CS276B

Text Information Retrieval, Mining, and Exploitation

Lecture 15
Bioinformatics I
March 6, 2003

(includes slides borrowed from R. Altman, J. Chang, L. Hirschman)

Bioinformatics Topics

- Today
  - Basic biology
  - Why text about biology is special
  - Text mining case studies
    - Microarray analysis
    - Abbreviation finding
    - Text-enhanced homology search
- Next week
  - Text mining in biological databases
  - KDD cup: Information extraction for bio-journals
  - Combining text mining and data mining

Basic Biology

Just Enough Molecular Biology

- *Entropy* (the tendency to disorder) always increases (cf. thermodynamics)
- Living organisms have low entropy compared with things like soil.
- They are relatively orderly...
- The most critical task is to maintain the distinction between inside and outside.

Just Enough Molecular Biology

- In order to maintain low entropy, living organisms must expend energy to keep things orderly.
- The functions of life, therefore, are meant to facilitate the acquisition and orderly expenditure of energy.

Just enough.

- The compartments with low entropy are separated from “the world.” Cells are the smallest unit of such compartments.
- Bacteria are single-cell organisms.
- Humans are multi-cell organisms.
- Low entropy compartments were difficult to get started *de novo*, and so have found ways to pass on the apparatus necessary to perpetuate themselves.

“Entropy-Fighting Apparatus:” Tasks

- Gather energy from environment
- Use energy to maintain inside/outside distinction
- Use extra energy to reproduce
- Develop strategies for being successful/efficient at the above tasks
  - develop ways to move around
  - develop signal transduction capabilities (e.g. vision)
  - develop methods for efficient energy capture (e.g. digestion)
  - develop ways to reproduce effectively

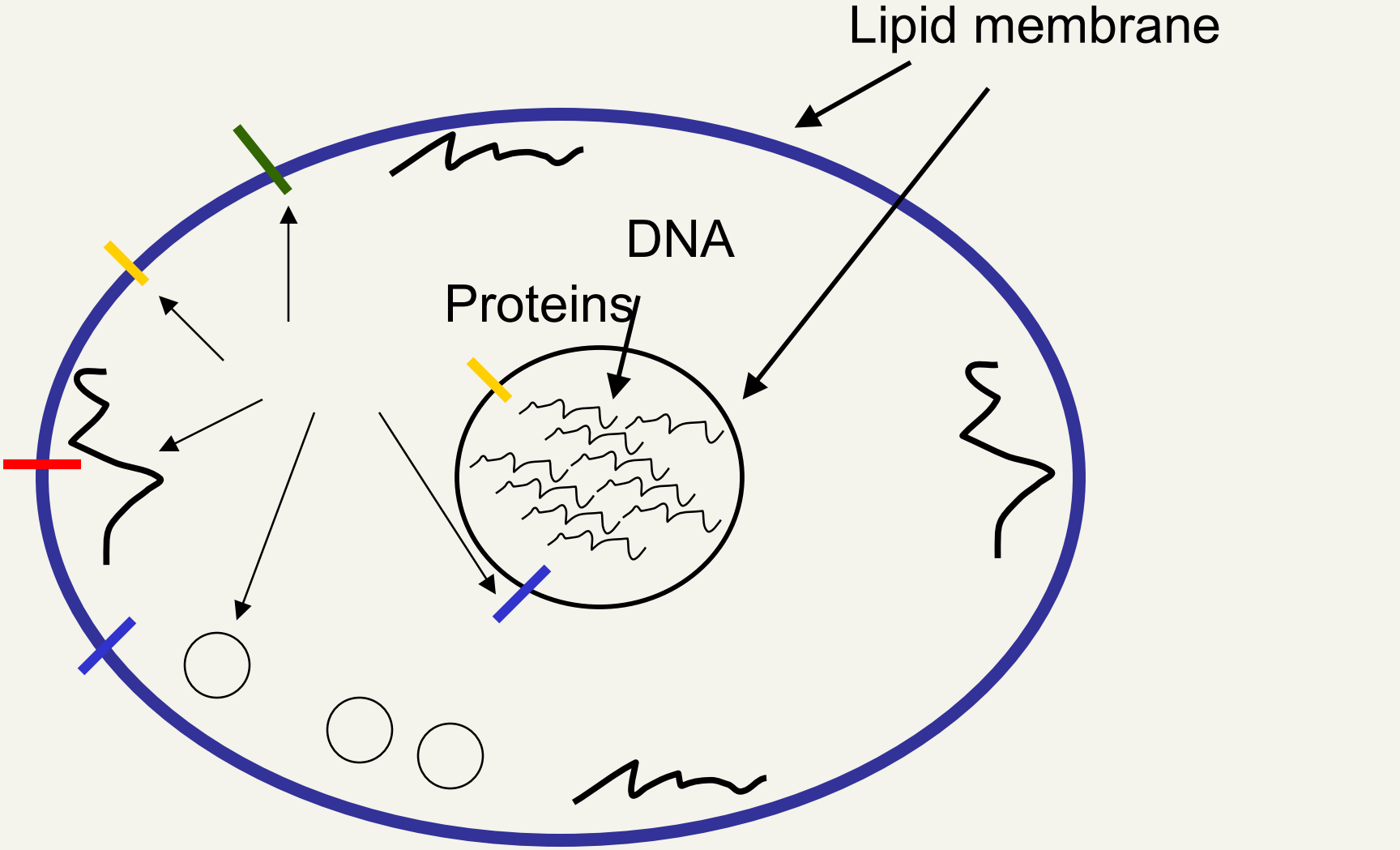
Just enough.

- In order to accomplish these tasks, living compartments on earth have developed three basic technologies
- 0. Ability to separate inside from outside (lipids)
- 1. Ability to build three-dimensional molecules that assist in the critical functions of life (proteins).
- 2. Ability to compress the information about how (and when) to build these molecules in a linear code (DNA).

Broad Generalization

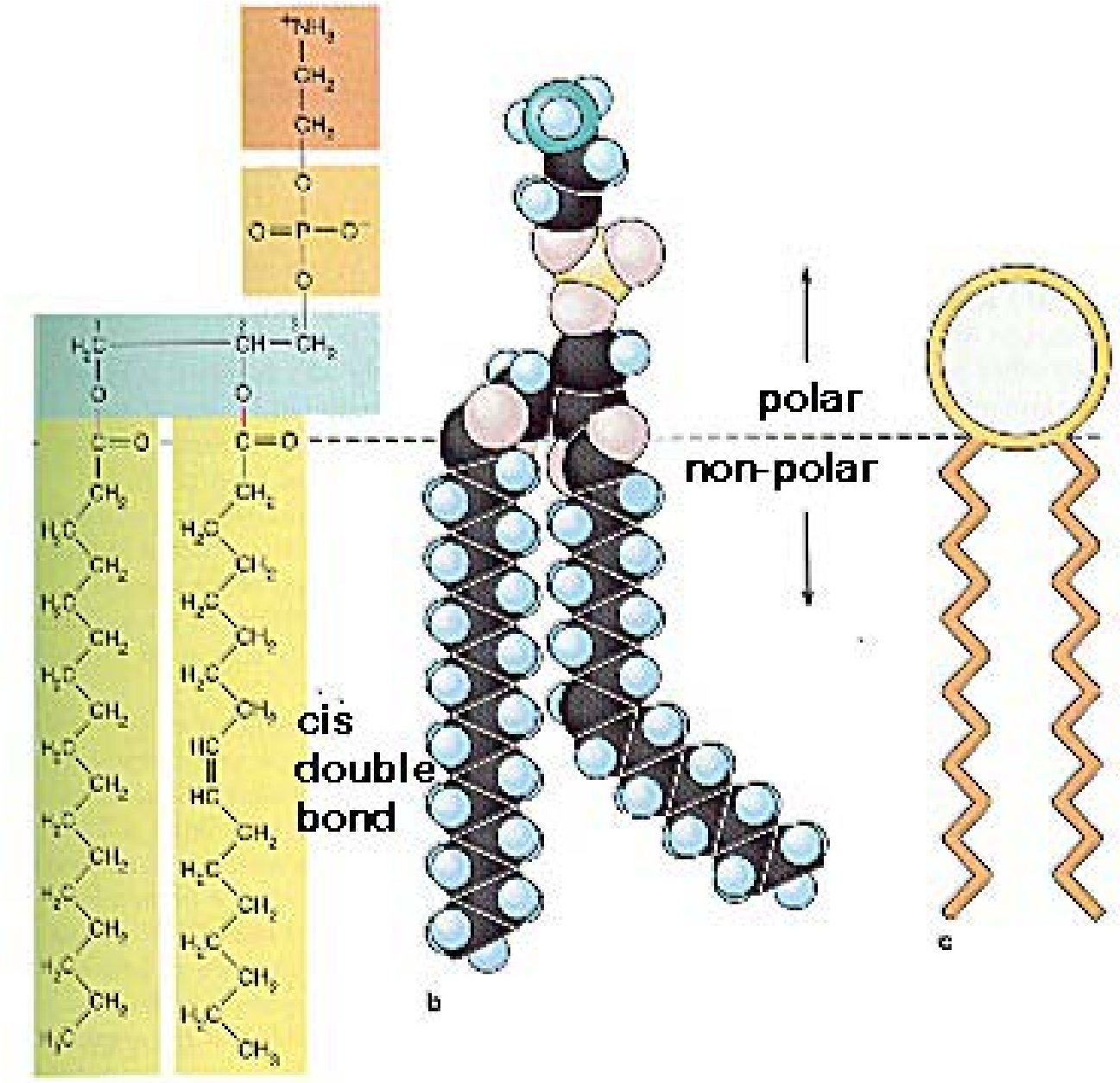
1. Lipid molecules: create compartments that separate inside/outside.
2. Protein molecules: do the work, and their 3D structure is critical.
3. DNA molecules: store the information

Bioinformatics Schematic of a Cell

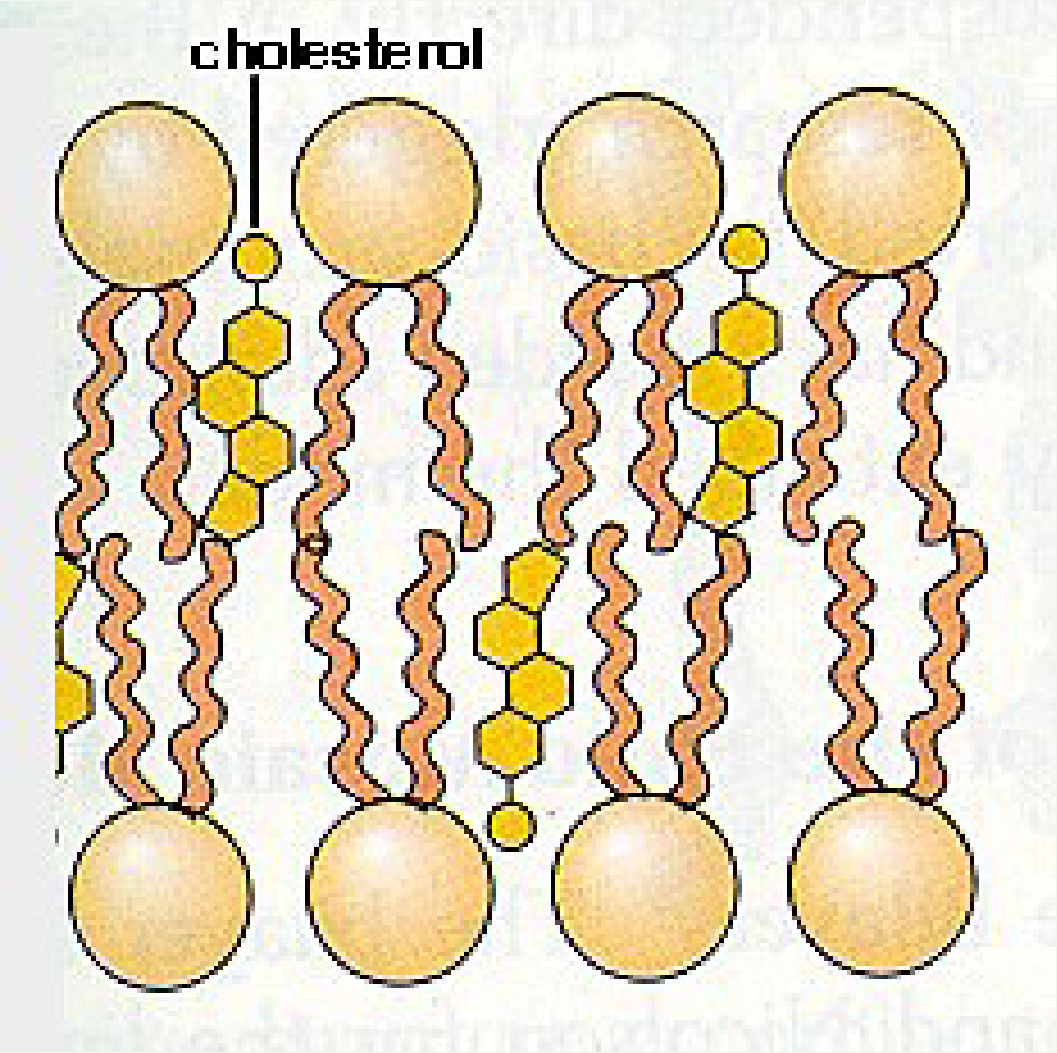


Lipids

- Hydrophilic (water loving) molecular fragment connected to hydrophobic fragment.
- Spontaneously form sheets (lipid bilayers, membranes) with hydrophilic ends on the outside, and hydrophobic ends on the inside.
- Create a very stable separation, not easy to pass through except for water and a few other small atoms/molecules.

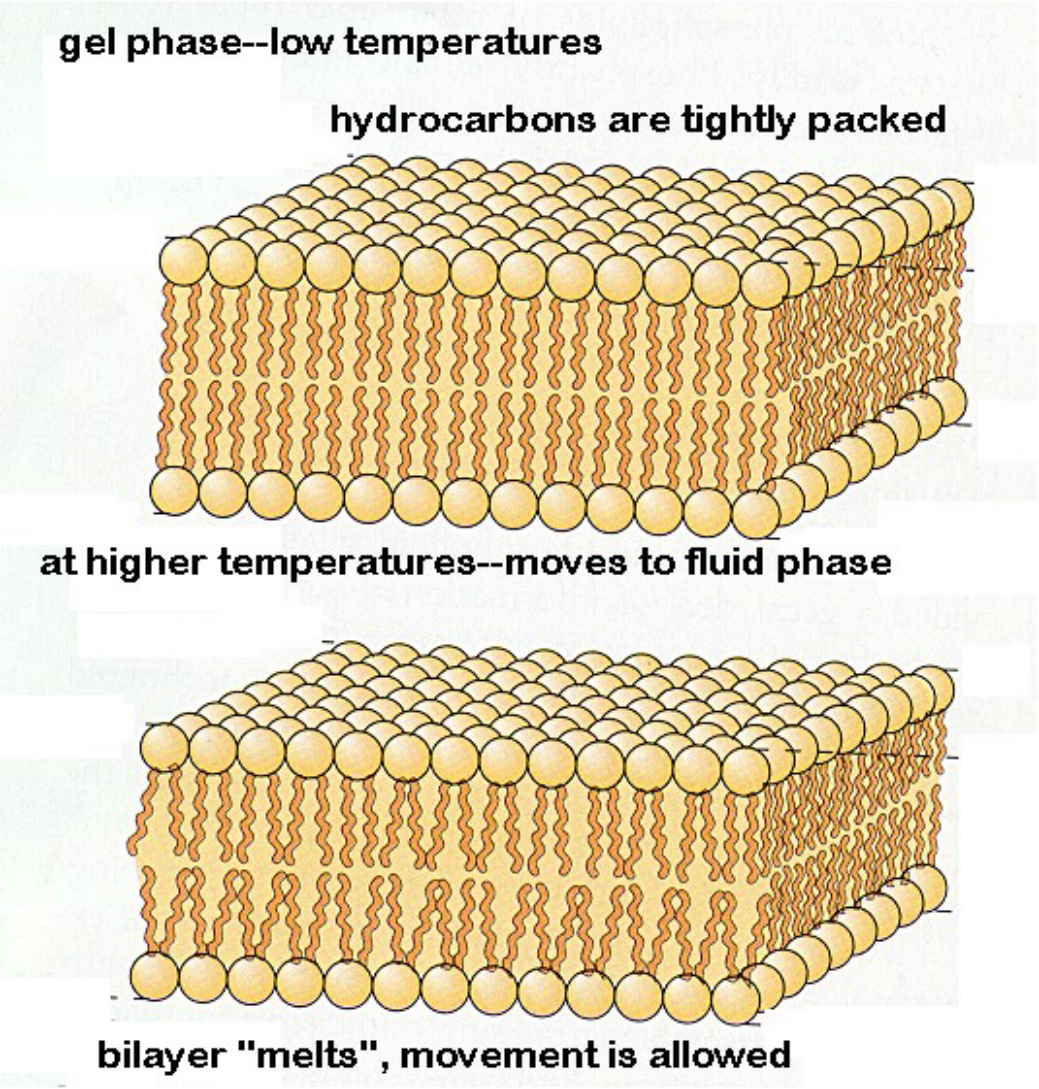

polar
non-polar
cis
double
bond
b
c

Lipid bilayer (hydrophobic in, hydrophilic out)

