Sequence Comparison

Mike Cherry

Genomics

Genetics 211 - Winter 2014

Goals of Sequence Comparison:

- Find similarity such that an inference of homology is justified.
  - Similarity = observed with sequence alignment
  - Homology = shared evolutionary history (ancestry)
- Find a new sequence (gene) of interest
- Provide biologically appropriate results.
  - Substitutions, insertions and deletions
- Compare as many sequences as fast as possible.

Local vs. Global Alignment

- Global Alignment

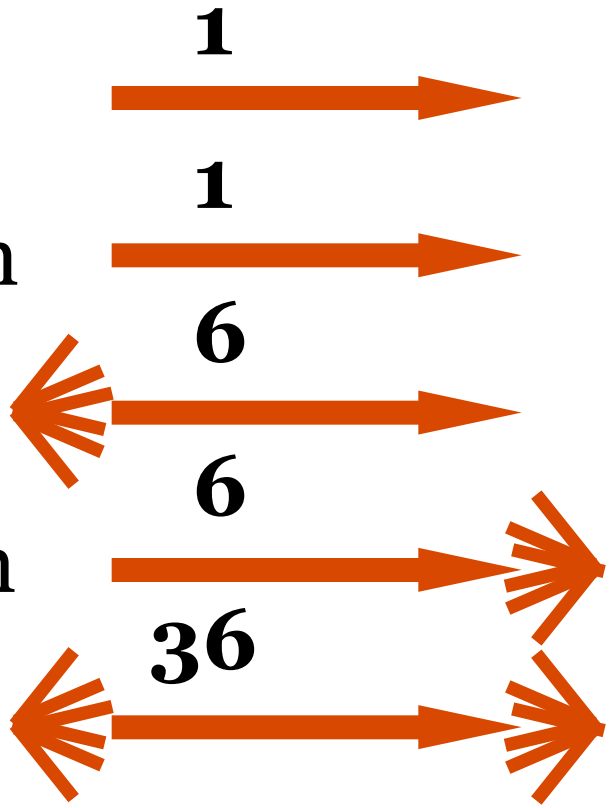
```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || | | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

- Local Alignment—better alignment to find conserved segment

```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Five flavors of BLAST

Program	Query		DB type
BLASTN	DNA	1	DNA
BLASTP	protein	1	protein
BLASTX	DNA	6	protein
TBLASTN	protein	6	DNA
TBLASTX	DNA	36	DNA

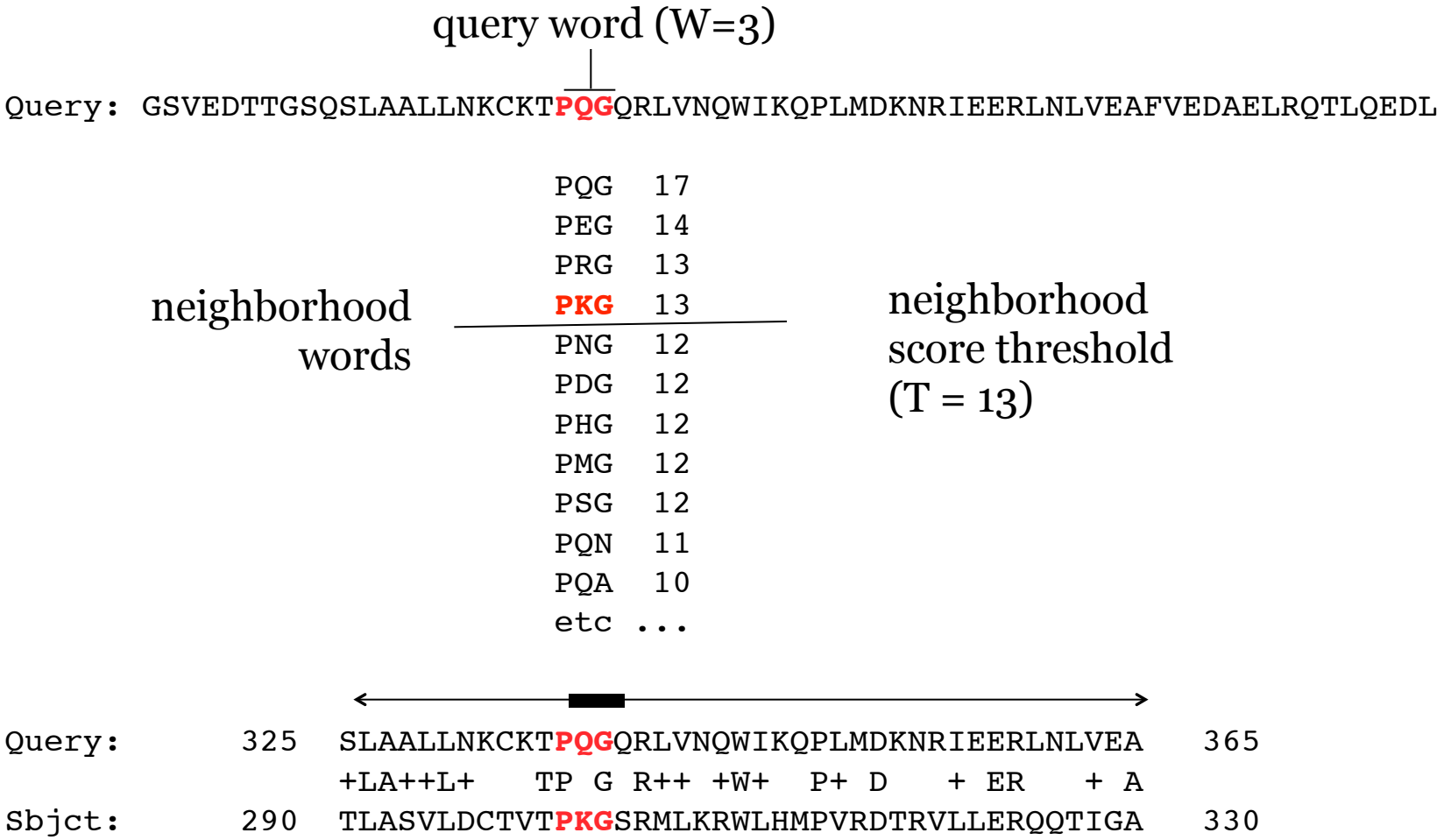


Objective of BLAST

"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length W with a score of at least T."

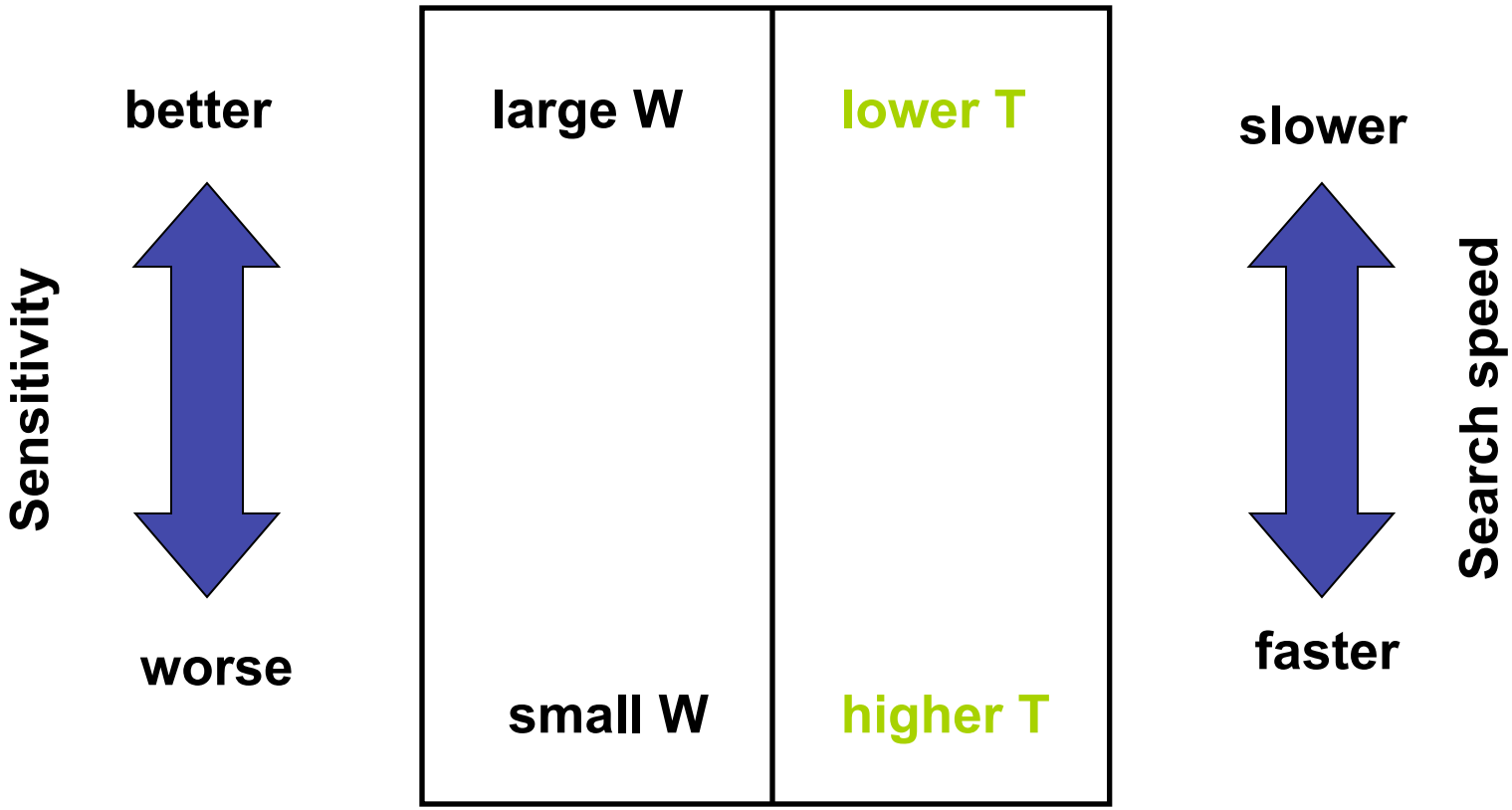
Altschul, Gish, et al. (1990)

The BLAST Search Algorithm



High-scoring Segment Pair (HSP)

(from NCBI Web site)



For proteins, default word size is 3.
(This yields a more accurate result than 2.)

Raw Scores (*S* values) from an Alignment

$$S = (\Sigma M_{ij}) - cO - dG,$$

where

M = score from a similarity matrix for a particular pair of amino acids (ij)

c = number of gaps

O = penalty for the existence of a gap

d = total length of gaps

G = per-residue penalty for extending the gap

Kerfeld and Scott, PLoS Biology 2011

Dynamic Programming Basics for sequence alignment. Smith-Waterman method.

		A	C
			←
A		1	1-1=0
	↑	↖	
C	1-1=0	1+1=2	

Use scores to complete the matrix, row by row. Add, or subtract, from neighboring cell with the highest score using this order:

1) diagonal ↖

2) up ↑

3) right ←

Scoring for nucleotides:

Match = 1

Gap = -1

Mismatch = -1

	T	A	C	T	A	A	C	G	C
A		1			1	1			
C			2	1			2	1	1
A		1	1	1	2	1	1	1	
C			2	1	1	1	2	1	
G				1			1	3	2
C			1					2	4
T	1			2	1			1	3

Match = 1
Gap = -1
Mismatch = -1

Local Alignment
Smith & Waterman

```
TACTAACGC
||  |  |||
AC-A-CGCT
```

BLAST Parameters and Constants

S = raw score (scoring matrix derived)
S' = bit score
E = chance of finding zero HSPs with score >= S
λ = constant based on scoring matrix
K = constant based on gap penalty
n = effective length of database
m = effective length of query

BLAST Scoring System

Raw score (S): Sum of scores for each aligned position and scores for gaps

$S = \lambda(\text{matches}) - \lambda(\text{mismatches}) - \lambda(\text{gap penalties})$

note: this score varies with the scoring matrix used and thus may not be meaningfully compared for different searches

Bit score (S'): Version of the raw score that is normalized by the scale of the scoring matrix (λ) and the scale of the gap penalty (K)

$S' = (\lambda S - \ln(K)) / \ln(2)$

note: because it is normalized the bit score can be meaningfully compared across searches

E value: Number of alignments with bit score S' or better that one would expect to find by chance in a search of a database of the same size

$E = mn2^{-S'}$

m = effective length of database

n = effective length of query sequence

note: E values may change if databases of different sizes are searched

BLAST output

```
>gi|14164377|dbj|BAB55676.1| (AB042828) Type II membrane protein of ER~mouse gene similar to
          alpha-mannosidase [Mus musculus]
          Length = 652

 Score =  177 bits (449), Expect = 2e-44
 Identities = 173/546 (31%), Positives = 250/546 (45%), Gaps = 96/546 (17%)

Query: 179 PEQGTELPPRRAEVPTKPPLPPARTQGTPVHLNY--------------RQKGVI------ 218
           P +GTE    R E P +P   P      P H  Y              R  G
Sbjct: 67  PRRGTE---GRLETPPEPGPTPGPGVCGPAHWGYALGGGGCGPDEYERRYSGAFPPQLRA 123
```

S’ S E

```
  Database: All non-redundant GenBank CDS
  translations+PDB+SwissProt+PIR+PRF
    Posted date:  Jan 31, 2002 12:59 AM
  Number of letters in database: 270,667,070
  Number of sequences in database:  0

Lambda     K      H
   0.319    0.137    0.420

Gapped
Lambda     K      H
   0.267    0.0410    0.140
```

λ K

```
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 26,665,471
Number of Sequences: 0
Number of extensions: 1297690
Number of successful extensions: 5012
Number of sequences better than 1.0e-05: 6
Number of HSP's better than  0.0 without gapping: 6
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 4968
Number of HSP's gapped (non-prelim): 10
length of query: 663
length of database: 18,444,546
effective HSP length: 108
effective length of query: 555
effective length of database: 12,037,230
effective search space: 6680662650
effective search space used: 6680662650
```

n

m

E-value or P-values

- NCBI-BLAST reports E-values because of the ease of understanding the difference between 5 and 10. The respective P-values would be 0.993 and 0.99995.

$$P = 1 - e^{-E}$$

- When E < 0.01, P-values and E-value are nearly identical.

E values or *p* values

Very small *E* values are very similar to *p* values.
E values of about 1 to 10 are far easier to interpret than corresponding *p* values.

E	*p*
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258 (about 0.1)
0.05	0.04877058 (about 0.05)
0.001	0.00099950 (about 0.001)
0.0001	0.0001000

Sometimes a real match has an E value > 0.5

Sequences producing significant alignments:	Score (bits)	E Value
gi|5803139|ref|NP_006735.1| retinol-binding protein 4, inte...	378	e-105
gi|230284|pdb|1RBP| Retinol Binding Protein >gi|493897|p...	371	e-103
gi|88364|pir||A27786 plasma retinol-binding protein - human	370	e-103
gi|4558179|pdb|1QAB|E Chain E, The Structure Of Human Retin...	363	e-100
gi|7770173|gb|AAF69622.1|AF119917_30 (AF119868) PRO2222 [Ho...	324	5e-89
gi|13645517|ref|XP_005907.2| retinol-binding protein 4, int...	233	9e-62
gi|296672|emb|CAA26553.1| (X02775) RBP [Homo sapiens]	207	8e-54
gi|5419892|emb|CAB46489.1| (X02824) RBP (aa 101-172) [Homo ...	149	2e-36
gi|2895204|gb|AAC02945.1| (AF025334) mutant retinol binding...	90	2e-18
gi|2895206|gb|AAC02946.1| (AF025335) mutant retinol binding...	73	2e-13
gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Hom...	55	4e-08
gi|619383|gb|AAB32200.1| apolipoprotein D, apoD [human, pla...	55	5e-08
gi|1246096|gb|AAB35919.1| (S80440) apolipoprotein D, apoD (...	43	3e-04
gi|223373|prf||0801163A complex-forming glycoprotein HC [Ho...	37	0.011
gi|4884164|emb|CAB43305.1| (AL050169) hypothetical protein ...	35	0.043
gi|13639329|ref|XP_005360.3| 61620 [Homo sapiens] >gi|13639...	35	0.043
gi|4502067|ref|NP_001624.1| alpha-1-microglobulin/bikunin p...	35	0.068
gi|14735821|ref|XP_029964.1| progestagen-associated endomet...	35	0.070
gi|4557393|ref|NP_000597.1| complement component 8, gamma p...	34	0.14
gi|4505583|ref|NP_002562.1| progestagen-associated endometr...	32	0.49
gi|13639651|ref|XP_005430.2| complement component 8, gamma ...	31	1.1

… try a reciprocal BLAST to confirm

Assessing whether proteins are homologous

```
>gi|4505583|ref|NP_002562.1| progestagen-associated endometrial protein (pla
           pregnancy-associated endometrial alpha-2-globulin, alpha
           uterine protein); Progestagen-associated endometrial
           protein (placental protein 14) [Homo sapiens]
 gi|190215|gb|AAA60147.1| (J04129) placental protein 14 [Homo sapiens]
          Length = 162

 Score = 32.0 bits (71), Expect = 0.49
 Identities = 26/107 (24%), Positives = 48/107 (44%), Gaps = 11/107 (10%)

Query: 26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
           + K++ +  + +GTW++MA      + L   + A   V  T  +          +L+ W+
Sbjct: 5   QTKQDLELPKLAGTWHSMAMAT-NNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 63

Query: 85  -VCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTY 130
             C +       T +P KFK+ Y  VA       ++  ++DTDYD +
Sbjct: 64  NSCVEKKVLGEKTGNPKKFKINY-TVA-------NEATLLDTDYDNF 102
```

Retinol Binding Protein 4 vs Progestagen-associated endometrial protein:
Low bit score, E value 0.49, 24% identity ("twilight zone").
But they are indeed homologous.
Try a BLAST search with PAEP as a query, and find many other lipocalins.

