Schedule Week 3
4 – 6 September

- ***Wednesday***
  - Quiz
  - Alignments /Scoring Systems
- ***Friday –***
  - Searching for homologs

Scoring Systems

PAM Matrices

- ***PAM means "percent accepted mutations"***
  - accepted means fixed in the population and is therefore a more complex process than simply mutation
- ***PAM-1 therefore is a scoring system for sequences in which 1% of the residues have undergone mutation***
- ***PAM-250 represents 250% mutation, i.e., an average of 2.5 accepted mutation per residue - a very distant relationship***
- ***PAM tries to model what happens at long evolutionary distances based on a simple Markov model derived from closely related sequences.***

Scoring Systems

PAM Matrices

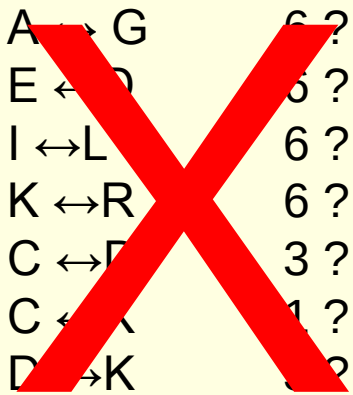
- ***Accepted point mutations - tabulate actual mutations by looking at proteins that are sufficiently closely related that there is no ambiguity in alignment***
  - Relying on actual observed mutations is why we call it empirical
  - Sequences no more than 15% different so that changes can be thought of as a single evolutionary step – no multiple mutations or back mutations
  - 1572 changes in 71 families
  - Consider a tree to correctly count changes

Scoring Systems

PAM Matrices

- *Counting mutations. How many changes in these sequences?*

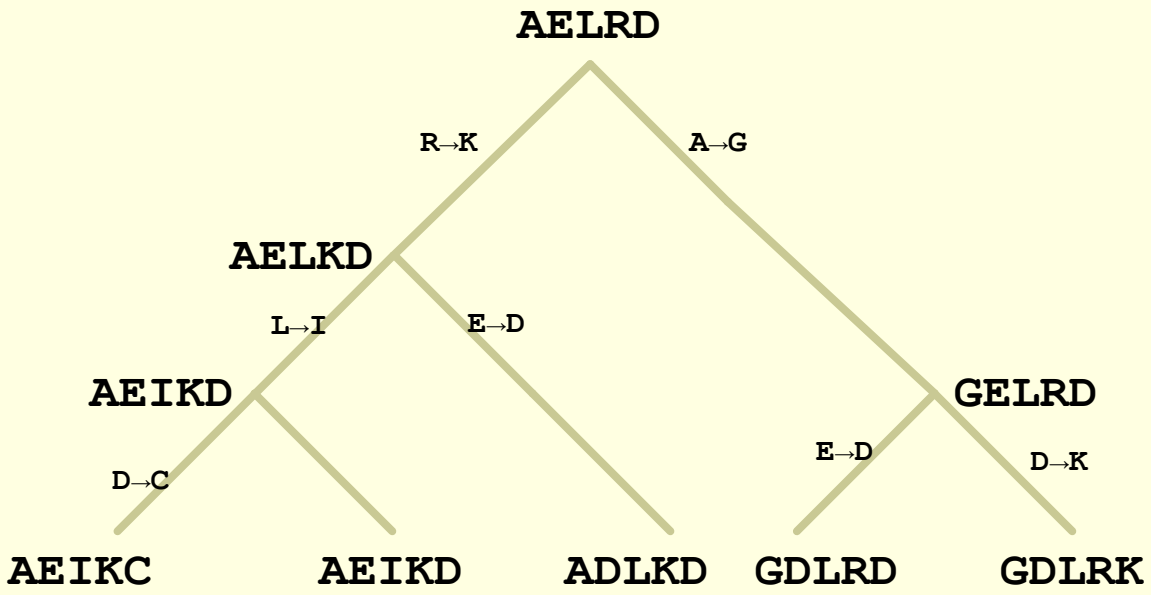
AEIKC
AEIKD
ADLKD
GDLRD
GDLRK



Scoring Systems

PAM Matrices

- ***Counting mutations (correctly)***



A→G	1
D →C	1
D →K	1
E →D	2
L →I	1
R →K	1

From Previous page

A ↔ G	6 ?
D ↔C	3 ?
D ↔K	3 ?
E ↔D	6 ?
L ↔I	6 ?
K ↔R	6 ?
C ↔K	1 ?

See also fig 5.1 in text

Scoring Systems

PAM matrices

- ***The goal is to make a matrix that reflects 1% change***
  - data were collected for amounts of change up to 15% so they have to be scaled
  - first calculate *mutability* – how likely is it that each kind of residue undergoes a change
  - then use eq. 5.6 to find a constant Λ so that the total change is 1%

$$0.01 = \sum_b f_b(1 - \Lambda m_b)$$

where f_b is the frequency of residue b and m_b is its mutability

  - A mutation probability matrix can then be calculated by multiplying this 1 PAM matrix by itself some number of times. This is a *Markov process*, we assume that the only thing that affects the observed mutation is the current residue

Scoring Systems

PAM Matrices

- ***Mutation probability matrix - probability that residue in column b will be replaced by residue in row a after some amount of evolution***

$$M_{bb} = 1 - \Lambda m_b \qquad M_{ab} = \frac{\Lambda m_b A_{ab}}{\sum_a A_{ab}}$$

m_b = mutability of residue b (probability of mutating)
A_{ab} = number of accepted point mutations
Λ = proportionality constant

Scoring Systems

PAM Matrices

- ***Mutation Probability Matrix***
  - Probability that the residue represented by the column will mutate into the residue represented by the row after a specified amount of mutation, for instance, 1 PAM
  - Not symmetric. The probability of A→S ≠ S → A
- ***Relatedness odds matrix***
  - Log-odds form of the mutation probability matrix
    - Log-odds matrix compares what is expected of homologous sequences to what is expected of unrelated (random) sequences.
  - Remember that a log-odds matrix compares two models, in this case a model of relationship by homology (the mutation probability matrix) and relationship by chance

$$R_{ij} = log(M_{ij} / f_i)$$

  - Where f_i is the frequency of amino acid residue *i* at random

Scoring Systems

PAM Matrices

- *1 PAM matrix can be extended to simulate longer distances by multiplying it by itself.*
- *Dayhoff recommended PAM 250*
- *More modern recommendations suggest PAM 125 is a better general choice*
- *Which is best depends on the sequences!*