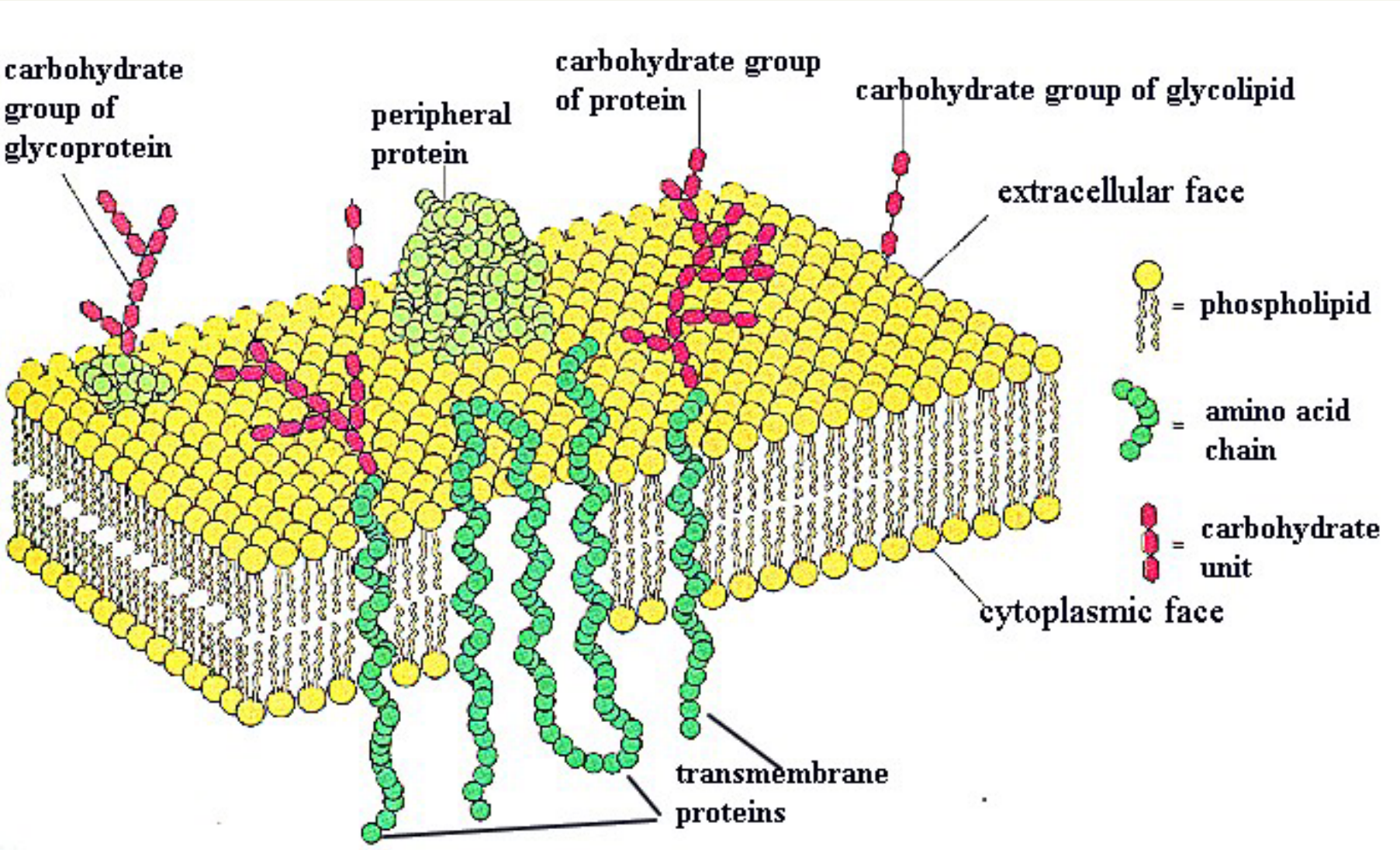


Basics of Lipid structure

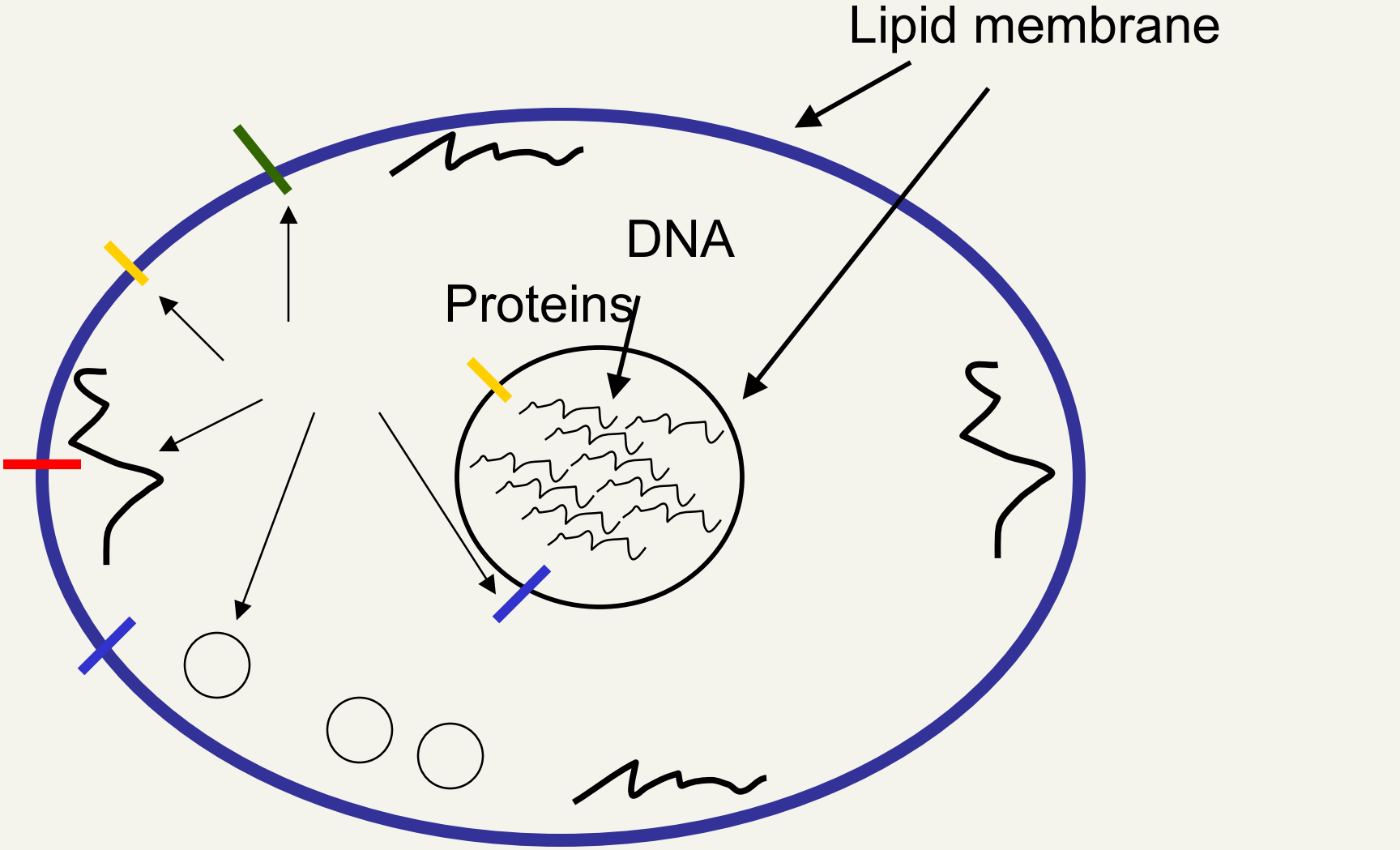
- Main goal: separate aqueous compartments effectively.



From http://cellbio.utmb.edu/cellbio/membrane_intro.htm


carbohydrate group of glycoprotein
peripheral protein
carbohydrate group of protein
carbohydrate group of glycolipid
extracellular face
= phospholipid
= amino acid chain
= carbohydrate unit
cytoplasmic face
transmembrane proteins

Bioinformatics Schematic of a Cell

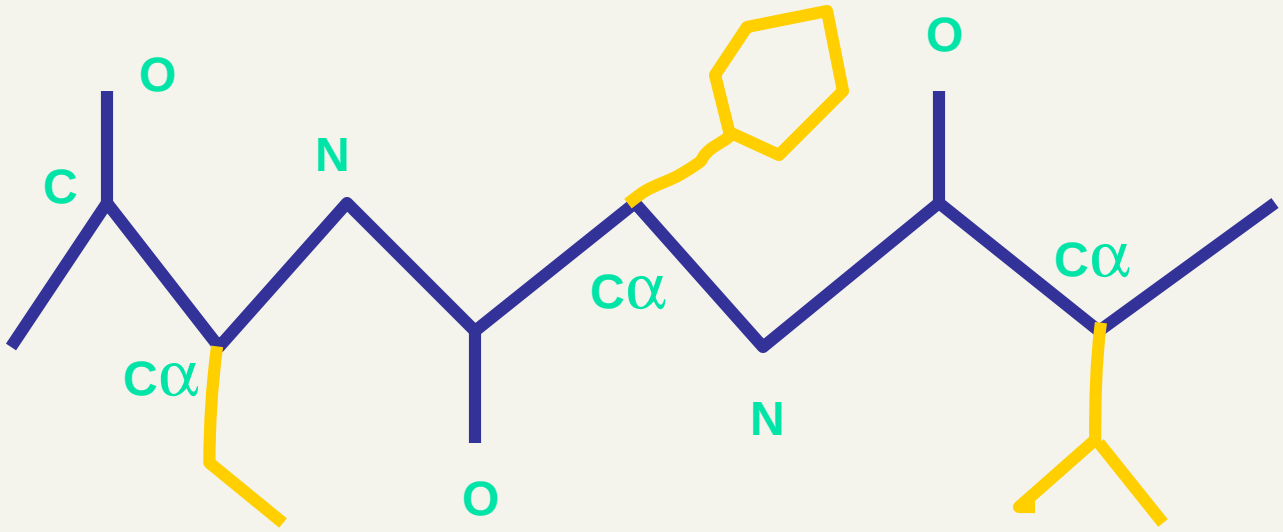


Protein molecules begin as a sequence of linked subunits

- These subunits are amino acids (also called residues).
- There are 20 different amino acids with different physical and chemical properties.
- The interaction of these properties allows a chain of the amino acids (up to 1000's long) to fold into a unique, reproducible 3D shape.

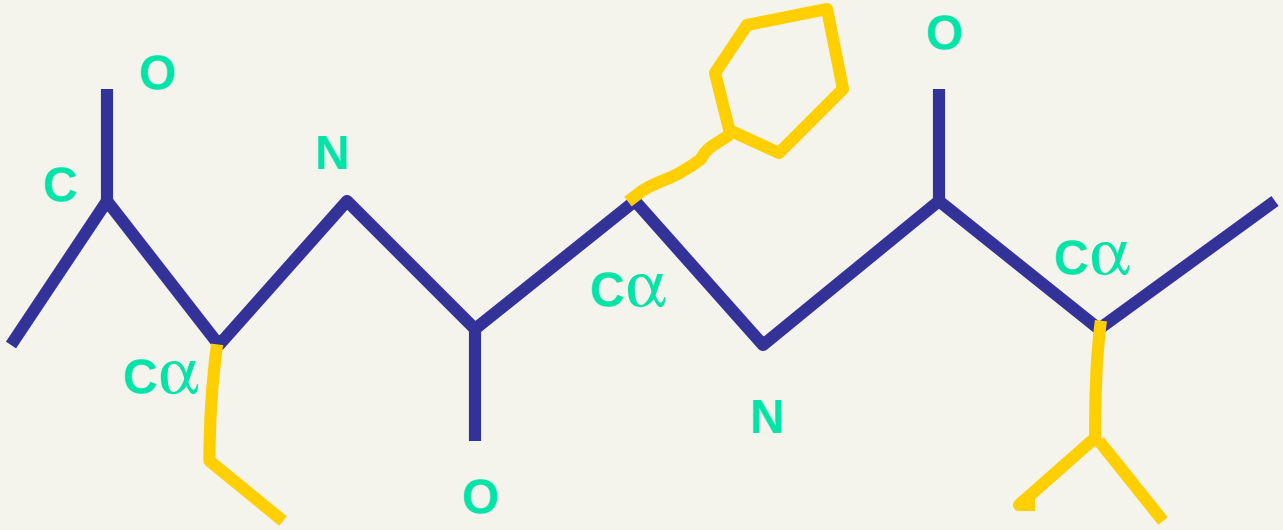
20 Amino Acids

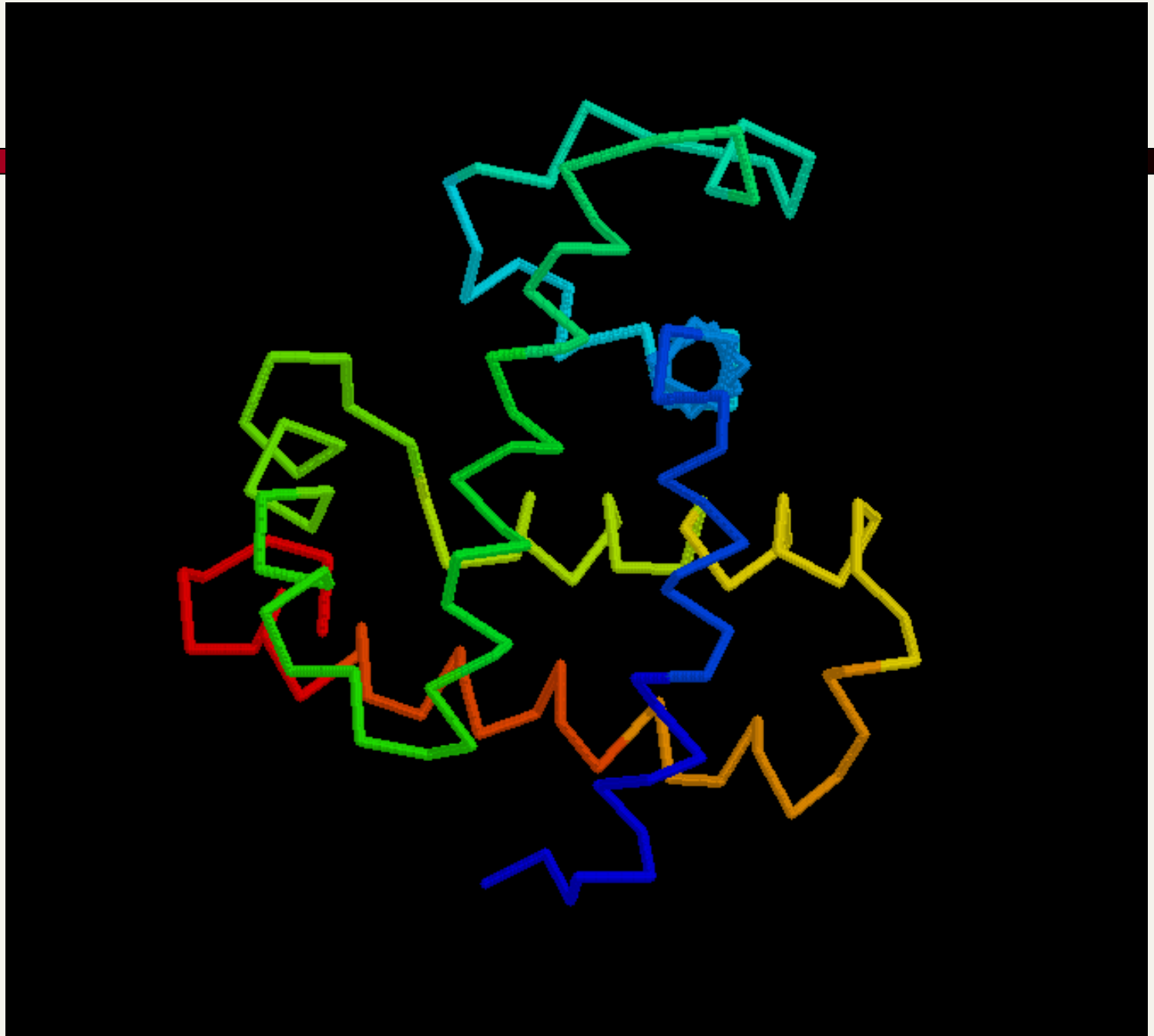
- Common, repeating backbone (blue)
- Unique sidechains (yellow)

