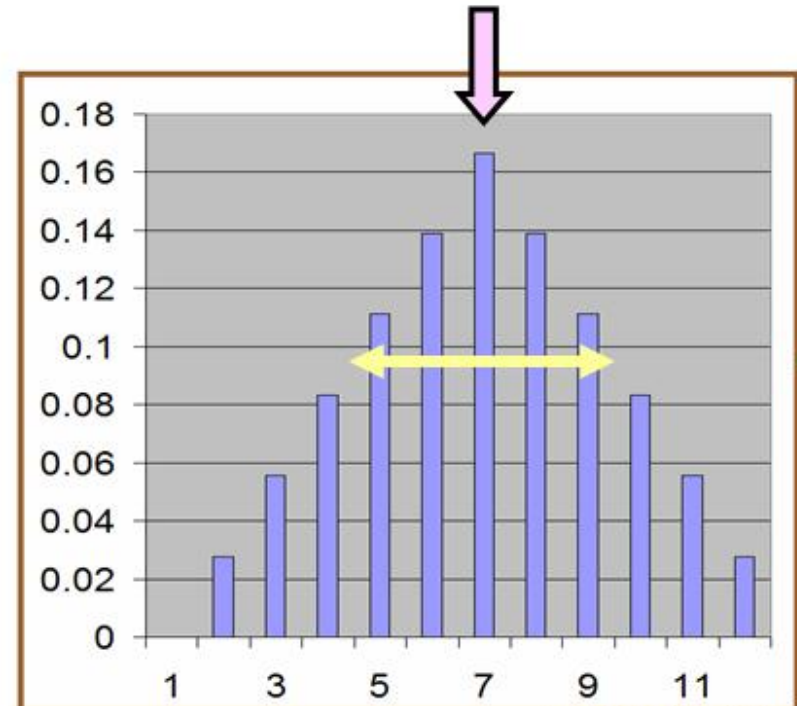
$$= 1974 / 36 = 54.83$$

Mean Squared

(Mean)²

$$\text{Variance} = (54.83 - 7^2) = 5.83$$

$$\begin{aligned} \text{SD} &= \text{sqrt}(\text{Variance}) \\ &= \text{sqrt}(5.83) \\ &= 2.415 \end{aligned}$$



UNIFORM DISTRIBUTION

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

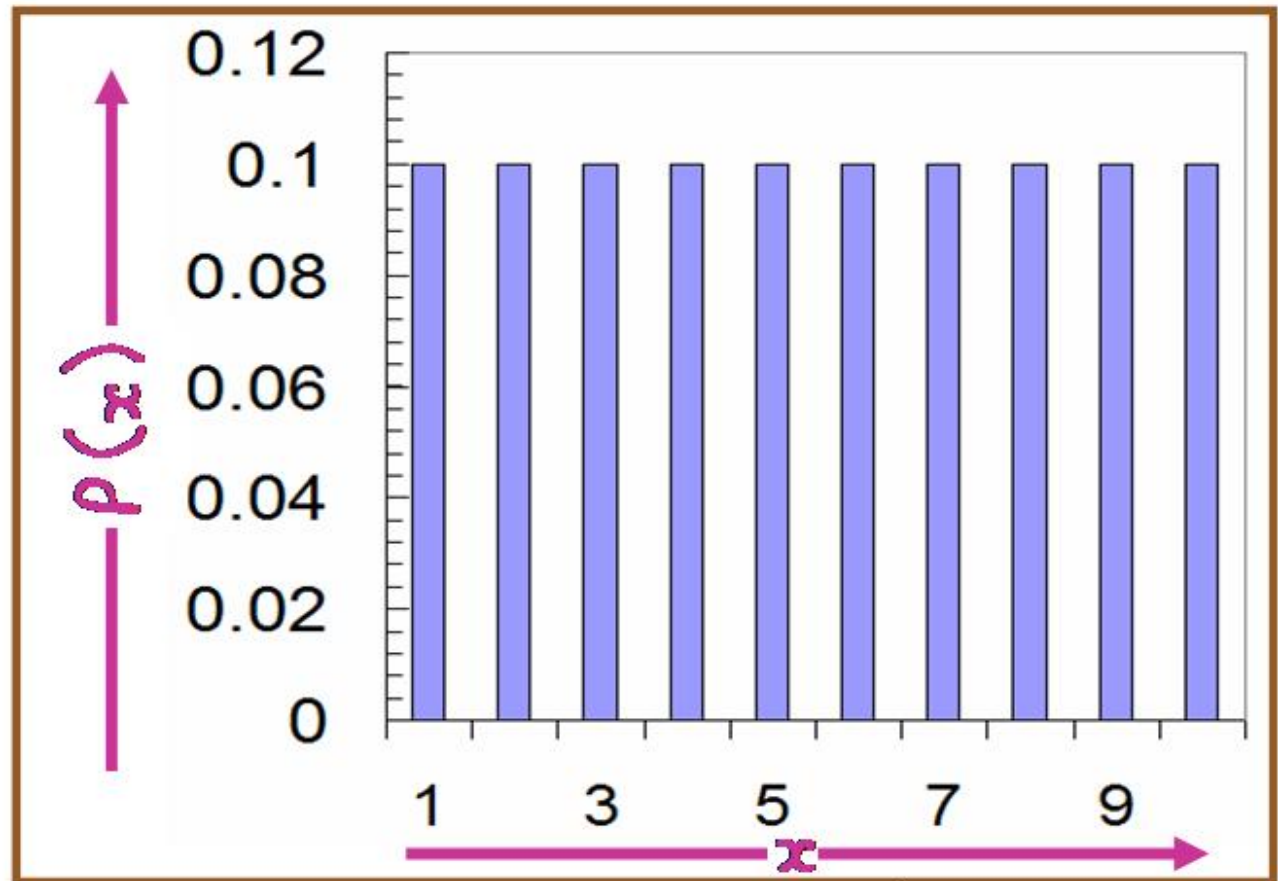
$$= 0 \text{ otherwise}$$

$$\text{Mean} = \min + \max / 2$$

$$= 5.5$$

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

$$= 1.5$$



BINOMIAL DISTRIBUTION

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

$$\text{Mean} = Np$$

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

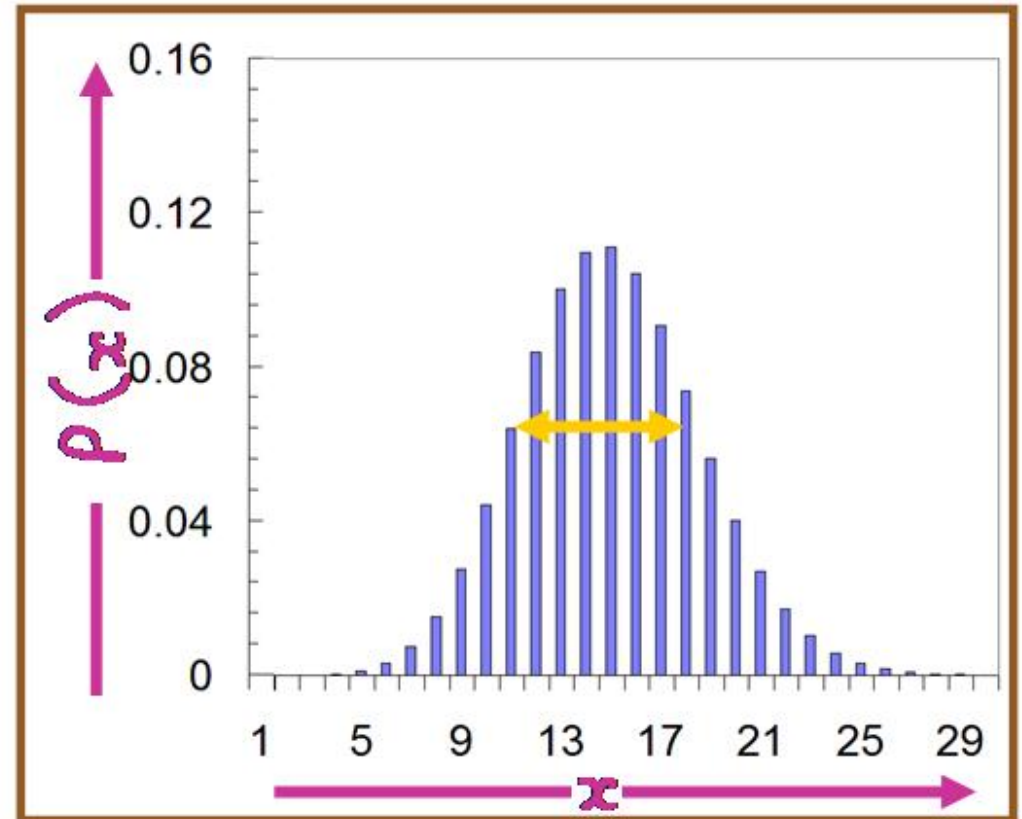
$$= \text{Mean} \times (1-p)$$

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	6
1	1	0	0	1	0	1	1	0	0	0	4
1	1	1	0	0	0	0	1	0	0	0	5
1	1	1	1	1	1	1	0	1	0	1	8
1	1	1	0	1	1	0	1	0	0	0	5
0	0	0	1	0	0	0	1	1	1	1	5
1	0	1	0	0	1	1	1	1	0	0	7
1	0	1	0	1	0	0	0	0	1	0	5
0	0	1	0	1	0	0	0	1	1	0	4

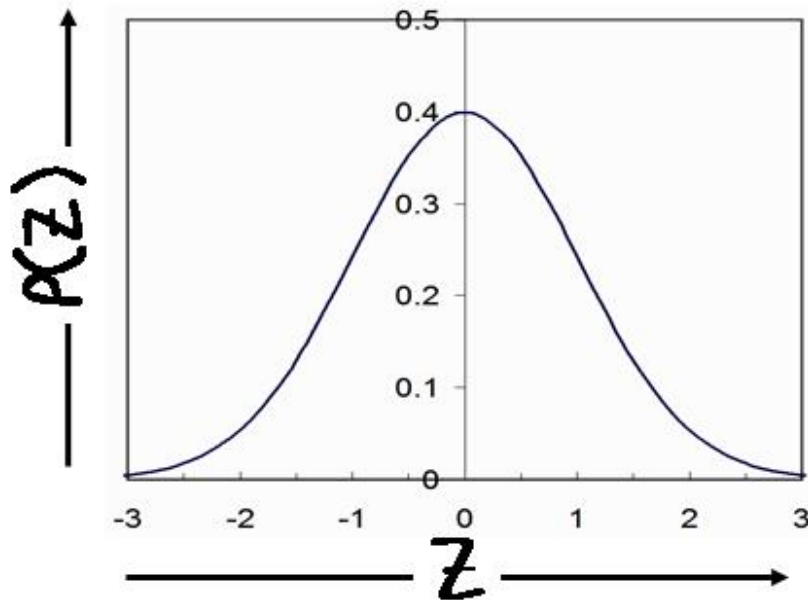
For $N=30, p=0.5,$
 Mean = 15, $\text{SD}^2 = 7.5$