Scoring Matrix

- Modeled Change in Protein Sequences
  - PAM (Accepted Point Mutations)
  - Schwartz & Dayhoff (1978)
- Experimentally Derived Matrix
  - BLOSUM (**BLO**CKS **Su**bstitution **M**atrix)
  - Henikoff & Henikoff (1992)

Accepted Point Mutations (PAM) or Percent Accepted Mutations

- Number of individual amino acid changes occurring per 100 aa residues as a result of evolution. PAM1 = unit of evolutionary divergence in which 1% of the amino acids have been changed.
- PAM of 250, or PAM250, represents $[PAM1]^{250}$. The PAM1 matrix multiplied against itself 250 times.

Creating the PAM1
Schwartz & Dayhoff (1978)

- Studied 34 protein super-families and grouped them into 71 phylogenetic trees. There were 1,572 changes observed. All sequences were at least 85% identical. Alignments were scanned with a 100 amino acid window.
- These are observed mutations thus the term accepted point mutations, accepted by natural selection and thus the dominant allele in the species.
- Normalized probability of change:

$$P_{ij} = (C_{ij} / T) \times (1 / F_i)$$

C_{ij} = number of changes from aa_i to aa_j

F_i = freq of aa_i in that group of sequences

T = total number of all aa changes in 100 sites

PAM250 log odds scoring matrix

A	2																		
R	-2	6																	
N	0	0	2																
D	0	-1	2	4															
C	-2	-4	-4	-5	12														
Q	0	1	1	2	-5	4													
E	0	-1	1	3	-5	2	4												
G	1	-3	0	1	-3	-1	0	5											
H	-1	2	2	1	-3	3	1	-2	6										
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5									
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6								
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5							
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6						
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9					
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6				
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2			
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3		
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

Protein family	**1 x 10**
Immunoglobulin (Ig) kappa chain	37
Kappa casein	33
Luteinizing hormone b	30
Lactalbumin	27
Complement component 3	27
Collagen	1.7
Troponon C, skeletal muscle	1.5
Alpha crystallin B chain	1.5
Glucagon	1.2
Glutamate dehydrogenase	0.9
Histone H2B, member Q	0.9
Ubiquitin	0

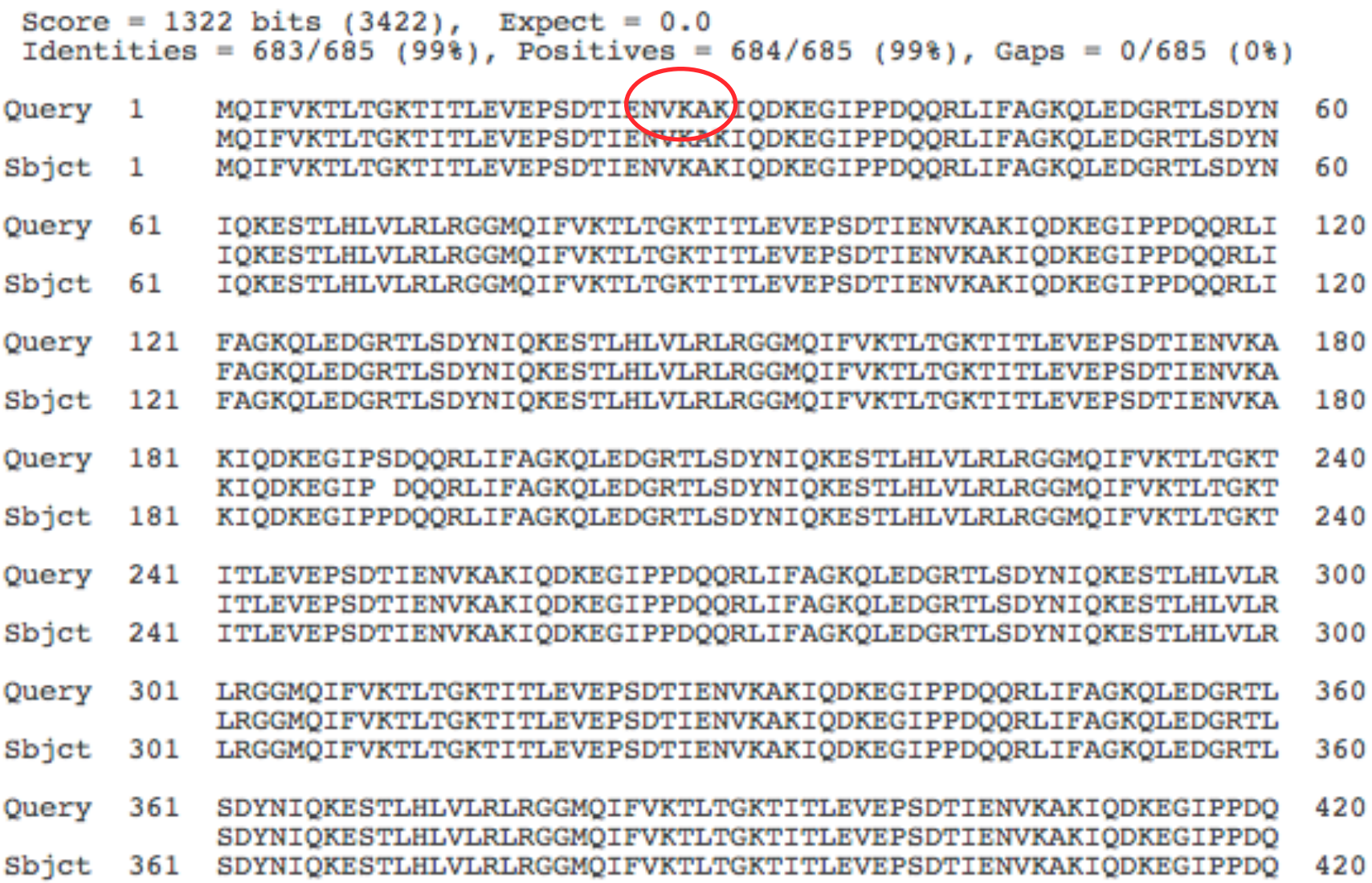
From Dayhoff (1978)

Kappa-casein, milk protein

cow (NC_007304) versus human (NP_005203)

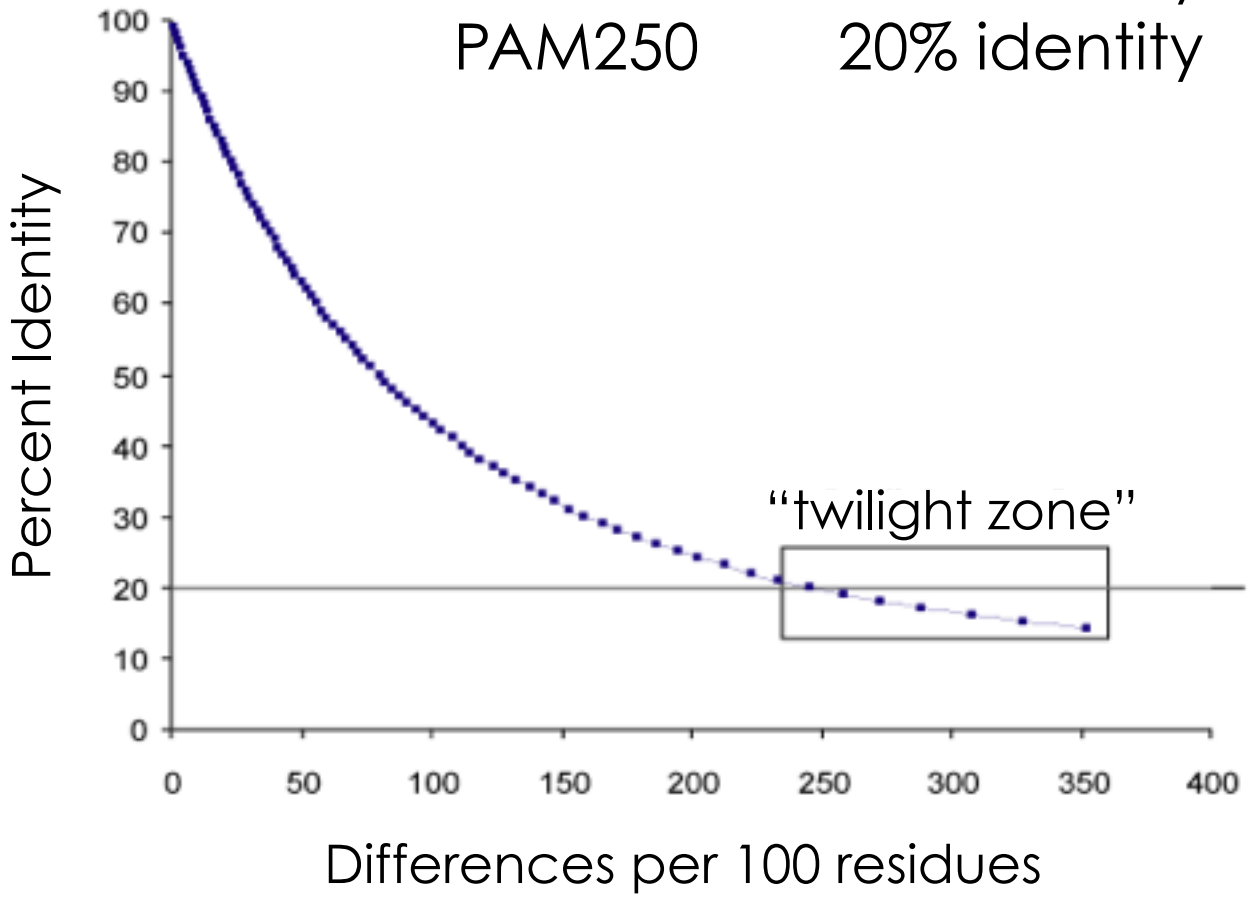
```
 Score =  174 bits (442),  Expect = 3e-44
 Identities = 99/186 (53%), Positives = 115/186 (61%), Gaps = 9/186 (4%)

Query  2    MKSFFLVVTILALTLPFLGAQEQNQEQPIRCEKDERFFSDKIAKYIPIQYVLSRYPSYGL  61
            MKSF LVV  LALTLPFL  + QNQ+QP   E DER F  K A Y+P+ YV + YP YG
Sbjct  1    MKSFLLVVNALALTLPFLAVEVQNQKQPACHENDERPFYQKTAPYVPMYYVPNSYPYYGT  60

Query  62   NYYQQKPVALINNQFLPYPYYAKPAAVRSPAQILQWQVLSNTVPAKSCQAQPTTMARHPH  121
            N YQ++P   INN ++P  YYA PA VR  AQI Q Q L N        + P T+ R P+
Sbjct  61   NLYQRRPAIAINNPYVPRTYYANPAVVRPHAQIPQRQYLPN--------SHPPTVVRRPN  112

Query  122  PHLSFMAIPPKKNQDKTEIPTINTIASGEPTSTPTTEAVESTVATLED-SPEVIESPPEI  180
             H SF+AIPPKK QDK  IPTINTIA+ EPT  P TE    +V T E  S  +I S PE
Sbjct  113  LHPSFIAIPPKKIQDKIIIPTINTIATVEPTPAPATEPTVDSVVTPEAFSESIITSTPET  172
```

human (NP_066289) vs mouse (NP_062613) ubiquitin C

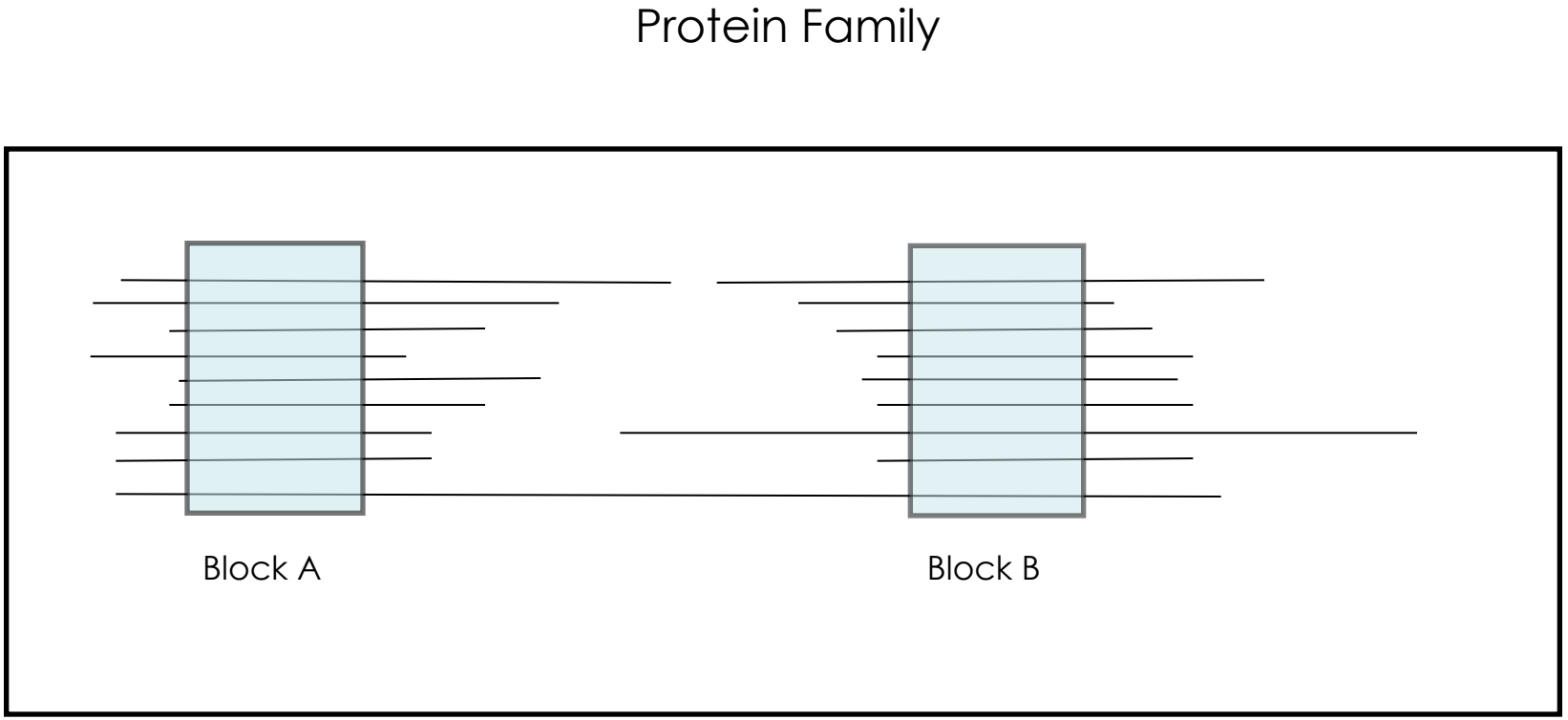


```
 Score = 1322 bits (3422),  Expect = 0.0
 Identities = 683/685 (99%), Positives = 684/685 (99%), Gaps = 0/685 (0%)

Query  1    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN  60
            MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN
Sbjct  1    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN  60

Query  61   IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI  120
            IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI
Sbjct  61   IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI  120

Query  121  FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA  180
            FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA
Sbjct  121  FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA  180

Query  181  KIQDKEGIPSDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT  240
            KIQDKEGIP DQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT
Sbjct  181  KIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT  240

Query  241  ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR  300
            ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR
Sbjct  241  ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR  300

Query  301  LRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTL  360
            LRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTL
Sbjct  301  LRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTL  360

Query  361  SDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQ  420
            SDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQ
Sbjct  361  SDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQ  420
```

PAM1	99% identity
PAM10.7	90% identity
PAM80	50% identity
PAM250	20% identity



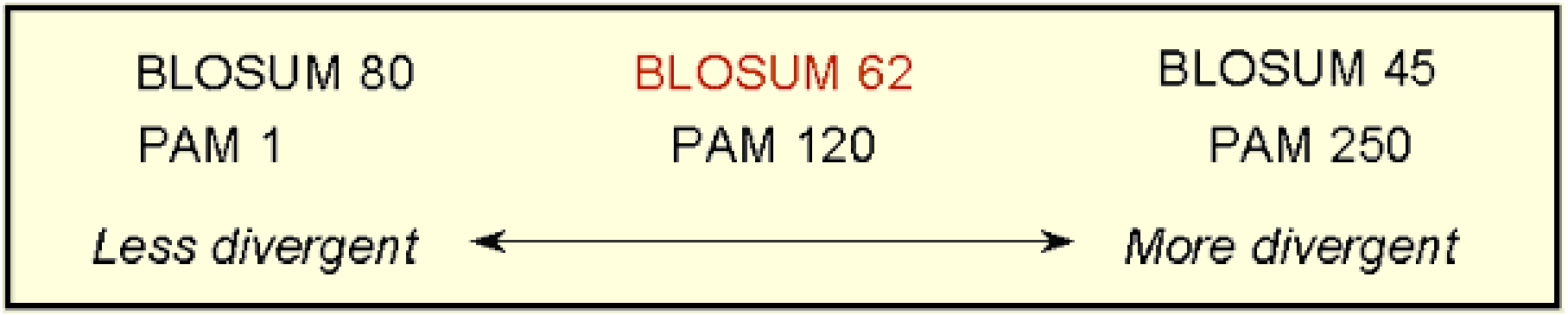
Deriving Substitution Scores BLOSUM Henikoff & Henikoff, 1992



