

CAP 5510: Introduction to Bioinformatics  
CGS 5166: Bioinformatics Tools

**Giri Narasimhan**

ECS 254; Phone: x3748

[giri@cis.fiu.edu](mailto:giri@cis.fiu.edu)

[www.cis.fiu.edu/~giri/teach/BioinfS11.html](http://www.cis.fiu.edu/~giri/teach/BioinfS11.html)

# BLAST Parameters and Output

- Type of sequence, nucleotide/protein
- Word size,  $w$
- Gap penalties,  $p_1$  and  $p_2$
- Neighborhood Threshold Score,  $T$
- Score Threshold,  $S$
- E-value Cutoff,  $E$
- Number of hits to display,  $H$
- Database to search,  $D$
- Scoring Matrix,  $M$
- Score  $s$  and E-value  $e$ 
  - E-value  $e$  is the expected number of sequences that would have an alignment score greater than the current score  $s$ .

# BLAST algorithm: Phase 1

Phase 1: get list of word pairs ( $w=3$ ) above threshold  $T$

Example: for a human RBP query

...FS**GTW**YA...

**GTW** is a word in this query sequence

A list of words ( $w=3$ ) is:

FSG SGT GTW TWY WYA

YSG TGT ATW SWY WFA

FTG SVT GSW TWF WYS

## Phase 1: Find list of similar words

□ Find list of words of length  $w$  (here  $w = 3$ ) and distance at least  $T$  (here  $T = 11$ )

● GTW	22
● GSW	18
● ATW	16
● NTW	16
● GTY	13
● GNW	10
● GAW	9

# Use BLOSUM to score word hits

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

## BLAST: Phases 2 & 3

□ Phase 2: Scan database for exact hits of similar words list and find **HotSpots**

□ Phase 3:

● Extend good hit in either direction.

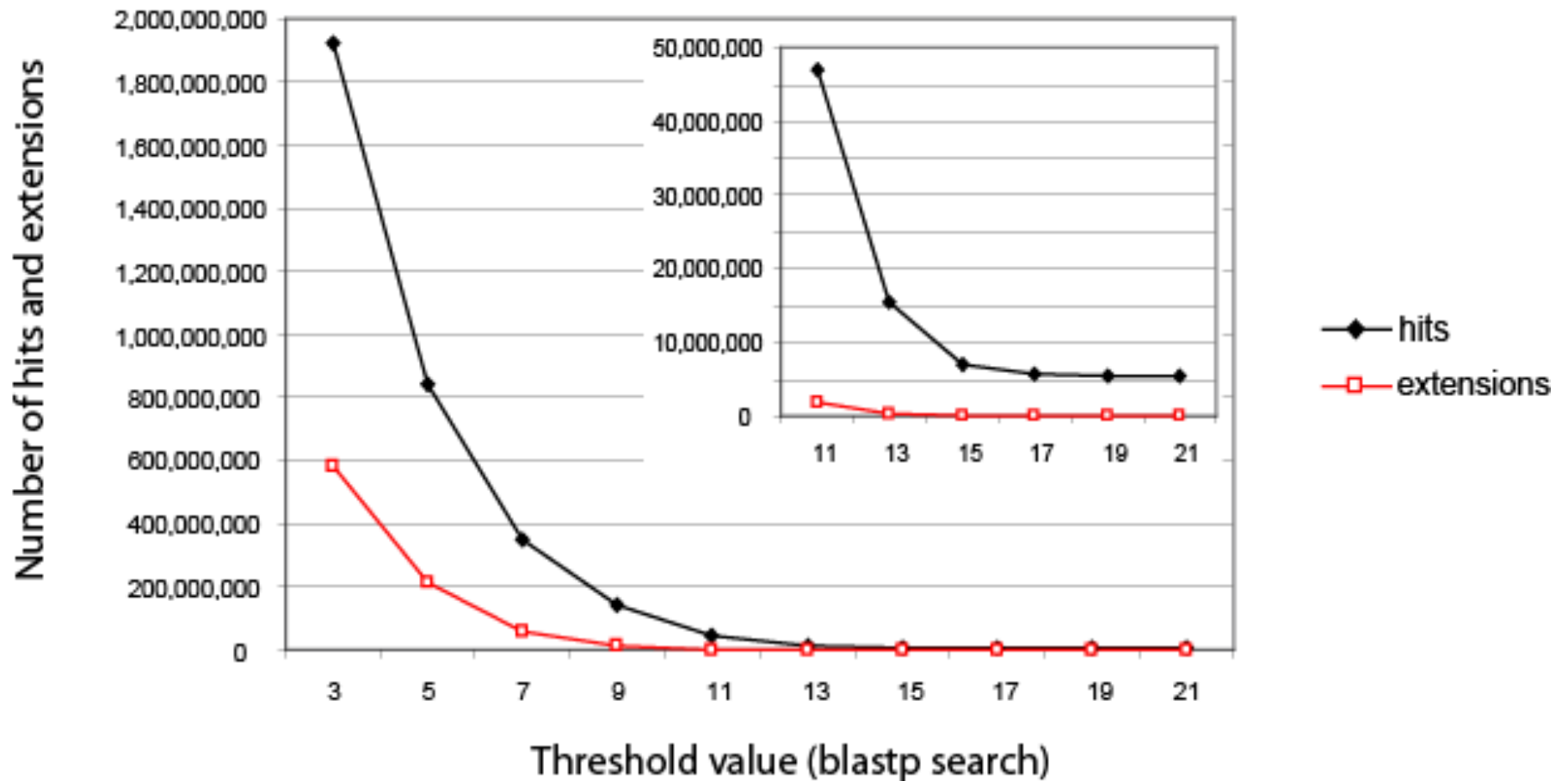
● Keep track of the score (use a scoring matrix)