Scoring Systems

PAM Matrices

- *Original PAM 250 matrix (compare fig 5.3)*

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

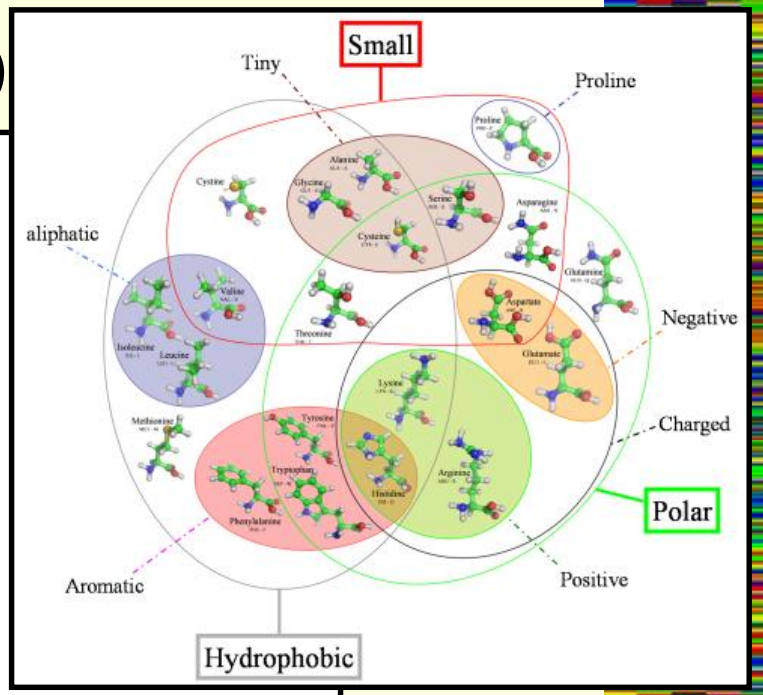
Small

Acid & Amide

Basic

Hydrophobic

Aromatic



Scoring Systems

PAM Matrices

- ***Problems with the PAM approach:***
  - Not all positions are the same, e.g., internal vs external
    - Evolutionary rates vary greatly within a sequence
    - Each position has a unique three dimensional environment
    - Environment changes over evolutionary time as surrounding residues change
  - The most mutable positions were inadvertently selected as the basis for the calculation
    - proteins change more rapidly at the least constrained positions and most slowly at buried “core” positions

Scoring Systems

BLOSUM (BLOcks SUbstitution Matrix)

- ***Based on PROSITE signatures***
  - Signatures are short expressions like C-X-X-C-X-X-X-S-T
- ***Locally align sequences to each signature to get "blocks“***
- ***Blocks are locally conserved regions, i.e., more constrained regions likely to be related to structure/function***
- ***Blocks contain sequences at all different evolutionary distances and may be highly biased (e.g., many identical sequences)***

Scoring Systems

BLOSUM Matrices (Ch 5.1)

- ***Dealing with bias and distance***
  - Cluster all sequences with less than X% identities
    - Clustered sequences count as 1 sequence
    - if X is 100% it simply removes identical sequences
    - if X < 100% it reduces the weight on closely related sequences
  - Calculate substitution frequencies and log-odds matrix
- ***This gives a BLOSUM X table***
  - BLOSUM 62 - sequences greater than 62% identical are clustered
  - BLOSUM 80 - sequences greater than 80% identical are clustered

Scoring Systems

BLOSUM Matrices

- *METHYLTRANSFERASE BI*

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47
CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

Unweighted (BLOSUM 100) count of transitions for column 1, total $(n^2 - n)/2$ transitions

$c_{FF} = 0$ $c_{FI} = 1$ $c_{FL} = 4$ $c_{FV} = 2$

$c_{II} = 0$ $c_{IL} = 4$ $c_{IV} = 2$

$c_{LL} = 6$ $c_{LV} = 8$

$c_{VV} = 1$

Scoring Systems

BLOSUM Matrices

- ***Unweighted (BLOSUM 100) count (c_{ij}) of transitions for column 1***

N=28	F	I	L	V
F	0	1	4	2
I		0	4	2
L			6	8
V				1

- ***N = 28 transitions, $f_{ii} = c_{ij} / N$ (foreground frequencies)***

	F	I	L	V
F	0.000	0.036	0.143	0.071
I		0.000	0.143	0.071
L			0.214	0.286
V				0.036

- ***Log-Odds $s_{ij} = \log_2(f_{ij} / p_i p_j)$ - Background frequencies, p_i, from database***
  - $p_F = 0.0397$ $p_I = 0.0529$ $p_L = 0.0917$ $p_V = 0.0649$

	F	I	L	V
F	undef	3.1	4.3	3.8
I		undef	3.9	3.4
L			4.7	4.6
V				3.1

Scoring Systems

BLOSUM Matrices

- ***Background frequencies are calculated from the overall frequencies of the letters – a random background model***
  - in the database (as on previous slide)
  - or in the alignment (if the alignment is large enough)

$p_{ii} = p_i^2$ $\quad$ $p_{ij} = 2\, p_i\, p_j$

- ***if the probabilities of the letters, p_i, are***
  - $p_F = 0.0397$ $\quad$ $p_I = 0.0529$ $\quad$ $p_L = 0.0917$ $\quad$ $p_V = 0.0649$

	F	I	L	V
F	0.0016	0.0042	0.0072	0.0052
I		0.0027	0.0097	0.0069
L			0.0084	0.0119
V				0.0042

Scoring Systems

BLOSUM Matrices

- ***Avoiding undefined logs***
- ***Undefined logs occur when either the foreground or background frequencies are zero***
- ***The problem arises because the sample is too small – rare items (residues) or events (mutations) are so uncommon they do not occur in the sample***
- ***Solution: use a pseudocount***
  - add one to every count (AKA plus one prior)
  - add a total of one distributed according to background

Scoring Systems

BLOSUM Matrices

- ***Unweighted (BLOSUM 100) with plus one prior***

N=38	F	I	L	V
F	1	2	5	3
I		1	5	3
L			7	9
V				2

- ***N = 38 transitions, $f_{ij} = c_{ij} / N$ (foreground frequencies)***

	F	I	L	V
F	0.026	0.053	0.132	0.079
I		0.026	0.132	0.079
L			0.184	0.237
V				0.053

	F	I	L	V
F	0.000	0.036	0.143	0.071
I		0.000	0.143	0.071
L			0.214	0.286
V				0.036

without pseudocount

- ***Log-Odds $s_{ij} = log2(f_{ij} / p_i p_j)$ - Background frequencies, p_i, from database***