BLOSUM 62 scoring matrix

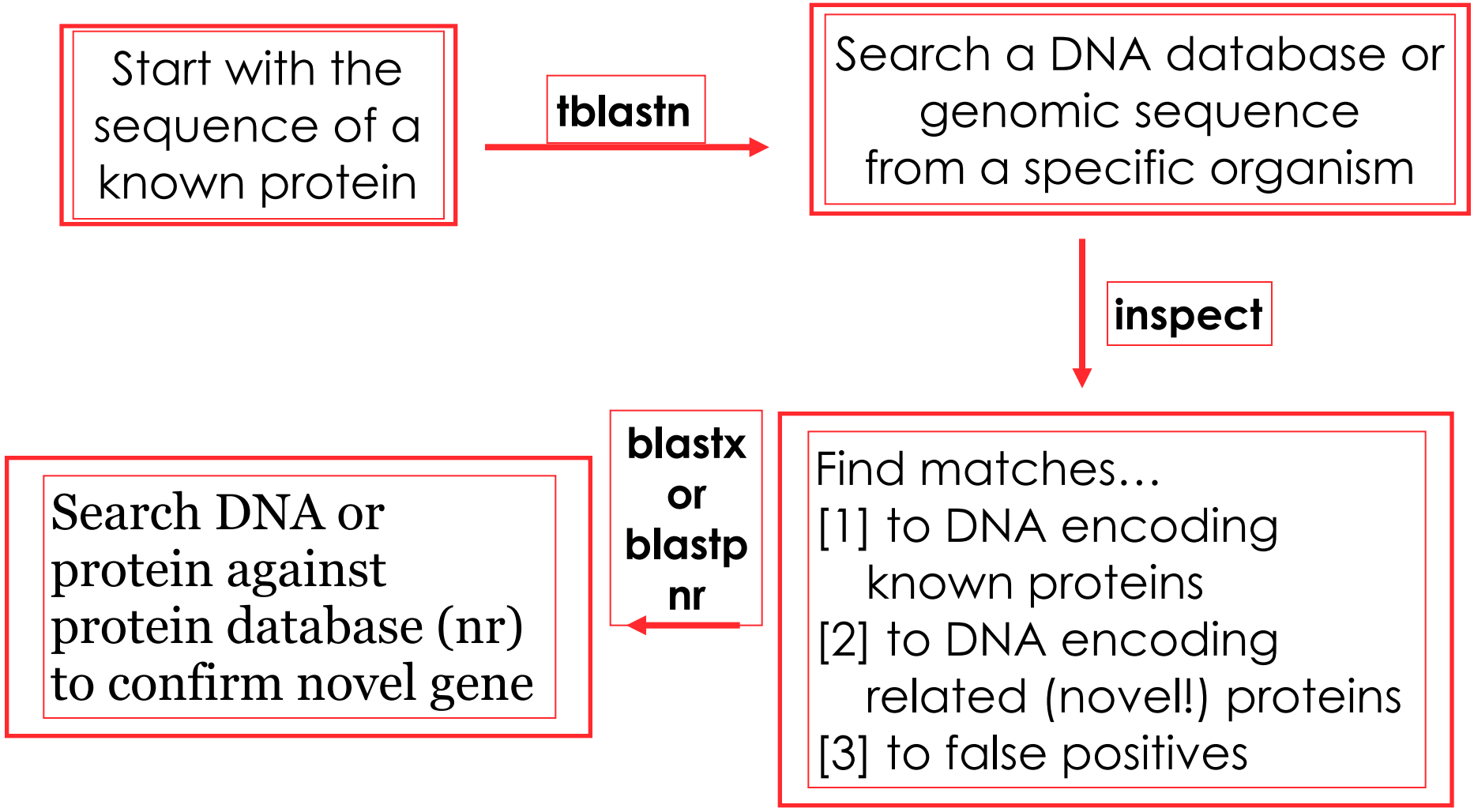
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Which Scoring Matrix should you use?



Depends on what you want to find?

Potential Flow for Sequence Identification



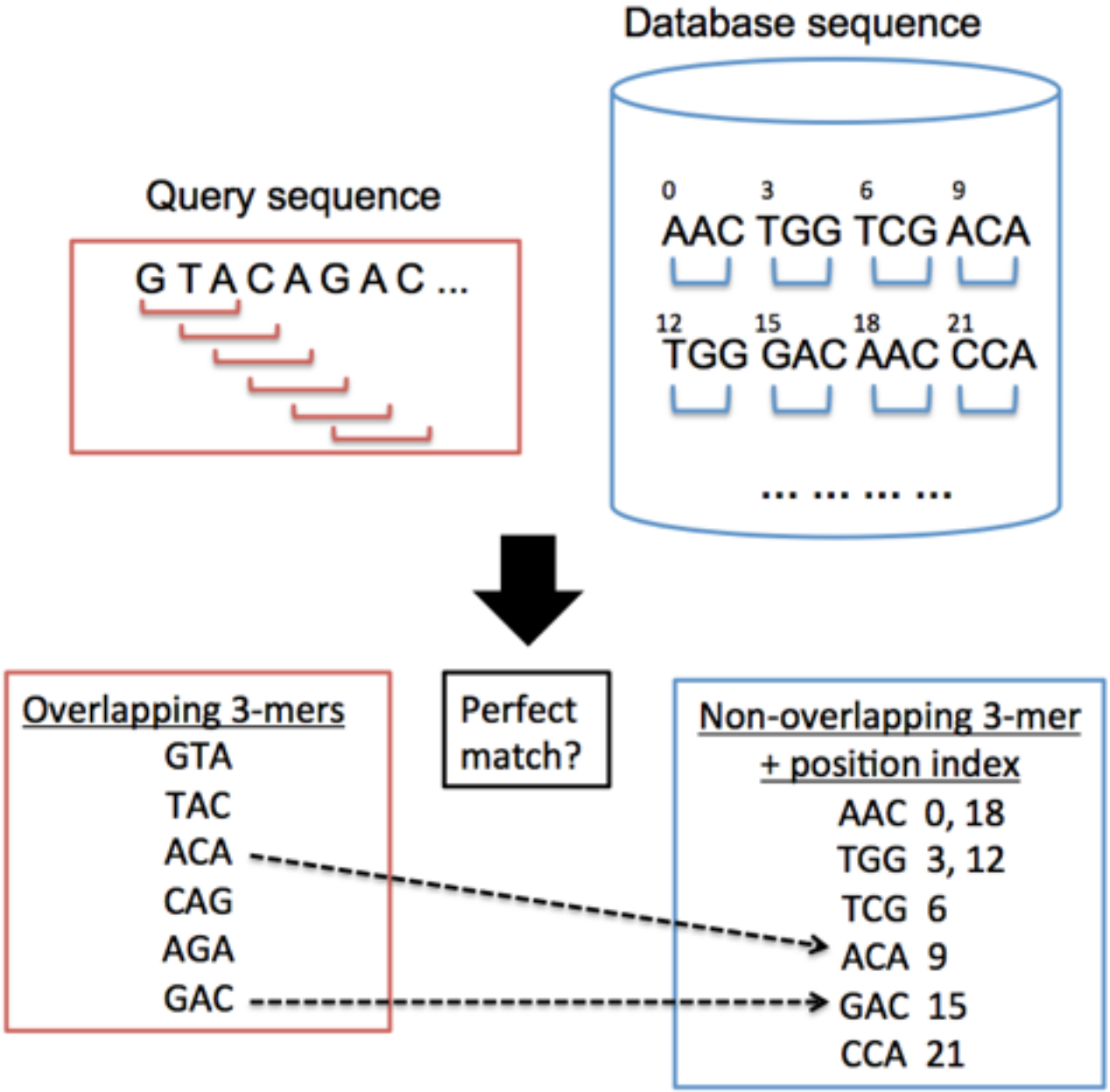
BLAT -- BLAST-Like Alignment Tool

By Jim Kent, UCSC

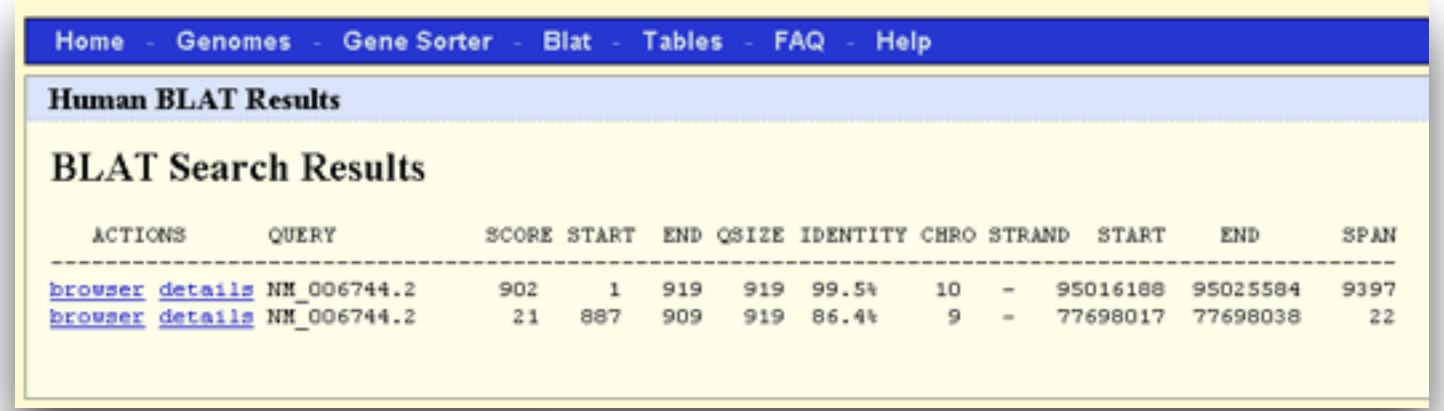
http://genome.ucsc.edu/cgi-bin/hgBlat

- BLAT is designed to find sequences of >95% similarity of length >40 bases. Perfect sequence matches of >33 bases are identified.
- Protein BLAT finds sequences of >80% similarity of length >20 amino acids.
- DNA BLAT works by keeping an index of the entire genome. The index consists of all non-overlapping 11-mers except for those in repeats.
- Protein BLAT works in a similar manner, except with 4-mers rather than 11-mers.
- The index is used to find areas of probable similarity. Then the sequence for the area of interest is read into memory for a detailed alignment.

BLAT Indexing



BLAT output includes browser and other formats



Home - Genomes - Gene Sorter - Blat - Tables - FAQ - Help

Human BLAT Results

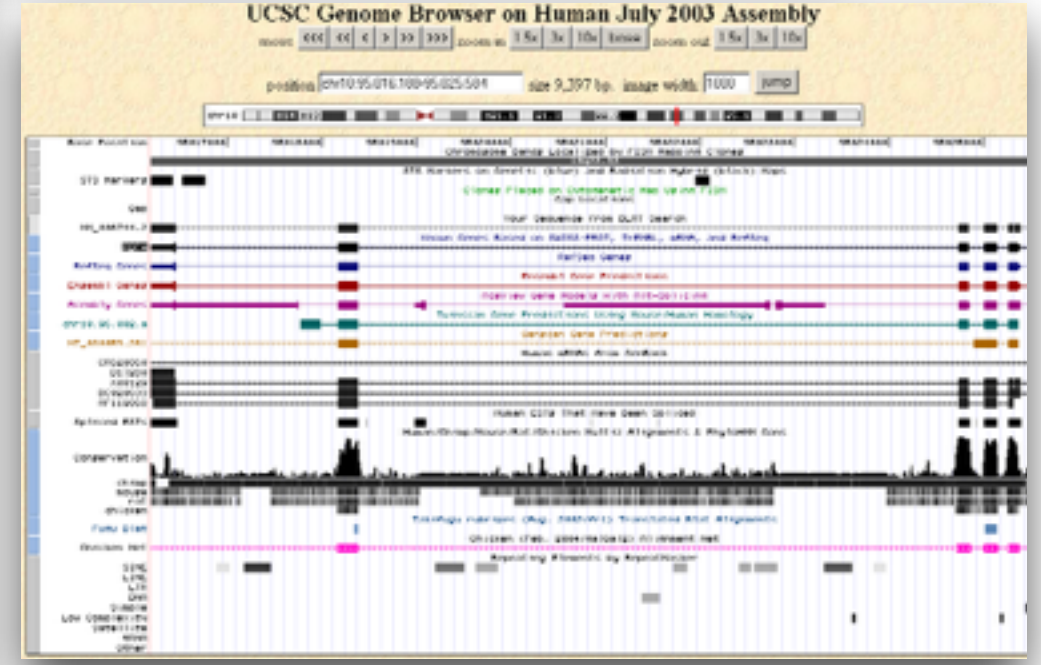
BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	NM_006744.2	902	1	919	919	99.5%	10	-	95016188	95025584	9397
browser details	NM_006744.2	21	887	909	919	86.4%	9	-	77698017	77698038	22

Alignment of NM_006744.2

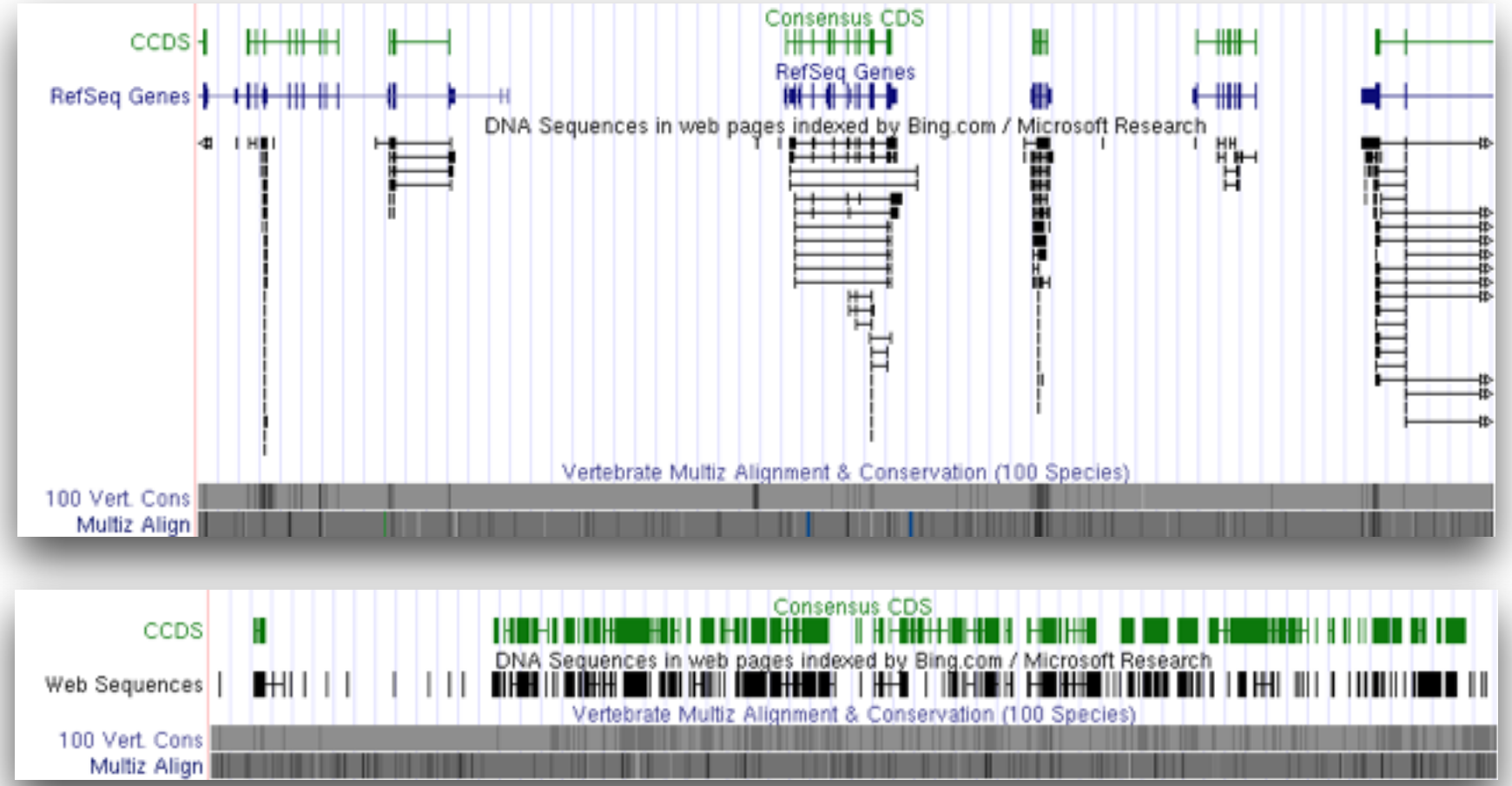
NM_006744.2
Human chr10
block1
block2
block3
block4
block5
block6
together

Side by Side Alignment



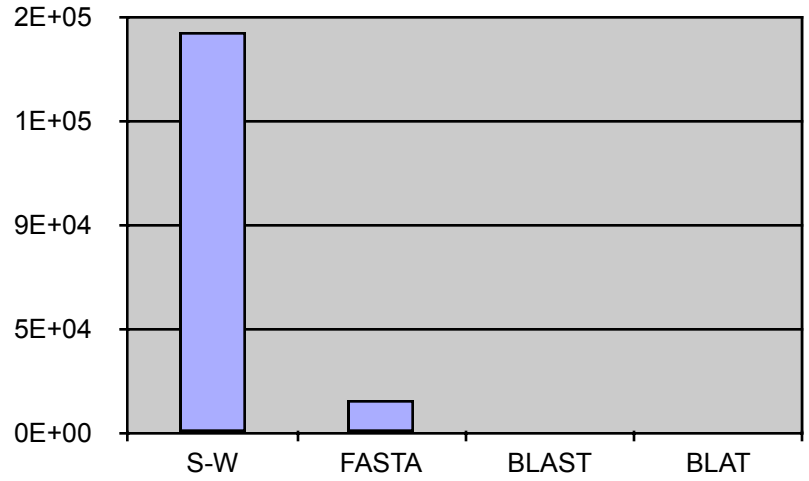
UCSC Genome Browser on Human July 2003 Assembly

Indexing all DNA Sequences on the web using BLAT. bing.com / Microsoft Research

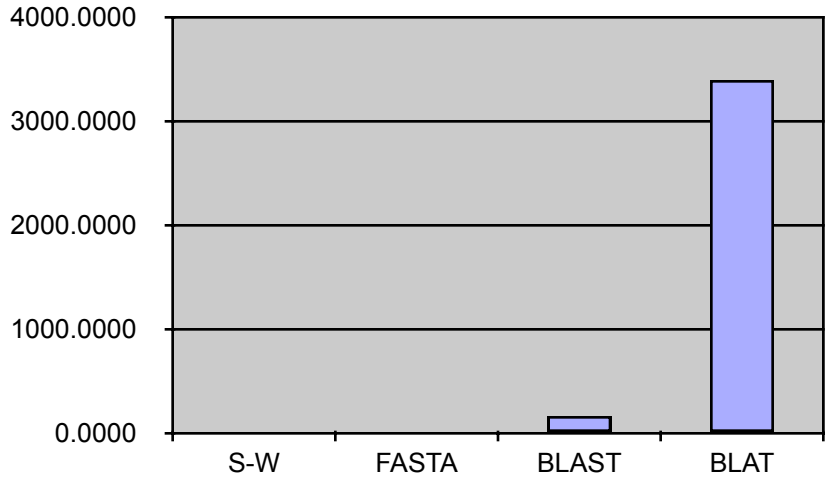


Relative Speed of Search Algorithm

Comparison of all *yeast* proteins by all *yeast* proteins (3 x 10^6 aa).