● Stop when the score drops below some cutoff.

```
KENFDKARFSGTWTYAMAKKDPEG 50 RBP (query)  
MKGLDIQKVAGTWTYSLAMAASD. 44 lactoglobulin (hit)
```

**extend** ← **Hit!** → **extend**

# BLAST: Threshold vs # Hits & Extensions



# Word Size

□ **Blastn**:  $w = 7, 11, \text{ or } 15$ .

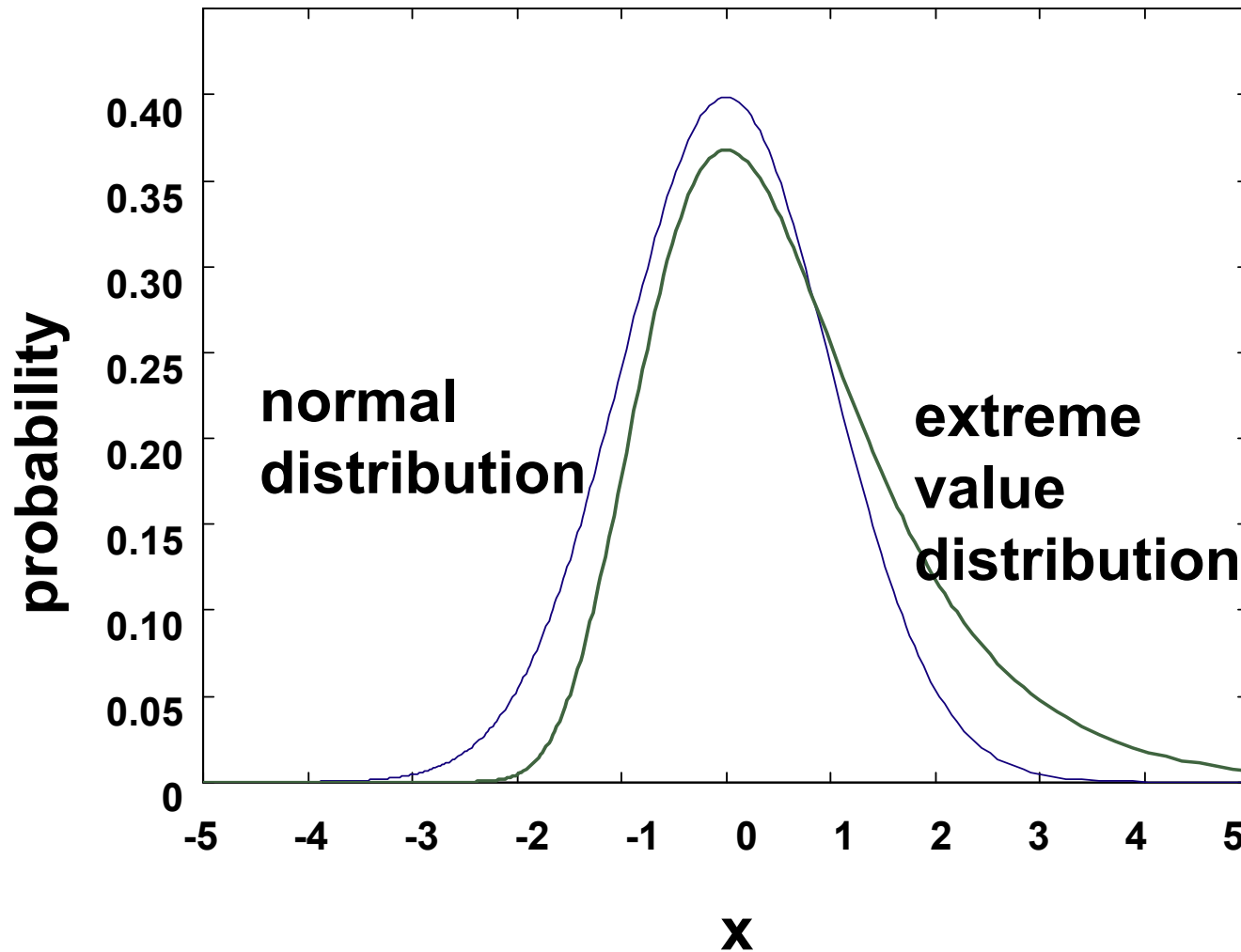
●  $w=15$  gives fewer matches and is faster than  $w=11$  or  $w=7$ .