	F	I	L	V
F	4.0	3.6	4.2	3.9
I		3.3	3.8	3.5
L			4.5	4.3
V				3.7

	F	I	L	V
F	undef	3.1	4.3	3.8
I		undef	3.9	3.4
L			4.7	4.6
V				3.1

without pseudocount

Scoring Systems

BLOSUM Matrices – BLOSUM 80

- *Looking for sequences > 80% identical*

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47
CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

Percent Identity

		1	2	3	4	5	6	7
0	TCMN_STRGA	39.4	33.3	36.4	39.4	36.4	48.5	36.4
1	TCMO_STRGA		33.3	30.3	30.3	30.3	36.4	45.5
2	ZRP4_MAIZE			48.5	45.5	48.5	42.4	42.4
3	CHMT_POPTM				81.8	93.9	48.5	45.5
4	COMT_EUCGU					78.8	51.5	48.5
5	COMT_MEDSA						48.5	45.5
6	CRTF_RHOSH							36.4
7	OMTA_ASPPA							

Scoring Systems

BLOSUM Matrices – BLOSUM 80

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68
CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47

CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47
```

each sequence counts as one

three sequences together count as one

N=15	F	I	L	V
F	0	1	2.67	1.33
I		0	2.67	1.33
L			2.33	3.33
V				0.33

N=25	F	I	L	V
F	1	2	3.67	2.33
I		1	3.67	2.33
L			3.33	4.33
V				1.33

with pseudocount

- ***Equivalent to 6 sequences so number of transitions should equal 15 (6*5/2)***

Scoring Systems

BLOSUM Matrices – BLOSUM80

- *foreground pseudocounts from previous slide*

N=25	F	I	L	V
F	1	2	3.67	2.33
I		1	3.67	2.33
L			3.33	4.33
V				1.33

N=25	F	I	L	V
F	0.040	0.080	0.147	0.093
I		0.040	0.147	0.093
L			0.133	0.173
V				0.053

- *Log-odds (same background as before)*

	F	I	L	V
F	4.6	4.3	4.4	4.2
I		3.9	3.9	3.8
L			4.0	3.9
V				3.7

	F	I	L	V
F	4.0	3.6	4.2	3.9
I		3.3	3.8	3.5
L			4.5	4.3
V				3.7

BLOSUM100 (ungrouped)

Scoring Systems

BLOSUM Matrices

- ***Derived from a very large set of conserved sequence motifs.***
- ***Represents more core of protein and less surface than PAM***