CPU Seconds for Search



Fold Faster than Smith-Waterman

Smith-Waterman	173,078 sec	(2 day)	1X
FASTA	14,223 sec	(4 hr)	12X
BLASTN	1,100 sec	(18 min)	157X
BLAT	51 sec	(51 sec)	3394X

UniProt (www.uniprot.org)

"United Protein Databases can be summarized as the creation, maintenance and provision of a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces."

Swiss Institute for Bioinformatics

European Bioinformatics Institute

Georgetown University

In part funded by the US NIH (NHGRI). Grant started Sept 2002.

NCBI's protein nr <===> UniRef100

Identical sequence of letters, from any organism, are combined into one entry.

UniRef90 & UniRef50

UniRef100 is processed and sequences that are 90% or 50% identical, respectively, are merged. These datasets are 40% and 65% smaller that UniRef100, respectively.

SwissPROT

Gold standard for protein sequences, extensively reviewed through manual curation

Metagenomics

The "great plate-count anomaly" : the vast majority of microbial cells that can be seen in a microscope and shown to be living with various staining procedures cannot be induced to produce colonies in cultures. It is estimated that only 0.1-1.0% of the living bacteria present in soils can be cultured under standard conditions; the culturable fraction of bacteria from aquatic environments is ten to a thousand times lower still.

Phylotyping gives some reliable information about "Who is there?" but because of within-species genomic diversity, only imperfect guesses as to "What are they doing?"

Phylotyping has been done using 16 S rRNA sequences for many used. One of the most common Genbank entry is 16S rRNA.

Environmental Communities

THE METAGENOMICS PROCESS

Extract all DNA from microbial community in sampled environment

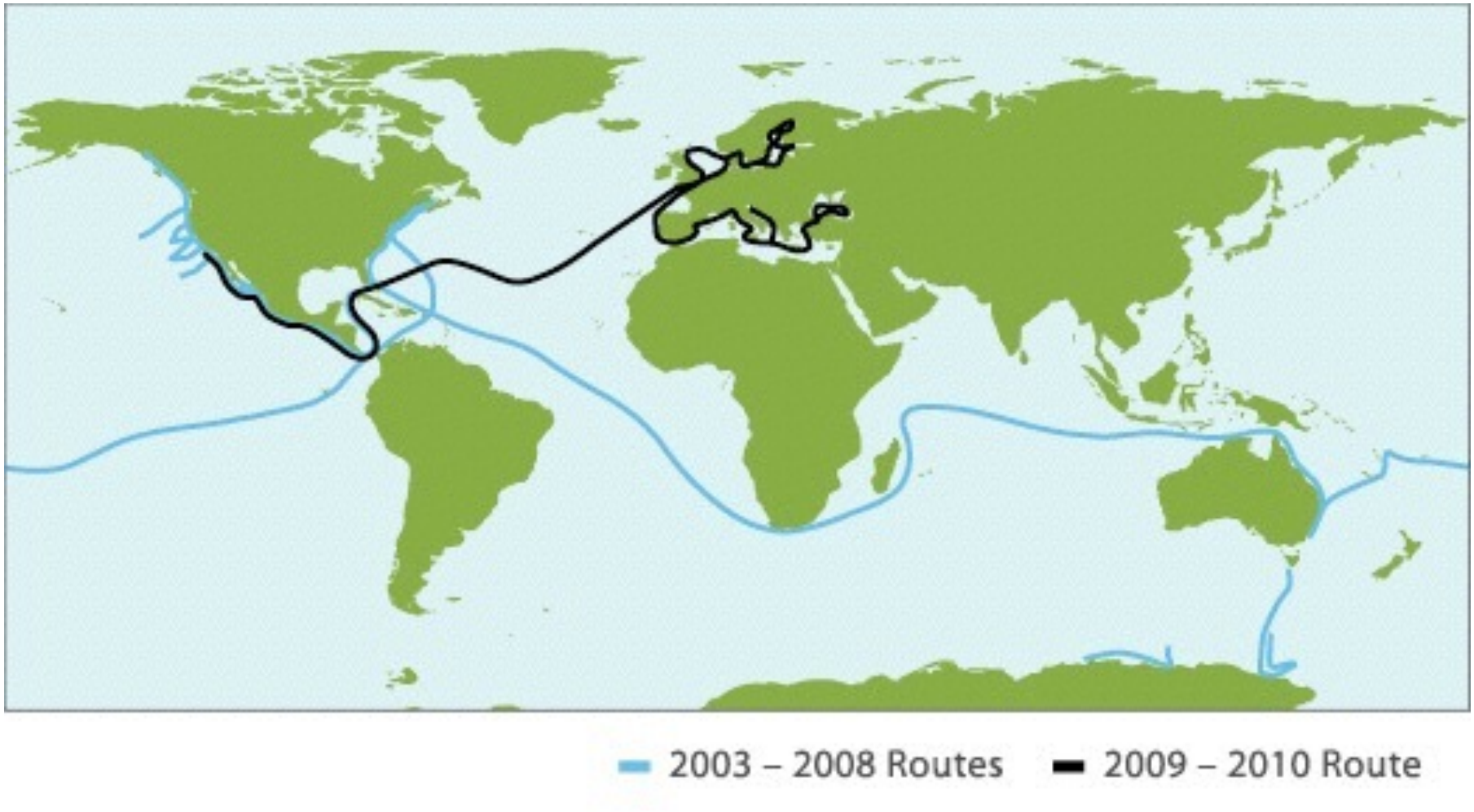
DETERMINE WHAT THE GENES ARE
(Sequence-based metagenomics)

- Identify genes and metabolic pathways
- Compare to other communities
- and more…

DETERMINE WHAT THE GENES DO
(Function-based metagenomics)

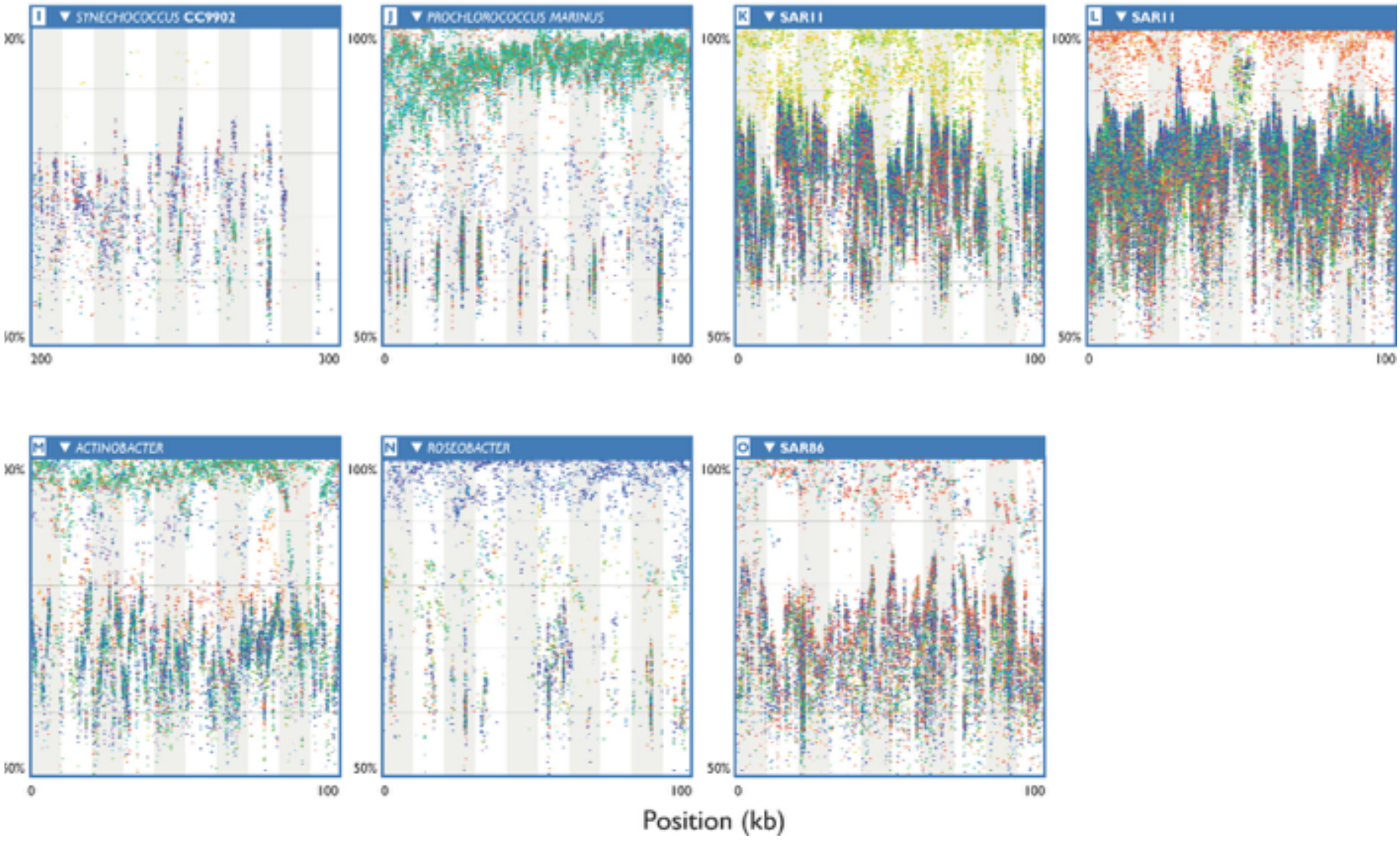
- Screen to identify functions of interest, such as vitamin or antibiotic production
- Find the genes that code for functions of interest
- and more…