□ **Megablast**:  $w = 28 \text{ to } 64$ .

● Megablast is VERY fast for finding closely related DNA sequences!



# Scores: Follow Extreme Value Distribution



$$E = Kmn e^{-\lambda S}$$

$m, n$  = seq length  
 $S$  = Raw Score  
 $K \approx$  Search space

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

$S'$  = Bit Score

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

$p$  = p-value

# E-value versus P-value

E-value	P-value
10	0.9999546
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001

**E-values are easier to interpret;**

**If query is short aa sequence, then use very large E-value;  
Sometimes even meaningful hits have large E-values.**

# Assessing whether proteins are homologous

```
>gi|4505583|ref|NP\_002562.1 progestagen-associated endometrial protein (placental protein 14, 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  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFVDETGQMSATAKGRVRLNNWD- 84
          + K++ + + +GTW++MA      + L  + A  V  T  +          +L+ W+
Sbjct: 5   QTKQDLELPKLAGTWHSMAMAT-NNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 63

Query: 85  -VCADMVGTFTDTEPAKFKMKYWGVASFLQKGNDDHWIVD TDYD TY 130
          C +      T +P KFK+ Y  VA      ++  ++DTDYD +
Sbjct: 64  NSCVEKKVLGEKTGNPKKFKINY-TVA-----NEATLLD TDYDNF 102
```

RBP4 and PAEP:

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.

## Difficulties with BLAST

- ❑ Use human beta globin as a query against human RefSeq proteins, and blastp does not “find” human myoglobin. This is because the two proteins are too distantly related. PSI-BLAST at NCBI as well as hidden Markov models easily solve this problem.
- ❑ How can we search using 10,000 base pairs as a query, or even millions of base pairs? Many BLAST-like tools for genomic DNA are available such as PatternHunter, Megablast, BLAT, and BLASTZ.

# Related Tools

## Megablast

- For long, closely-related sequences
- Uses large  $w$  and is very fast

## BLAT

- UCSC tool
- DB broken into words; query is searched

## PatternHunter

- Generalized seeds used instead of words

## BLASTZ, Lagan, SSAHA

# Rules of Thumb

- ❑ Most sequences with significant similarity over their entire lengths are homologous.
- ❑ Matches that are > 50% identical in a 20-40 aa region occur frequently by chance.
- ❑ Distantly related homologs may lack significant similarity. Homologous sequences may have few absolutely conserved residues.
- ❑ A homologous to B & B to C  $\Rightarrow$  A homologous to C.
- ❑ Low complexity regions, transmembrane regions and coiled-coil regions frequently display significant similarity without homology.
- ❑ Greater evolutionary distance implies that length of a local alignment required to achieve a statistically significant score also increases.

# Rules of Thumb

- Results of searches using different scoring systems may be compared directly using normalized scores.
- If  $S$  is the (raw) score for a local alignment, the **normalized** score  $S'$  (in bits) is given by

$$S' = \frac{\lambda - \ln(K)}{\ln(2)}$$

The parameters depend on the scoring system.

- **Statistically significant normalized score,**

$$S' > \log\left(\frac{N}{E}\right)$$

where E-value =  $E$ , and  $N$  = size of search space.