- ***BLOSUM is always a blend of evolutionary distances***

PAM Matrices

- ***Derived from a relatively small set of closely related, small, and mostly globular proteins***
  - Note that JTT update (discussed in text) does not consider trees
- ***Biased towards surface***
- ***Cover entire evolutionary range from identical to random***

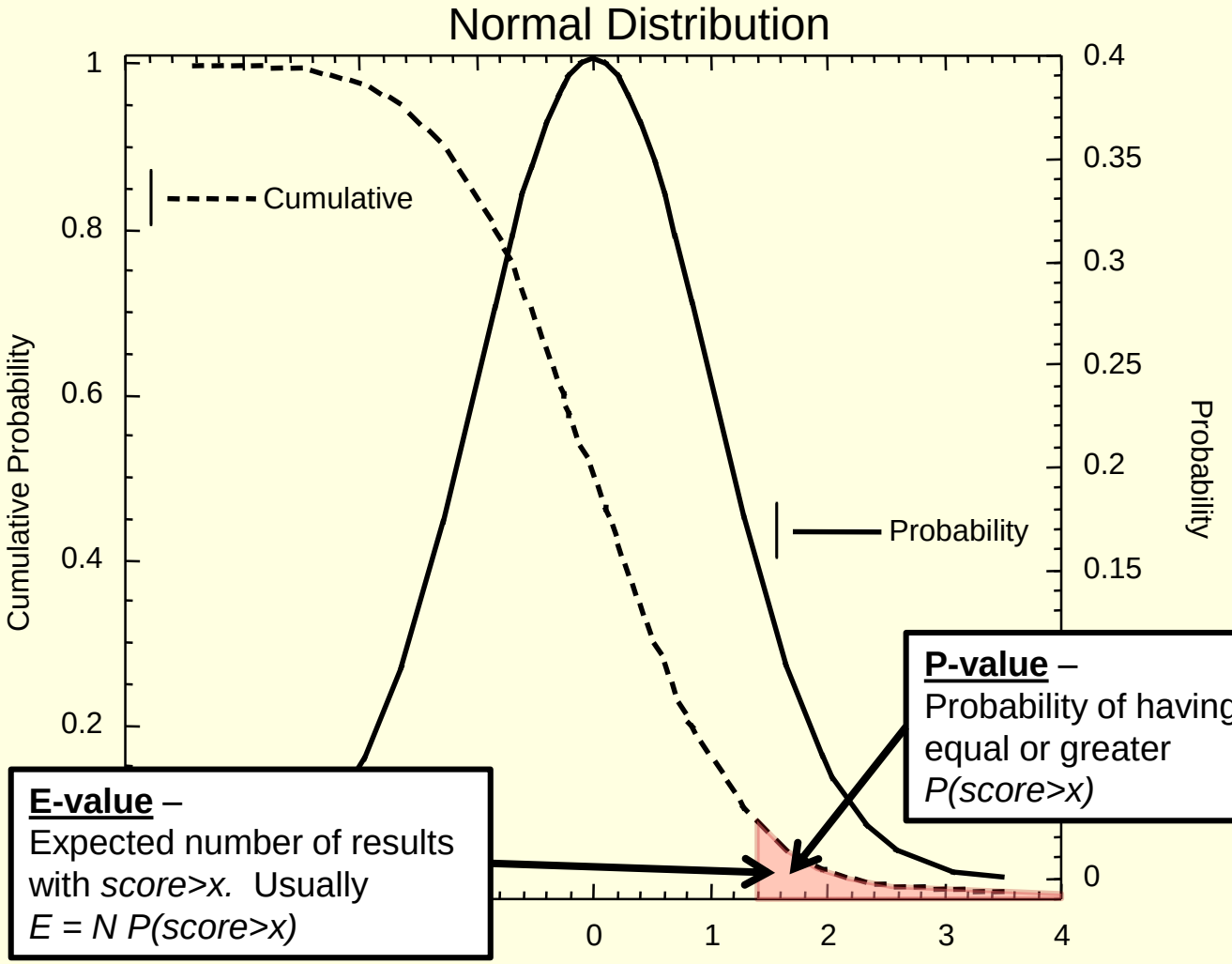
Scoring Systems

Target Frequencies

- *Alignment programs "try" to produce alignments in which the letter-pairs occur such that the highest possible score is produced*
- *EVERY scoring matrix implies a set of target frequencies*
- *The best alignments are produced when the target frequencies implied by the scoring matrix are the same as the correct alignment*
- *Disagreement between the scoring matrix target frequencies, and the biologically correct pair frequencies is one of the main reasons that the mathematically optimal alignment is not necessarily the biologically correct one (approximation made for gaps is the other)*

Sequence Comparison

Alignment Significance

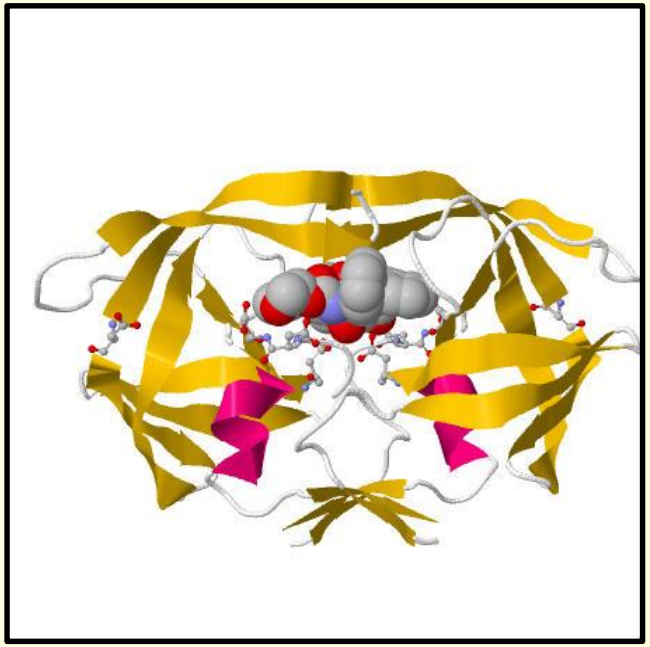


E-value –
Expected number of results with *score>x.* Usually *E = N P(score>x)*

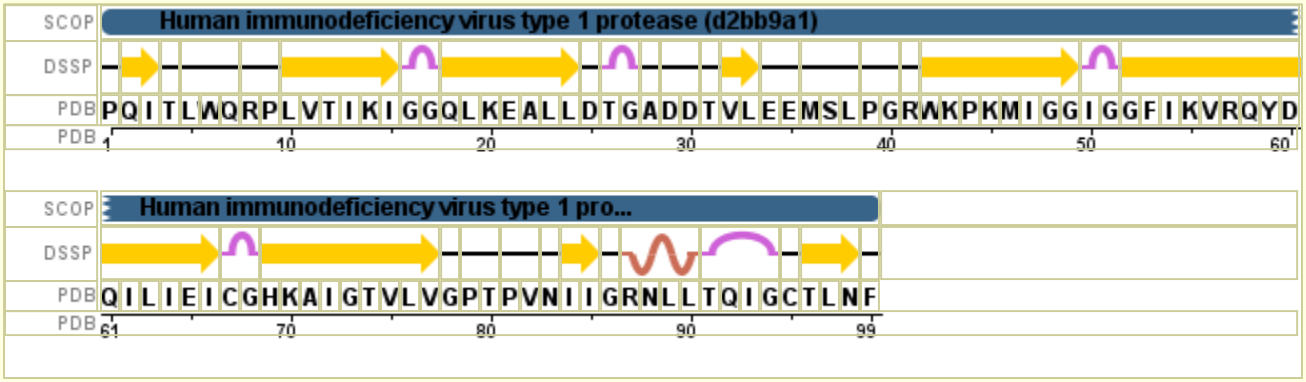
P-value –
Probability of having a score equal or greater *P(score>x)*

Sequence Comparison

Practical Alignments



- ***HIV-1 protease***
  - Cleaves viral polyprotein into active proteins
  - 99 residues
  - Aspartic protease
  - Active site residues 25-27
  - Active site flap: 46..50,52..56



Sequence Comparison

Practical Alignments

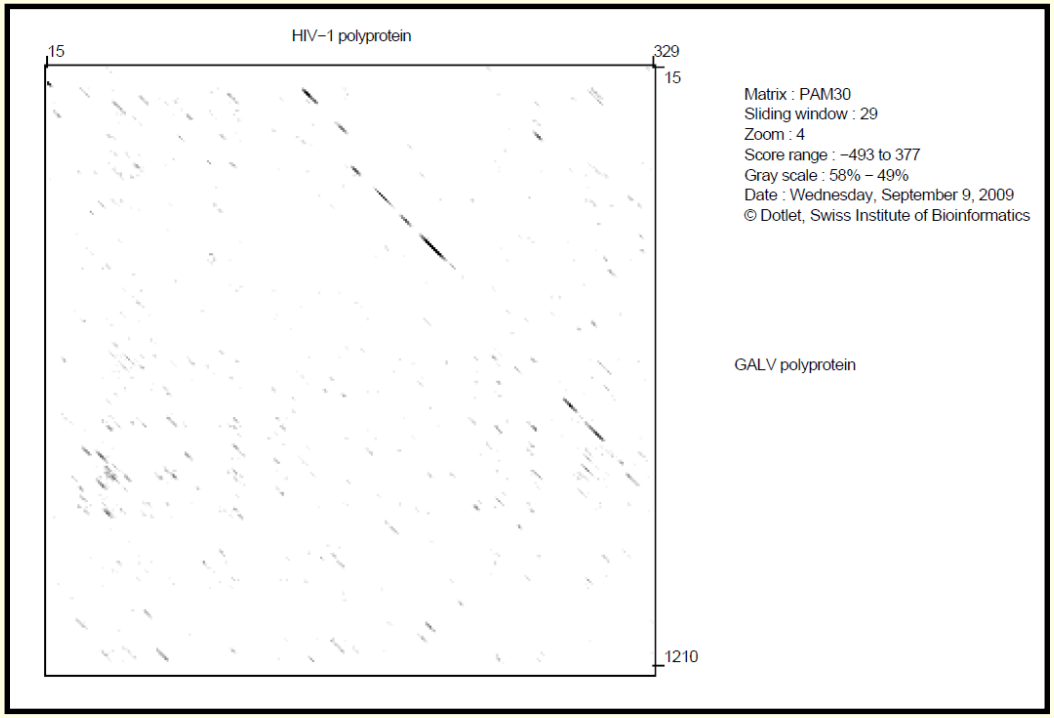
- ***GALV – Gibbon Ape Leukemia Virus***
  - Does it have the same mechanism (viral protease cleaves polyprotein)?
  - Where is protease active site?
  - Does it recognize the same cleavage site?