Global Ocean Survey - - JCVI

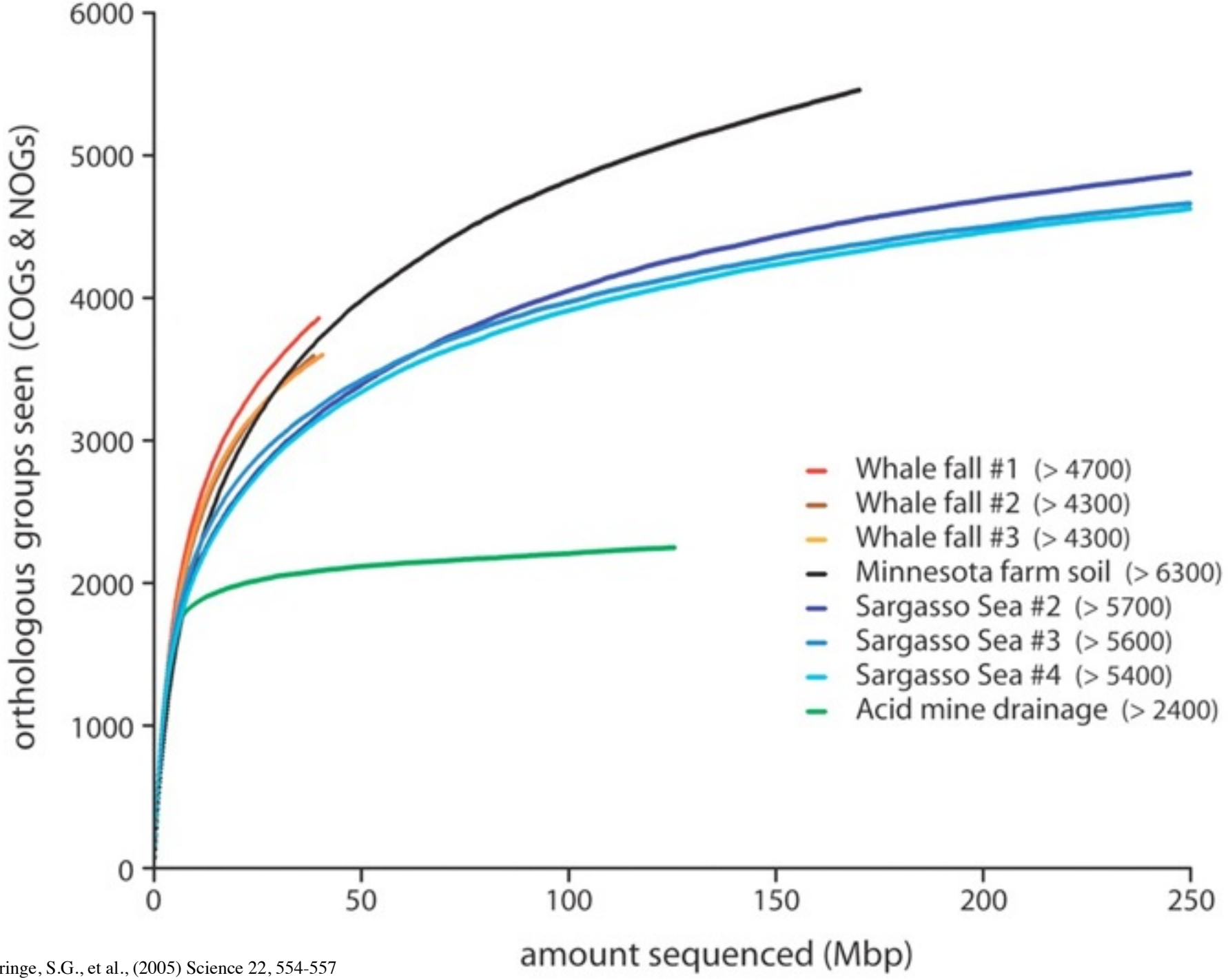


Kunin, V., *et al*., Micro. and Mol. Biol. Rev., (2008) **72**, 557-578

Comparison to alternate reference sequences

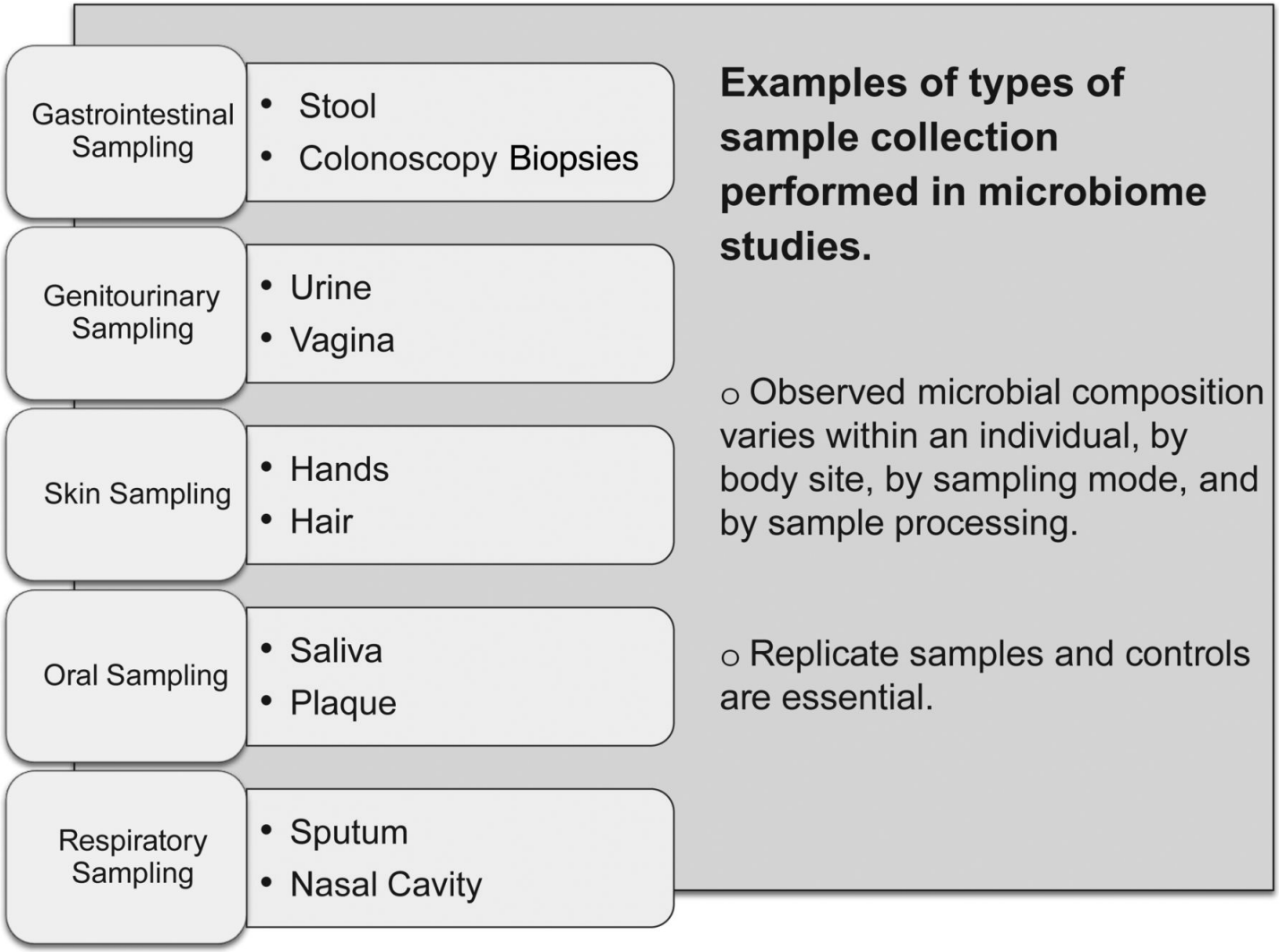


Rusch, D. et al. 2007 PLoS Biology



Tringe, S.G., et al., (2005) Science 22, 554-557

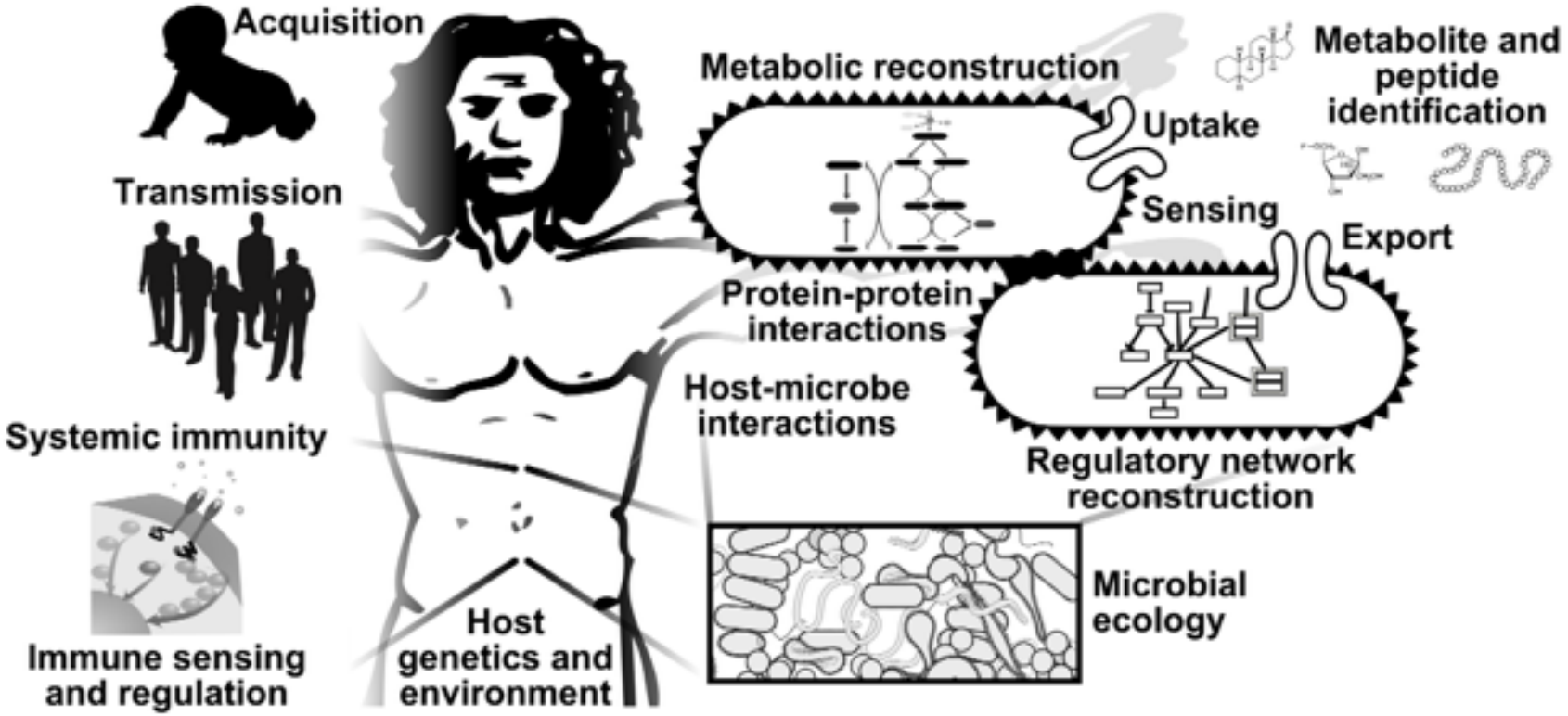
Human microbiome sampling examples.



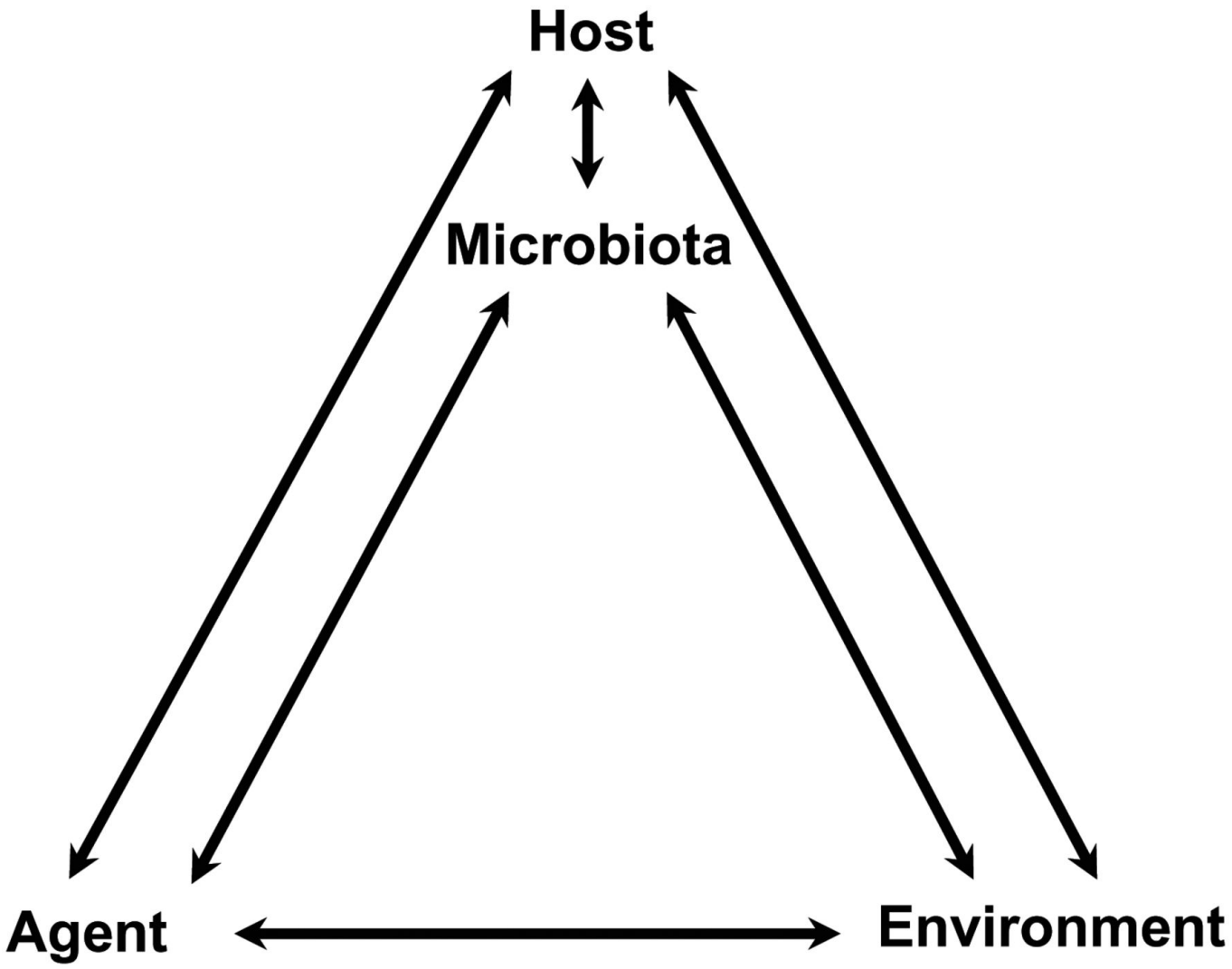
Foxman B , and Rosenthal M Am. J. Epidemiol. 2013;177:197-201

American Journal of EPIDEMIOLOGY

Human Microbiome Project

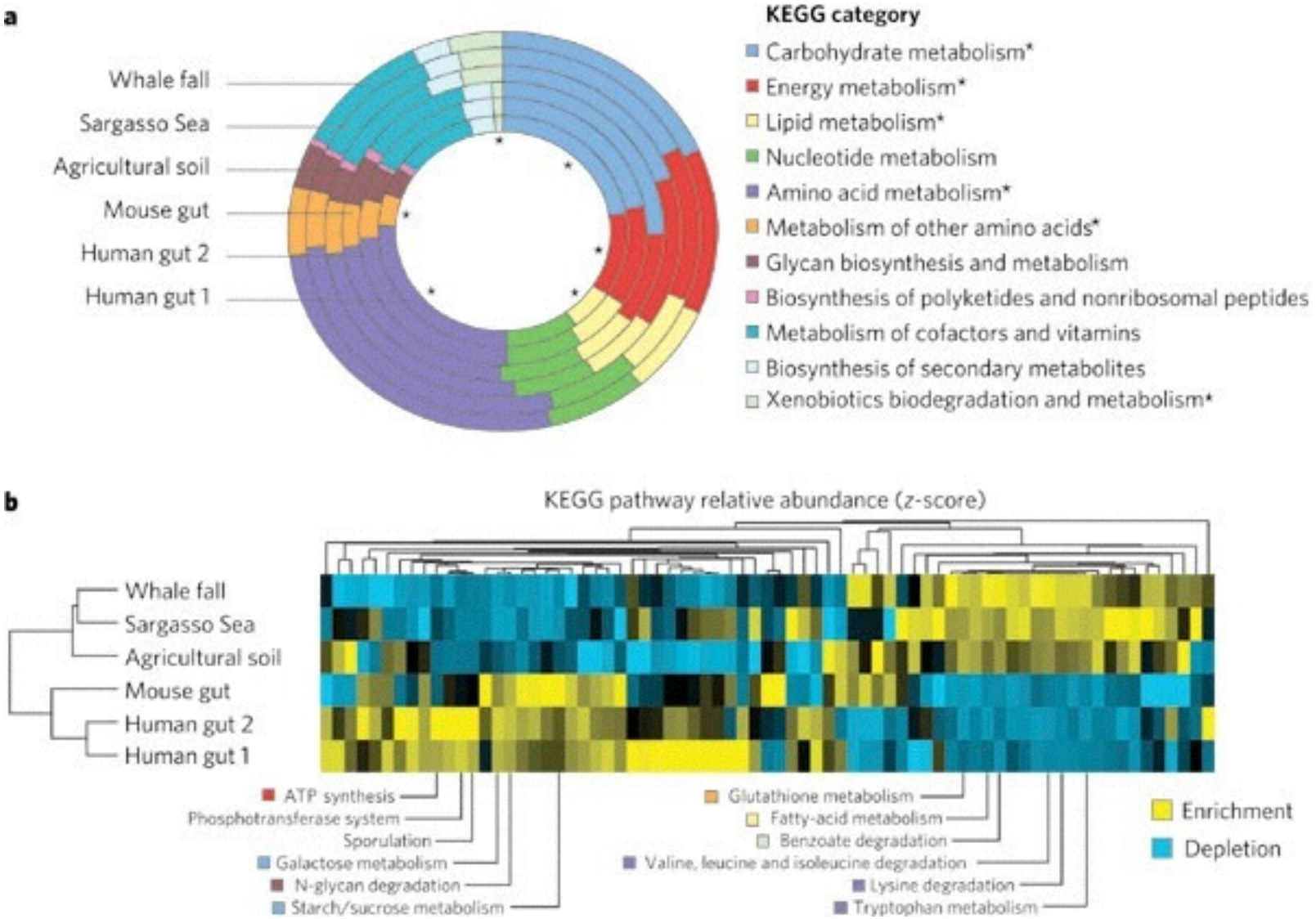


Microbiota can modify the effects of agent and environment on the host and indicate host changes in response to agent and environmental exposures.



Foxman B , and Rosenthal M Am. J. Epidemiol. 2013;177:197-201

American Journal of
EPIDEMIOLOGY



Turnbaugh, P.J., et al., (2007) Nature 449, 804-810

Mammoth Metagenomics

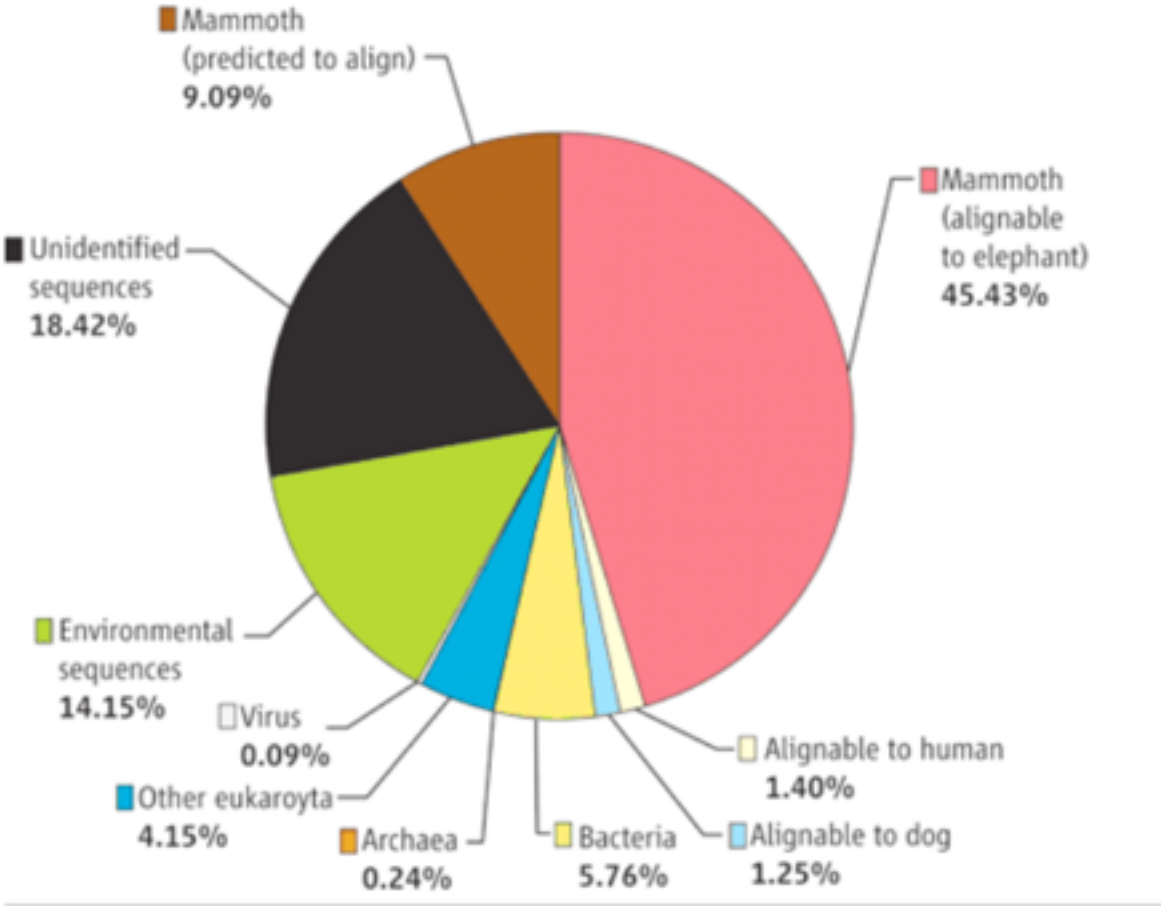
reads from mammoth tooth sample aligned to three genome sequences

	Elephant	Human	Dog
Total no. reads	302,692 (100%)	302,692 (100%)	302,692 (100%)
Aligned reads	137,527 (45.4%)	4,237 (1.4%)	3,775 (1.2%)
Uniquely aligning reads	44,442 (14.7%)	3,901 (1.3%)	3,548 (1.2%)
Multiply aligned reads	93,085 (30.8%)	336 (0.1%)	227 (0.1%)
Reads with at least 95% identity	90,507 (30.0%)	1,184 (0.4%)	1,140 (0.4%)
Reads with 100% identity	21,952 (7.3%)	116 (0.04%)	142 (0.05%)
Uniquely aligning base pairs	4,332,350	318,966	291,714
Identity in unique alignments	98.55%	92.68%	92.91%
Mitochondrial reads	209	-	-
Identity in mitochondrial reads	95.93%	-	-
Mitochondrial base pairs	16,419	-	-