For $N=100, p = 0.15,$
 Mean = 15, $\text{SD}^2 = 15 \times 0.85$

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



NORMAL DISTRIBUTION



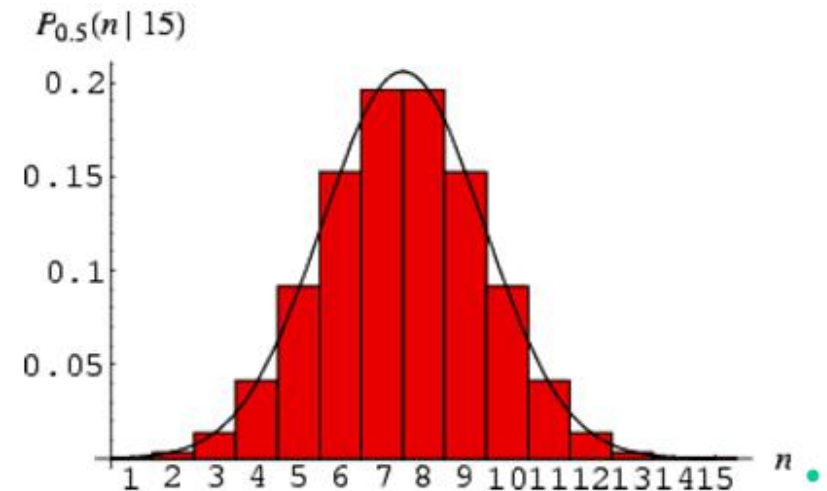
Gaussian:

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

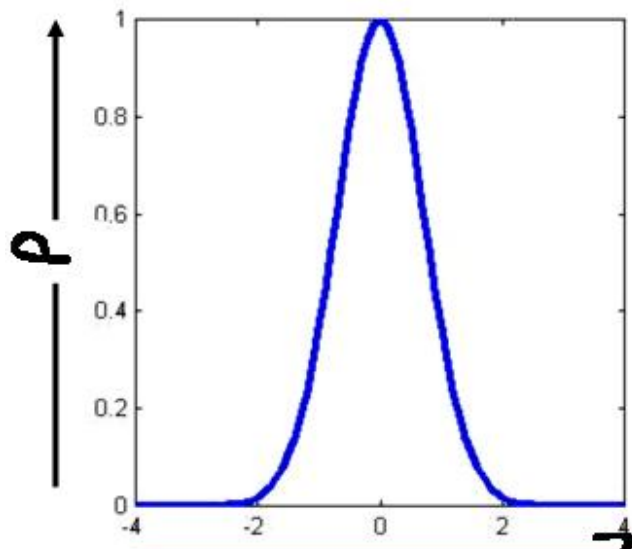
$$\text{Mean} = 0$$

$$\text{SD} = 1$$

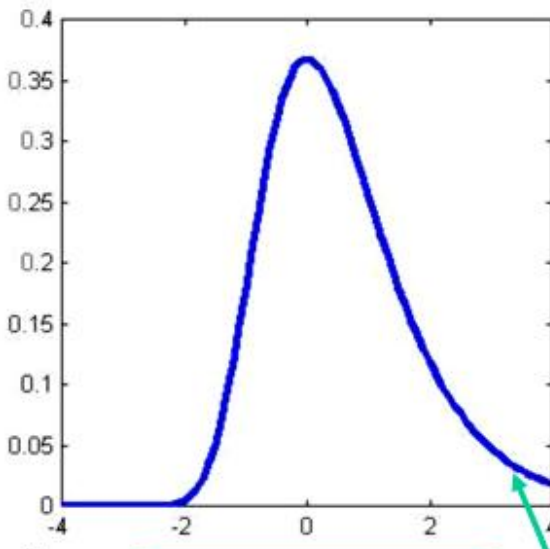
The Binomial distribution
Becomes like a Normal
Distribution as the sample
size increases as the sample
Size increases.



EXTREME VALUE DISTRIBUTIONS



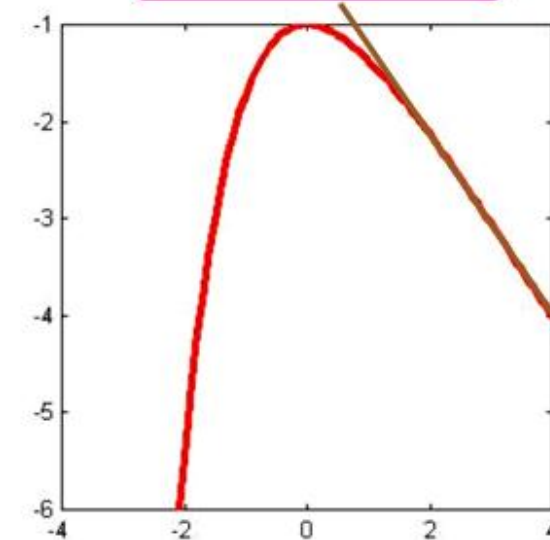
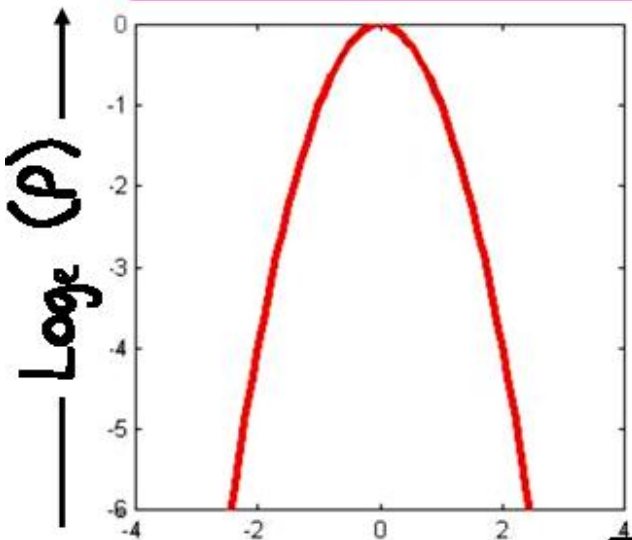
Normal (or Gaussian)



Extreme Value

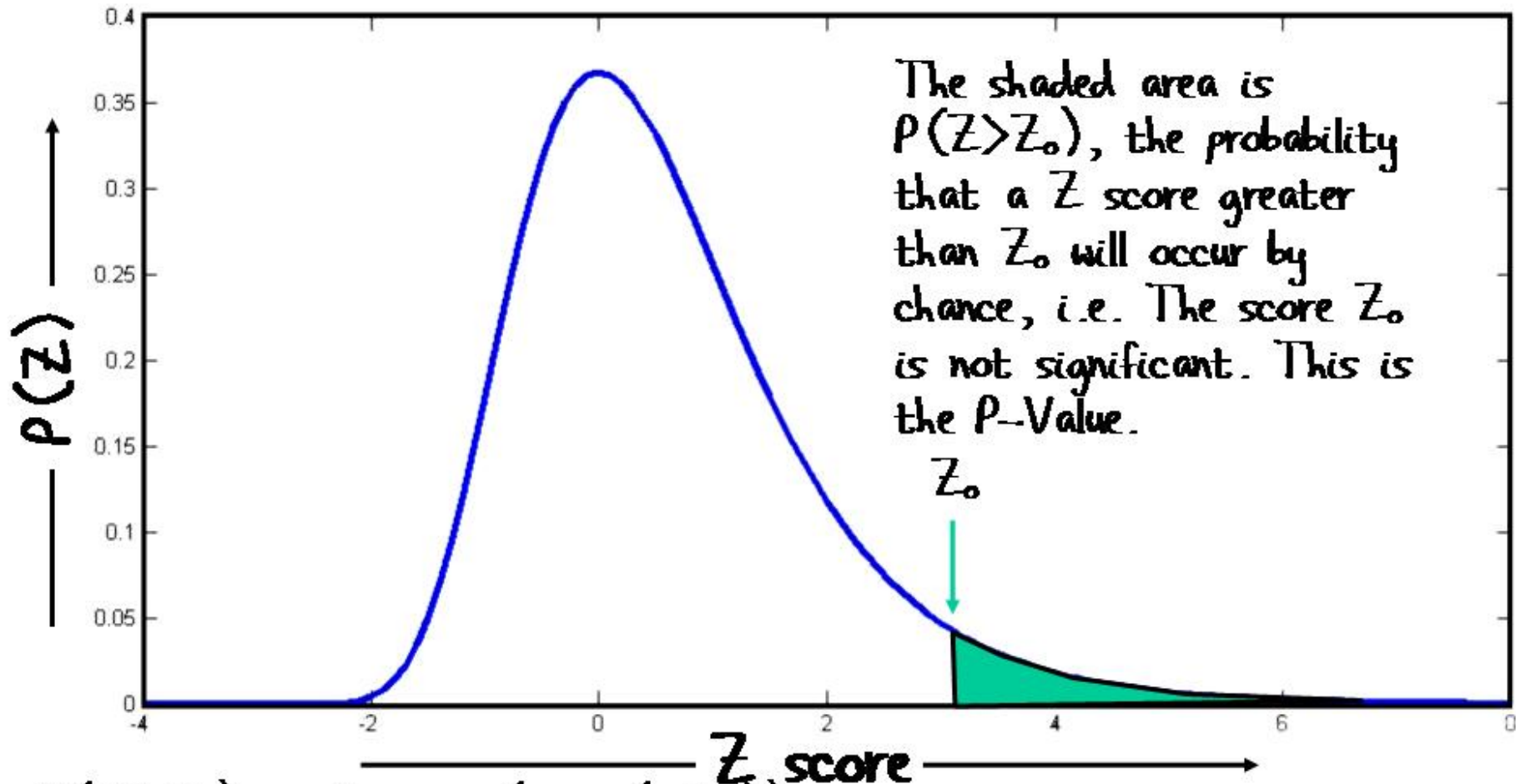
- Normal:
 $P(Z) = \exp(-Z^2)$
 $\log_e(P(Z)) = -Z^2$

- Extreme Value:
 $P(Z) = \exp(-Z - \exp(-Z))$
 $\log_e(P(Z)) = -Z - \exp(-Z)$
where $Z = (\text{score} - \text{mean}) / \text{SD}$



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, $E(Z > Z_0)$, to indicate whether the score, S_0 , is significant.

Note, $Z_0 = (S_0 - \text{Mean})/SD$, 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.