```
>P21414 Gibbon Ape Leukemia virus (GALV) entire polyprotein
gsqgsdplpe prvtltvegt pieflvdtga ehsvltqpmg kvgsrrtvve gatgskvypw
ttkrllkigh kqvthsflvi pecpapllgr dlltklkaqi qfsaegpqvt wgerptmclv
lnleeeyrlh ekpvpssidp swlqlfptvw aeragmglan qvppvvvelr sgaspvavrq
ypmskeareg irphiqkfld lgvlvpcrsp wntpllpvkk pgtndyrpvq dlreinkrvq
dihptvpnpy nllsslppsy twysvldlkd affclrlhpn sqplfafewk dpekgntgql
twtrlpqgfk nsptlfdeal hrdlapfral npqvvllqyv ddllvaapty edckkgtqkl
lqelsklgyr vsakkaqlcq revtylgyll kegkrwltpa rkatvmkipv pttprqvref
lgtagfcrlw ipgfaslaap lypltkesip fiwteehqqa fdhikkalls apalalpdlt
kpftlyider agvargvltq tlgpwrrpva ylskkldpva sgwptclkav aavalllkda
dkltlgqnvt viashslesi vrqppdrwmt narmthyqsl llnervsfap pavlnpatll
pveseatpvh rcseilaeet gtrrdledqp lpgvptwytd gssfitegkr ragapivdgk
rtvwasslpe gtsaqkaelv altqalrlae gkniniytds ryafatahih gaiykqrgll
tsagkdiknk eeilalleai hlprrvaiih cpghqrgsnp vatgnrrade aakqaalstr
vlagttkpqe piepaqektr preltpdrgk efikrlhqlt hlgpekllql vnrtsllipn
lqsavrevts qcqacamtna vttyretgkr qrgdrpgvyw evdfteikpg rygnkyllvf
idtfsgwvea fptktetali vckkileeil prfgipkvlg sdngpafvaq vsqglatqlg
inwklhcayr pqssgqverm nrtiketltk laletggkdw vtllplallr arntpgrfgl
tpyeilyggp ppilesgetl gpddrflpvl fthlkaleiv rtqiwdqike vykpgtvtip
hpfqvgdqvl vrrhrpssle prwkgpylvl lttptavkvd giaawvhash lkpappsapd
eswelektdh plklrirrrr desak
```

Sequence Comparison

Practical Alignments

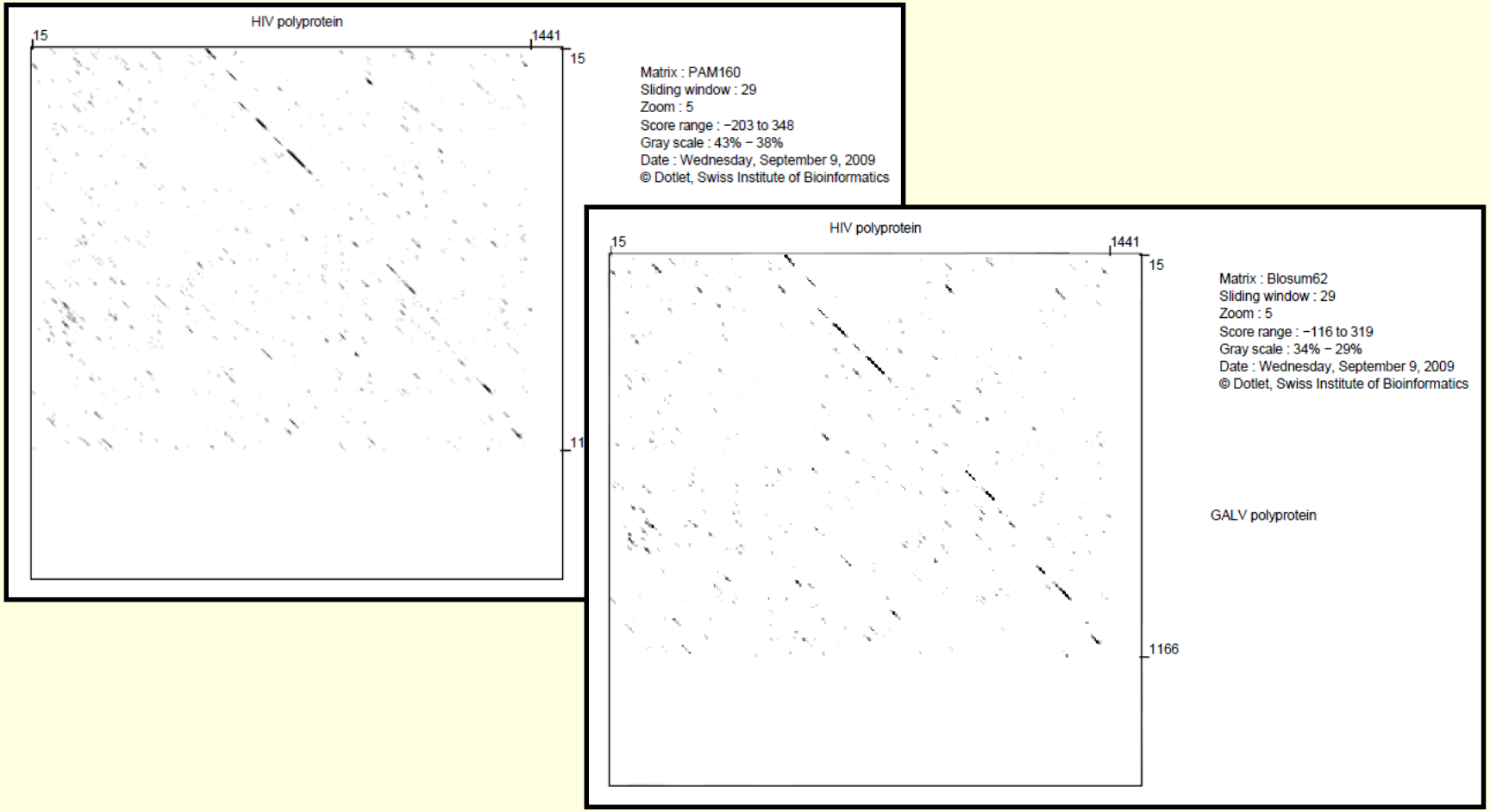
- *HIV-1 polyprotein vs GALV polyprotein*



Sequence Comparison

Practical Alignments

- *HIV-1 polyprotein vs GALV polyprotein*



Sequence Comparison

Practical Alignments

- **Global alignment of HIV and GALV polyprotein**