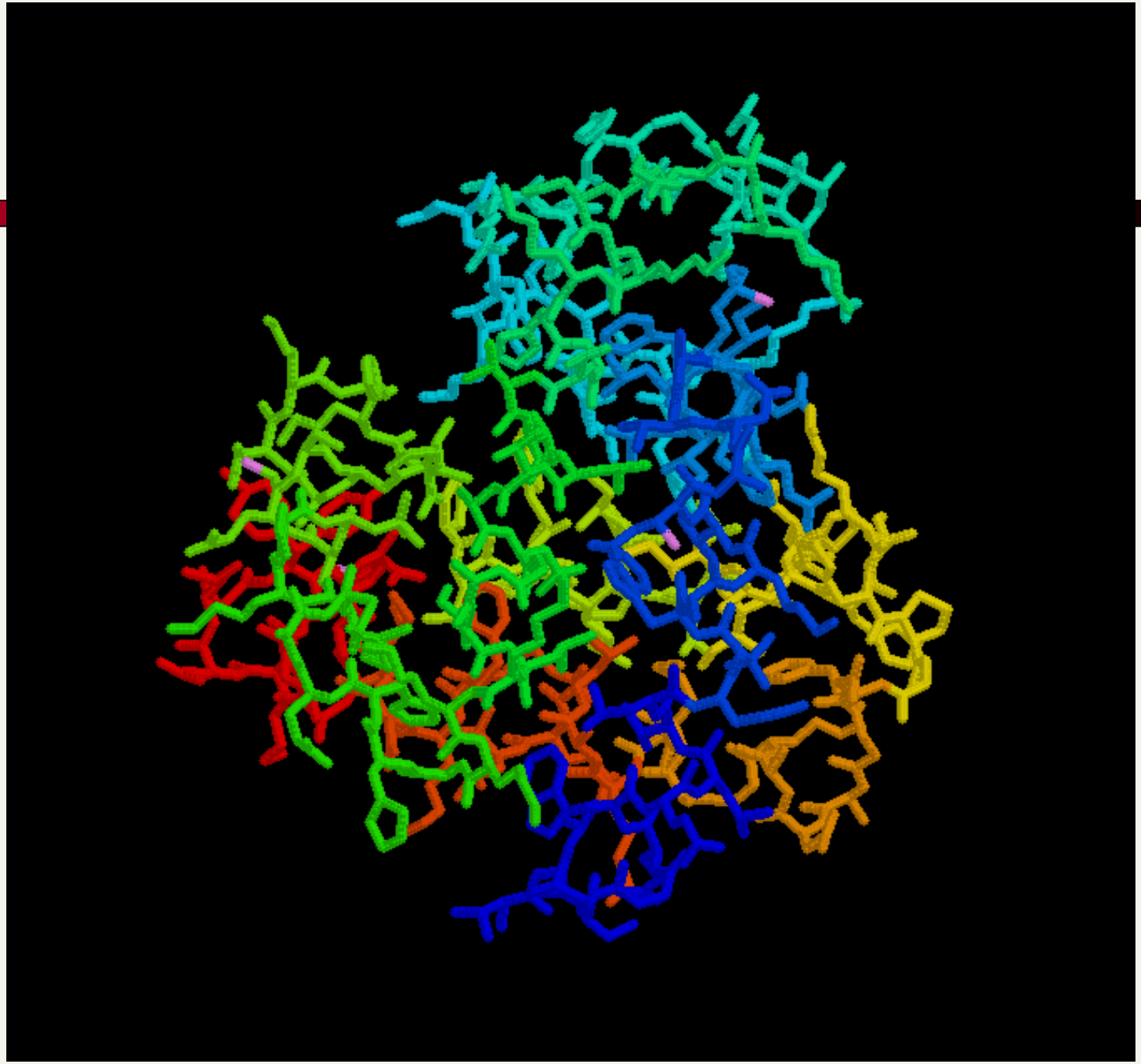


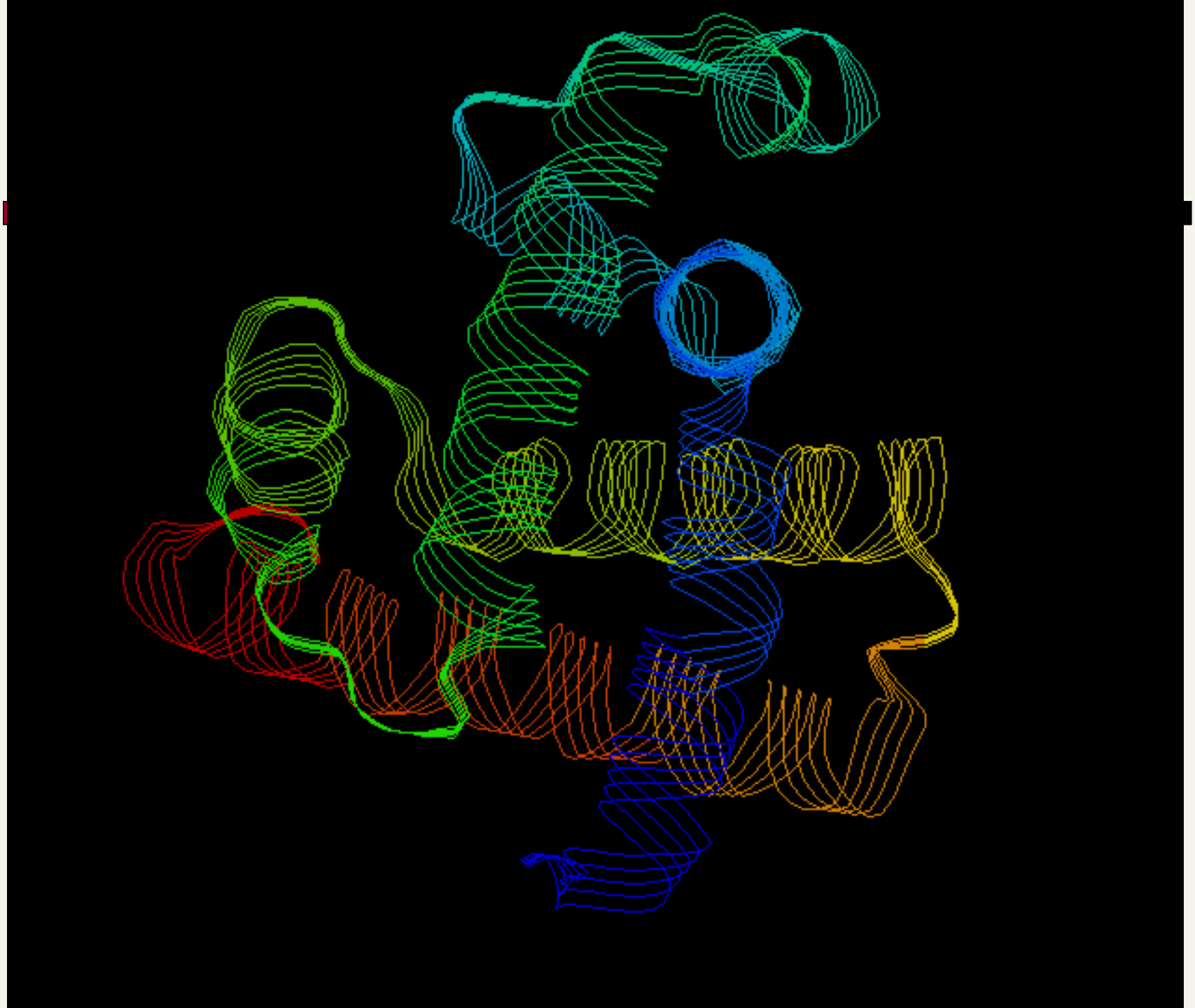
Shorthand for Protein Sequence

- Specify the sequence of amino acids:
  - Alanine-Tyrosine-Valine
  - ALA-TYR-VAL
  - A-Y-V