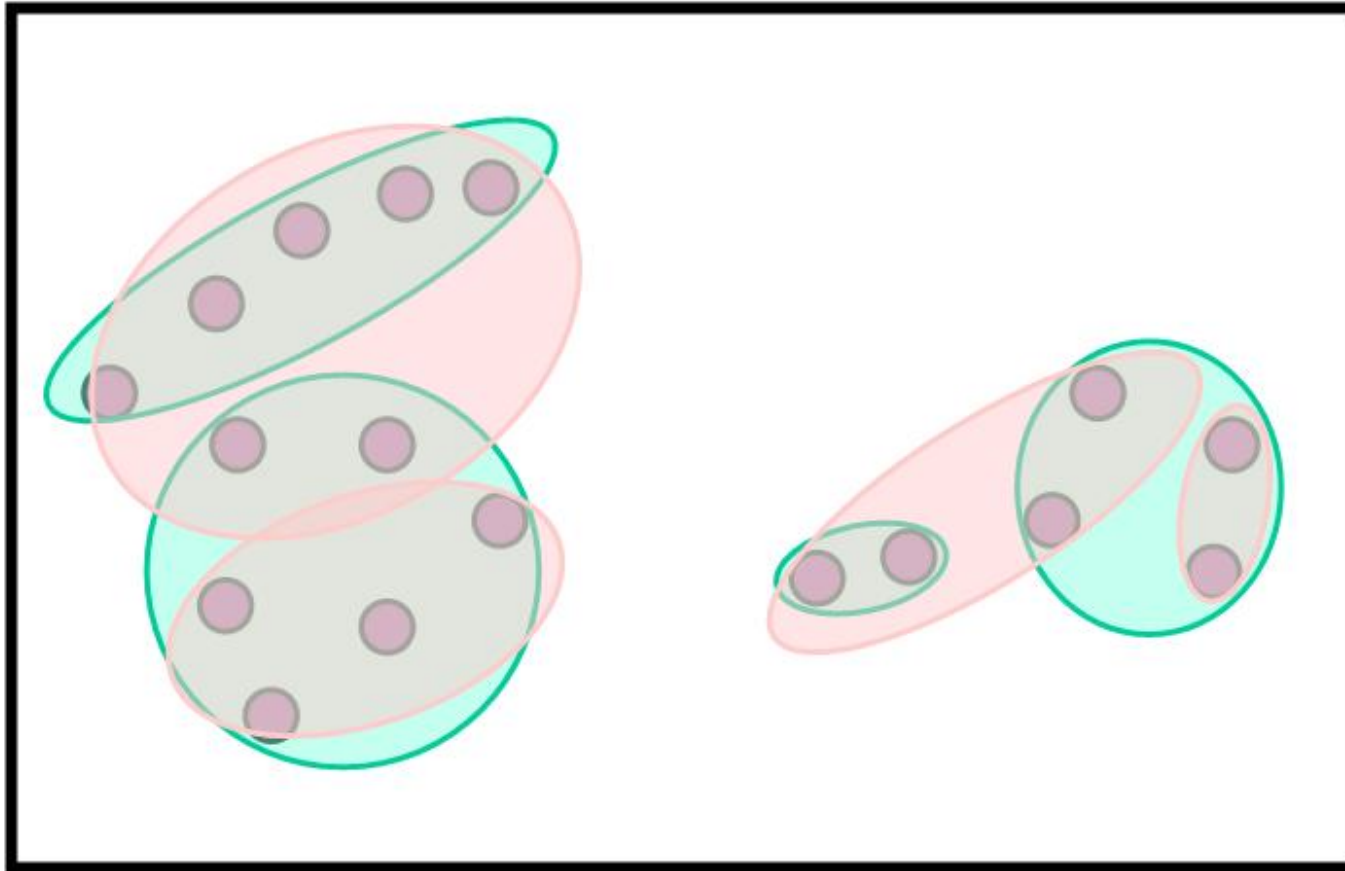
K-Means Clustering.

ROC Curves.

Views of Structure Space.

Multidimensional Scaling.

INTUITIVE CLUSTERING



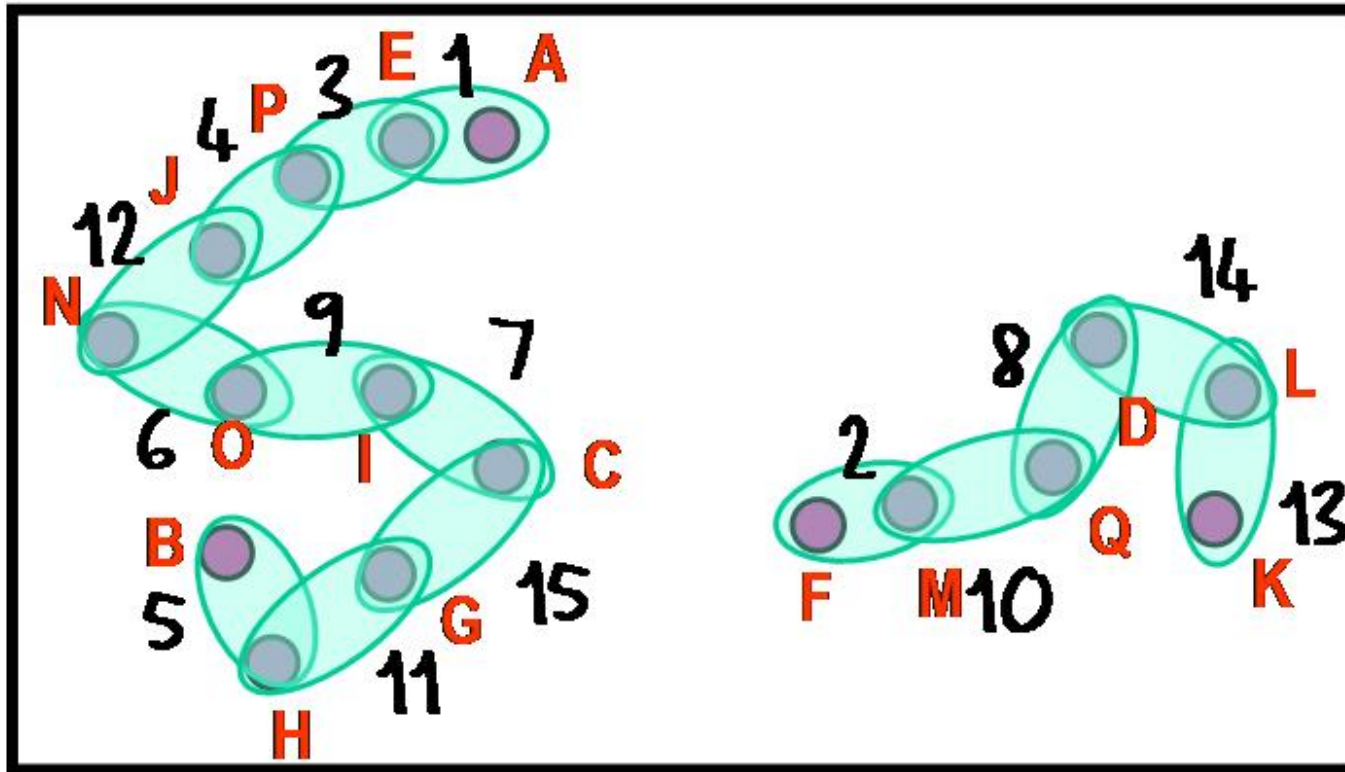
- Many possibilities.

- Not so easy.

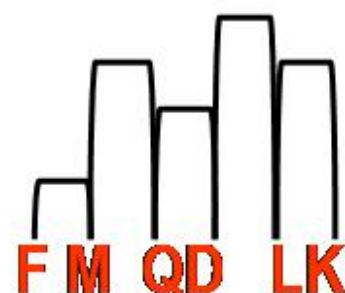
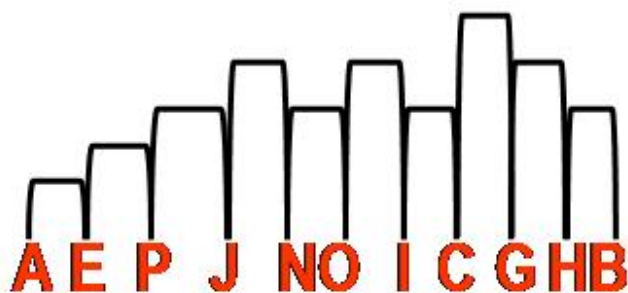
- What is best clustering?