```
ALIGN calculates a global alignment of two sequences
 version 2.2u
Please cite: Myers and Miller, CABIOS (1989) 4:11-17
hiv-1 protease 99 bp                                  99 aa vs.
GALV polyprotein 1165 bp                            1165 aa
 using matrix file: BLOSUM50, gap open/ext: -14/-4
   3.7% identity in 1165 aa overlap;                        Global score: -4163


hiv-1  ----------PQITL-----------------------------------------------
                 :..::
GALV   GSQGSDPLPEPRVTLTVEGTPIEFLVDTGAEHSVLTQPMGKVGSRRTVVEGATGSKVYPW

hiv-1  ------------------------------------------------------------
GALV   TTKRLLKIGHKQVTHSFLVIPECPAPLLGRDLLTKLKAQIQFSAEGPQVTWGERPTMCLV

hiv-1  ------------------------------------------------------------
GALV   LNLEEEYRLHEKPVPSSIDPSWLQLFPTVWAERAGMGLANQVPPVVVELRSGASPVAVRQ

hiv-1  ------------------------------WQRPLVTVKIGG------------------
                                     :. ::. ::  :
GALV   YPMSKEAREGIRPHIQKFLDLGVLVPCRSPWNTPLLPVKKPGTNDYRPVQDLREINKRVQ

hiv-1  ------------------------------------------------------------
GALV   DIHPTVPNPYNLLSSLPPSYTWYSVLDLKDAFFCLRLHPNSQPLFAFEWKDPEKGNTGQL

hiv-1  ------------------------------------------------------------
GALV   TWTRLPQGFKNSPTLFDEALHRDLAPFRALNPQVVLLQYVDDLLVAAPTYEDCKKGTQKL

hiv-1  ------------------------------------------------------------
GALV   LQELSKLGYRVSAKKAQLCQREVTYLGYLLKEGKRWLTPARKATVMKIPVPTTPRQVREF
20
hiv-1  ------------------------------------------QLKEALLDT---------
                                                  ..:.::::..
GALV   LGTAGFCRLWIPGFASLAAPLYPLTKESIPFIWTEEHQQAFDHIKKALLSAPALALPDLT

hiv-1  ------------------------------------------------------------
GALV   KPFTLYIDERAGVARGVLTQTLGPWRRPVAYLSKKLDPVASGWPTCLKAVAAVALLLKDA

hiv-1  -----GADDTV-------------------------------------------------
            : . ::
GALV   DKLTLGQNVTVIASHSLESIVRQPPDRWMTNARMTHYQSLLLNERVSFAPPAVLNPATLL

hiv-1  -------------------------LEDINLPG----------------------------
                                :::  :::
GALV   PVESEATPVHRCSEILAEETGTRRDLEDQPLPGVPTWYTDGSSFITEGKRRAGAPIVDGK

hiv-1  ------------------------------------------------------------
GALV   RTVWASSLPEGTSAQKAELVALTQALRLAEGKNINIYTDSRYAFATAHIHGAIYKQRGLL

hiv-1  ------------------------------------------------------------
GALV   TSAGKDIKNKEEILALLEAIHLPRRVAIIHCPGHQRGSNPVATGNRRADEAAKQAALSTR

hiv-1  -----------------KWKPKMIGGIGG--FIK---------------------------
                        : .:. .    :  :::
GALV   VLAGTTKPQEPIEPAQEKTRPRELTPDRGKEFIKRLHQLTHLGPEKLLQLVNRTSLLIPN

hiv-1  ----VRQYD-------------------------------------------------------
           ::.
GALV   LQSAVREVTSQCQACAMTNAVTTYRETGKRQRGDRPGVYWEVDFTEIKPGRYGNKYLLVF

hiv-1  ----------------------QILIEICGKKAIGTVL---VGPTPVNIIGRNMLTQIG
                             .:: ::  . .:  ::    ::. :  ..... ::.:
GALV   IDTFSGWVEAFPTKTETALIVCKKILEEILPRFGIPKVLGSDNGPAFVAQVSQGLATQLG

hiv-1  CTLNF------------------------------------------------------------
        . ..
GALV   INWKLHCAYRPQSSGQVERMNRTIKETLTKLALETGGKDWVTLLPLALLRARNTPGRFGL

hiv-1  ------------------------------------------------------------
GALV   TPYEILYGGPPPILESGETLGPDDRFLPVLFTHLKALEIVRTQIWDQIKEVYKPGTVTIP

hiv-1  ------------------------------------------------------------
GALV   HPFQVGDQVLVRRHRPSSLEPRWKGPYLVLLTTPTAVKVDGIAAWVHASHLKPAPPSAPD

hiv-1  -------------------------
GALV   ESWELEKTDHPLKLRIRRRRDESAK
```

Sequence Comparison

Practical Alignments

- HIV polyprotein vs GALV polyprotein

```
using matrix file: BL50 (15/-5), gap-open/ext: -14/-4 E(limit)    0.05

  29.0% identity in 93 aa overlap (501-592:10-102); score:  107 E(10000):  0.029

              510       520       530       540       550       560
hiv-1  RPVVTVRVGGQLKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQILIEIC
       .: ::. : :    : :.::::. .:: .        .  .. :  :         . :..:
GALV   EPRVTLTVEGTPIEFLVDTGAEHSVLTQPMGKVGSRRTVVEGATGSKVYPWTTKRLLKIG
      10        20        30        40        50        60

              570       580       590
hiv-1  GKKAIGTILVGPT-PVNIIGRNMLTQIGCTLNF
        :..  ..:: :  :. ..::..::..   ..:
GALV   HKQVTHSFLVIPECPAPLLGRDLLTKLKAQIQF
      70        80        90       100
```

A

```
27.9% identity in 247 aa overlap (595-835:160-399); score:  296 E(10000): 2.6e-21

          600       610       620       630       640       650
hiv-1  SPIETVPVKLKPGMDGPRVKQWPLTEEKIKALTEICKDMEKEGKILKIGPENPYNTPVFA
       . .  : :.:. : .   :.:.:...:  ...    . .   : ..    ..:.:::..
GALV   NQVPPVVVELRSGASPVAVRQYPMSKEAREGIRPHIQKFLDLGVLVPC--RSPWNTPLLP
     160       170       180       190       200         210