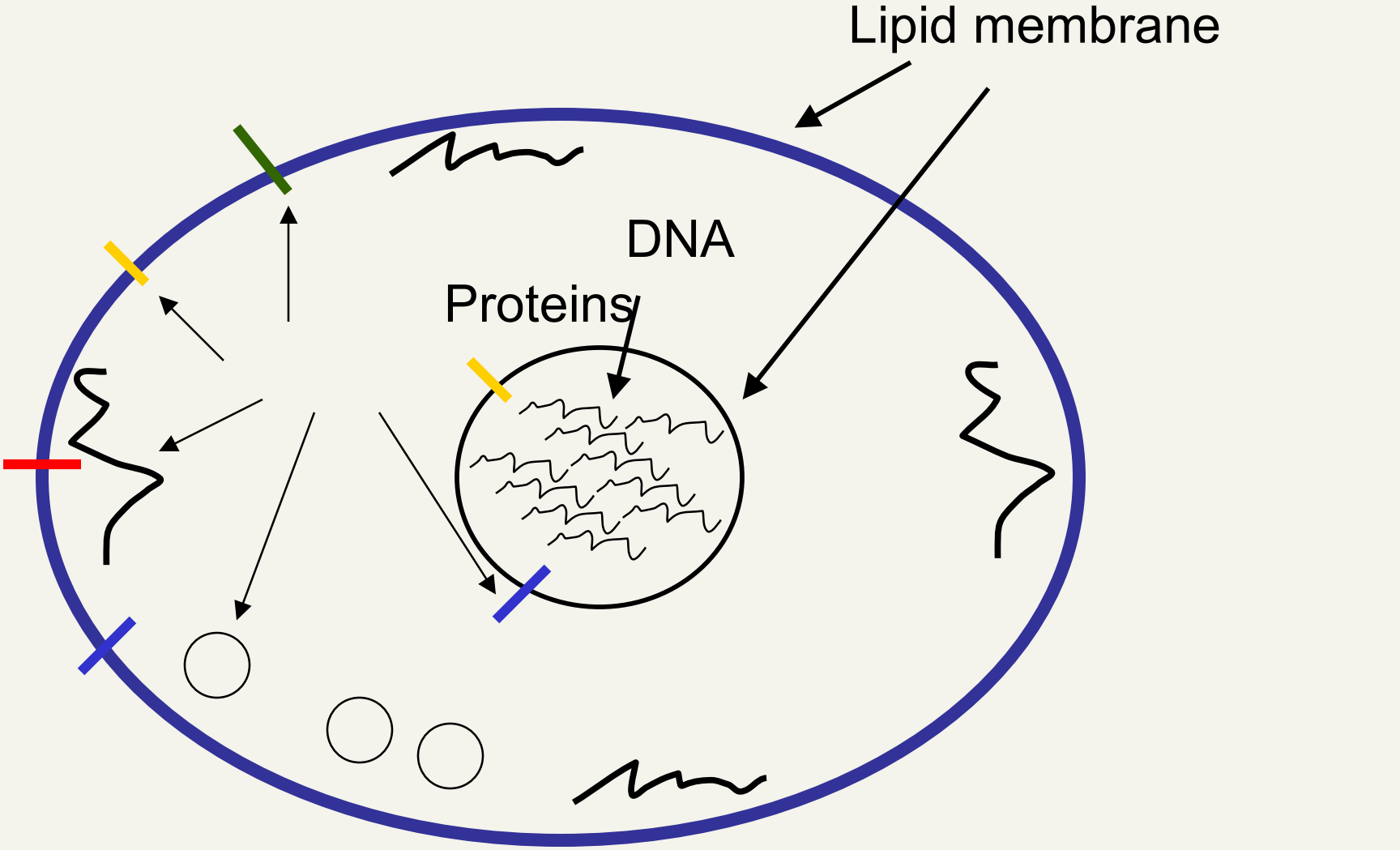




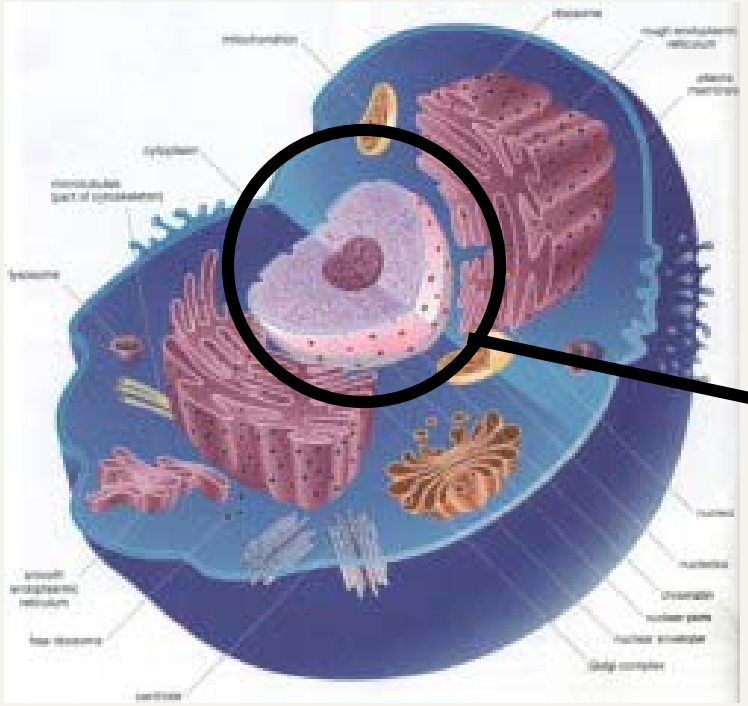


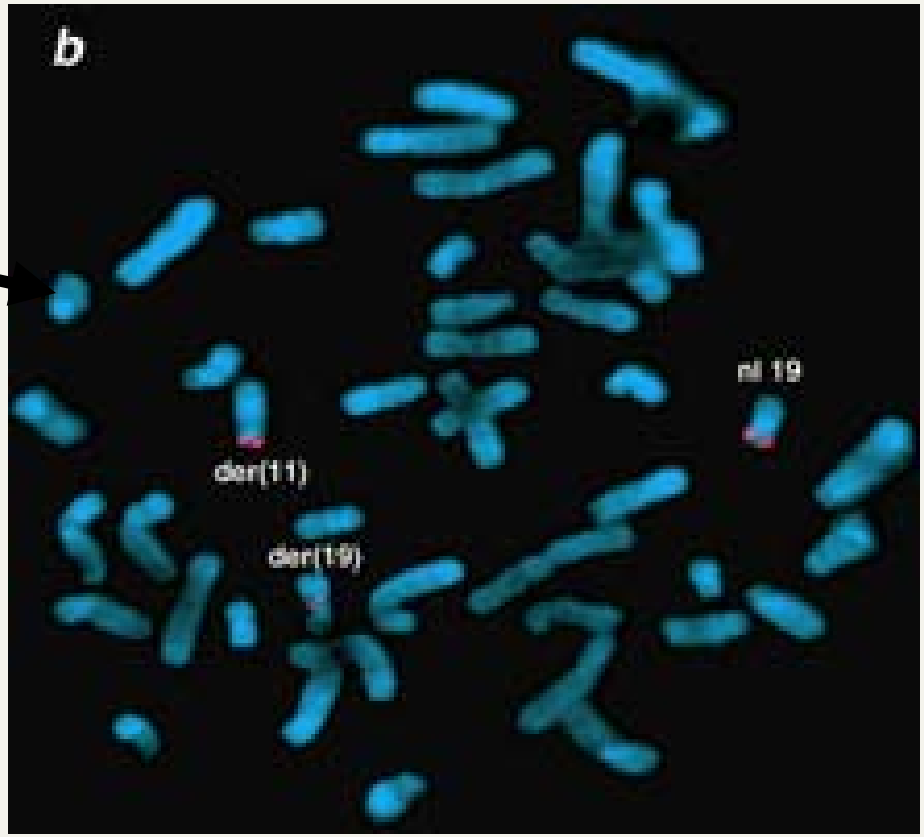


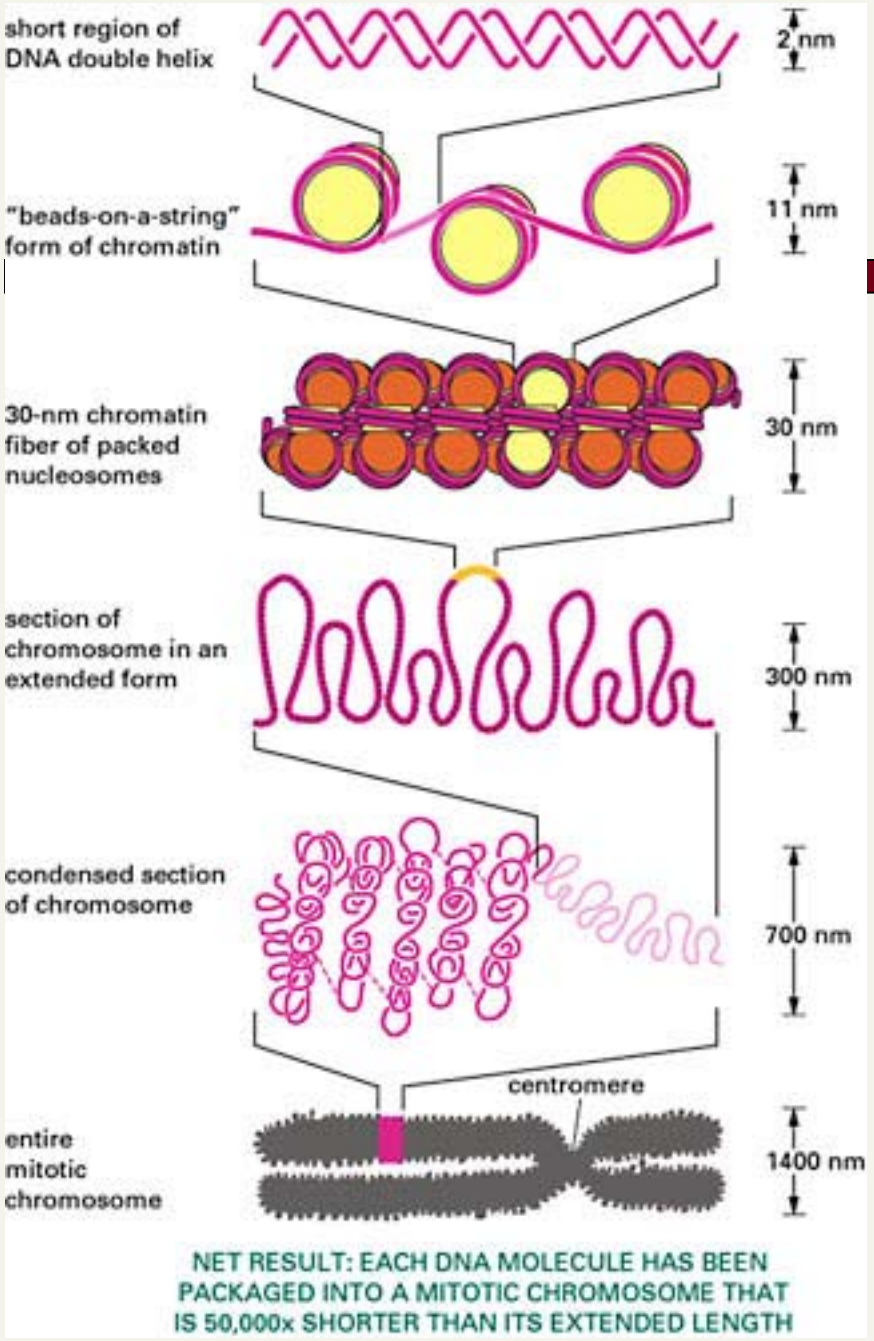
Bioinformatics Schematic of a Cell



Human DNA







DNA packs in the nucleus to form chromosome

The sequence of amino acids in a protein is specified by DNA

- DNA uses an alphabet of 4 letters (ATCG), more commonly called bases.
- Although the 4 letters have interesting chemical structure, for our purposes they are just information carriers.
- Long sequences of these 4 letters are linked together to create GENES and CONTROL INFORMATION.

DNA is a sequence too

- It also has a common backbone, and then specialized sidechains. But there are only 4 specialized sidechains: Adenine, Cytosine, Guanine and Thymidine = A, C, G, and T.
- A sequence of these subunits is also specified as a string:
- e.g., ACTTAGGACATTTTTAG
- This is a shorthand for the chemical structure, which is not important right now.

DNA encodes Protein (and RNA)

- Each of the twenty protein amino acids can be specified by 3 consecutive DNA bases.
- The Ribosome "reads" a sequence of DNA bases (three at a time) and creates the corresponding protein chain—which folds itself based on the amino acid properties.
- See: http://ntri.tamuk.edu/cell/ribosomes.html
- The 64 mappings of 3 bases to 1 amino acid is called the GENETIC CODE and is universal (on earth...).

Genetic Code (T=U here) (e.g. Tyrosine = UAU or UAC)

One-letter code	Amino acid	Three-letter code	Genetic code
A	Alanine	Ala	GC*
C	Cysteine	Cys	UGU, UGC
D	Aspartic Acid	Asp	GAU, GAC
E	Glutamic Acid	Glu	GAA, GAG
F	Phenylalanine	Phe	UUU, UUC
G	Glycine	Gly	GG*
H	Histidine	His	CAU, CAC
I	Isoleucine	Ile	AUU, AUC, AUA
K	Lysine	Lys	AAA, AAG
L	Leucine	Leu	UUA, UUG,CU*
M	Methionine	Met	AUG
N	Asparagine	Asn	AAU, AAC
P	Proline	Pro	CC*
Q	Glutamine	Gln	CAA, CAG
R	Arginine	Arg	CG*, AGA, AGG
S	Serine	Ser	UC*, AGU, AGC
T	Threonine	Thr	AC*
V	Valine	Val	GU*
W	Tryptophan	Trp	UGG
Y	Tyrosine	Tyr	UAU, UAC

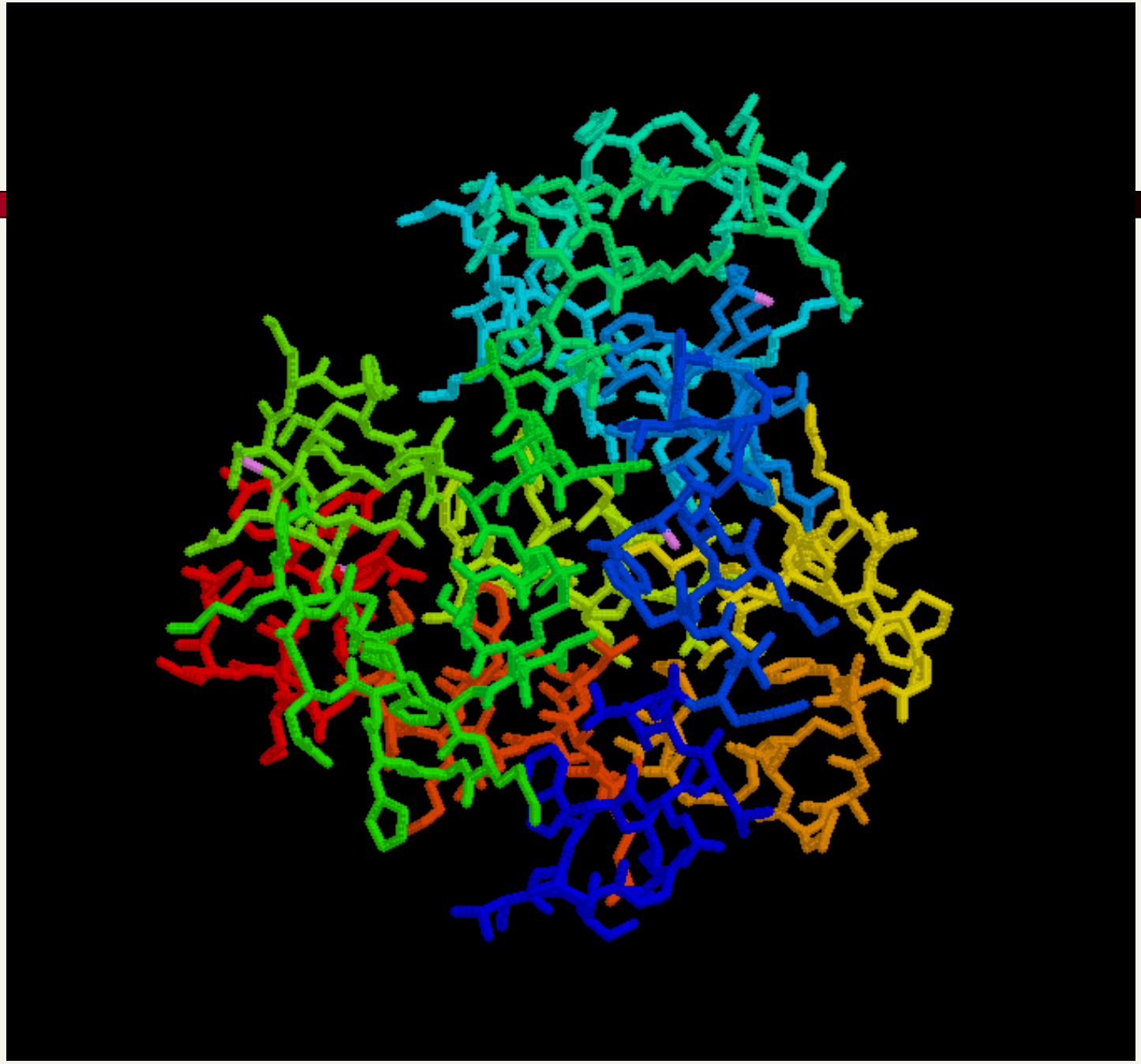
Myoglobin: Gene and Protein

Gene

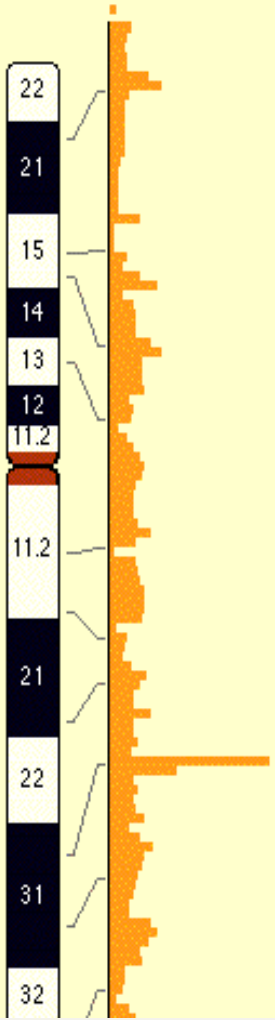
```
ctgcagataa ctaactaaag gagaacaaca acaatggttc tgtctgaagg
tgaatggcag ctggttctgc atgtttgggc taaagttgaa gctgacgtcg
ctggtcatgg tcaggacatc ttgattcgac tgttcaaatc tcatccggaa
actctggaaa aattcgatcg tttcaaacat ctgaaaactg aagctgaaat
gaaagcttct gaagatctga aaaaacatgg tgttaccgtg ttaactgccc
taggtgctat ccttaagaaa aaagggcatc atgaagctga gctcaaaccg
cttgcgcaat cgcatgctac taaacataag atcccgatca aatacctgga
attcatctct gaagcgatca tccatgttct gcattctaga catccaggta
acttcggtgc tgacgctcag ggtgctatga acaaagctct cgagctgttc
cgtaaagata tcgctgctaa ctgggttacc agggttaatg aggtacc

BASE COUNT      155 a    108 c    115 g    129 t
```

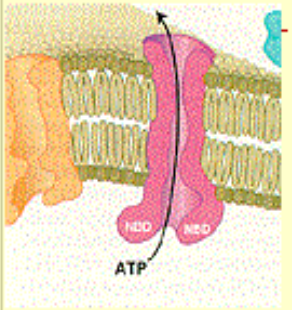
Protein

```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPG
NFGADAQGAMNKALELFRKDIAAKYKELGYQG
```



Why We Care: Diseases





CFTR. The gene encoding a chloride ion channel is defective in patients with cystic fibrosis

IMAGE CREDIT: Q. Alawqati, Columbia University, NY, USA. Adapted by K. Sutliff, SCIENCE.

OBS. The obese (Ob) mutation in the mouse provides a useful model system for studying human obseity

IMAGE CREDIT: Jeff Friedman, Rockfeller University, New York, NY, USA. Reprinted from SCIENCE.

Genes: Statistics

- The set of all genes required for an organism is the organism's GENOME.
- The human genome has 3,000,000,000 bases divided into 23 linear segments (chromosomes).
- A gene has on average 1340 DNA bases, thus specifying a protein of about 447 amino acids.
- Humans have about 35,000 genes = 40,000,000 DNA bases = 3% of total DNA in genome.
- Humans have another 2,960,000,000 bases for control information. (e.g. when, where, how long, etc...)

Computational Molecular Biology

- Main focus used to be
  - Sequence analysis (human genome project)
  - Structure analysis (what is 3d structure of proteins?)
- Increasingly, the focus is:
  - Function analysis

This is where text mining can help.

Biological Structure and Function

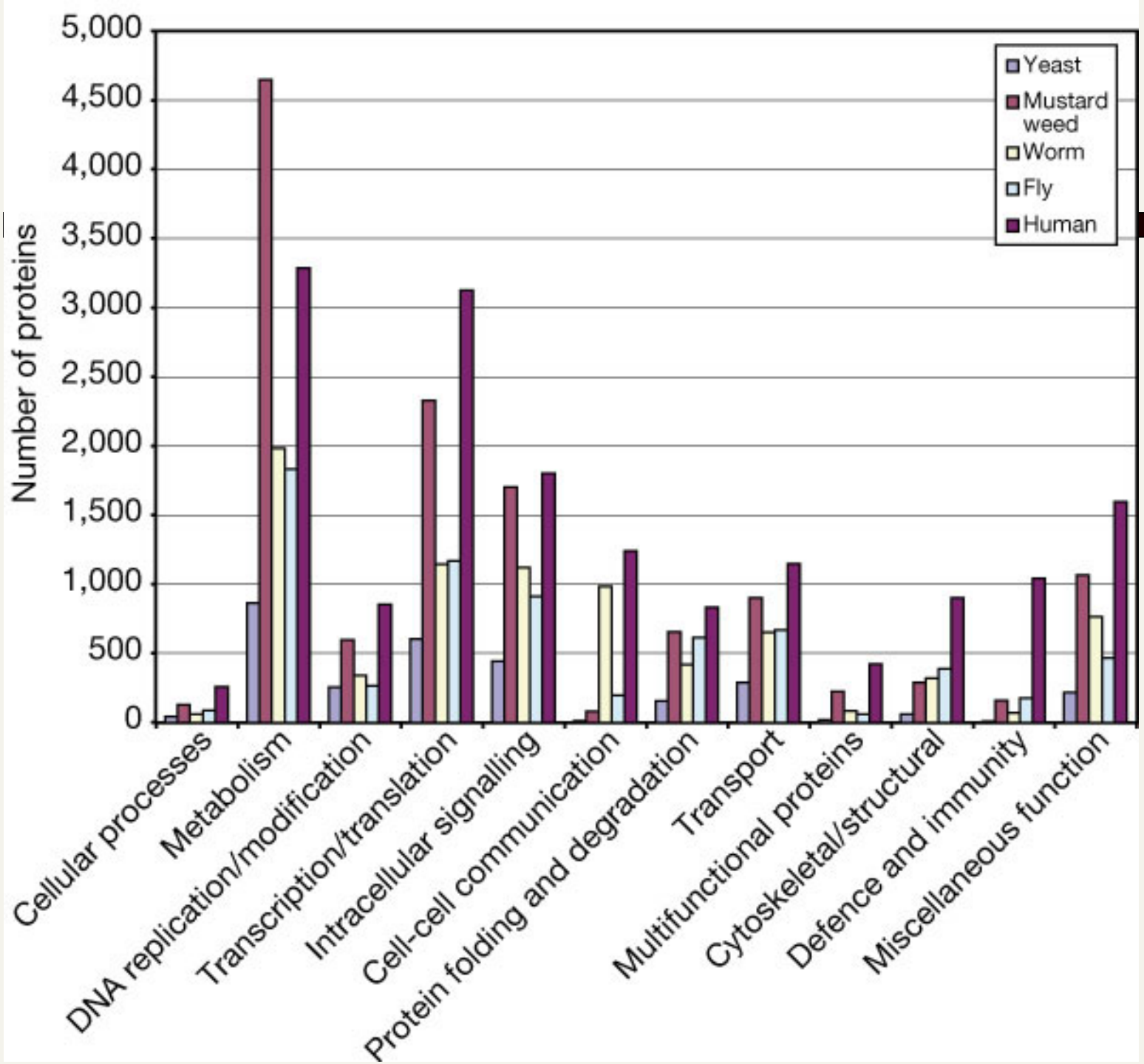
- Sequence & Structure
  - Precise representation as 1D and 3D objects.
- Function: somewhat fuzzy
  - Often represented as text

What are Functions of Genes?