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

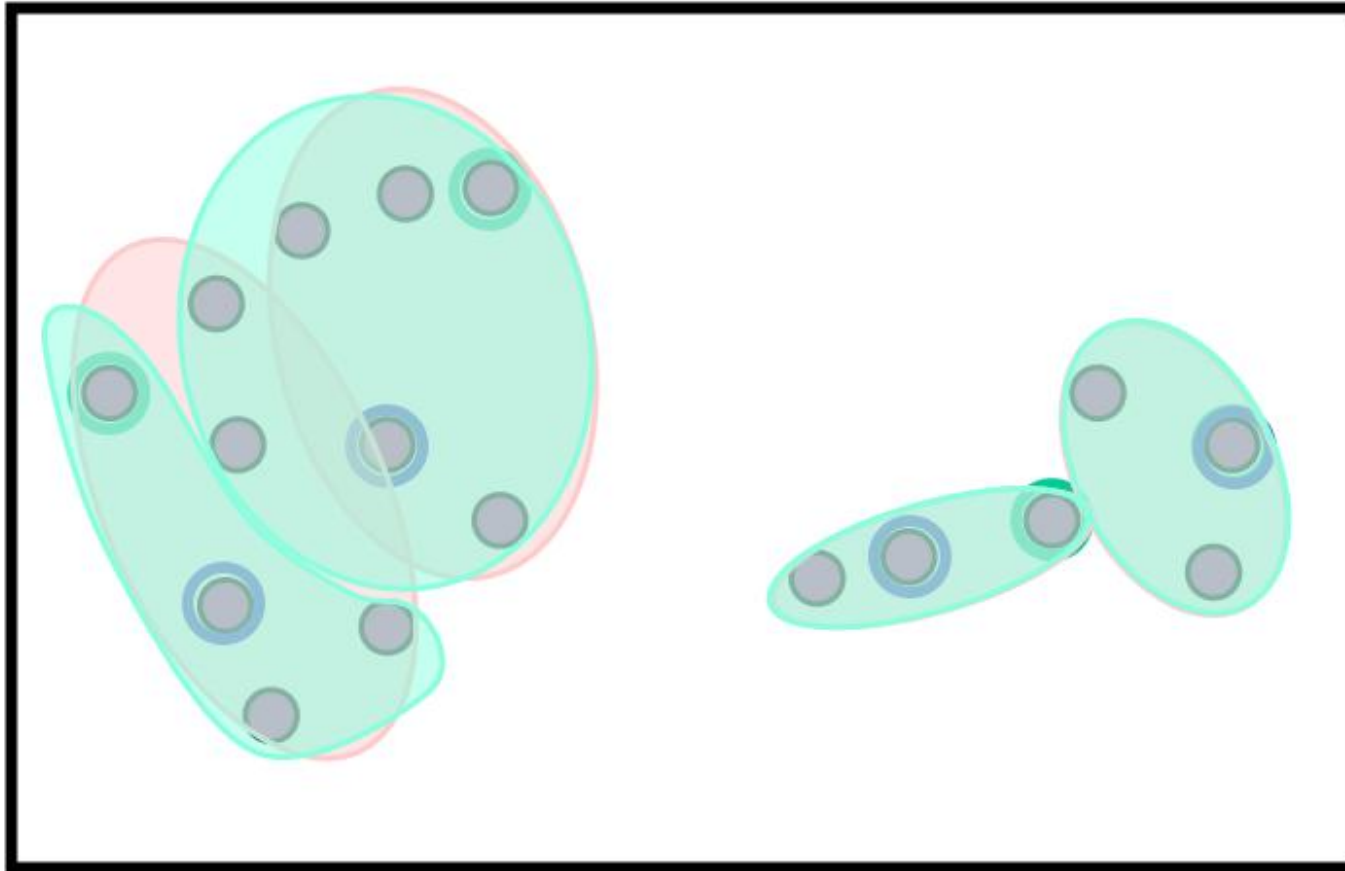
HIERARCHICAL CLUSTERING



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



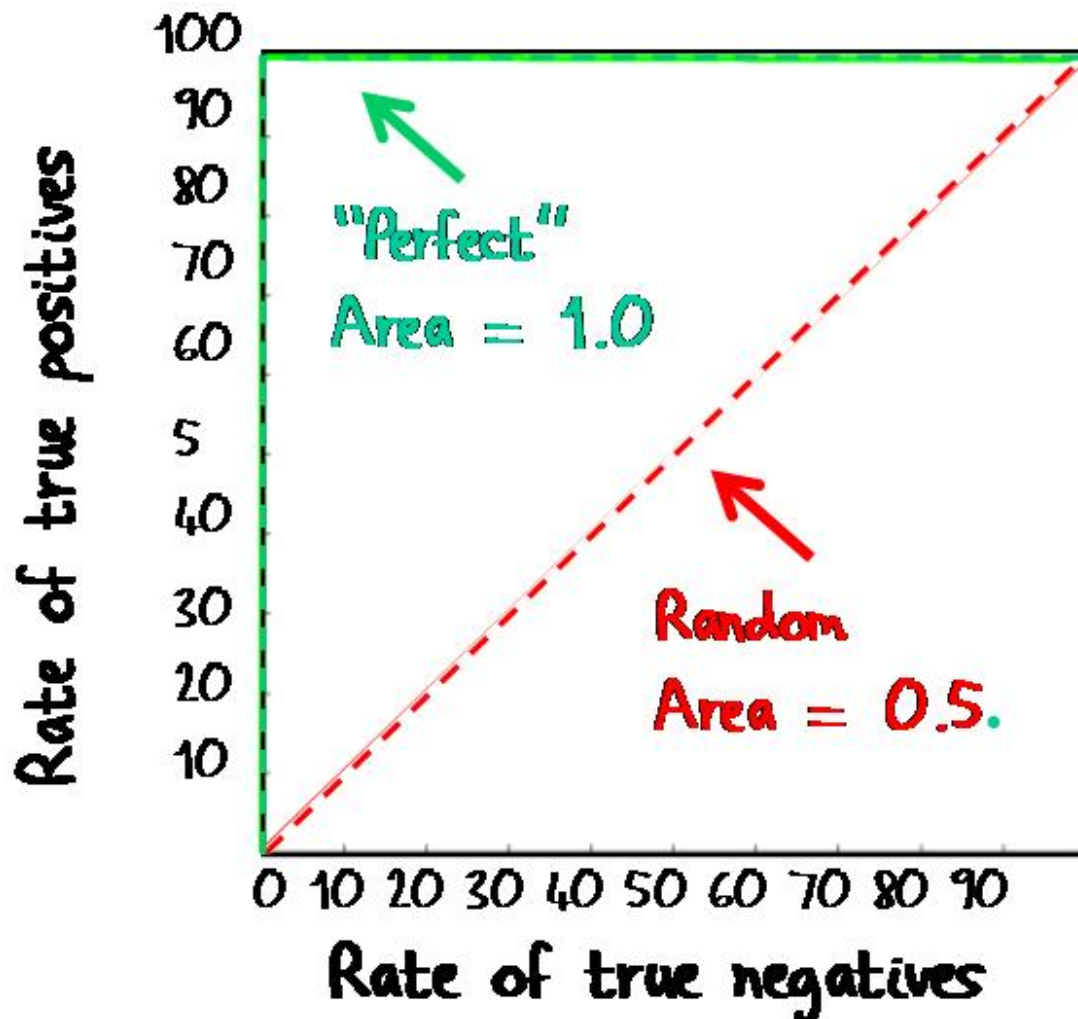
K-MEANS CLUSTERING



- 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.

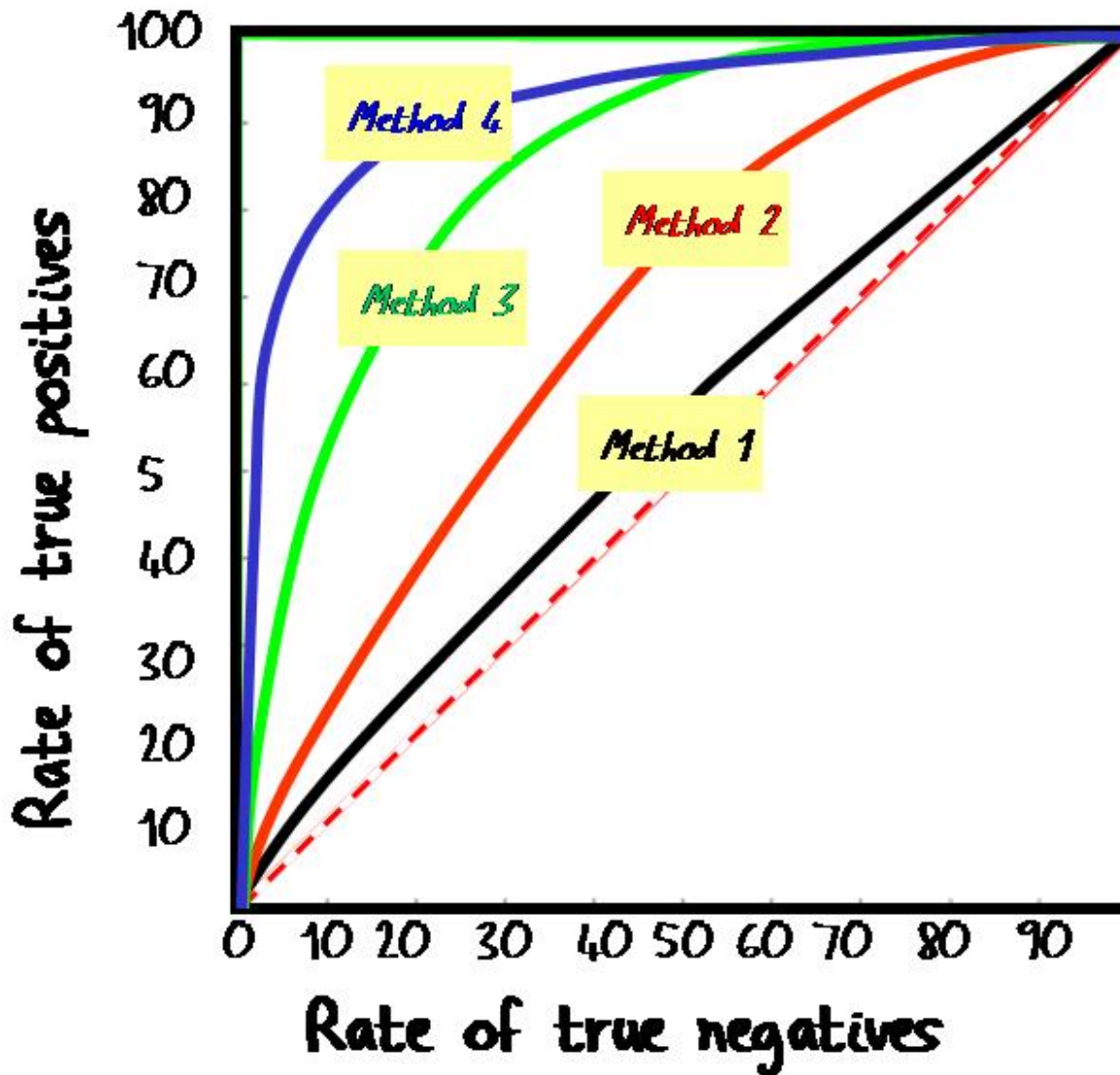
ROC ANALYSIS

(Receiver Operating



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

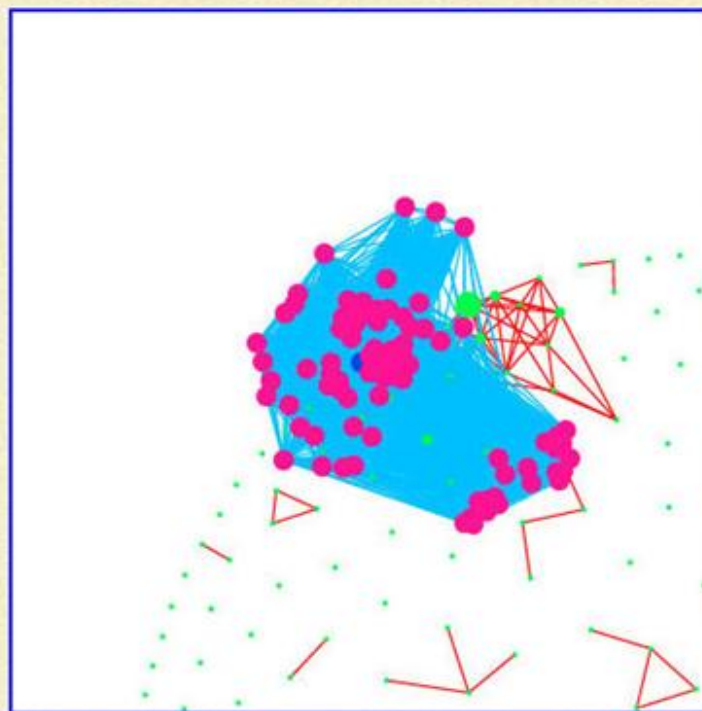
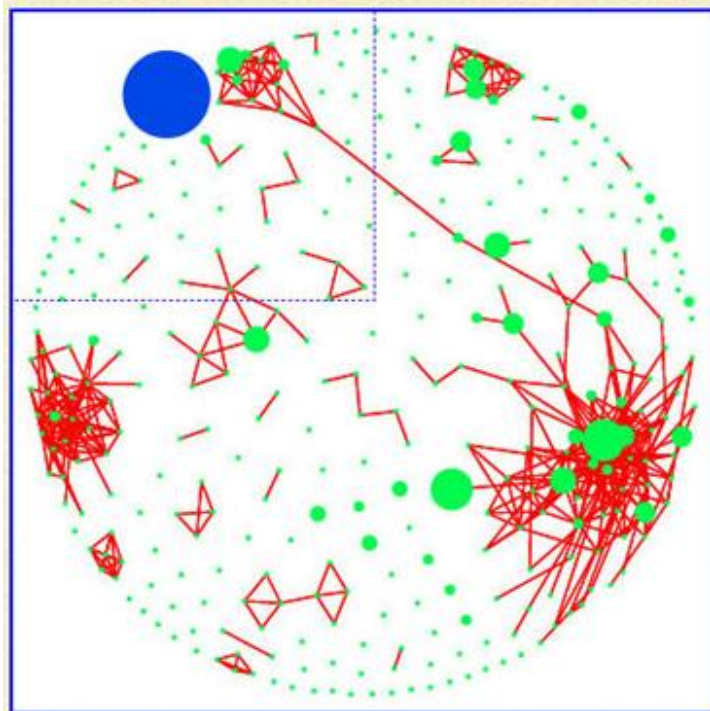
ROC CURVE EXAMPLES