# Multiple Alignments

- Global
  - ClustalW, ClustalX
  - MSA
  - T-Coffee
- Local
  - BLOCKS
  - eMOTIF
  - GIBBS
  - HMMER
  - MACAW
  - MEME
- Other
  - Profile Analysis from msa (UCSD)
  - SAM HMM (from msa)



# MSA of glyceraldehyde 3-phosphate dehydrogenases: example of high conservation

---

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

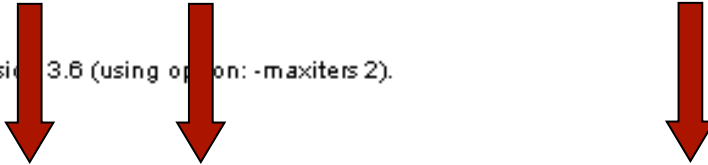
# MSA may need gaps

1: HomoloGene:1330. Gene conserved in Euteleostomi

Downlo

## Multiple Sequence Alignment

Generated by MUSCLE [\[see reference\]](#) version 3.6 (using option: -maxiters 2).



<a href="#">NP_001744.2</a>	1	MSGGKYVDS---EGHLYTVP	IREQ	NIYKPN	NKAM-	ADELSEKQVYDAHT	46																		
<a href="#">XP_519325.2</a>	1	MSGGKYVDS---EGHLYTVP	IREQ	NIYKPN	NKAM-	ADELSEKQVYDAHT	46																		
<a href="#">NP_001003296.1</a>	1	MSGGKYVDS---EGHLYTVP	IREQ	NIYKPN	NKAM-	AEEMSEKQVYDAHT	46																		
<a href="#">NP_776429.1</a>	1	MSGGKYVDS---EGHLYTVP	IREQ	NIYKPN	NKAM-	AEEMNEKQVYDAHT	46																		
<a href="#">NP_031642.1</a>	1	MSGGKYVDS---EGHLYTVP	IREQ	NIYKPN	NKAM-	ADEVTEKQVYDAHT	46																		
<a href="#">NP_113744.1</a>	1	MSGGKYVDS---EGHLYTVP	IREQ	NIYKPN	NKAM-	ADEVNEKQVYDAHT	46																		
<a href="#">XP_001234148.1</a>	1	---MEYFQ---	EAF	LYAAP	VREQ	NIYKPN	NKMM-	ADELSEKAVHDVHT	42																
<a href="#">NP_997816.1</a>	1	MTSG-	YKDG	TPEEE	YAH	SPIR	KQGN	IYKPN	NKEMD	NDS	INEK	TLQ	DVHT	49											
<a href="#">NP_001744.2</a>	47	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAEPE	GTHS	FDGI	WKAS	F	T	T	F	T	V	T	K	96					
<a href="#">XP_519325.2</a>	47	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAEPE	GTHS	FDGI	WKAS	F	T	T	F	T	V	T	K	96					
<a href="#">NP_001003296.1</a>	47	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAEPE	GTHS	FDGI	WKAS	F	T	T	F	T	V	T	K	96					
<a href="#">NP_776429.1</a>	47	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAEPE	GTHS	FDGI	WKAS	F	T	T	F	T	V	T	K	96					
<a href="#">NP_031642.1</a>	47	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAEPE	GTHS	FDGI	WKAS	F	T	T	F	T	V	T	K	96					
<a href="#">NP_113744.1</a>	47	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAEPE	GTHS	FDGI	WKAS	F	T	T	F	T	V	T	K	96					
<a href="#">XP_001234148.1</a>	43	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAEPE	GTHS	FDGI	WKAS	F	T	T	F	T	V	T	K	92					
<a href="#">NP_997816.1</a>	50	KEIDL	VNRD	PKHL	NDDV	VKID	FEDV	IAE	PAG	TYS	FDG	V	W	K	A	S	F	T	T	F	T	V	T	K	99

# Yet another example

```

NP 061485.1      1  -----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
XP 855587.1      1  -----MQAIKCVVVEDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
NP 776588.1      1  -----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
NP 033033.1      1  -----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
NP 599193.1      1  -----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
NP 990348.1      1  -----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
NP 956065.1      1  -----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
NP 648121.1      1  -----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM 45
XP 366655.1      1  MAAPGVQSLKCVVTGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSASVM 50
XP 329350.1      1  MLTGEMLTLDLFL-----TCLLISYTTNAFPGEYIPTVFDNYSASVM 43
NP 195320.1      1  --MSASRF IKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV 48
NP 179371.1      1  --MSASRF IKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV 48
NP 190698.1      1  --MSASRFVKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV 48
NP 195228.1      1  --MSASRF IKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVI 48
NP 001048639.1  1  --MSASRF IKCVTVGDGAVGKTCMLISYTSNTFPTDYVPTVFDNFSANVV 48