          660       670       680       690            700
hiv-1  IKKKDSTKWRKLVNFRELNKRTQDFWEVQLGIPHPAGLKKKKS-----VTVLDVGDAYFS
       .::  .. .: . ..::.:::.::   ..  .:.: .: ..         .:::. ::.:
GALV   VKKPGTNDYRPVQDLREINKRVQD---IHPTVPNPYNLLSSLPPSYTWYSVLDLKDAFFC
       220       230       240           250       260       270

     710       720       730       740       750       760
hiv-1  VPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRTKNPE
       . :  . .   ::   . .. . : .  .. ::::.:.::..:. .. . : :::. ::.
GALV   LRLHPNSQPLFAFEWKDPEKGNTG-QLTWTRLPQGFKNSPTLFDEALHRDLAPFRALNPQ
          280       290       300       310       320       330

     770       780       790       800       810       820
hiv-1  IVIYQYMDDLYVGSDLEIGQHRTKIEELREHLLKWGFTTPDKKHQK-EPPFLWMGYELHP
       .:. ::.::: :..     . .   ..: ..: : :. .  :: :  .    ..:: :.
GALV   VVLLQYVDDLLVAAPT-YEDCKKGTQKLLQELSKLGYRVSAKKAQLCQREVTYLGYLLKE
          340       350       360       370       380       390

      830
hiv-1  DKWTVQP
        :  . :
GALV   GKRWLTP
```

B

```
40.0% identity in 80 aa overlap (1020-1095:625-704); score:  120 E(10000): 0.0014

    1020      1030      1040      1050      1060      1070
hiv-1  QLETEPIVGAETFYVDGAANRETKKGKAGYVTDRGRQKV--VSLTE-TTNQKTELQAIHL
       .:: .:. :. :.:.::..     : .::     :.. :   :: : :. ::.:: :.
GALV   DLEDQPLPGVPTWYTDGSSFITEGKRRAGAPIVDGKRTVWASSLPEGTSAQKAELVALTQ
          630       640       650       660       670       680

       1080      1090
hiv-1  ALQDS-GSEVNIVTDSQYAL
       ::. . :...:: :::.::.
GALV   ALRLAEGKNINIYTDSRYAF
          690       700
```

C

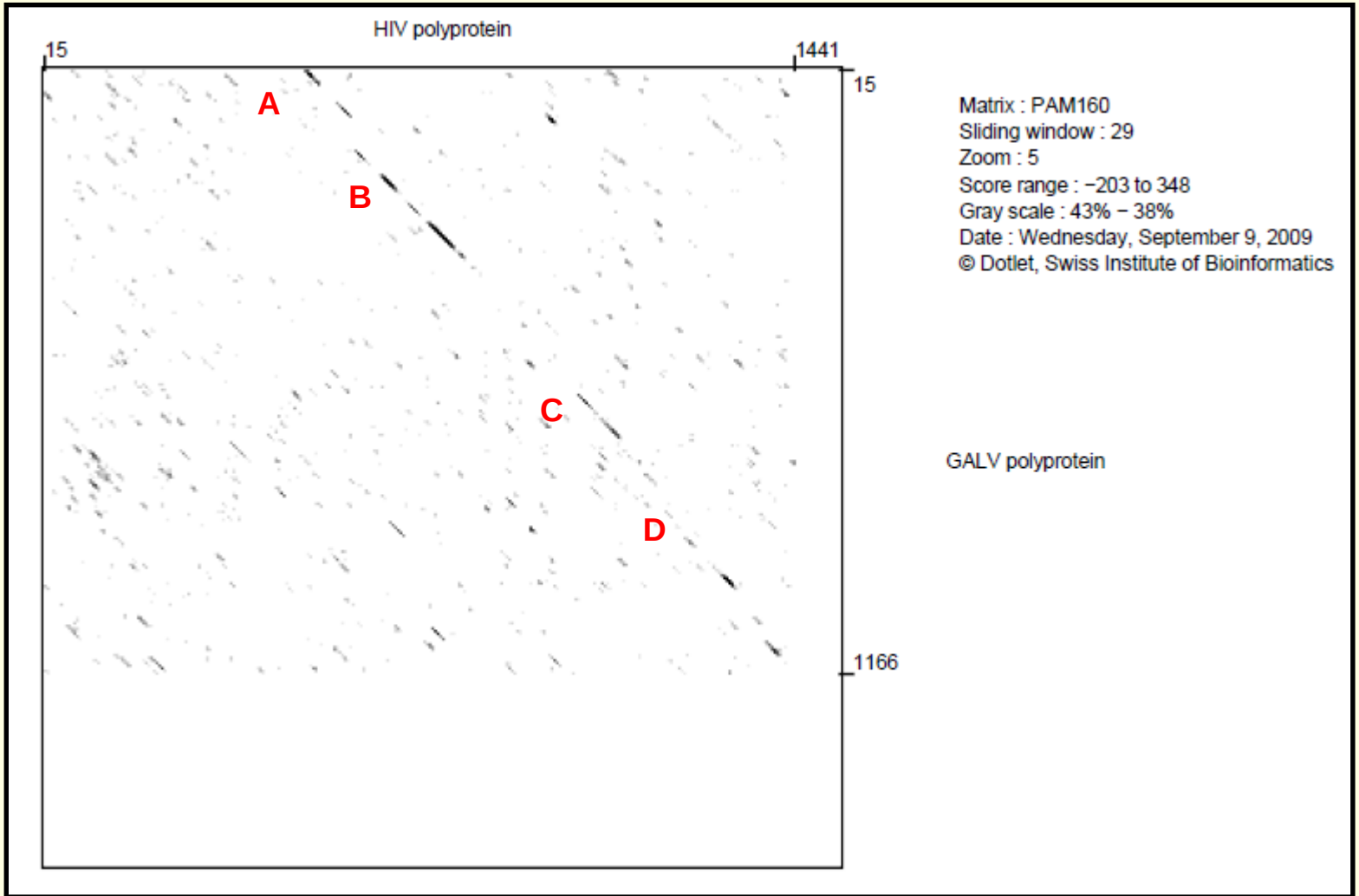
```
 25.7% identity in 148 aa overlap (1184-1317:844-991); score:  120 E(10000): 0.0014

         1190            1200      1210      1220           1230
hiv-1  VAKEIVASCDKCQLKG------EAMHGQVDCSPGI-WQLDCTHLE-----GKIIIVAVHV
       ...:....:. : . .      :. . :    ::. :..: :...     .: ..: . .
GALV   AVREVTSQCQACAMTNAVTTYRETGKRQRGDRPGVYWEVDFTEIKPGRYGNKYLLVFIDT
          850       860       870       880       890       900