Poinar et al (2006) Science 311:5759

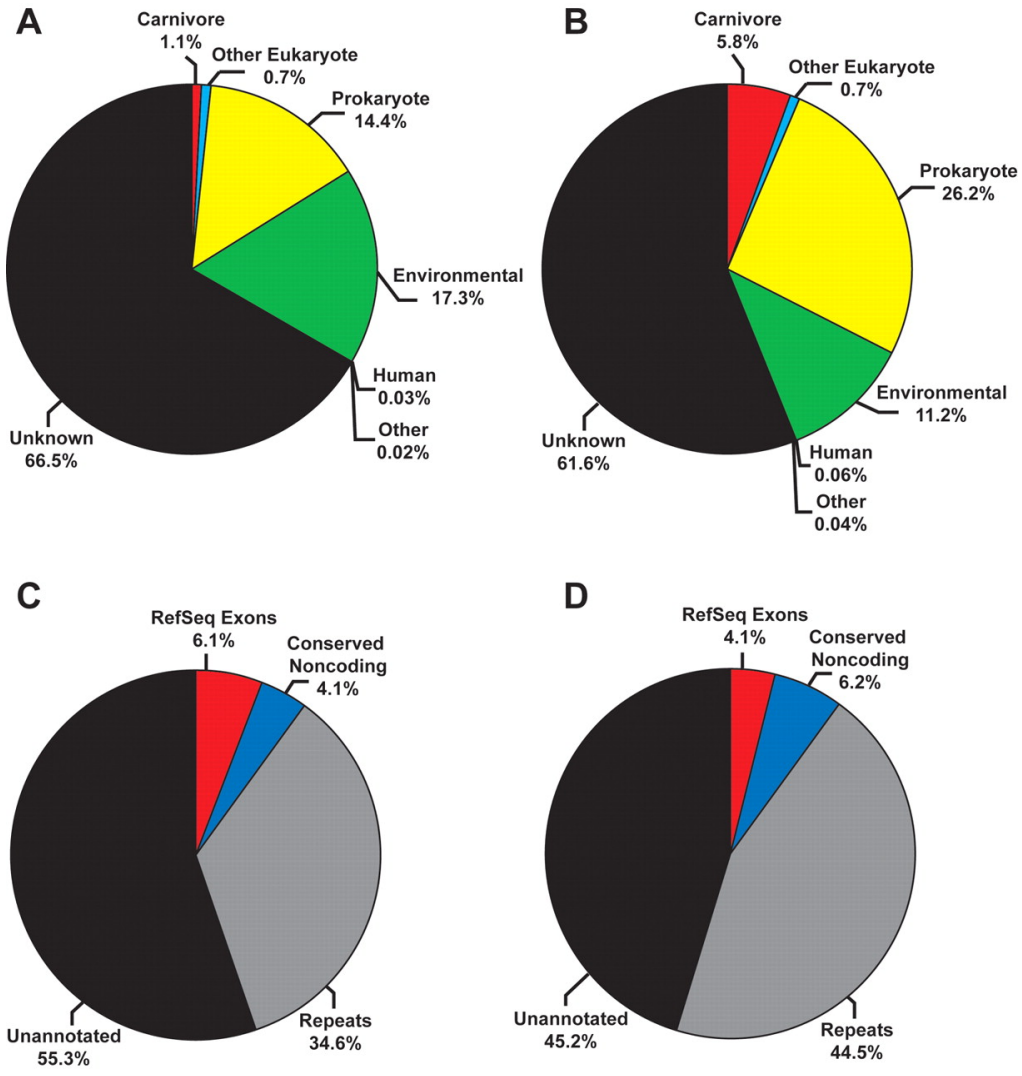
Mammoth Metagenomics

Mapping 13M reads modern elephant and GenBank



Poinar et al (2006) Science 311:5759

Characterization of two independent cave bear genomic libraries.