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

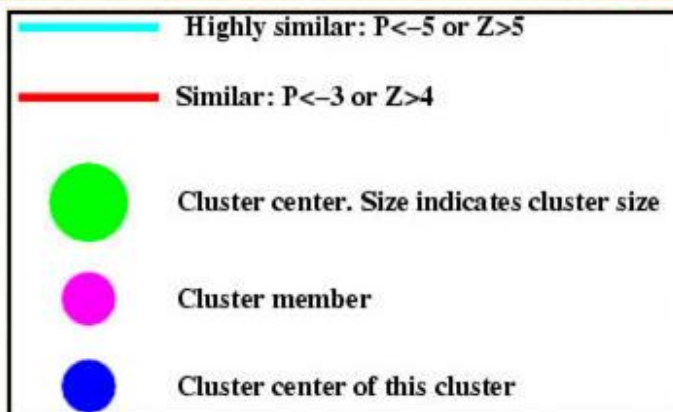
VIEWS OF STRUCTURE SPACE

(A:) Chimeric hemoglobin beta-alpha {Synthetic, based on }



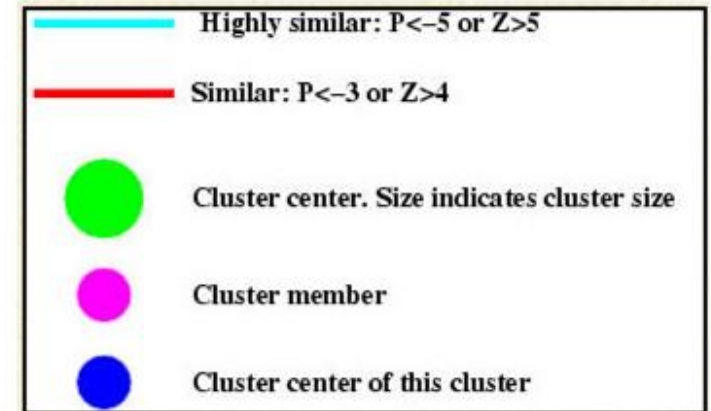
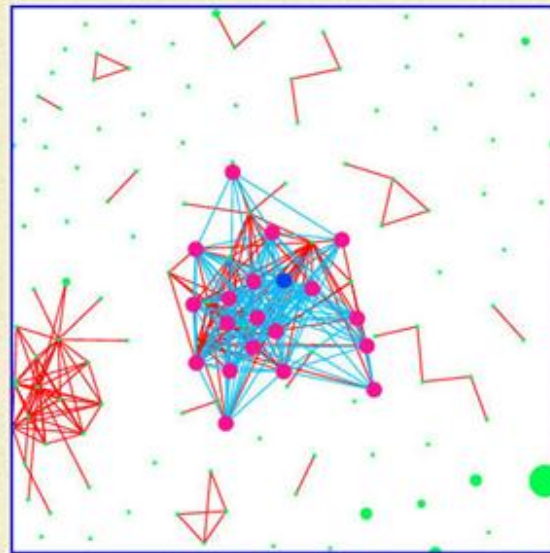
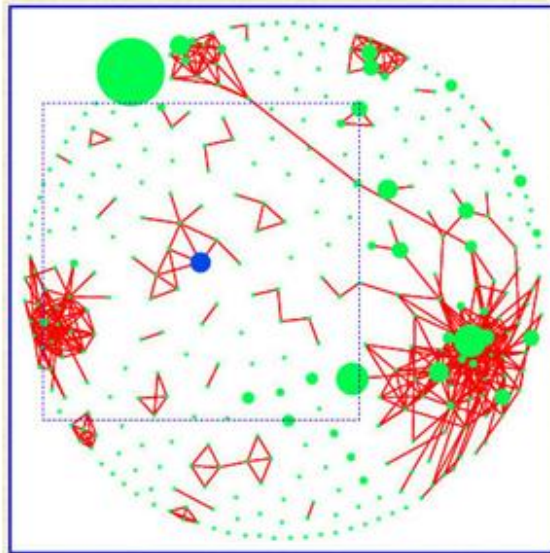
How can one see
5000 structure
space?

Erik

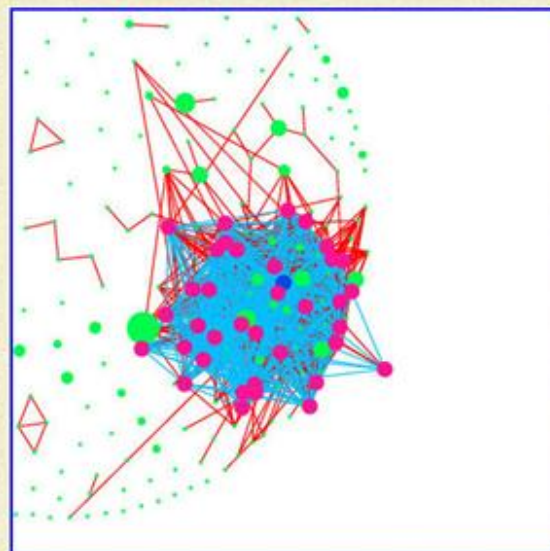
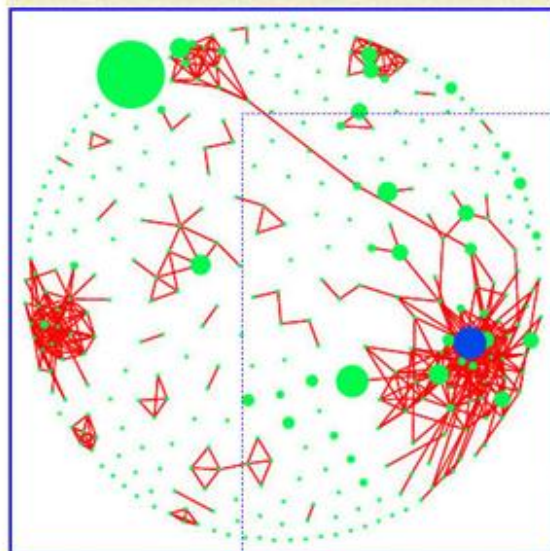


VIEWS OF STRUCTURE SPACE

(A:) Mating type protein A1 Homeodomain {Baker's yeast (Sa



(A:) Aspartate receptor, ligand-binding domain {Salmo



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.

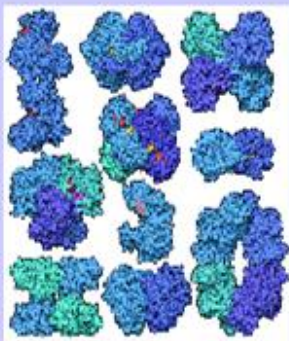
PROTEIN DATABASE RCSB

<http://www.rcsb.org/pd>

[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?

[ABOUT PDB](#) | [NEW FEATURES](#) | [USER GUIDES](#) | [FILE FORMATS](#) | [DATA UNIFORMITY](#) | [STRUCTURAL GENOMICS](#) | [SOFTWARE](#) | [PUBLICATIONS](#) | [EDUCATION](#)

Search the Archive

Enter a [PDB ID](#) or keyword

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-I 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

©Michael Levitt 04

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

RNABase.org
The RNA Structure Database

PDB or NDB ID Code

[RNABase Home](#) [Search RNABase](#) [Help/About](#) [Contact RNABase](#)

[Listing of RNABase Entries](#) [Search RNABase](#) [Analyze Your Structure](#) [RNABase Meta-Analysis](#) [Reference & Education](#) [About RNABase](#)

Listing of RNABase Entries

Complete Listing	A listing of all entries in RNABase with links to detailed records for each entry. All Entries
Technique Listing	A listing of all entries in RNABase by experimental technique. x-ray crystallography - NMR spectroscopy - all other methods
Category Listing	A listing of all entries in RNABase by structural or functional category. transfer RNAs - ribosomal RNAs - messenger RNAs - transcription-related RNAs - introns - splicing-related RNAs - signal recognition particle RNAs - ribozymes - RNase P - aptamers - pseudoknots - tetraloops - bulges - DNA-RNA hybrids - PNA-RNA hybrids - drug-RNA complexes - viral & phage RNAs
Outlier Rate Listing	A tabulation of error rate for each structure in RNABase organized by technique and category. All Entries

Search RNABase

Basic Search	Find the entry you are looking for by PDB or NDB code, author name, classification, experimental technique, resolution, or keywords.
Advanced Search	For those seeking more precise search capabilities.