- Signal transduction: sensing a physical signal and turning into a chemical signal
- Structural support: creating the shape and pliability of a cell or set of cells
- Enzymatic catalysis: accelerating chemical transformations otherwise too slow.
- Transport: getting things into and out of separated compartments

What are the Functions of Genes?

- Movement: contracting in order to pull things together or push things apart.
- Transcription control: deciding when other genes should be turned ON/OFF
- Trafficking: affecting where different elements end up inside the cell

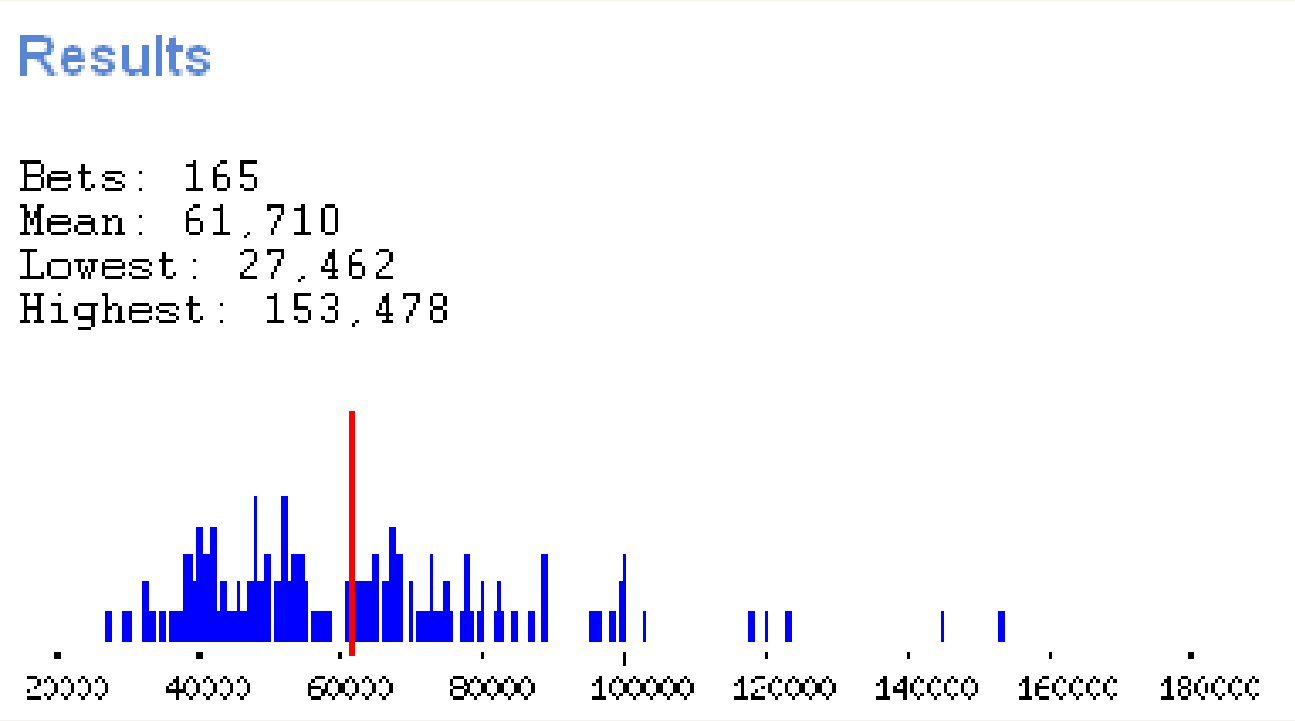

Number of proteins
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Yeast
Mustard weed
Worm
Fly
Human
Cellular processes
Metabolism
DNA replication/modification
Transcription/translation
Intracellular signalling
Cell-cell communication
Protein folding and degradation
Transport
Multifunctional proteins
Cytoskeletal/structural
Defence and immunity
Miscellaneous function

Why So Few Human Genes?

- Complexity is not a function of the number of genes.
  - Control information critical.
- Complexity is a function of the number of genes, and mustard weed is more complex than we are.
- Number of genes is not estimated correctly.

How Many Genes Do You Have?

- http://www.ensembl.org/Genesweep/
- Bet how many human genes there are
- Winner to be decided May 2003?



Basic Biology: Summary

- Three “technologies”: lipids, proteins, DNA
- Biology needs text mining / NLP
- Biology is an information-intensive science.
  - A lot of the information is in text.
  - Biology is a natural application area for text mining/processing.
- Function is key for understanding biology.
  - There are formal and precise representations for sequence and structure.
  - Text is still the main representation for function.

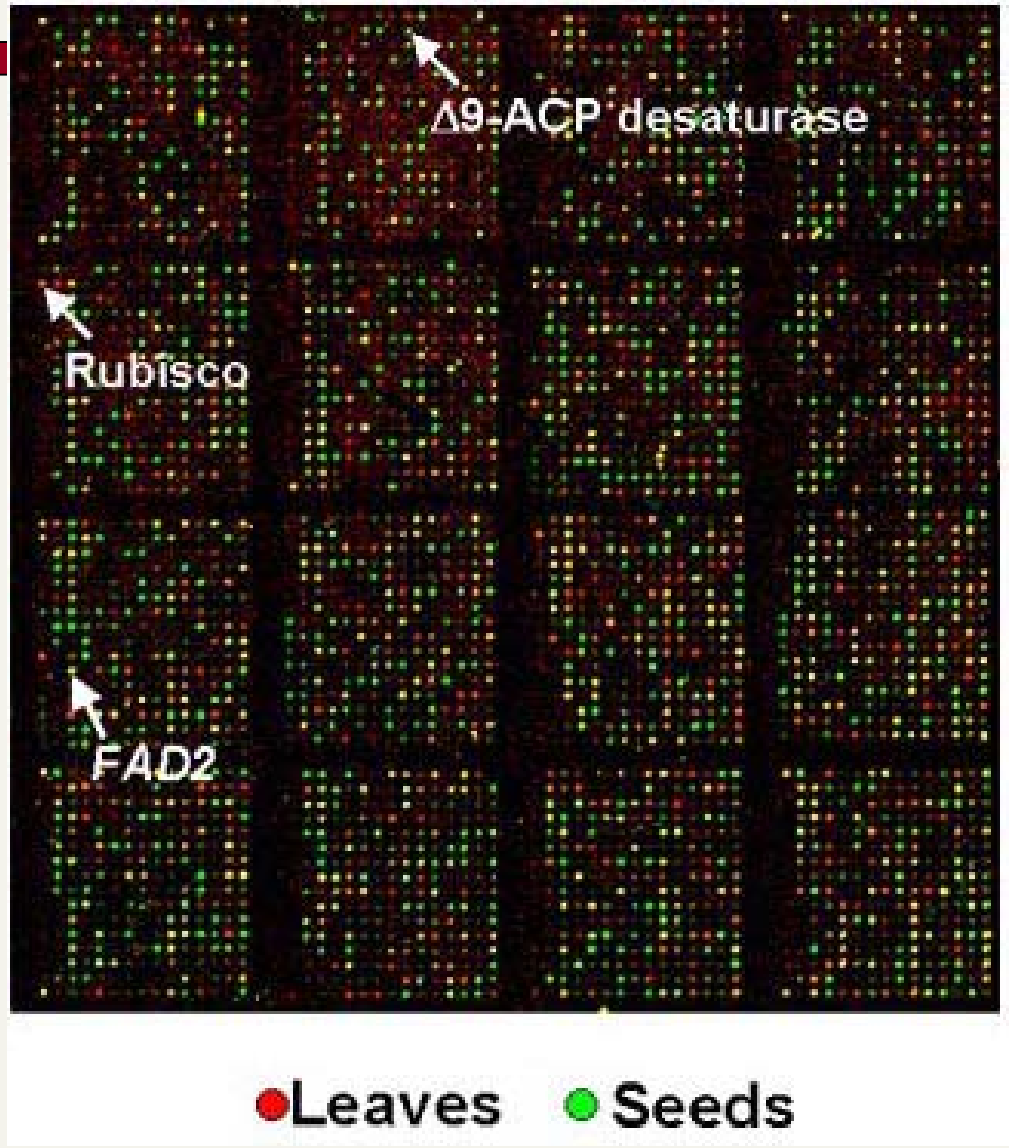
Microarray Analysis

Microarrays

- Measure the expression of genes
- 2-color arrays compare 2 conditions, control and experimental
- Upregulated = red, downregulated = green
- Example Application: clinical diagnosis

A cDNA Microarray

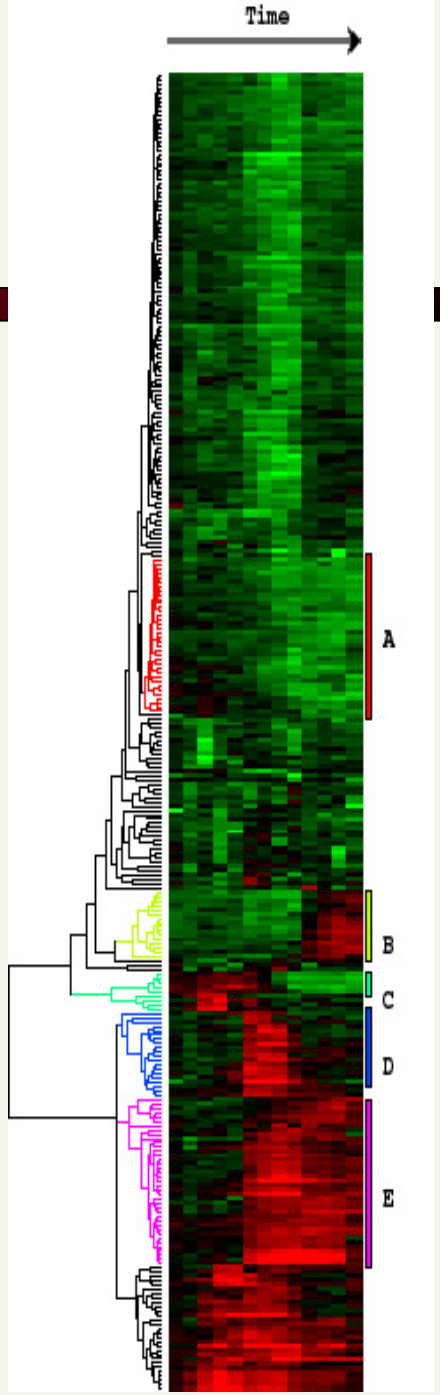
(Source: C. Benning)



Common Analysis Procedure

- Quality control (did the experiment work?)
- Cropping (select affected genes)
- Clustering (group genes)
- Manual exploration of data
- Sense making

Clustering: Example (Eisen et al.)



Text in Microarray Analysis

- Each biologist only know a few genes well.
- Wading through search results is tedious and time consuming.
- Relating measurements with existing knowledge is a key part of microarray analysis.

Two Approaches

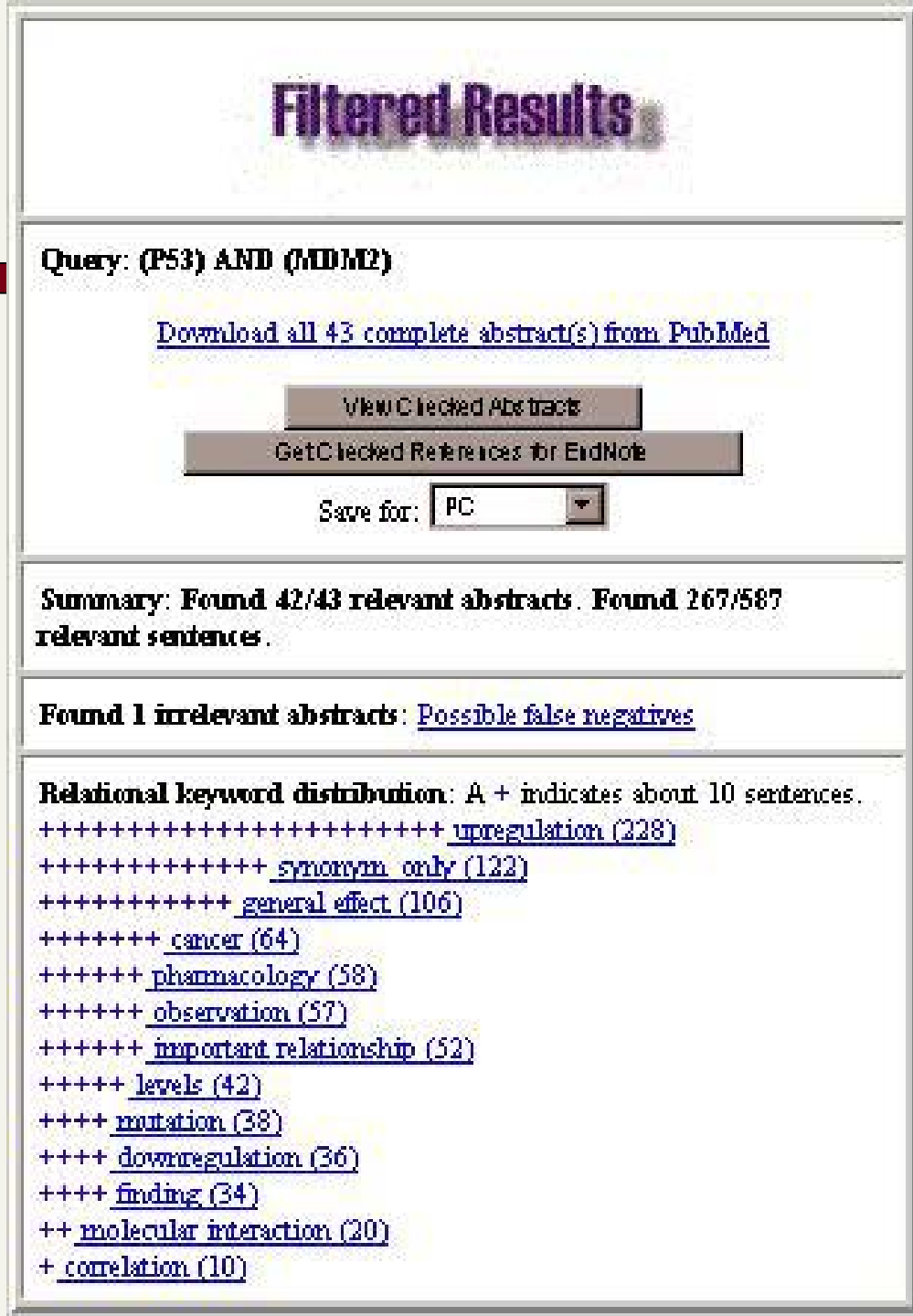
- Cluster on numeric data, then interpret textually
- Cluster on textual data, then interpret numerically

MedMiner:
First Numbers, then Text

- Identify group of genes based on experimental data
- MedMiner
  - Identifies significant keywords
  - Creates a list of relevant contexts

MedMiner (Tanabe et al.)

Key words



MedMiner (Tanabe et al.), cont.