```

```

NP 061485.1      46  VDGKPVNLGLWDTAGQEDYDRLRPLSYPQTVGETYGKDITSRGKDKPIAD 95
XP 855587.1      46  VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D 76
NP 776588.1      46  VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D 76
NP 033033.1      46  VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D 76
NP 599193.1      46  VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D 76
NP 990348.1      46  VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D 76
NP 956065.1      46  VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D 76
NP 648121.1      46  VDAKPINLGLWDTAG-----D 76
XP 366655.1      51  VDGKPIISLGLWDTAG-----D 76
XP 329350.1      44  VDGKPVSLGLWDTAG-----D 76
NP 195320.1      49  VNGATVNLGLWDTAG-----D 76
NP 179371.1      49  VNGATVNLGLWDTAG-----D 76
NP 190698.1      49  VNGSTVNLGLWDTAGQEDYNRLRPLSYRGA-----D 79
NP 195228.1      49  VDGNTINLGLWDTAGQEDYNRLRPLSYRGA-----D 79
NP 001048639.1  49  VDGSTVNLGLWDTAGQEDYNRLRPLSYRGA-----D 77

```



This insertion could be due to alternative splicing

# Multiple Alignments: CLUSTALW

- \* identical
- : conserved substitutions
- . semi-conserved substitutions

```

gi|2213819          CDN-ELKSEAIIEHLCASEFALR-----MKIKEVKKENGDKK 223
gi|12656123        -----ELKSEAIIEHLCASEFALR-----MKIKEVKKENGD-   31
gi|7512442          CKNKNDNDNDIMETLCKNDFALK-----IKVKEITYINRDTK 211
gi|1344282          QDECKFDYVEVYETSSSGAFSLLGRFCGAEPPLVSSHHELAVLFRTDH 400
  
```

: . : \* . . \*:\* . :\*:

- Red: AVFPMLW (Small & hydrophobic)
- Blue: DE (Acidic)
- Magenta: RHK (Basic)
- Green: STYHCNGQ (Hydroxyl, Amine, Basic)
- Gray: Others

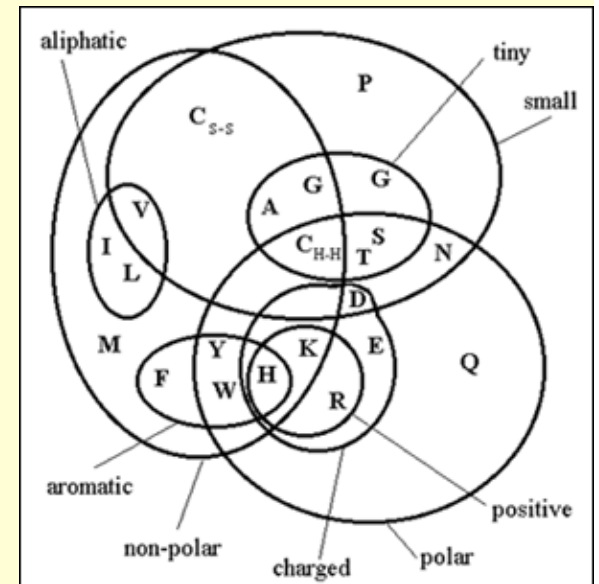


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

# MSA: Progressive Method

- Perform global pairwise alignments
- Build guide tree
- Progressively align the sequences

# How to Score Multiple Alignments?

## □ Sum of Pairs Score (SP)

- Optimal alignment:  $O(d^N)$  [Dynamic Prog]
- Approximate Algorithm: **Approx Ratio 2**
  - Locate Center:  $O(d^2N^2)$
  - Locate Consensus:  $O(d^2N^2)$

**Consensus char**: char with min distance sum

**Consensus string**: string of consensus char

**Center**: input string with min distance sum

# Multiple Alignment Methods

- Phylogenetic Tree Alignment (NP-Complete)
  - Given tree, task is to label leaves with strings
- Iterative Method(s)
  - Build a MST using the distance function
- Clustering Methods
  - Hierarchical Clustering
  - K-Means Clustering

## Multiple Alignment Methods (Cont'd)

### □ Gibbs Sampling Method

- Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science*, 1993

### □ Hidden Markov Model

- Krogh, Brown, Mian, Sjolander, Haussler, *JMB*, 1994



# Multiple Sequence Alignments (MSA)

## □ Choice of Scoring Function

- Global vs local
- Gap penalties
- Substitution matrices
- Incorporating other information
- Statistical Significance

## □ Computational Issues

- Exact/heuristic/approximate algorithms for optimal MSA
- Progressive/Iterative/DP
- Iterative: Stochastic/Non-stochastic/Consistency-based

## □ Evaluating MSAs

- Choice of good test sets or benchmarks (BALiBASE)
- How to decide thresholds for good/bad alignments