183 Backward

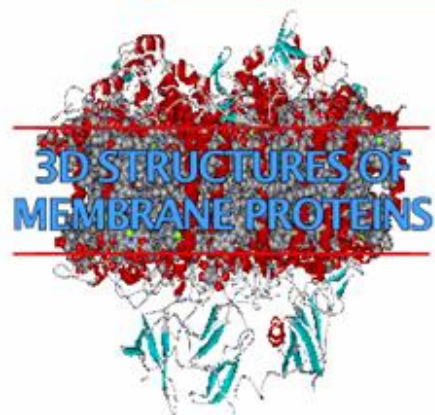
<http://www.rnabase.org>

©Michael Levitt 04

MEMBRANE PROTEIN DATA BASE

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

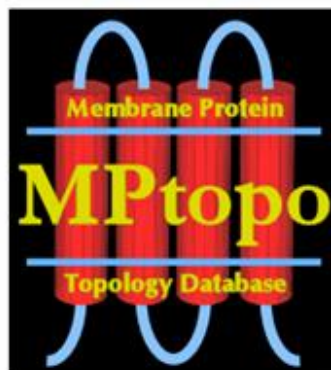
MEMBRANE PROTEIN RESOURCES



[Membrane Protein Structures]

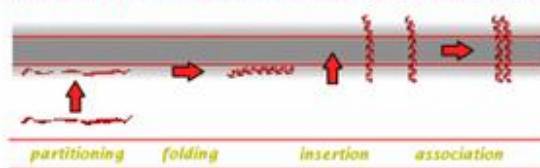
Additional Resources
[Energetics of Protein-Bilayer Interactions](#)
[Experiment-Based Hydrophobicity Scales](#)
[Structure of Fluid Lipid Bilayers](#)

[Transport Protein Database from UCSD](#)

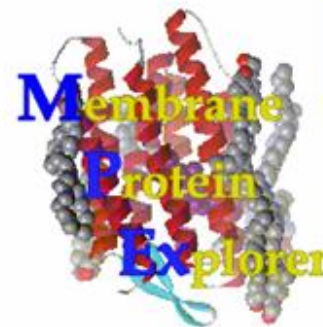


[Membrane Protein Topology Database]

Principles of Membrane Protein Folding and Stability



[Membrane Protein Folding and Stability]



[Membrane Protein Explorer]

Additional Resources
[MSB Bibliographic Database](#)
[Structural Biophysics Web Resources](#)
[Meetings of Interest to MP biophysicists](#)

5 Backward

from the [Stephen White Laboratory](#) at UC Irvine

Author: [Stephen White](#)
copyrighted (c) 2001-2003. All rights reserved.



Page last updated: 7 Nov 2003

©Michael Levitt 04

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

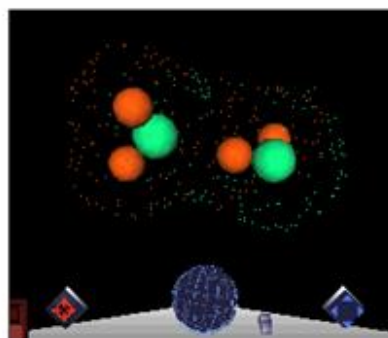
Back to

52 Backward

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

We welcome feedback
and comments at
ms23@nyu.edu

Library of 3-D Molecular Structures



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

[[About the Database](#)]

To enter the library, click on the appropriate buttons below.

Water and Ice

Carbons

Hydrocarbons

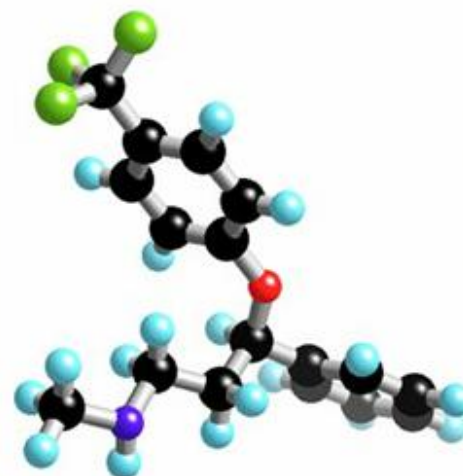
Molecules of Life

Drugs

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>

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

©Michael Levitt 04

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/chemistry>

©Michael Levitt 04

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 understand 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 describe 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**



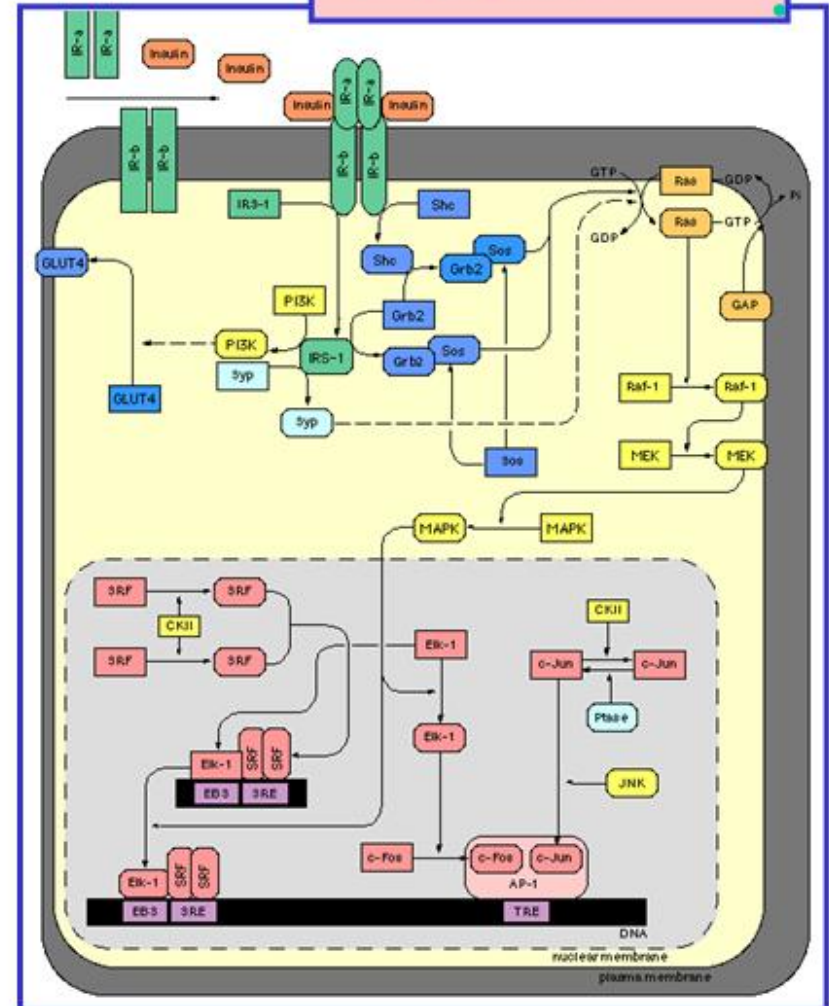
This WWW service "SPAD" is still under development.

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/pdf.html>

©Michael Levitt 04

PATHWAY DATABASES



KEGG - Table of Contents

KEGG2 PATHWAY GENES LIGAND EXPRESSION BRITE XML API DBGET

1. KEGG Databases

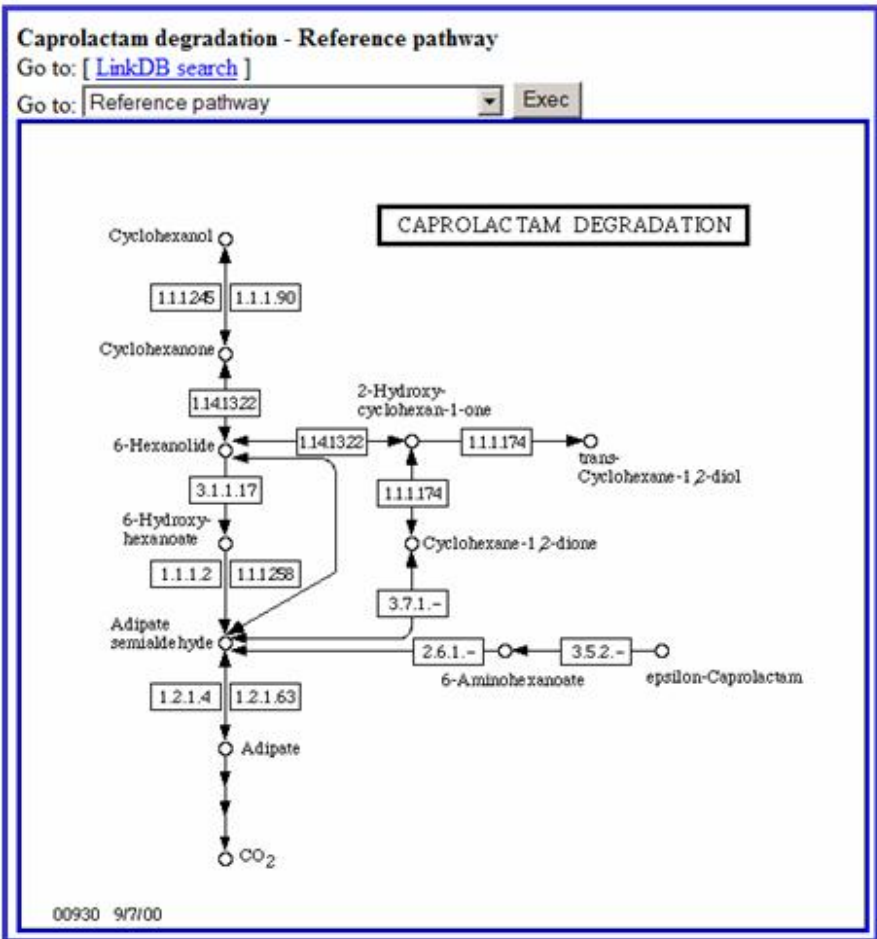
Category	Database	Search & Compute	DBGET Search
Pathway information	KEGG PATHWAY Database	XML Search objects in KEGG pathways Color objects in KEGG pathways Generate possible reaction paths	PATHWAY
Genomic information	KEGG GENES Database	KO Search similar GENES sequences Search similar GENOME sequences	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	OGENES
	Complete plastid genomes	
	Complete nucleomorph genomes	

238 Backward



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

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

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 Fuqu v2.0	03 Mar 2003
Mosquito	2a	MOZ 2	01 Oct 2003
Fruitfly	3a	BGDP 3.1	02 Jul 2003
<i>C. elegans</i>	102	WS 102	02 Jul 2003
<i>C. briggsae</i>	25	cb25_ago8	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 Help Desk

Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site) Documentation

Have you tried?

Ensembl Chimp Preview Browser

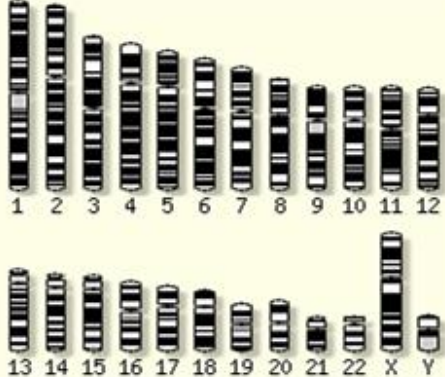
A preliminary Chimp assembly is now available at <http://pre.ensembl.org>



Click for more information

http://www.ensembl.org

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

©Michael Levitt 04

http://localhost/SB228_Lec_6_2004/Slide43.JPG [2/11/2004 9:01:59 PM]

ENSEMBL EBI SANGER GENOME VIEWER

Ensembl Human ContigView

The Wellcome Trust Sanger Institute EBI

Home Human What's New TextSearch BlastSearch MartSearch Export Data Download Disease Browser Docs

Find Sequence Y Lookup [e.g. AC067852, AP003171] Help

Chromosome Y

Chr Y p11.2

Overview

Chromosome band

DNA(contigs)

Markers

Ensembl Genes

Gene legend

- ENSEMBL PREDICTED GENES (KNOWN)
- ENSEMBL PREDICTED GENES (NOVEL)
- ENSEMBL PSEUDOGENES

• Huge amount of

• Amazing

Detailed View

Jump to Chromosome: Y bp 8877590 to 8977589 Refresh

<< 2 Mb < 1 Mb Zoom

Features DAS Sources Repeats Decorations Export Jump to Image

Length

Human proteins

Proteins

Genscans

Basepair View

Zoom Window

Length Genscans

EST trans.