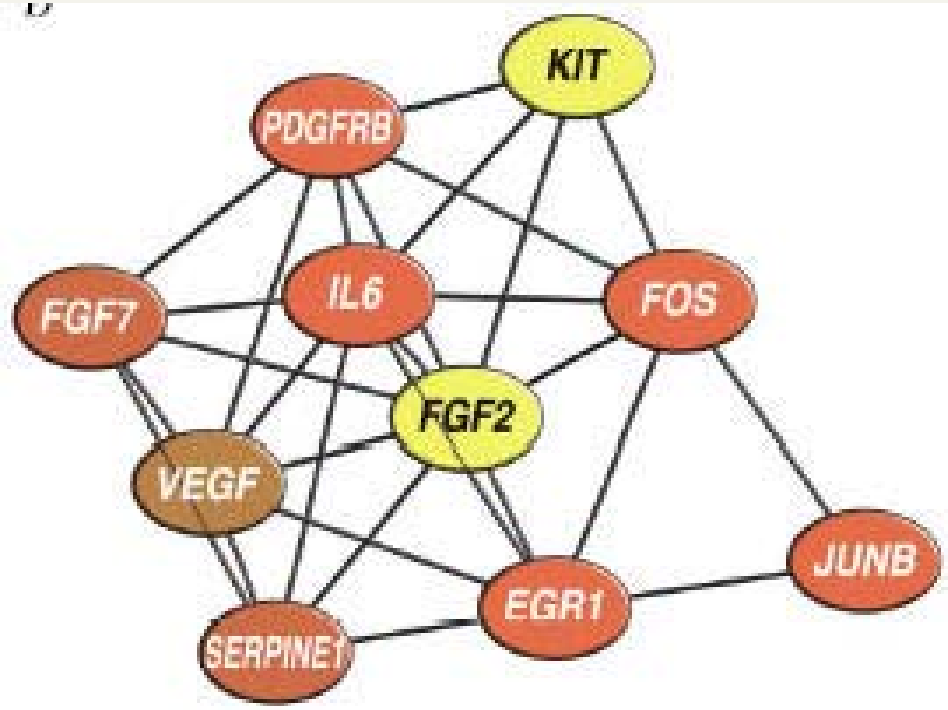
Contexts

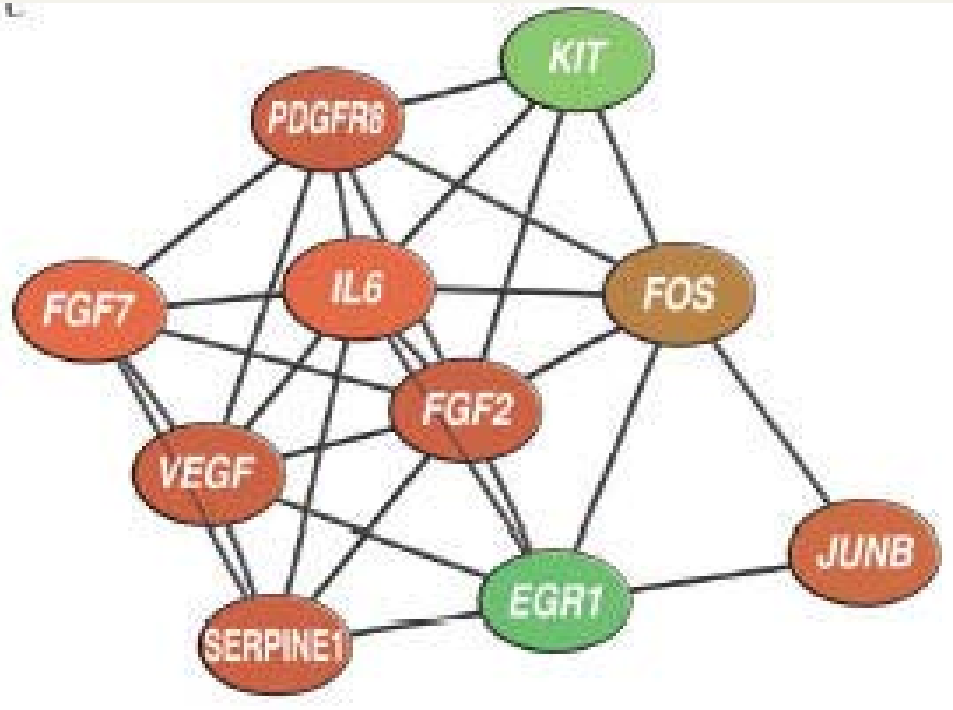
correlation	Link to Abstract
correlat	top of page
Overexpression of MDM2 (>/=10-fold) was significantly correlated with adriamycin resistance and decreased duration of CR1. (2000)	PMID 10637478
The chi2 test was performed to describe the correlation between the Ki-67 index and p53, MDM2, and p21 protein expression. (1999)	PMID 10632343
A strong correlation was observed between the Ki-67 index > 10% and both MDM2 and p21 proteins. (1999)	PMID 10632343
Accumulation of p53 and MDM2 overexpression correlated with the grade of malignancy. (1999)	PMID 10631716
No correlation was found between p53 accumulation and the histopathology of gastric cancer. (1999)	PMID 10631716
p53 accumulation and MDM2 overexpression did not correlate with tumor size, nodal status, presence of metastases, age or survival. (1999)	PMID 10631716

PubGene:
First Text, then Numbers

- Compile a list of all genes
- Compute co-occurrence of genes in medline articles
- Display network(s) of selected genes
- Color-code nodes to indicate degree of up/downregulation

Text Cluster Analysis (Jenssen et al.)





1H expression levels

8H expression levels

Highly upregulated at 1H

Why Text about Biology is Special

Biological Terminology: A Challenge

- Large number of entities (genes, proteins etc)
- Evolving field, no widely followed standards for terminology -> Rapid Change, Inconsistency
- Ambiguity: Many (short) terms with multiple meanings (eg, CAN)
- Synonymy: ARA70, ELE1alpha, RFG
- High complexity -> Complex phrases

What are the concepts of interest?

- Genes (D4DR)
- Proteins (hexosaminidase)
- Compounds (acetaminophen)
- Function (lipid metabolism)
- Process (apoptosis = cell death)
- Pathway (Urea cycle)
- Disease (Alzheimer's)

Complex Phrases

- Characterization of the repressor function of the nuclear orphan receptor retinoid receptor-related testis-associated receptor/germ nuclear factor

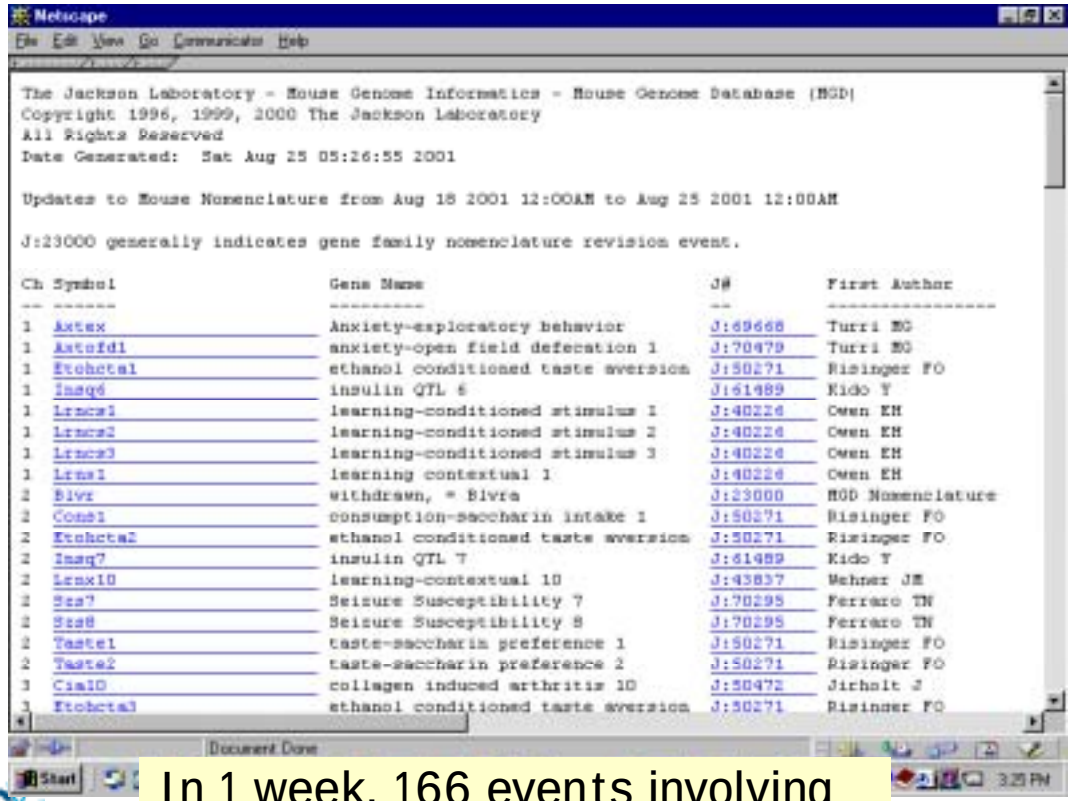
Inconsistency

- No consistency across species

	Protease	Inhibitor	signal
Fruit fly	Tolloid	Sog	dpp
Frog	Xolloid	Chordin	BMP2/BMP4
Zebrafish	Minifin	Chordino	swirl

Rapid Change

Mouse Genome Nomenclature Events 8/25



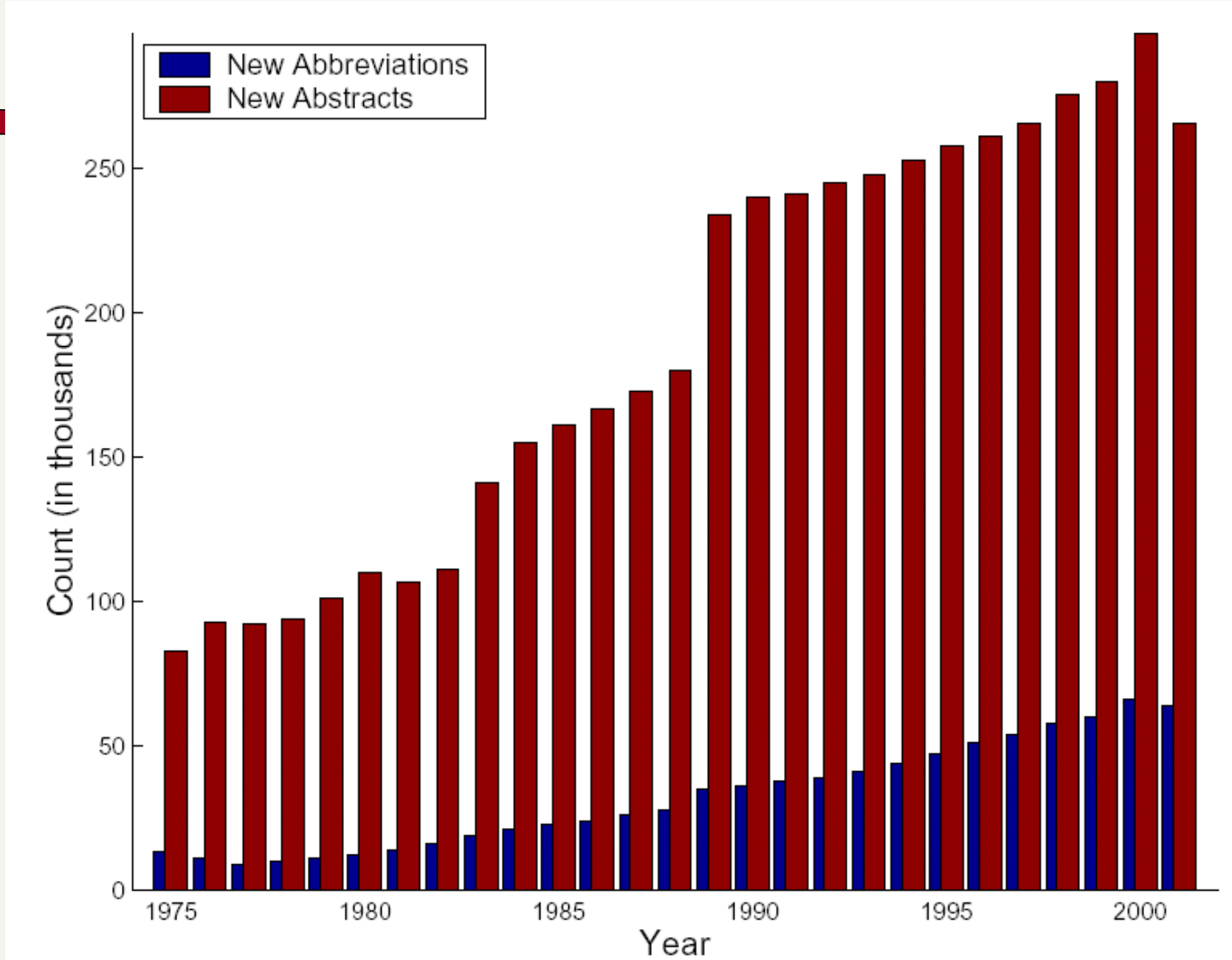
In 1 week, 166 events involving change of nomenclature

Abbreviation Mining (Chang,Schütze&Altman)

Abbreviations in Biology

- Two problems
  - “Coreference”/Synonymy
    - What is PCA an abbreviation for?
  - Ambiguity
    - If PCA has >1 expansions, which is right here?
- Only important concepts are abbreviated.
- Effective way of jump starting terminology acquisition.

Frequency of Abbreviations



Ambiguity Example
PCA has >60 expansions

"p-chloroamphetamine" "p-chloroaniline" "p-coumaric acid" "p.rothrombin c.omplex a.ctivity" "para-chloramphetamine" "parietal cell antibodies" "parietal cell autoantibodies" "paroxysmal cerebellar ataxia" "passive cutaneous anaphylactic" "patient care appraisal" "patient controlled analgesia" "patient controlled anesthesia" "pca" "pentachloroanisole" "percent cortical area" "perchloracetic acid" "perchloric acid" "percutaneous coronary angioplasty" "percutaneous coronary atherectomy" "pericallosal artery" "peritoneal carcinomatosis" "peritoneal carcinosis" "personal care attendant" "phenazine-1-carboxylic acid" "phenylciclopentylacetic acid" "phenylcyclohexylamine" "physical capacity assessment" "pig coronary artery" "plate count agar" "pneumococcal capsular antigen" "pole climbing avoidance" "polyclonal activator" "polyclonal antibody" "polyclonal antisera" "polycyclic aromatic content" "porous coated anatomic" "porous coated total hip arthroplasty" "porous-coated hip arthroplasties" "porous-coated patellar component" "porta-caval anastomosis" "portable clinical analyzer" "portacaval anastomosis" "portacaval shunt" "post chigger attachment" "postconceptional age" "posterior cerebral arteries" "posterior communicating artery" "posterior cortical atrophy" "posterior crico-arytenoid" "potassium channels activators" "presence of parietal cell" "primary cardiac arrest" "primary congenital aphakia" "principal component analyses" "procoagulant" "procoagulant activities" "procoagulant cellular activity" "procoagulatory activity" "prostatic carcinoma" "protein c activator" "prothrombin complex activity" "protocatechuate" "protocatechuic acid" "protocaval anastomosis" "pulmonary corpora amylacea" "pyroglutamic acid" "pyrrolidone carboxylic acid" "pyrrolidone-2-carboxylic acid" "pyrrolidone-5-carboxylic acid" "pyrroline-5-carboxylate"

Problem 1: Ambiguity

- “Senses” of an abbreviation are usually not related.
- Long form often occurs at least once in a document.
- Disambiguating abbreviations is easy.

Problem 2: “Coreference”

- Goal: Establish that abbreviation and long form are coreferring.
- Strategy:
  - Treat each pattern w*(c*) as a hypothesis.
  - Reject hypothesis if well-formedness conditions are not met.
  - Accept otherwise.

Dynamic Programming

- Align the abbreviation with the preceding text using dynamic programming.
- Associate costs with each alignment that reflect well-formedness of the abbreviation.

Example

- Medline excerpt: According to a system proposed by the European group for the immunological classification of leukemia (EGIL) ....
- Align: “EGIL” with preceding text

```
E........G.............I.................................L.......
European group for the immunological classification of leukemia
```

Dynamic Programming Alignment costs

long form	abbreviation	cost
ε	character c	100.0
c_1	c_2 (c_1!=c_2)	100.0
non-initial c	first c	100.0
initial c	ε	5.0
initial c	c	0.0
non-initial c	ε	0.1
non-initial c	c	1.0

Evaluation: Precision

- Algorithm tested on a dictionary of abbreviations available from the China Medical Tribute (452)
- 406 (90%) correct
- Error analysis:
  - Syllable boundaries
  - “Morphology”
  - Semantics
  - Suboptimal length/wellformedness trade-off

Errors: Syllable Boundaries

P-------I------------M--------------------s
phosphatidylinositol manno-oligosaccharides

a-------E------------E----------------G-----
amplitude-integrated electroencephalography

Errors: “Morphology”

pr---------o--M------M------P-------s-
precursors of matrix metalloproteinase

N------A---P------R------T-a-s-e----
nicotinate phosphoribosyltransferase

C--------I---------------N ------I-
cervical intraepithelial n-eoplasia

Errors: Semantics

a---P------L----------------------A---------
antiphospholipid anticardiolipin antibodies

G-------6-P---------D----------------------
glucose-6-phosphate dehydrogenase-deficient

Errors: Incorrect Tradeoff Length vs. Well-Formedness

P___O_____P__C_____
pulmonary complications
P___O_____P_____C_____
Postoperative pulmonary complications

P_____P__R__O_____M_____
premature rupture of the membranes
P_____P_____R_____O_____M_____
Preterm premature rupture of the membranes

Recall

- Analyze all of Medline (37 gigabytes)
- Identify all possible candidates
- 375 correctly identified out of 452 (83%)
- Errors:
  - Precision errors
  - Abbreviation not in Medline
  - Narrow scope of defining context

Errors: Abbreviations not in Medline

- VATS: video assisted thorascopy (vs. video assisted thorascopy surgery)
- VVR: ventricular volume reduction

Errors: Narrow Scope of Defining Context

- We only mine text segments for abbreviations that match regular expression.
- This regular expression was too narrowly defined.