J P Noonan et al. Science 2005;309:597-599

13,000 year old human coprolite

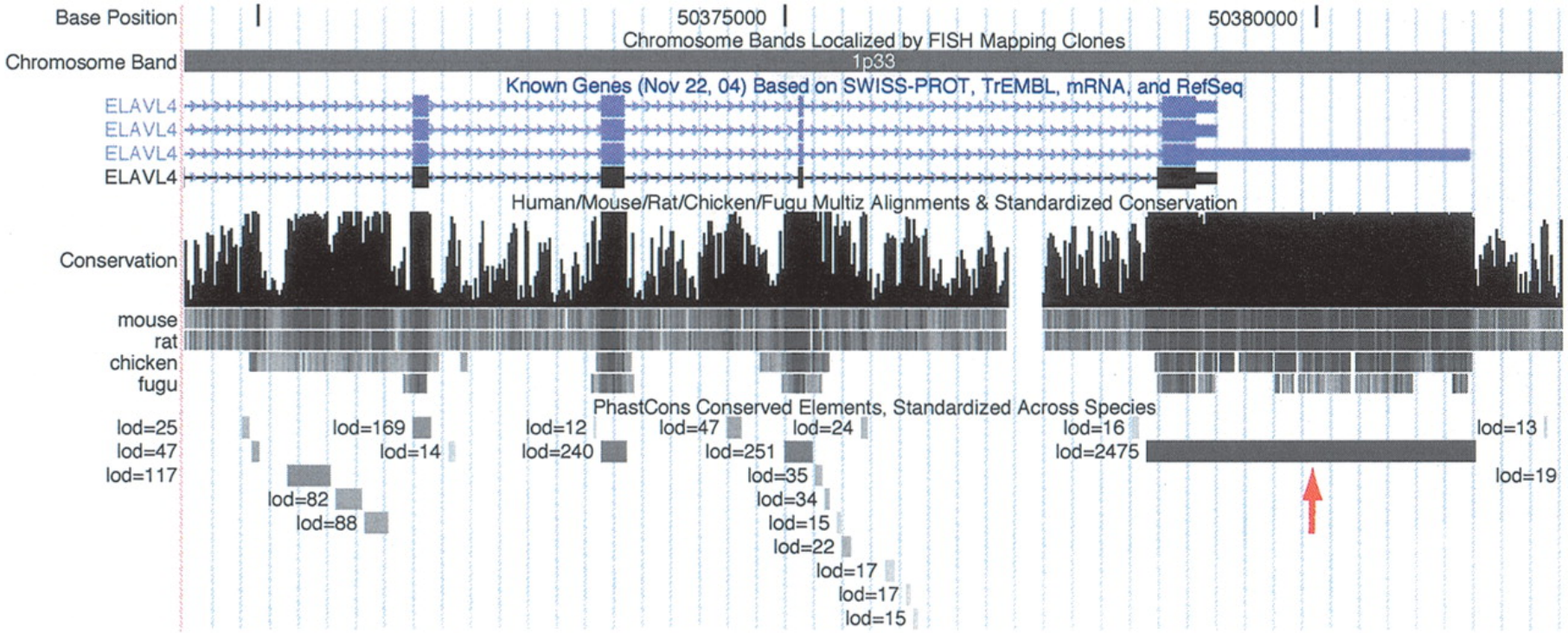
Multiple Sequence Alignment

Multiple sequence alignment of glyceraldehyde 3-phosphate dehydrogenases

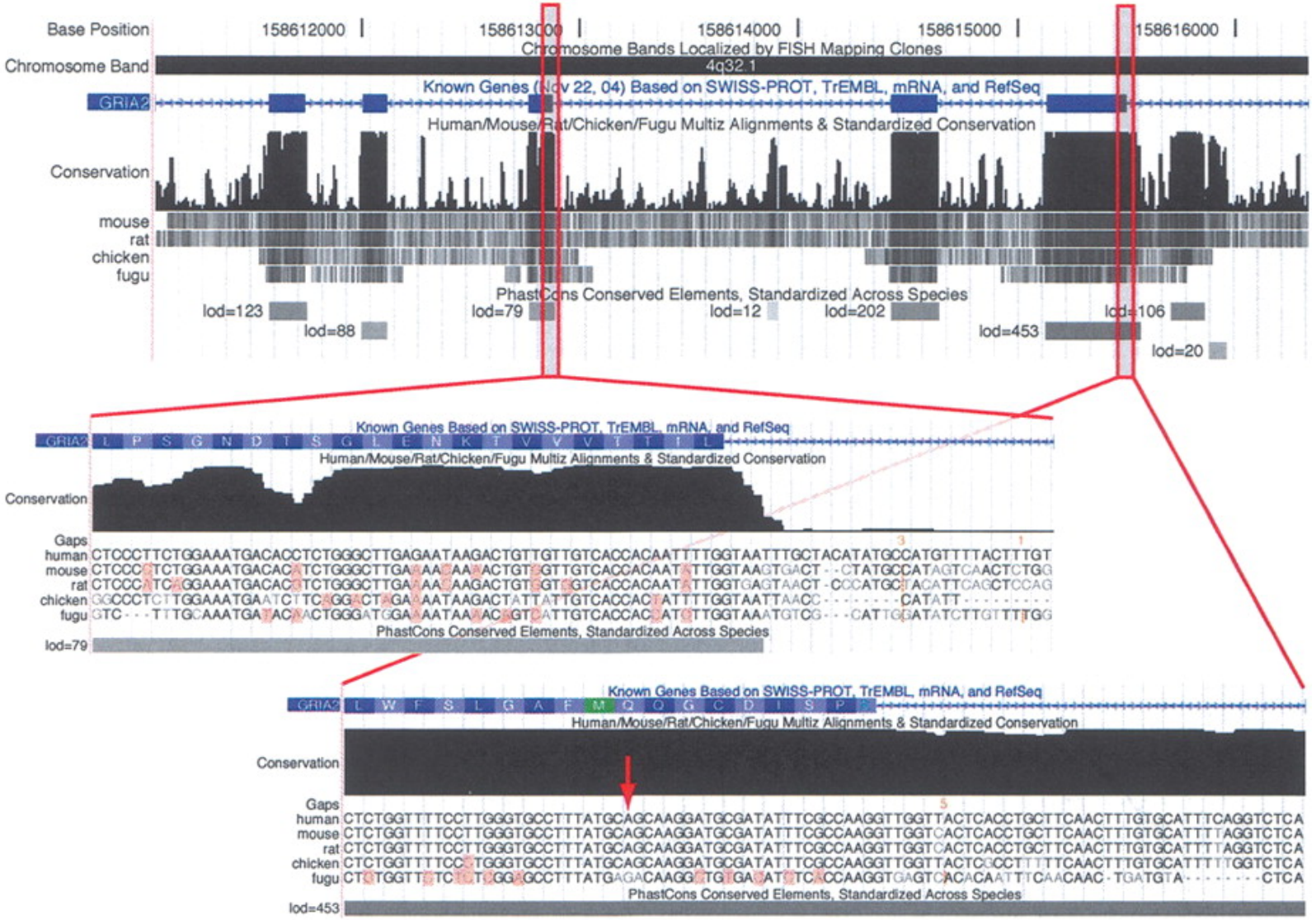
```
fly        GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human      GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant      GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium  GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast      GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon   GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly        KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human      KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant      KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium  KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast      KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon   KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly        GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human      GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant      GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium  GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast      GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon   GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

Bioinformatics and Functional Genomics, J. Pevsner.

Extreme conservation between human, mouse, rat, chicken and fugu



Siepel A et al. Genome Res. 2005;15:1034-1050



Siepel A et al. Genome Res. 2005;15:1034-1050

Aligned Families of Proteins

Statistical Model for Five Related Proteins

12345
CCGTL
CGHSV
GCGSL
CGGTL
CCGSS

	1	2	3	4	5
Prob(C)	0.8	0.6	-	-	-
Prob(G)	0.2	0.4	0.8	-	-
Prob(H)	-	-	0.2	-	-
Prob(S)	-	-	-	0.6	0.2
Prob(T)	-	-	-	0.4	-
Prob(L)	-	-	-	-	0.6
Prob(V)	-	-	-	-	0.2

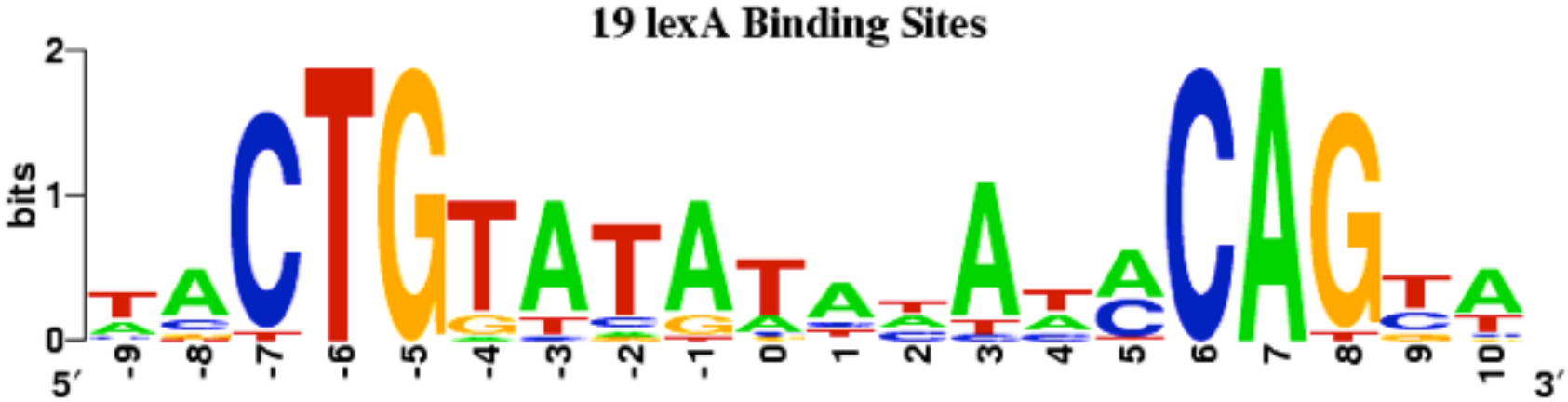
CGGSV = 0.8 * 0.4 * 0.8 * 0.6 * 0.2 = 0.031

Log Odds: $\log_e(0.8)+ \log_e(0.4)+ \log_e(0.8)+ \log_e(0.6)+ \log_e(0.2)=-3.48$

Position Specific Scoring Matrix (PSSM) for LDL (LPB000033) from BLOCKS

Position of Match	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z	*	-
1.	-27	-28	-30	-30	-4	-30	-33	-24	6	19	-29	-1	-26	-36	1	25	-8	7	-25	31	-14	-27	-1	0	0
2.	7	-65	28	-64	6	-53	-67	-64	37	-64	-45	-45	-67	-69	-63	-66	-60	-56	-36	-66	-42	-60	-33	0	0
3.	6	6	-31	11	26	-40	7	-28	-31	-3	-38	-34	-1	-37	-23	-30	2	-28	4	-42	-13	-40	-1	0	0
4.	13	-5	-26	11	-26	-35	-30	-27	-22	13	-30	-27	-27	-36	-25	21	0	-25	16	-39	-13	-35	-25	0	0
5.	24	7	-29	10	-34	-38	5	-36	-34	-37	7	-32	3	-41	-36	-39	12	-30	-32	-45	-17	-42	-36	0	0
6.	-24	-5	-28	-3	-15	-34	16	-32	-30	-11	8	-27	-6	2	-32	-5	20	8	-30	-41	-10	-38	-25	0	0
7.	-58	-23	-52	8	-62	-44	-67	-61	-38	-63	31	-27	-63	-69	-60	-62	-64	-57	-43	-60	-44	-57	-61	0	0
8.	-13	23	-33	24	-21	-34	-5	28	-41	-5	-41	-35	22	-39	4	-27	-28	-30	-39	31	-18	-25	-7	0	0
9.	-33	0	-42	1	-41	-51	33	-40	-53	-37	-53	-47	-2	-4	-42	10	7	-39	-50	-50	-26	-51	-42	0	0
10.	-4	-15	-18	-25	-24	-23	6	-24	-4	14	9	-15	-3	-31	-22	-19	10	9	8	-32	-7	-28	-23	0	0
11.	5	11	-23	23	8	8	-7	12	-26	9	-12	-23	-5	-29	-4	-6	-9	-21	-25	-29	-7	-19	1	0	0
12.	-37	-42	-41	-44	-39	-42	-45	-34	-3	-2	8	-33	-39	-4	0	33	9	-38	-20	-48	-19	-44	-17	0	0
13.	-6	-5	-18	-18	-17	-18	-22	6	-18	14	-14	-4	12	-27	9	8	4	11	0	-23	-5	11	-3	0	0
14.	-14	-37	-26	-39	-36	7	-37	-36	-17	-36	7	-23	-35	8	-6	-38	15	8	22	-37	-14	-31	-19	0	0
15.	-47	-57	-40	-59	-55	-35	-58	-52	24	-8	27	-26	-56	-58	-49	-46	-54	-46	-28	-51	-34	-46	-52	0	0
16.	0	-18	-19	-33	-29	17	-34	-25	13	-29	5	-14	1	-37	12	-30	-29	-24	21	-25	-12	4	-6	0	0
17.	-22	-5	-26	-24	-10	-32	8	-1	-34	-25	-21	-30	19	-36	13	-3	18	16	-31	-36	-12	1	3	0	0
18.	-8	-7	-21	4	12	15	-27	-21	-21	0	1	-20	-21	-32	-1	-23	16	4	-7	-25	-6	8	5	0	0
19.	-33	25	-37	32	-5	-44	-5	-29	-39	-29	-44	-39	17	-43	-1	10	4	-31	-3	-48	-18	-42	-3	0	0
20.	-43	-54	-36	-56	-52	-32	-54	-46	19	-42	27	-22	-52	-55	-46	12	-50	-42	5	-46	-29	-42	-49	0	0
21.	-19	-21	-22	-20	16	-29	-7	-21	-8	0	-11	-21	-22	-32	10	6	11	8	13	-34	-7	-30	13	0	0
22.	-27	-5	-30	-21	12	-24	-30	23	-32	-21	-16	-3	15	-37	13	6	-3	-26	-31	45	-14	-17	12	0	0
23.	-81	-91	-91	-89	-87	-96	-91	-88	-94	-86	-95	-90	-93	45	-88	-89	-86	-87	-91	-99	-83	-97	-88	0	0
24.	15	-4	-35	-30	-7	-47	-35	-30	-46	-28	-47	-41	31	-45	-33	23	5	-32	-41	-49	-23	-44	-21	0	0
25.	10	-34	-34	-36	-38	-45	32	-38	-46	9	-46	-41	-32	-43	-36	-33	8	-33	-41	-45	-25	-46	-37	0	0
26.	5	-58	-41	-59	-52	-42	-58	-51	34	-52	6	-33	-56	-59	16	-53	-52	-46	-3	-55	-31	-50	-14	0	0
27.	24	-40	-28	-42	-39	-39	-35	-40	-30	-39	4	-31	-38	11	-39	-40	-8	21	2	-46	-19	-42	-39	0	0
28.	-32	-45	-28	-47	-45	-24	6	-40	7	-43	21	-17	-43	-47	-41	-42	-39	-11	19	33	-21	-30	-43	0	0
29.	-55	4	-59	39	-39	-63	-53	-50	-58	10	-63	-59	-40	-61	-49	11	-52	-53	-3	-69	-39	-62	-44	0	0
30.	-24	-33	-24	-34	-33	19	-33	7	-11	-32	12	1	-31	23	-30	-32	8	-25	-8	-23	-13	17	-32	0	0
31.	-20	-3	-18	10	7	-12	-28	1	3	-18	11	18	-20	-9	-8	5	-13	-19	7	-21	-6	12	-2	0	0

Logos provide a simple visualization of a PSSM



Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Res., 14:1188-1190 [weblogo.berkeley.edu]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res. 18:6097-6100

Considerations when making a profile.

- How are missing sequences represented?
- Many sequences are needed to create a useful alignment, but not too many that are closely related.
- Where are the gaps located?

Position specific iterated BLAST: PSI-BLAST

The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query.

PSI-BLAST
Interative Protein-Protein BLAST

BLASTP (first iteration)

Analyze output and create PSSM

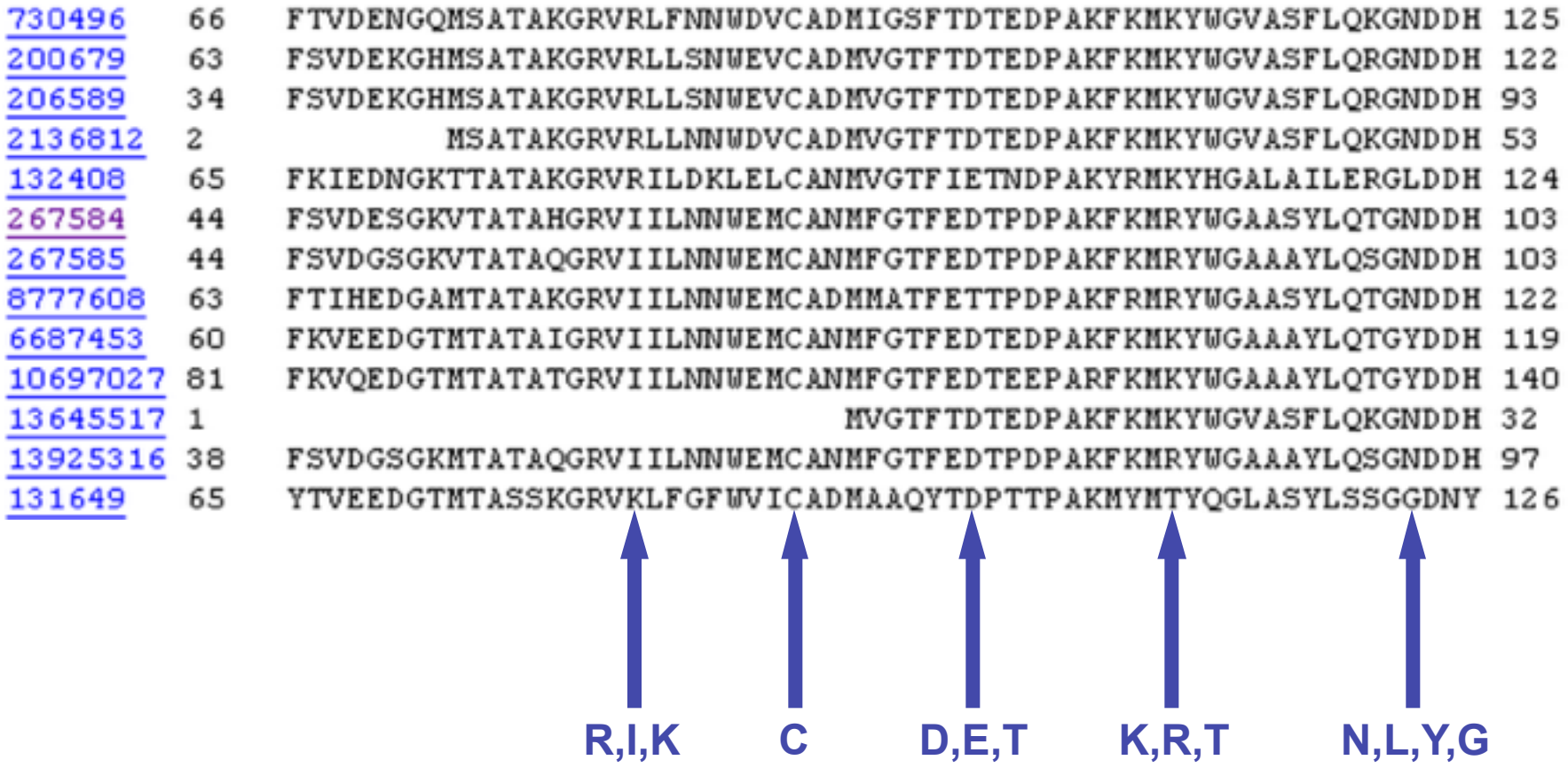
(Repeat until no change or iteration limit)

PSSM used to search database

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a "profile" or specialized position-specific scoring matrix (PSSM)


730496 66 FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125
200679 63 FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122
206589 34 FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 93
2136812 2 MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 53
132408 65 FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584 44 FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103
267585 44 FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 103
8777608 63 FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122
6687453 60 FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119
10697027 81 FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140
13645517 1 MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 32
13925316 38 FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 97
131649 65 YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126
R,I,K
C
D,E,T
K,R,T
N,L,Y,G

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8	L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
										...											
37	S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41	Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a "profile" or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

	gi|6978523|ref|NP_036909.1|	apolipoprotein D [Rattus norvegicus]...	147	4e-35
	gi|1542847|dbj|BAA13453.1|	(D87752) alpha1-microglobulin/bikunin...	144	6e-34
	gi|619383|gb|AAB32200.1|	apolipoprotein D, apoD [human, plasma, ...	143	8e-34
	gi|5419892|emb|CAB46489.1|	(X02824) RBP (aa 101-172) [Homo sapiens]	139	1e-32
	gi|4502163|ref|NP_001638.1|	apolipoprotein D precursor [Homo sap...	138	4e-32
	gi|584763|sp|P37153|APD_RABIT	APOLIPOPROTEIN D PRECURSOR >gi|482...	134	4e-31
	gi|1703341|sp|P51909|APD_CAVPO	APOLIPOPROTEIN D PRECURSOR >gi|11...	133	7e-31
	gi|2895204|gb|AAC02945.1|	(AF025334) mutant retinol binding prot...	80	9e-15
	gi|1246096|gb|AAB35919.1|	(S80440) apolipoprotein D, apoD (C-ter...	77	8e-14
	gi|2895206|gb|AAC02946.1|	(AF025335) mutant retinol binding prot...	67	8e-11
NEW	gi|1346419|sp|P49291|LAZA_SCHAM	LAZARILLO PROTEIN PRECURSOR >gi|...	63	1e-09
NEW	gi|2506821|sp|P00978|AMBP_BOVIN	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
NEW	gi|2497696|sp|Q07456|AMBP_MOUSE	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
NEW	gi|6680684|ref|NP_031469.1|	alpha 1 microglobulin/bikunin [Mus m...	62	2e-09
NEW	gi|12836446|dbj|BAB23659.1|	(AK004907) putative [Mus musculus]	62	3e-09
NEW	gi|6978497|ref|NP_037033.1|	alpha-1 microglobulin/bikunin [Rattu...	62	3e-09
NEW	gi|2507586|sp|P04366|AMBP_PIG	AMBP PROTEIN PRECURSOR [CONTAINS: ...	61	8e-09
NEW	gi|1085207|pir||JC2556	alpha-1-microglobulin/inter-alpha-trypsin...	60	1e-08
NEW	gi|2988354|dbj|BAA25305.1|	(AB006444) alpha-1-microglobulin/biku...	59	2e-08
NEW	gi|108233|pir||S13493	alpha-1-microglobulin - pig	59	2e-08
NEW	gi|1882|emb|CAA36306.1|	(X52087) precursor codes for two protein...	59	2e-08
NEW	gi|9181923|gb|AAF85707.1|AF276505_1	(AF276505) neural Lazarillo ...	59	3e-08
NEW	gi|7296083|gb|AAF51378.1|	(AE003586) NLaz gene product [Drosophi...	58	3e-08
NEW	gi|117330|sp|P80007|CRA2_HOMGA	CRUSTACYANIN A2 SUBUNIT >gi|10275...	57	8e-08
NEW	gi|2497695|sp|Q60559|AMBP_MESAU	AMBP PROTEIN PRECURSOR [CONTAINS...	57	1e-07
NEW	gi|102968|pir||S22400	insecticyanin A - tobacco hornworm >gi|971...	56	1e-07
NEW	gi|4502067|ref|NP_001624.1|	alpha-1-microglobulin/bikunin precur...	56	2e-07
NEW	gi|1146408|gb|AAA85089.1|	(L41641) gallerin [Galleria mellonella]	56	2e-07
NEW	gi|2497694|sp|Q62577|AMBP_MERUN	AMBP PROTEIN PRECURSOR [CONTAINS...	55	3e-07
NEW	gi|1213589|dbj|BAA12075.1|	(D83712) Prostaglandin D Synthase [Xe...	54	5e-07
	gi|539717|pir||A61233	retinol-binding protein - cat (fragment)	54	8e-07
NEW	gi|266472|sp|Q01584|LIPO_BUFMA	LIPOCALIN PRECURSOR >gi|104284|pi...	53	1e-06
	gi|265042|gb|AAB25283.1|	retinol-binding protein, RBP (N-termina...	52	3e-06
NEW	gi|1079295|pir||S52354	gene cpl-1 protein - African clawed frog ...	52	3e-06
NEW	gi|732003|sp|P39281|BLC_ECOLI	OUTER MEMBRANE LIPOPROTEIN BLC PRE...	51	9e-06

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a "profile" or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

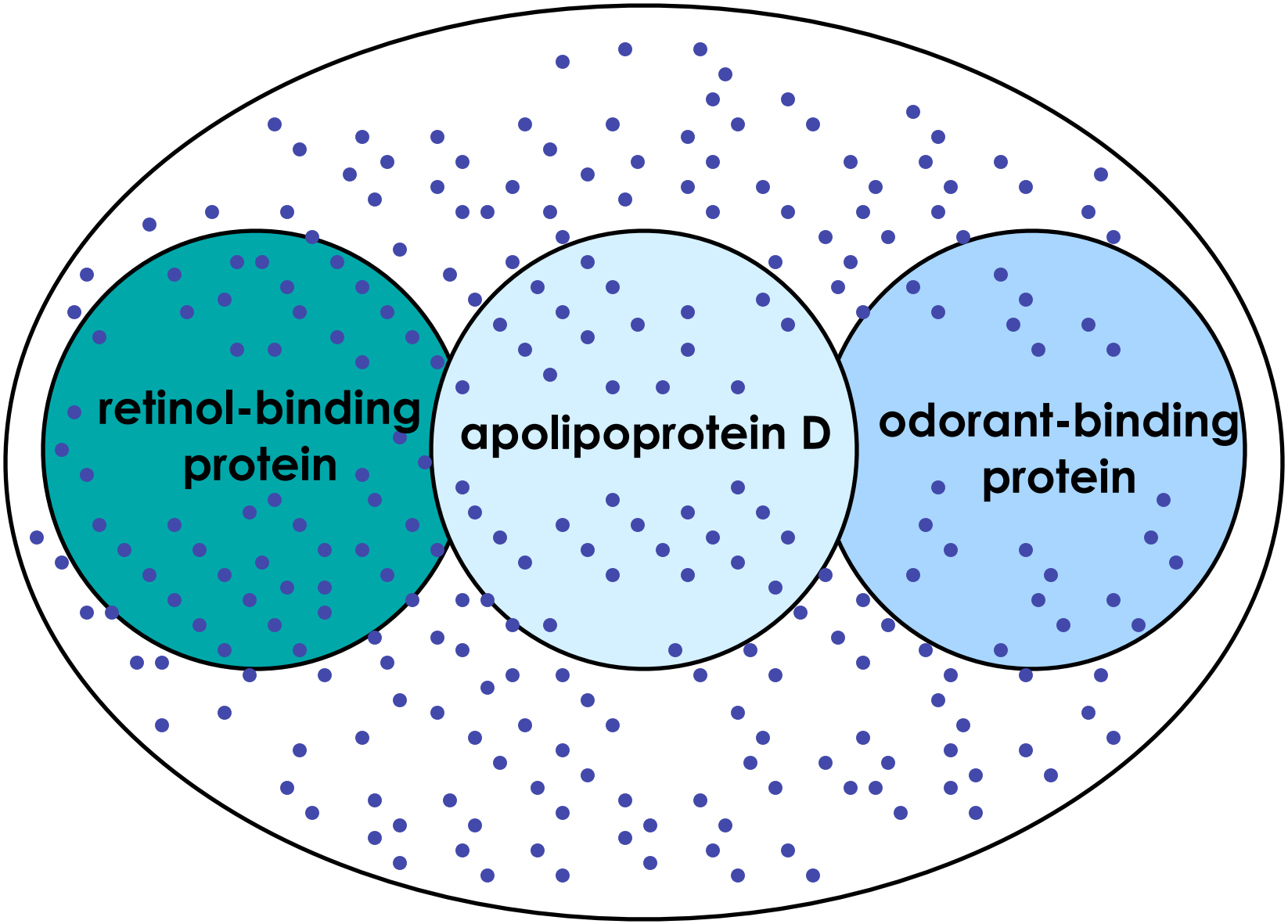
[4] PSI-BLAST estimates statistical significance (E values)

[5] Repeat steps [3] and [4] iteratively, typically 5 times. At each new search, a new profile is used as the query.

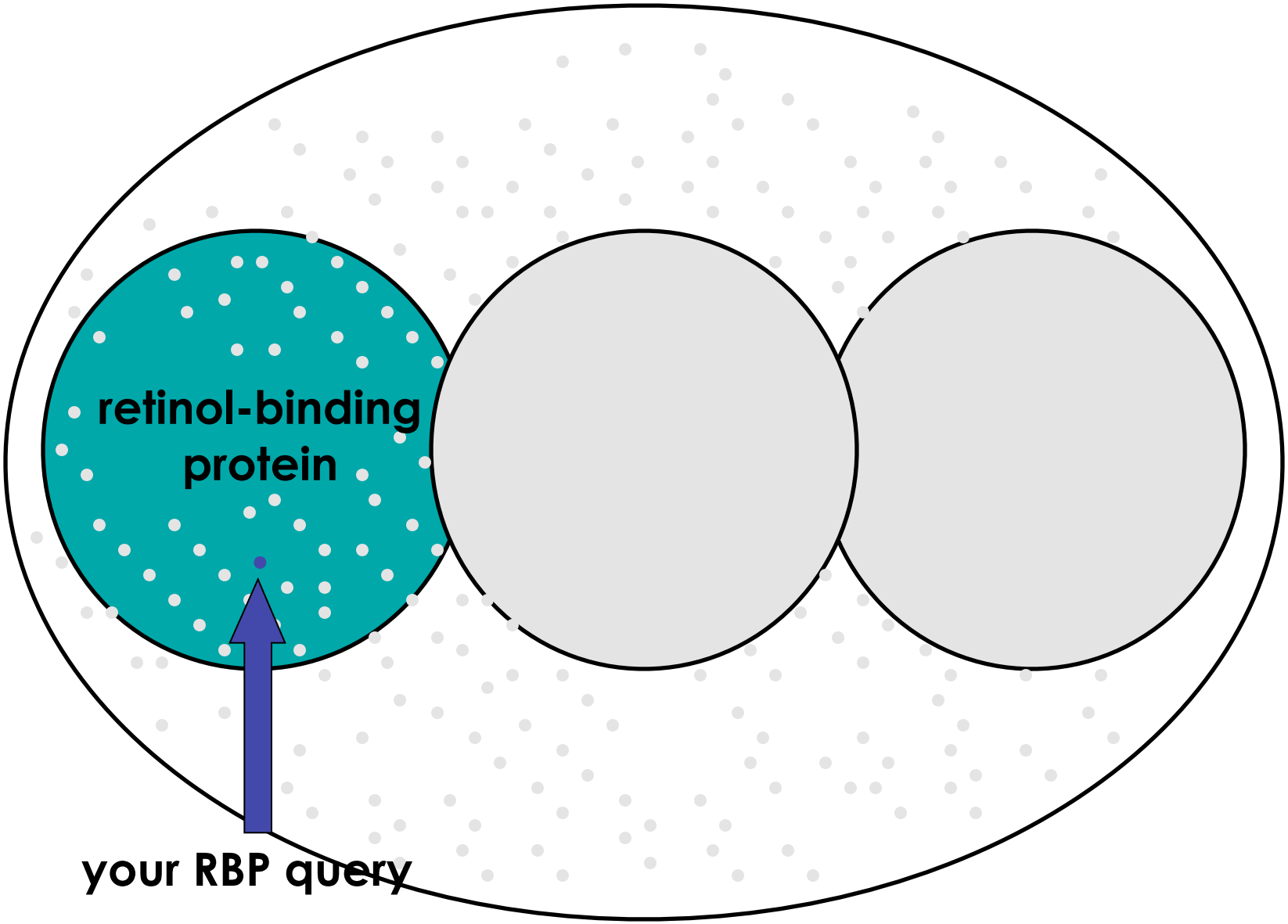
Results of a PSI-BLAST search

Iteration	# hits	# hits > threshold
1	104	49
2	173	96
3	236	178
4	301	240
5	344	283
6	342	298
7	378	310
8	382	320

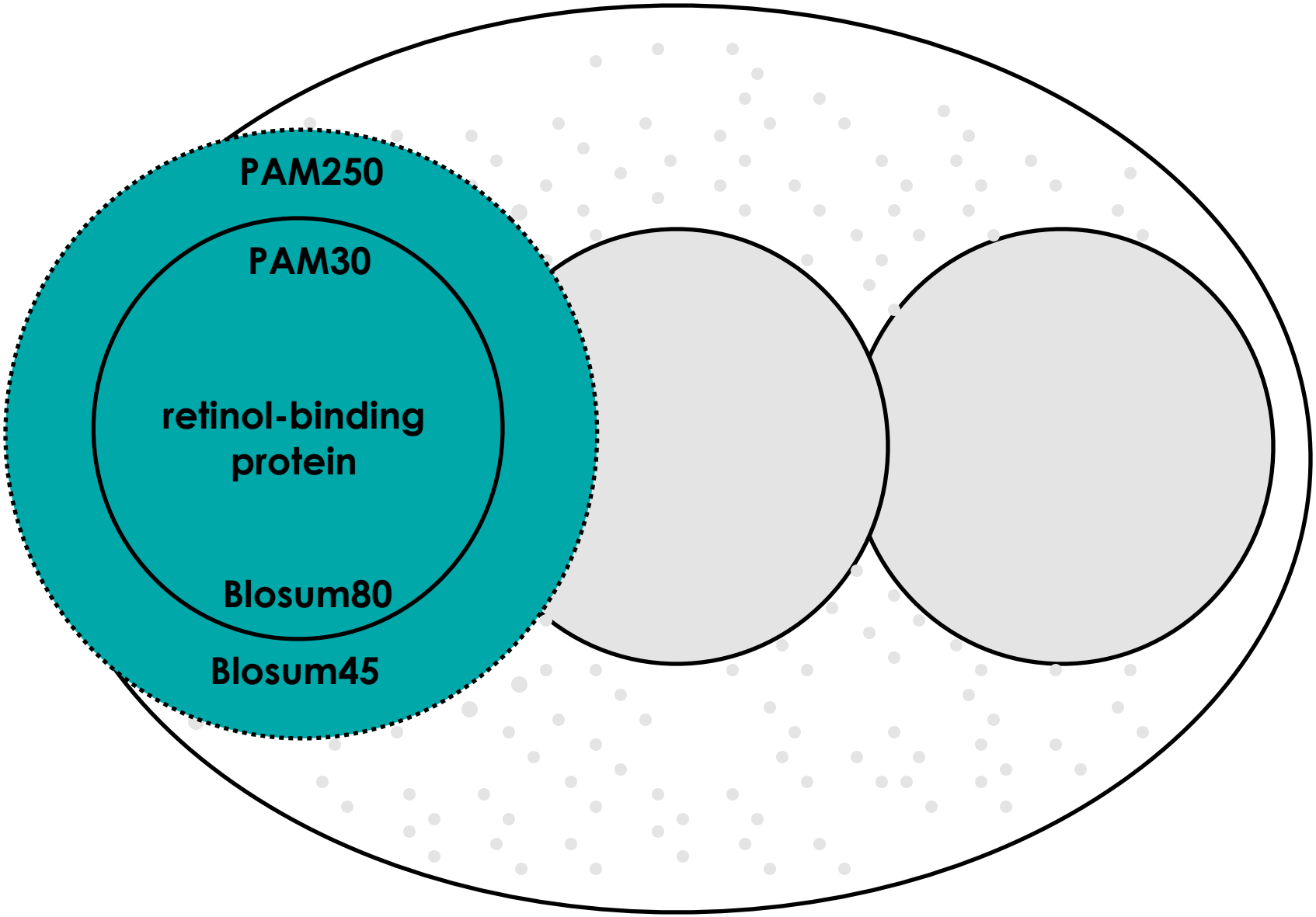
The universe of lipocalins (each dot is a protein)



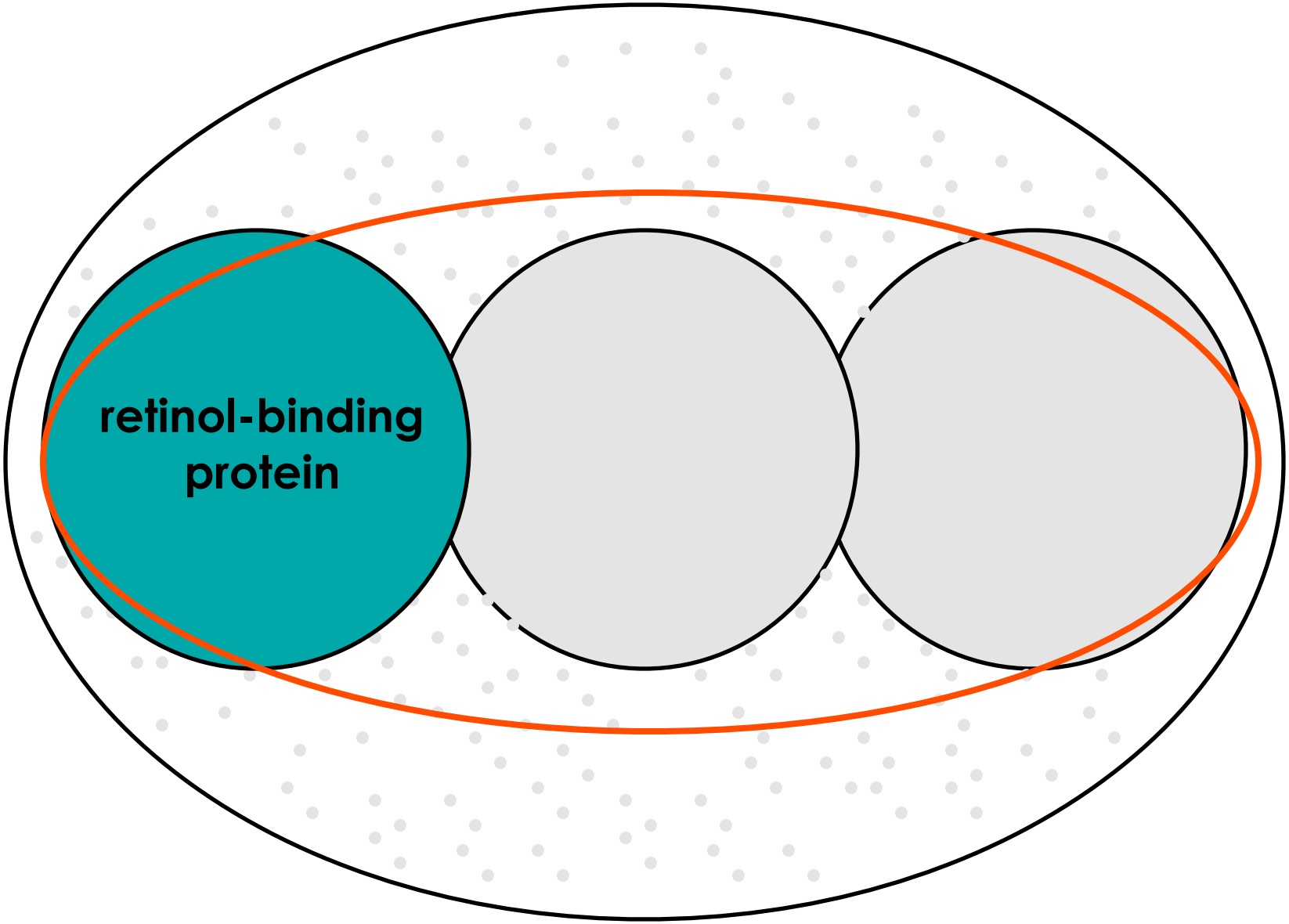
Scoring matrices focus on the big (or small) picture



Scoring matrices focus on the big (or small) picture



PSI-BLAST generates scoring matrices more powerful than PAM or BLOSUM



PSI-BLAST alignment of RBP and β-lactoglobulin: iteration 1