Ensembl trans.

Amino acids

Sequence

DNA(contigs)

ENSESTT0000040504

ENSESTT0000040500

ENSESTT0000040499

ENSESTT0000040498

ENSESTT0000040496

ENST0000036706

Ensembl novel trans

ENST0000022443

Ensembl novel trans

TSPY

Ensembl known trans

E E L P L P C P I A R P P Q S C R N L N P H S Q Q Y T S K S L R Q

A G A G G T C C T T T G C C A T G T G C C C C T R G C T C C C C G C C T C A C C A T C G T G C C C T H A C C T G G C C C T C R C H G T C G A C A C C A C T G A A C A G G C T C R G G

©Michael Levitt 04

NCBI GENOME VIEWER

NCBI | PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

Search: Find Find in This View Advanced Search

Homo sapiens Map View | [BLAST The Human Genome](#)
Build 34 Version 2

Chromosome: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#)
[16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) X [Y]

Master Map: [Genes On](#) | **Maps & Options**
 Total Genes On Chromosome: 247
 Region Displayed: 8,877K-8,977K bp
[Download/View Sequence/Evidence](#)

Genes Labeled: 5 Total Genes in Region: 5

Ideogram	Contig	HsUni6	Gene_seq	Symbol	LinkOut	E	Cyto	Description
				LOC392580	sv pr dl ev mm	P	Y	LOC392580
				TSPY	OMIM sv pr dl ev mm hm	C		Yp11.2 testis specific protein, Y-linked
				LOC392581	sv pr dl ev mm	P	Y	LOC392581

Jump from

Same position on

Region Shown: 8877591 - 8977590

out. zoom in

Yp11.3
Yp11.2
Yq11.1
Yq11.2
Yq12

default master

©Michael Levitt 04

NCBI GENOME VIEWER

Start with Chromosome

It has 1359

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

Symbol	LinkOut	E	Cyto	Description
C17orf31	sv pr dl ev mm hm		C 17p13.3	chromosome 17 open reading frame 31
STX8	OMIM sv pr dl ev mm hm		C 17p12	syntaxin 8
DNAH9	OMIM sv pr dl ev mm hm		C 17p12	dynein, axonemal, heavy polypeptide 9
SSH2	OMIM sv pr dl ev mm hm		C 17q11.2	slingshot 2
NF1	OMIM sv pr dl ev mm hm		C 17q11.2	neurofibromin 1 (neurofibromatosis, v
MYO1D	OMIM sv pr dl ev mm hm		C 17q11-q12	myosin ID
ACCN1	OMIM sv pr dl ev mm hm		C 17q11.2-q12	amiloride-sensitive cation channel 1, ne
ACACA	OMIM sv pr dl ev mm hm		C 17q21	acetyl-Coenzyme A carboxylase alpha
SCAP1	OMIM sv pr dl ev mm hm		C 17q21.32	src family associated phosphoprotein 1
CA10	OMIM sv pr dl ev mm hm		C 17q21	carbonic anhydrase X
FLJ38335	sv pr dl ev mm hm		C 17q23.1-q23.2	hypothetical protein FLJ38335
MSI2	OMIM sv pr dl ev mm hm		C 17q23.2	musashi homolog 2 (Drosophila)
PPM1E	sv pr dl ev mm hm		C 17q23.2	protein phosphatase 1E (PP2C domain
BCAS3	OMIM sv pr dl ev mm hm		C 17q23	breast carcinoma amplified sequence 3
DKFZP564D166	sv pr dl ev mm hm		C 17q24.1	putative ankyrin-repeat containing prote
MGC33887	sv pr dl ev mm hm		C 17q24.2	hypothetical protein MGC33887
PRKCA	OMIM sv pr dl ev mm hm		C 17q22-q23.2	protein kinase C, alpha
PITPNC1	OMIM sv pr dl ev mm hm		C 17q24.3	phosphatidylinositol transfer protein, cy
SLC39A11	sv pr dl ev mm hm		C 17q25.1	solute carrier family 39 (metal ion trans
raptor	OMIM sv pr dl ev mm hm		C 17q25.3	raptor

NCBI IS THE PLACE FOR SEQUENCE

The screenshot shows the NCBI homepage with the following elements:

- Header:** NCBI logo, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health.
- Navigation:** PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, Structure.
- Search:** Search Entrez for [] Go
- Left Sidebar:** Site Map, Guide to NCBI resources, About NCBI, GenBank, Literature databases, Molecular databases.
- Main Content:**
 - What does NCBI do?** Established in 1988 as a national resource for molecular biology information...
 - Hot Spots:** Clusters of orthologous groups, Electronic PCR, E-Utilities, Gene expression omnibus, Genes and disease, Human genome resources, Human-mouse homology maps, LocusLink, Malaria genetics & genomics, Map Viewer, MHC, Mouse genome resources, NCBI Handbook, ORF finder, Rat genome.
 - PubMed Central:** An archive of life sciences journals. Free fulltext, Over 100,000 articles from over 130 journals, Linked to PubMed and fully searchable.
 - dbMHC:** A new NCBI resource that provides a platform for genetic and clinical data related to the human Major Histocompatibility Complex (MHC).
 - NCBI Newsletter:** The Reference Human Genome at NCBI, The Human Genome Project has produced the first reference sequence for the human genome.

• PubMed Central (book

• Entrez

• Map

• 17,300 Backward

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

©Michael Levitt 04

EBI IS THE PLACE FOR TOOLS

The screenshot shows the EMBL-EBI website interface. At the top, there is a search bar with the text "Get Nucleotide sequences for" and a "Go" button. Below the search bar is the EMBL-EBI logo and the text "European Bioinformatics Institute". A navigation menu includes "EBI Home", "About EBI", "Research", "Services", "Toolbox", "Databases", "Downloads", and "Submissions". A "VIEW ALL SERVICES" link is also present. On the left, a tree diagram shows the site's structure: Databases, Toolbox, Submissions, Downloads, and Services Help. A "Services Overview" box is also visible. The main content area is divided into several sections: "Databases" (Database Browsing & Entry Retrieval via...), "Toolbox" (Homology & Similarity, Prot. Function. Analysis, Sequence Analysis, Structural Analysis, Tools Miscellaneous), "Submissions" (AEdb, ArrayExpress via MIAMExpress, EMBL via WEBIN, MGTHLA, PDB-AutoDep, UniProt via SPIN, Webin-Align), "Downloads" (EBI FTP Server, Help Files, Database Repository, Software Repository), and "Databases" (Literature Databases, Microarray Databases, Nucleotide Databases, Protein Databases, Structure Databases). A "WHATS 2can?" box is also present.

FAST

Clustal

Expression

DAL

7,200 Backward

<http://www.ebi.ac.uk/services/index.htm>

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

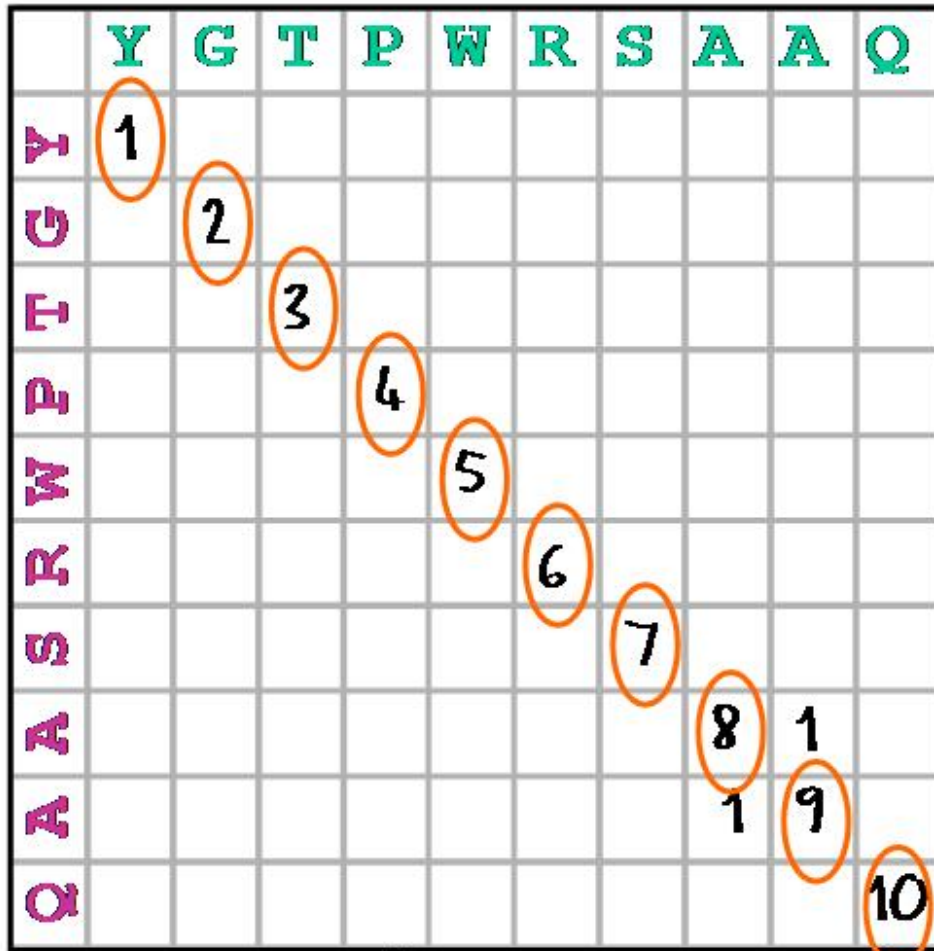
	Y	G	T	P	W	R	S	A	A	Q
Y	1									
G		1								
T			1							
P				1						
W					1					
R						1				
S							1			
A								1	1	
A								1	1	
Q										1