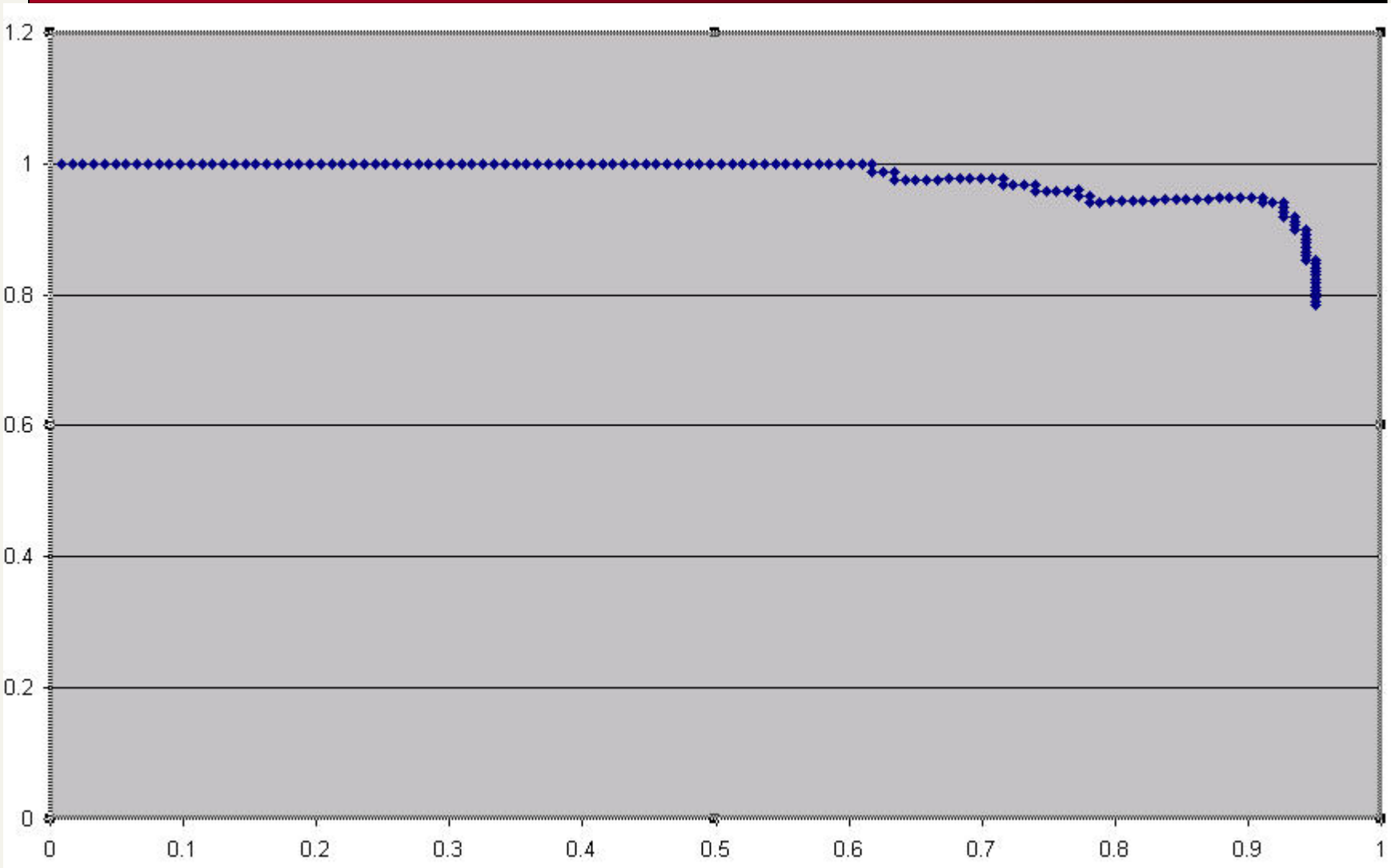
“Post”-definition

```
ACA2p (Arabidopsis Ca2+-ATPase, isoform 2 protein
```

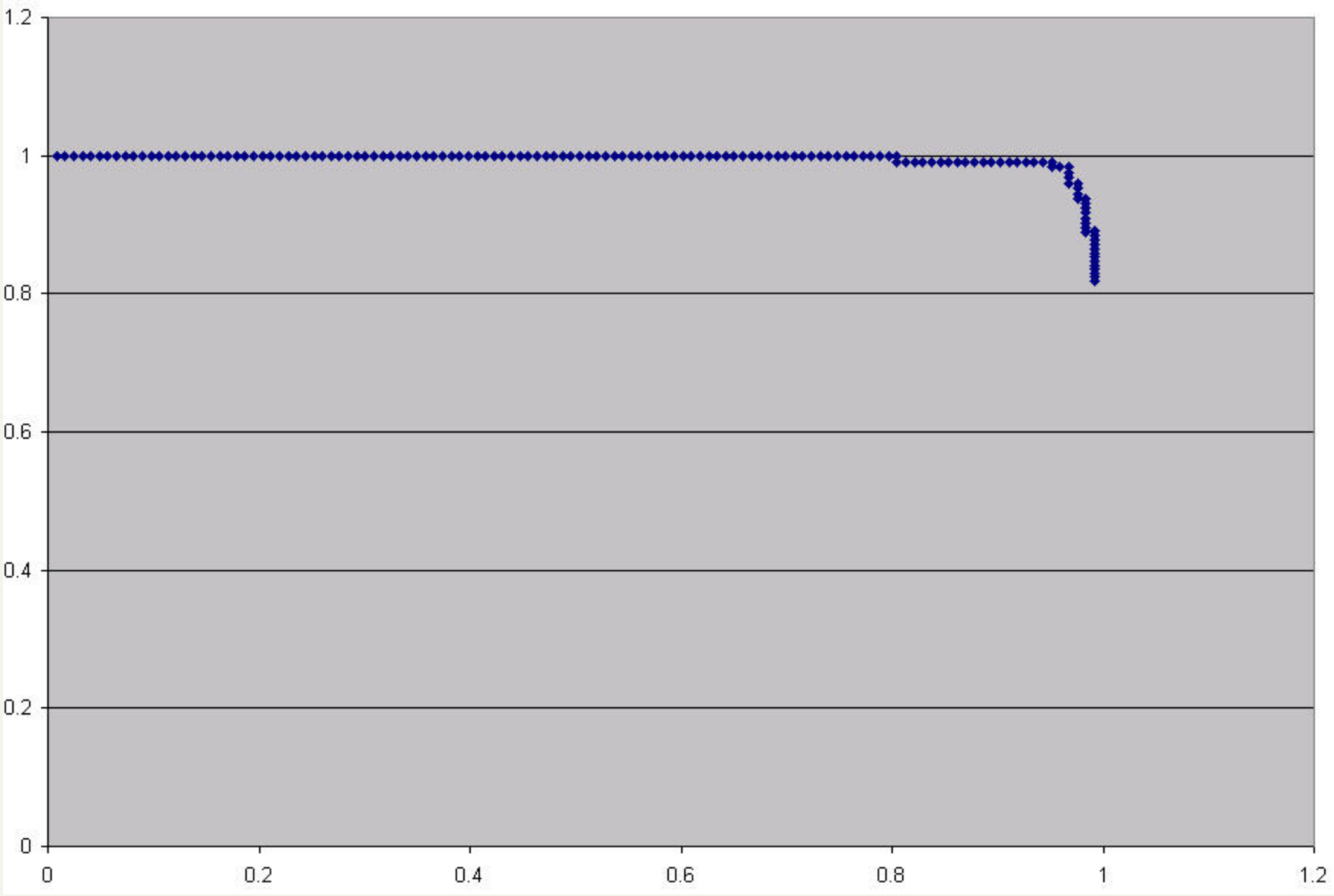
Non-standard term

```
benzodiazepine receptor (peripheral) (BZRP)
```

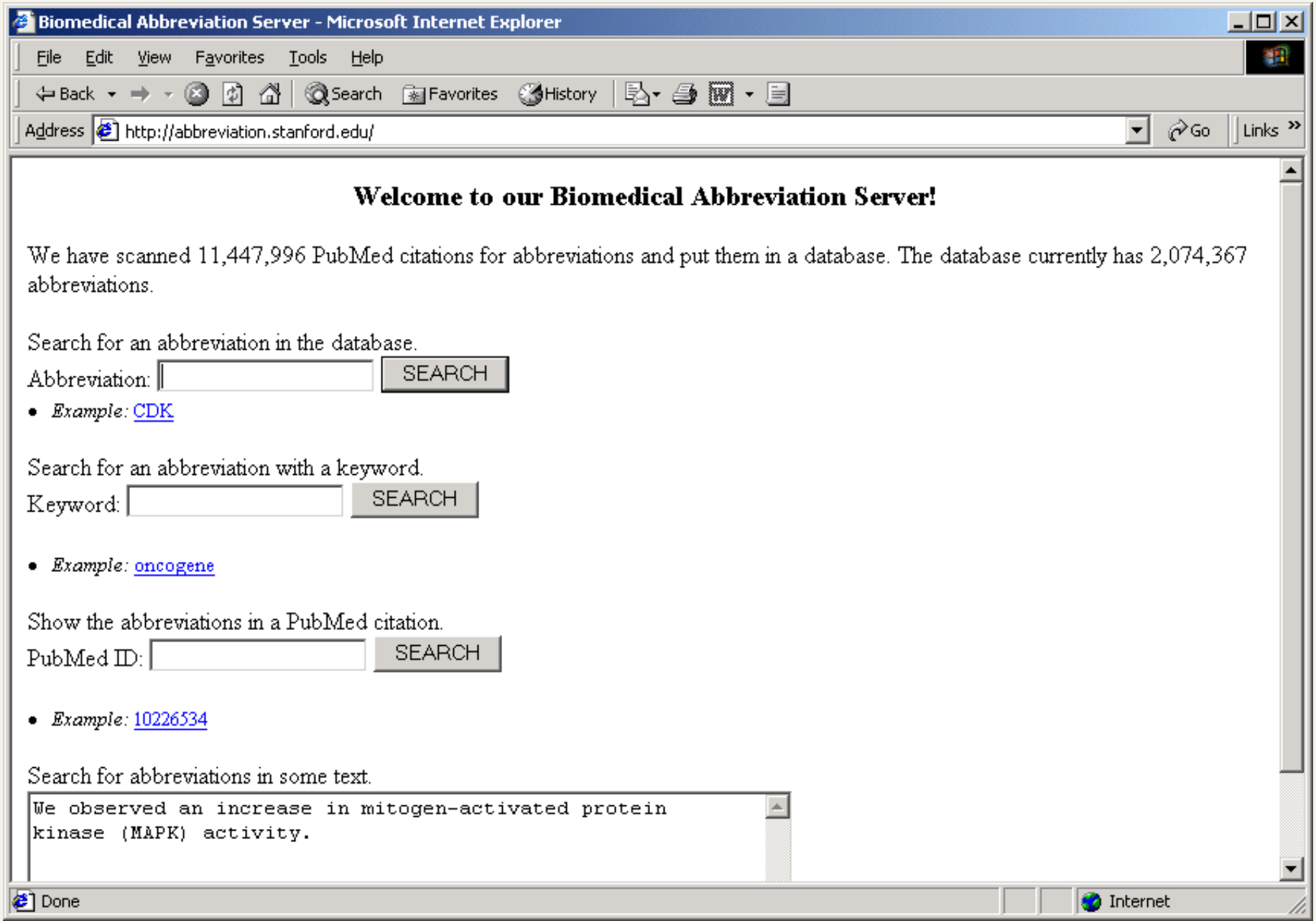
Evaluation: recall/precision
No syllable boundaries

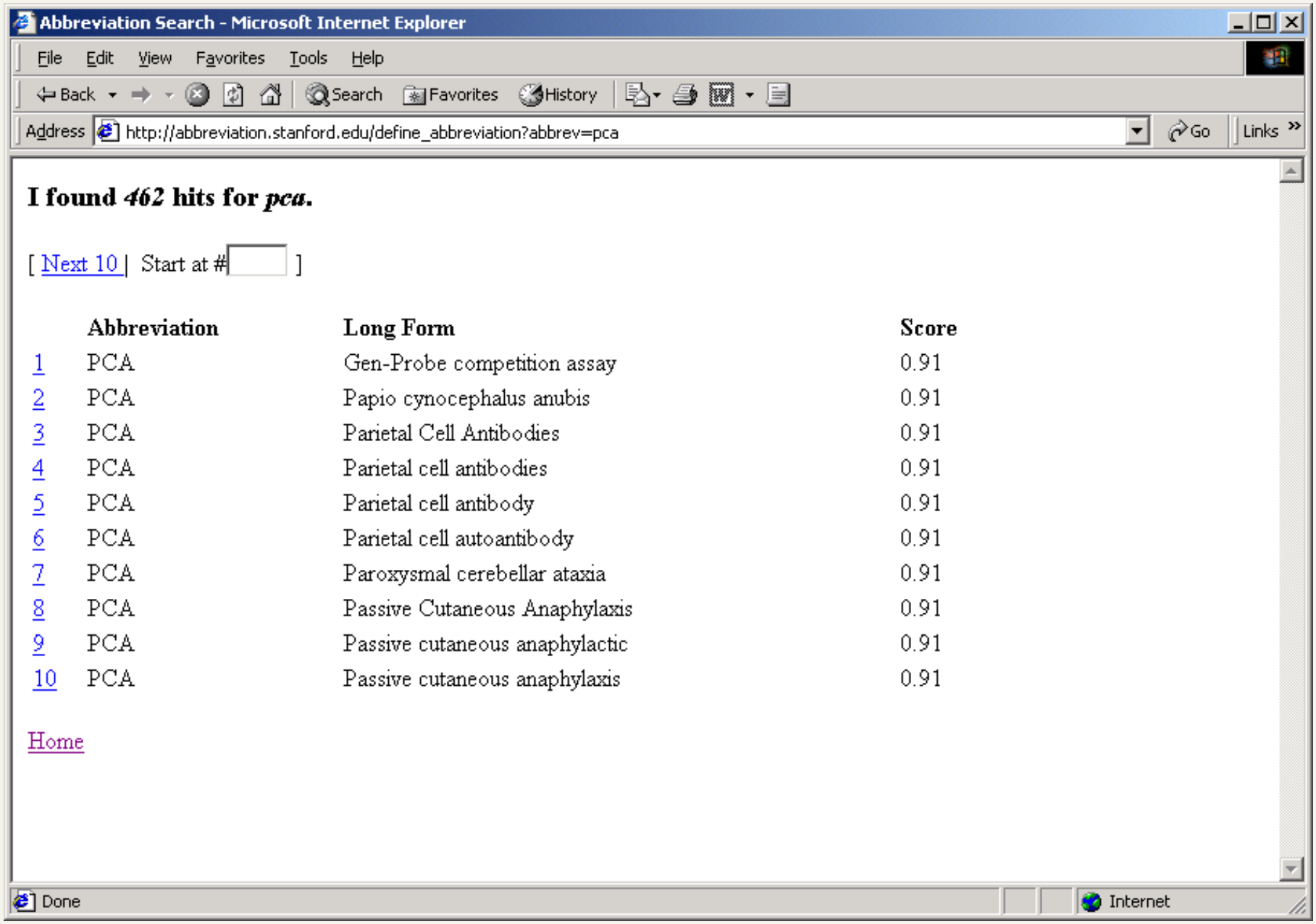


w/ syllable boundaries corrected



Jeff Chang's Abbreviation Server




Abbreviation Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://abbreviation.stanford.edu/define_abbreviation?abbrev=pca
Go Links
I found 462 hits for pca.
[Next 10 | Start at #]
Abbreviation Long Form Score
1 PCA Gen-Probe competition assay 0.91
2 PCA Papio cynocephalus anubis 0.91
3 PCA Parietal Cell Antibodies 0.91
4 PCA Parietal cell antibodies 0.91
5 PCA Parietal cell antibody 0.91
6 PCA Parietal cell autoantibody 0.91
7 PCA Paroxysmal cerebellar ataxia 0.91
8 PCA Passive Cutaneous Anaphylaxis 0.91
9 PCA Passive cutaneous anaphylactic 0.91
10 PCA Passive cutaneous anaphylaxis 0.91
Home
Done
Internet

Approach 2

- The algorithm shown only considers the best alignment. If (best score>θ) accept else reject.
- Alternative
  - Generate a set of good alignments
  - Build feature representation
  - Classify feature representation

Features for Classifier

- Describes the abbreviation.
  - Lower Abbrev
- Describes the alignment.
  - Aligned
  - Unused Words
  - AlignsPerWord
- Describes the characters aligned.
  - WordBegin
  - WordEnd
  - SyllableBoundary
  - HasNeighbor