```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
           V+ENFD  ++ G WY + +K P      + I A +S+ E G +    K          ++
Sbjct: 33  VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIEVLNK---------ELS 82

Query: 87  ADMVGTF---------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR 137
            D  GT          ++  +PAK +++++ +          +WI+ TDY+ YA+ YSC
Sbjct: 83  PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
               L ++D      + ++  R+P  LPPE
Sbjct: 136 TFFWLFHVD------FFWILGRNPY-LPPE 158
```

PSI-BLAST alignment of RBP and β-lactoglobulin: iteration 2

```
Score =  140 bits (353), Expect = 1e-32
Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)

Query: 4   VWALLLLAAWAAAERDCRVSSF--------RVKENFDKARFSGTWYAMAKKDPEGLFLQD 55
           V  L+ LA  A      +  +F          V+ENFD  ++ G WY + +K P      +
Sbjct: 2   VTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEI-EKIPASFEKGN 60

Query: 56  NIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMV---GTFTDTEDPAKFKMKYWGVASF 112
            I A +S+ E G +    K        + D   + V       ++  +PAK +++++ +
Sbjct: 61  CIQANYSLMENGNIEVLNKEL-----SPDGTMNQVKGEAKQSNVSEPAKLEVQFFPL--- 112

Query: 113 LQKGNDDHWIVDTDYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC     L ++D       + ++  R+P  LPPE
Sbjct: 113 --MPPAPYWILATDYENYALVYSCTTFFWLFHVD------FFWILGRNPY-LPPET 159
```

PSI-BLAST alignment of RBP and β-lactoglobulin: iteration 3

```
Score =  159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3   WVWALLLLAAWAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
            V  L+ LA  A               +  S  V+ENFD  ++ G WY + K
Sbjct: 1   MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55  DNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ 114
            + I A +S+ E G +    K          V  +     ++  +PAK +++++ +
Sbjct: 60  NCIQANYSLMENGNIEVLNKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115 KGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC             + ++  R+P  LPPE
Sbjct: 113 MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159
```

PSI-BLAST: the problem of corruption

PSI-BLAST is useful to detect weak but biologically meaningful relationships between proteins.

The main source of false positives is the spurious amplification of sequences not related to the query. For instance, a query with a coiled-coil motif may detect thousands of other proteins with this motif that are not homologous.

Once even a single spurious protein is included in a PSI-BLAST search above threshold, it will not go away.

Progressive alignment

- Advantages
  - Biologically reasonable search strategy
  - Relatively fast & efficient
- Disadvantages
  - Quality deteriorates when sequences are distantly related
  - Strongly dependent upon initial alignments since early errors are "locked in"

Slide from JD Wren

ClustalW
http://www.ebi.ac.uk/clustalw/

- Most popular multiple alignment tool
- 'W' stands for 'weighted' (different parts of alignment are weighted differently).
- Three-step process
  1) Construct pairwise alignments
  2) Build Guide Tree
  3) Progressive alignment built using the tree

Slide from JD Wren

Step 1: Pairwise Alignment

- Aligns each sequence against each other giving a similarity matrix
- Similarity = exact matches / sequence length (percent identity)

	v_1	v_2	v_3	v_4
v_1	–			
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–

(.17 means 17 % identical)

Slide from JD Wren

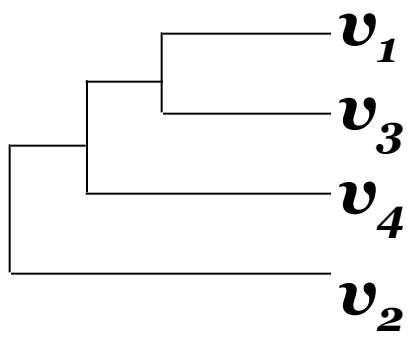
Step 2: Guide Tree

- Create Guide Tree using the similarity matrix
  - ClustalW uses the neighbor-joining method
  - Guide tree reflects evolutionary relations?

Slide from JD Wren

Step 2: Guide Tree (cont'd)

	v_1	v_2	v_3	v_4
v_1	–			
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–



Calculate:

$v_{1,3}$ = alignment (v_1, v_3)

$v_{1,3,4}$ = alignment$((v_{1,3}), v_4)$

$v_{1,2,3,4}$ = alignment$((v_{1,3,4}), v_2)$

Slide from JD Wren

Step 3: Progressive Alignment

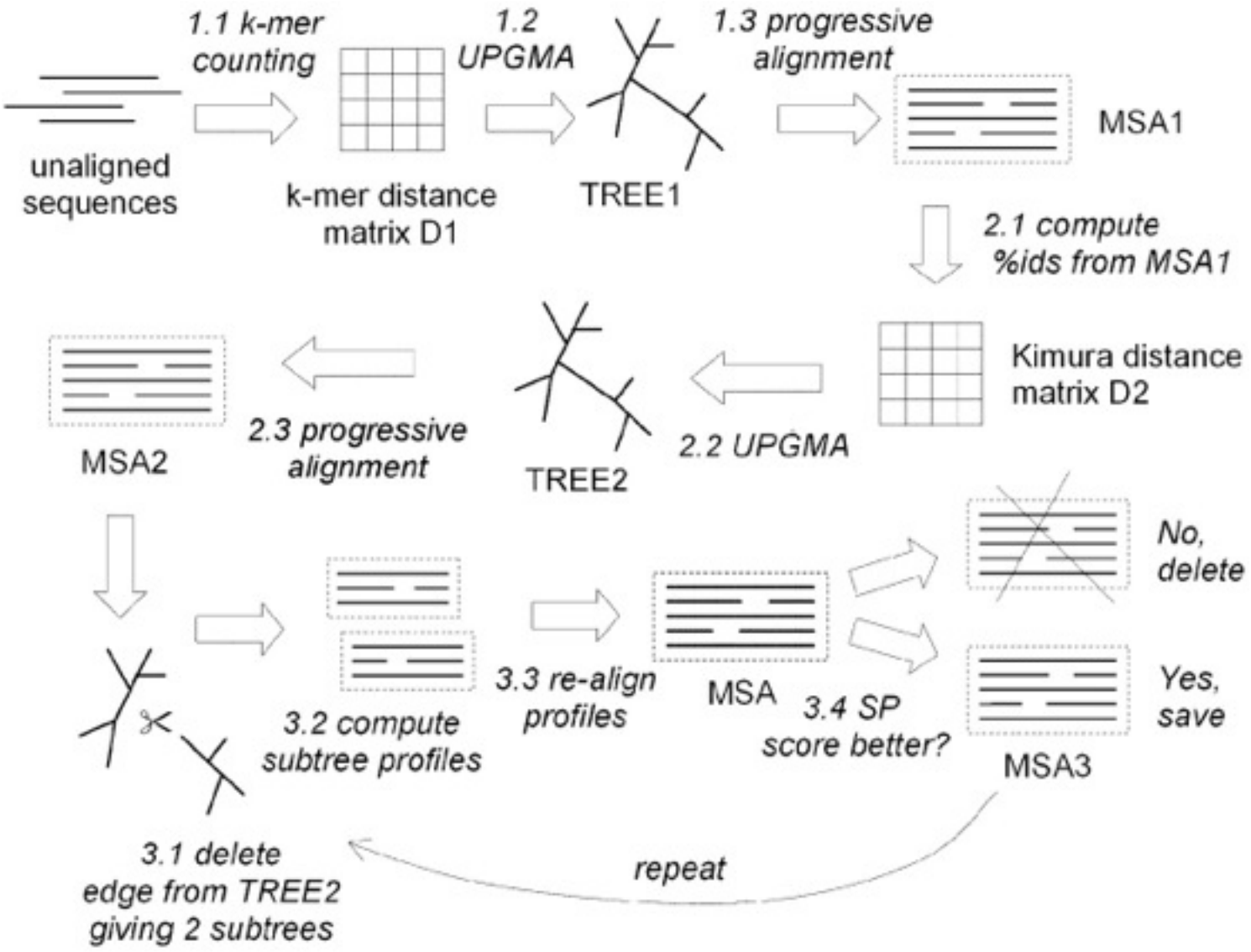
- Start by aligning the two most similar sequences
- Following the guide tree, add in the next sequences, aligning to the existing alignment
- Insert gaps as necessary

```
FOS_RAT          PEEMSVTS-LDLTGGLPEATTPESEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD
FOS_MOUSE        PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD
FOS_CHICK        SEELAAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD
FOSB_MOUSE       PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP-----------------LPFQ
FOSB_HUMAN       PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP-----------------LPFQ
                 .   . :   ** .     :..  *:.*   *   . *                   **:
```

Dots and stars show degree of conservation in a column.

Slide from JD Wren

MUSCLE Algorithm



Edgar, R. C. Nucl. Acids Res. 2004 32:1792-1797

Orthology/Homology – Protein Family Resources

TreeFAM www.treefam.org (Sanger)
PANTHER www.pantherdb.org (SRI)
ProPhyler www.prophyler.org (Stanford)

Clusters of Orthology Groups (KOGS)
InParanoid
HomoloGene
OrthoMCL-DB

Reading

- Korf, Yandell & Bedell (2003) BLAST: An Essential Guide to the Basic Local Alignment Search Tool. O'Reilly
- Jonathan Pevsner (2003) *Bioinformatics and Functional Genomics.* Wiley-Liss
- Mount (2004) Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press
- Baxevanis & Ouellette (2001) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley Interscience
- Jones & Pevzner (2004) An Introduction to Bioinformatic Algorithms. (MIT Press)
- Salzberg, Searls & Kasif (1998) Computational Methods in Molecular Biology. Elsevier
- Waterman (1995) Introduction to Computation Biology. Chapman & Hall