- 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	

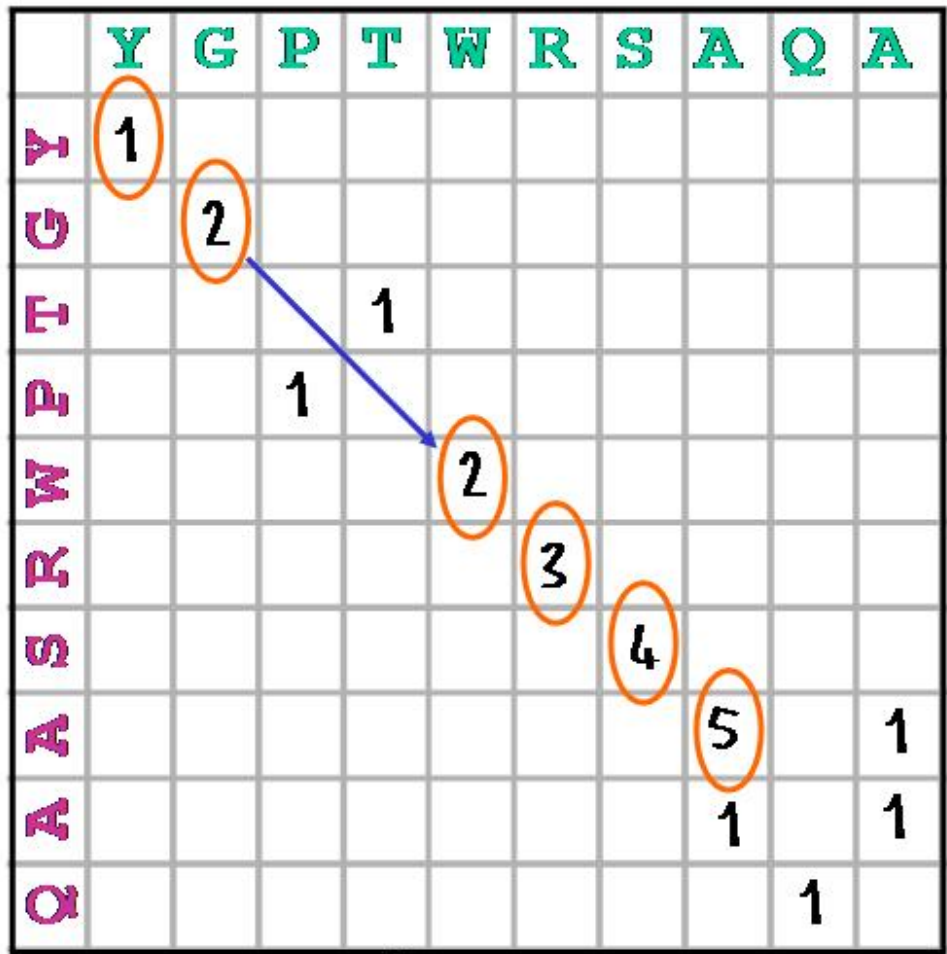
Similarity Matrix

- 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



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

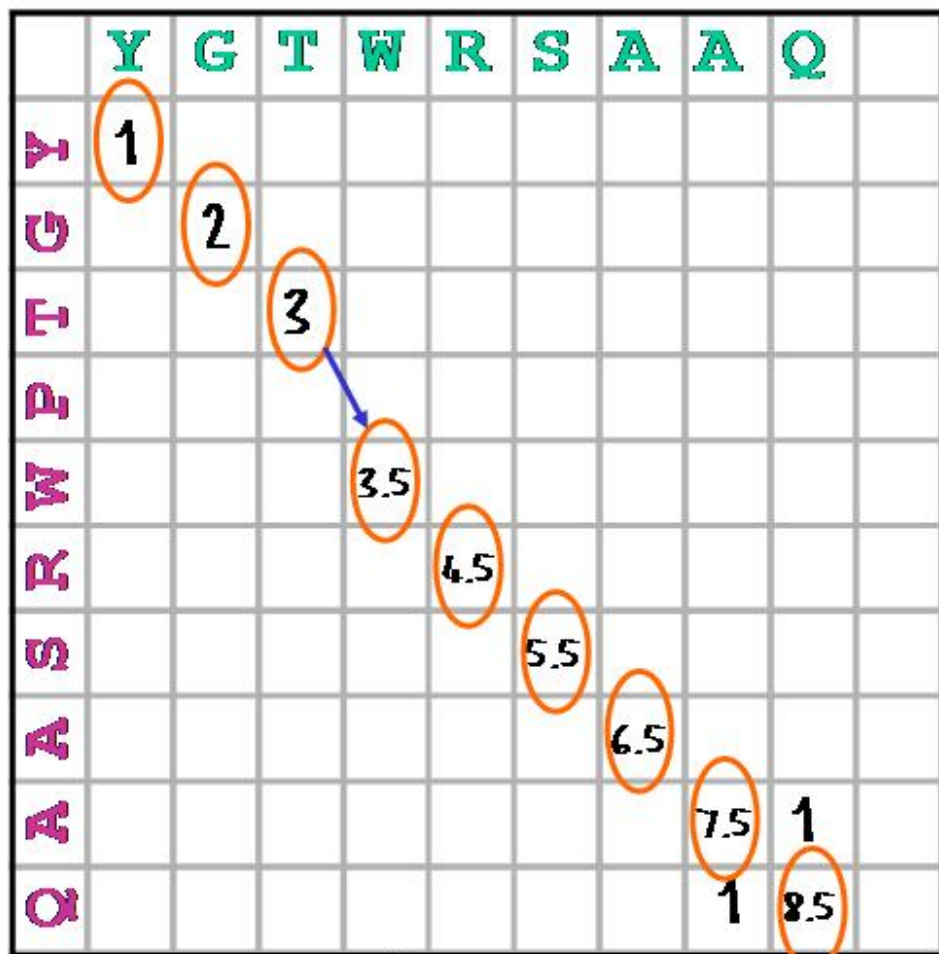
Similarity Matrix

- 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

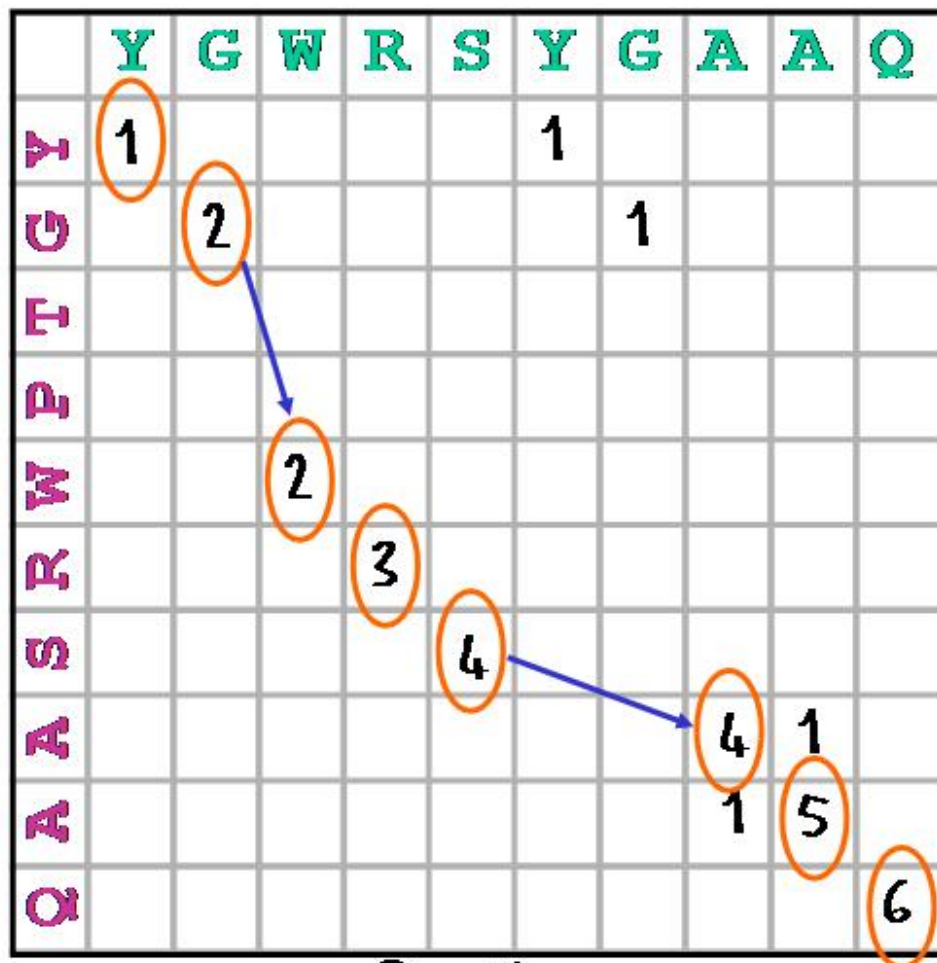
Similarity Matrix

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



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 (Hennikoff)

SCORING MATRIX

		Amino Acid j																									
		A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z			
Amino Acid i	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	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	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 Blosu62 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-1] + gap-J-to-ij
 where gap-J-to-ij = max{i-1, 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.

Sum scores along diagonals with identities.

	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

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

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

Similarity Matrix

Sum Matrix

BLAST AND PSI-BLAST

- Look for triplets that have high match scores.
For example with Blosun62: **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

The screenshot shows the NCBI BLAST website. At the top left is the NCBI logo. To the right, the word "BLAST" is written in large blue letters. Below this is a navigation bar with tabs for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area is divided into sections: Nucleotide, Protein, Translated, Genomes, Special, and Meta. Each section contains a list of search options. On the left side, there is a vertical menu with categories: Info (FAQs, News, References, Credits), Education (Program selection guide, Tutorial, URL API guide), Download (Executables, Databases, Source code), and Support (Helpdesk, Mailing list).

PubMed	Entrez	BLAST	OMIM	Taxonomy	Structure
	Nucleotide				
	<ul style="list-style-type: none">Discontiguous megablastMegablastNucleotide-nucleotide BLAST (blastn)Search for short, nearly exact matchesSearch trace archives with megablast or discontiguous megablast				
	Translated				
	<ul style="list-style-type: none">Translated query vs. protein database (blastx)Protein query vs. translated database (tblastn)Translated query vs. translated database (tblastx)				
	Special				
	<ul style="list-style-type: none">Align two sequences (bl2seq)Screen for vector contamination (VecScreen)Immunoglobulin BLAST (IgBlast)				
		Protein			
		<ul style="list-style-type: none">Protein-protein BLAST (blastp)PHI- and PSI-BLASTSearch for short, nearly exact matchesSearch the conserved domain database (rpsblast)Search by domain architecture (cdart)			
		Genomes			
		<ul style="list-style-type: none">Human, mouse, ratFugu rubripes, zebrafishInsects, nematodes, plants, fungi, malariaMicrobial genomes, other eukaryotic genomes			
		Meta			
		<ul style="list-style-type: none">Retrieve results by RIDGet this page with javascript-free links			

Blast is a
amazing
resource.

Play with it.

This is the
only way to
learn.

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

©Michael Levitt 04

THE END
of Lecture 6