Weights of Abbreviation Features

CONSTANT	-9.70
LowerAbbrev	-1.21
Aligned	3.67
UnusedWords	-5.82
AlignsPerWord	0.70
WordBegin	5.54
WordEnd	-1.40
SyllableBoundary	2.08
HasNeighbor	1.50

Discussion

- Overall an easy problem
- Could learn the parameters of dynamic programming from training set.
  - Minimize cost: α align-cost + (1-α) recognition-cost
- Related work: see resources

Text-Enhanced Homology Search (Chang, Raychaudhuri, Altman)

Sequence Homology Detection

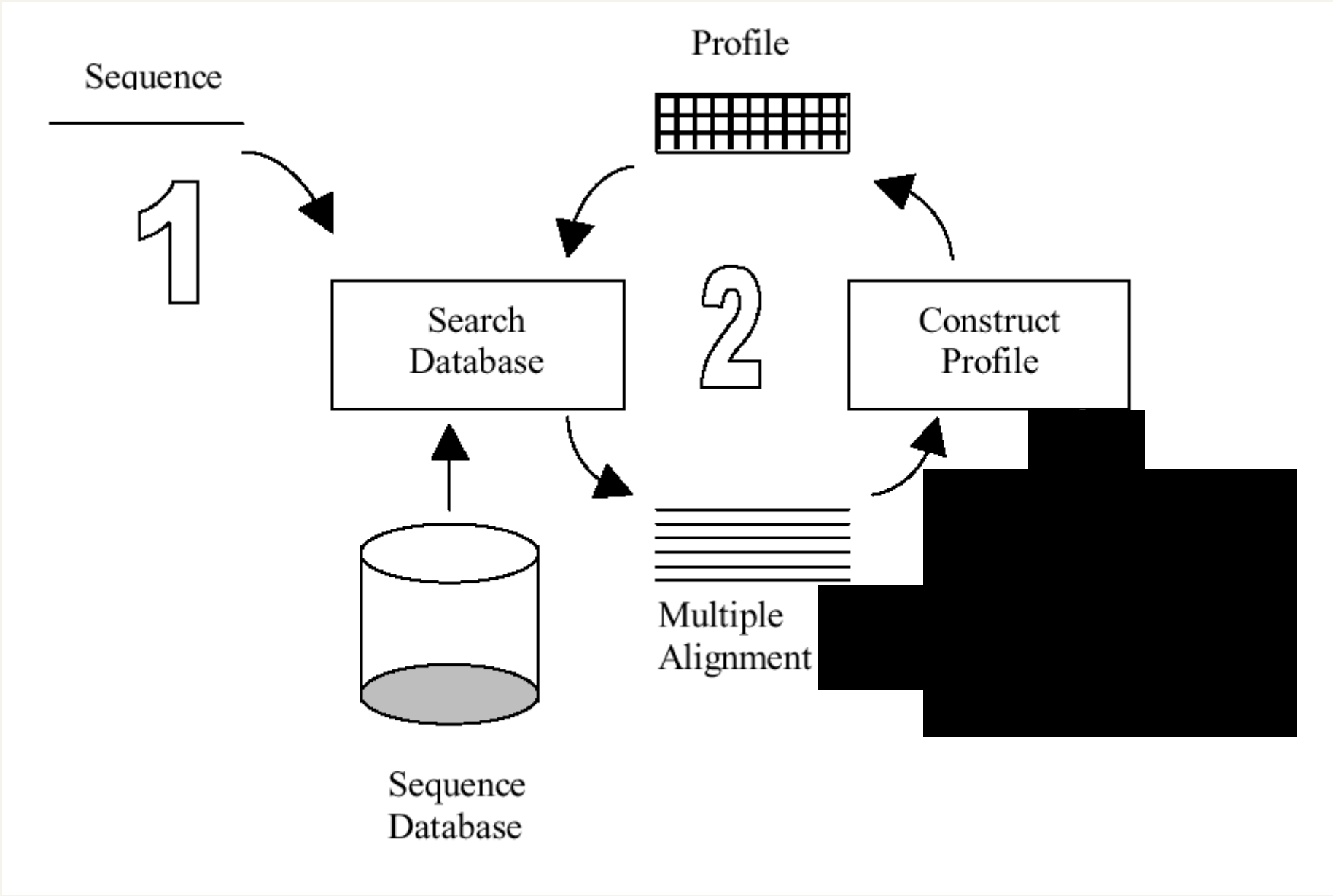
- Obtaining sequence information is easy; characterizing sequences is hard.
- Organisms share a common basis of genes and pathways.
- Information can be predicted for a novel sequence based on sequence similarity:
  - Function
  - Cellular role
  - Structure

Evaluation:
China Medical Tribune

- ▪List of 452 biomedical abbreviations with expansions
- ▪One model randomly picked from converged subset.
- ▪Evaluation of precision: Test algorithm on set of 452
- ▪Evaluation of recall: Run algorithm on medline

PSI-BLAST

- Used to detect protein sequence homology. (Iterated version of universally used BLAST program.)
- Searches a database for sequences with high sequence similarity to a query sequence.
- Creates a profile from similar sequences and iterates the search to improve sensitivity.

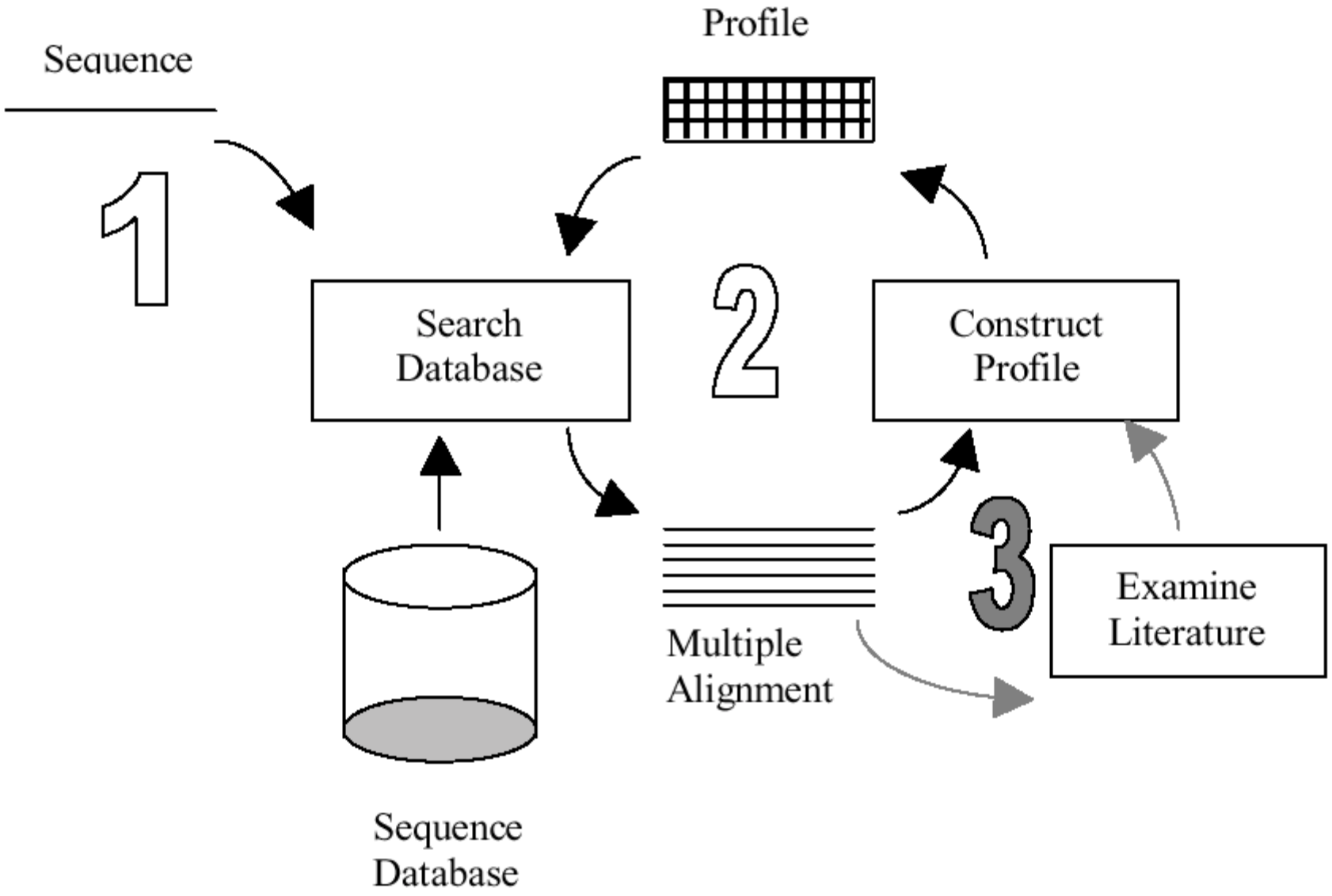

Profile
Sequence
1
Search
Database
2
Construct
Profile
Multiple
Alignment
Sequence
Database

PSI-BLAST Problem: Profile Drift

- At each iteration, could find non-homologous (false positive) proteins.
- False positives create a poor profile, leading to more false positives.

Addressing Profile Drift

- **PROBLEM**: Sequence similarity is only one indicator of homology.
  - More clues, e.g. protein functional role, exists in the literature.
- **SOLUTION**: we incorporate MEDLINE text into PSI-BLAST.


Sequence
Profile
1
Search Database
2
Construct Profile
3
Examine Literature
Multiple Alignment
Sequence Database

Modification to PSI-BLAST

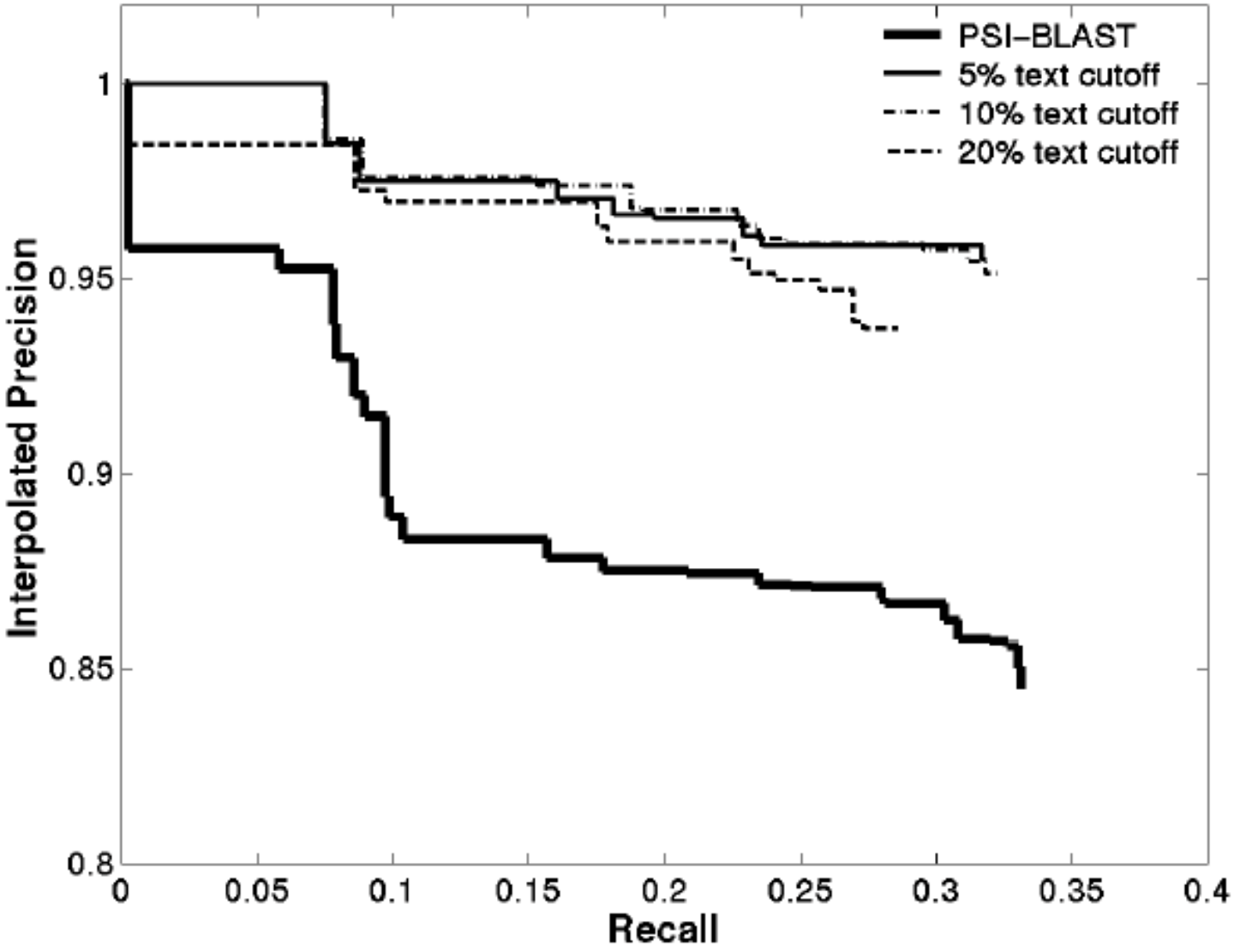
- *Before including a sequence, measure similarity of literature. Throw away sequences with least similar literatures to avoid drift.*
- Literature obtained from SWISS-PROT gene annotations to MEDLINE (text, keywords).
- Define domain-specific "stop" words (< 3 sequences or >85,000 sequences) = 80,479 out of 147,639.
- Use similarity metric between literatures (for genes) based on word vector cosine.

Evaluation

- Created families of homologous proteins based on SCOP (gold standard site for homologous proteins-- http://scop.berkeley.edu/)
- Select one sequence per protein family:
  - Families must have >= five members
  - Associated with at least four references
  - Select sequence with worst performance on a non-iterated BLAST search

Evaluation

- Compared homology search results from original and our modified PSI-BLAST.
- Dropped lowest 5%, 10% and 20% of literature-similar genes during PSI-BLAST iterations


PSI–BLAST
5% text cutoff
10% text cutoff
20% text cutoff
Interpolated Precision
1
0.95
0.9
0.85
0.8
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Recall

Results

- 46/54 families had identical performance
- 2 families suffered from PSI-BLAST drift, avoided with text-PSI-BLAST.
- 3 families did not converge for PSI-BLAST, but converged well with text-PSI-BLAST
- 2 families converged for both, with slightly better performance by regular PSI-BLAST.

				Convergence		Precision		Recall	
Superfamily	**Query Sequence**	**Words**	**# Seqs**	**PSI-BLAST**	**Text 10%**	**PSI-BLAST**	**Text 10%**	**PSI-BLAST**	**Text 10%**
EGF/Laminin	C1R_HUMAN	1661	5	yes	no	0.11	N/A	0.8	N/A
Acid proteases	POL_HV2RO	1271	22	yes	no	0.6	N/A	0.27	N/A
PLP-dependent transferases	GLYC_RABIT	1052	21	no	yes	N/A	1	N/A	0.1
Thioredoxin-like	CAQS_RABIT	1516	13	no	yes	N/A	1	N/A	0.38
Glucocorticoid receptor-like (DNA-binding domain)	CYSR_CHICK	1738	10	no	yes	N/A	0.8	N/A	0.4
EF-hand	SCP_NERDI	963	31	yes	yes	0.92	0.92	0.74	0.71
Glycosyl-transferases	CHLY_HEVBR	1007	20	yes	yes	1	1	0.2	0.15
Snake toxin-like	CD59_HUMAN	2435	23	yes	yes	1	1	0.13	0.09

Discussion

- Profile drift is rare in this test set and can sometimes be alleviated when it occurs.
- Overall PSI-BLAST precision can be increased using text information.

Resources

- http://www.smi.stanford.edu/projects/helix/psb01/chang.pdf
- Pac Symp Biocomput. 2001;:374-83. PMID: 11262956
- Blast: http://www.ncbi.nlm.nih.gov/BLAST/
- http://abbreviation.stanford.edu
- http://citeseer.nj.nec.com/chang02creating.html, J Am Med Inform Assoc 2002 Nov-Dec;9(6):612-20, Creating an online dictionary of abbreviations from MEDLINE, Chang JT, Schutze H, Altman RB.

- Medinfo 2001;10(Pt 1):371-5 Automatic extraction of acronym-meaning pairs from MEDLINE databases. Pustejovsky J, Castano J, Cochran B, Kotecki M,Morrell M.

- Pac Symp Biocomput 2003;:451-62 A simple algorithm for identifying abbreviation definitions in biomedical text. Schwartz AS, Hearst MA.
- http://www.hpl.hp.com/shl/people/eytan/srad.html