          1240      1250       1260      1270      1280
hiv-1  ASGYIEAEVIPAETGQETAYFILK-LAGRWPV-KVVHTDNGSNFTSAAVKAACWWANIKQ
        ::..::    .::. .   ::. .  :. . ::. .:::  :.. . ..     .:.
GALV   FSGWVEAFPTKTETALIVCKKILEEILPRFGIPKVLGSDNGPAFVAQVSQGLATQLGINW
          910       920       930       940       950       960

    1290      1300      1310
hiv-1  EFGIPYNPQSQGVVESMNKELKKIIGQV
       ..   : :::.: :: ::. .:. . ..
GALV   KLHCAYRPQSSGQVERMNRTIKETLTKL
          970       980       990
```

D

Sequence Comparison

Practical Alignments

- *HIV-1 polyprotein vs GALV polyprotein*



Sequence Comparison

Practical Alignments

```
Statistics: (shuffled [500]) MLE statistics: Lambda= 0.0782;  K= 213.3
Threshold: E() < 1 score: 192
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.0 Oct 2008)
Parameters: BL50 matrix (15:-5), open/ext: -1/-1
 Scan time:  0.100

>>>AAA21284, 99 aa vs TMP.q2 library

>>P63120  HERV-K endogenous retrovirus type K,  protease                        (156 aa)
 Waterman-Eggert score: 266;  22.3 bits; E(1) <  0.0031
36.3% identity (51.6% similar) in 124 aa overlap (6-97:1-106)
Entrez Lookup  Re-search database  General re-search
                10           20          30          40                50           60
AAA212 W--Q----RPLVTVK--IGG-QLKEALLDTGAD-DTVLEDINL-PGK-WKPKM--I-G--GIG--GFIKVRQYDQ-ILIE
       :  :    :: :  :  : : :. :.:.:::::  ...  .:  : : : ::.  . :  :::  .  .:  : :    :
P63120 WASQVSENRP-VC-KAIIQGKQF-EGLVDTGADV-SIIA-LNQWP-KNW-PKQKAVTGLVGIGTAS--EV--Y-QST--E
               10          20          30          40          50          60

                70          80          90
AAA212 I--C-G---KKAIGTVLVGP--T--PVNIIGRNMLTQIGC--TL
       :  : :   ...  ::   :  :  :.:. ::..: : :   :.
P63120 ILHCLGPDNQES--TVQ--PMITSIPLNLWGRDLLQQWGAEITM
```

Sequence Comparison

Practical Alignments

```
Statistics: (shuffled [500]) MLE statistics: Lambda= 0.1032;  K=0.0432
Threshold: E() < 1 score: 63
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.0 Oct 2008)
Parameters: BL50 matrix (15:-5), open/ext: -5/-2
 Scan time:  0.090

>>>AAA21284, 99 aa vs TMP.q2 library

>>P63120  HERV-K endogenous retrovirus type K,  protease                        (156 aa)
 Waterman-Eggert score: 166;  29.3 bits; E(1) <  2.4e-05
35.3% identity (57.8% similar) in 102 aa overlap (2-94:4-101)
Entrez Lookup  Re-search database  General re-search
              10        20        30         40             50        60         70
AAA212 QITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINL-PGKW-KPKMI-G--GIGGFIKVRQYDQILIEIC-GKKAI-GT
       :..  .::.  . : :.  :.:.::::: ...  .:  : .: : : . :  :::   .: :  .::   : :     .:
P63120 QVSE-NRPVCKAIIQGKQFEGLVDTGADVSIIA-LNQWPKNWPKQKAVTGLVGIGTASEVYQSTEILH--CLGPDNQEST
             10        20        30         40        50        60          70

             80        90
AAA212 V--LVGPTPVNIIGRNMLTQIG
       :  ..   :.:. ::..: : :
P63120 VQPMITSIPLNLWGRDLLQQWG
      80        90        100
```

Sequence Comparison

Practical Alignments

```
Statistics: (shuffled [500]) MLE statistics: Lambda= 0.1745;  K=0.05741
Threshold: E() < 1 score: 39
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.0 Oct 2008)
Parameters: BL50 matrix (15:-5), open/ext: -10/-2
 Scan time:  0.070

>>>AAA21284, 99 aa vs TMP.q2 library

>>P63120  HERV-K endogenous retrovirus type K,  protease                        (156 aa)
 Waterman-Eggert score: 137;  38.6 bits; E(1) <  3.7e-08
33.0% identity (57.4% similar) in 94 aa overlap (7-94:8-101)
Entrez Lookup  Re-search database  General re-search
         10        20        30        40            50        60        70          80
AAA212 QRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKW-KPKMIGG---IGGFIKVRQYDQILIEICGKKAIGTV--LVGPT
       .::.  . : :.  :.:.:::::: ...   . : .: : : . :   ::   .: :  .::  .   .  .::  ..
P63120 NRPVCKAIIQGKQFEGLVDTGADVSIIALNQWPKNWPKQKAVTGLVGIGTASEVYQSTEILHCLGPDNQESTVQPMITSI
        10        20        30        40        50        60        70        80

               90
AAA212 PVNIIGRNMLTQIG
       :.:. ::..: : :
P63120 PLNLWGRDLLQQWG
        90       100
```

Sequence Comparison

Practical Alignments

```
Statistics: (shuffled [500]) MLE statistics: Lambda= 0.1904;  K=0.06576
Threshold: E() < 1 score: 36
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.0 Oct 2008)
Parameters: BL50 matrix (15:-5), open/ext: -12/-2

>>>AAA21284, 99 aa vs TMP.q2 library

>>P63120  HERV-K endogenous retrovirus type K,  protease                      (156 aa)
 Waterman-Eggert score: 131;  39.9 bits; E(1) <  1.5e-08
33.0% identity (57.4% similar) in 94 aa overlap (7-94:8-101)
Entrez Lookup  Re-search database  General re-search
         10        20        30        40            50        60        70          80
AAA212 QRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKW-KPKMIGG---IGGFIKVRQYDQILIEICGKKAIGTV--LVGPT
       .::.  . : :.  :.:.::::: ...   . : .: : : . :   ::   .: :  .::  .   .  .::  ..
P63120 NRPVCKAIIQGKQFEGLVDTGADVSIIALNQWPKNWPKQKAVTGLVGIGTASEVYQSTEILHCLGPDNQESTVQPMITSI
        10        20        30        40        50        60        70        80

               90
AAA212 PVNIIGRNMLTQIG
       :.:. ::..: : :
P63120 PLNLWGRDLLQQWG
        90       100
```

Sequence Comparison

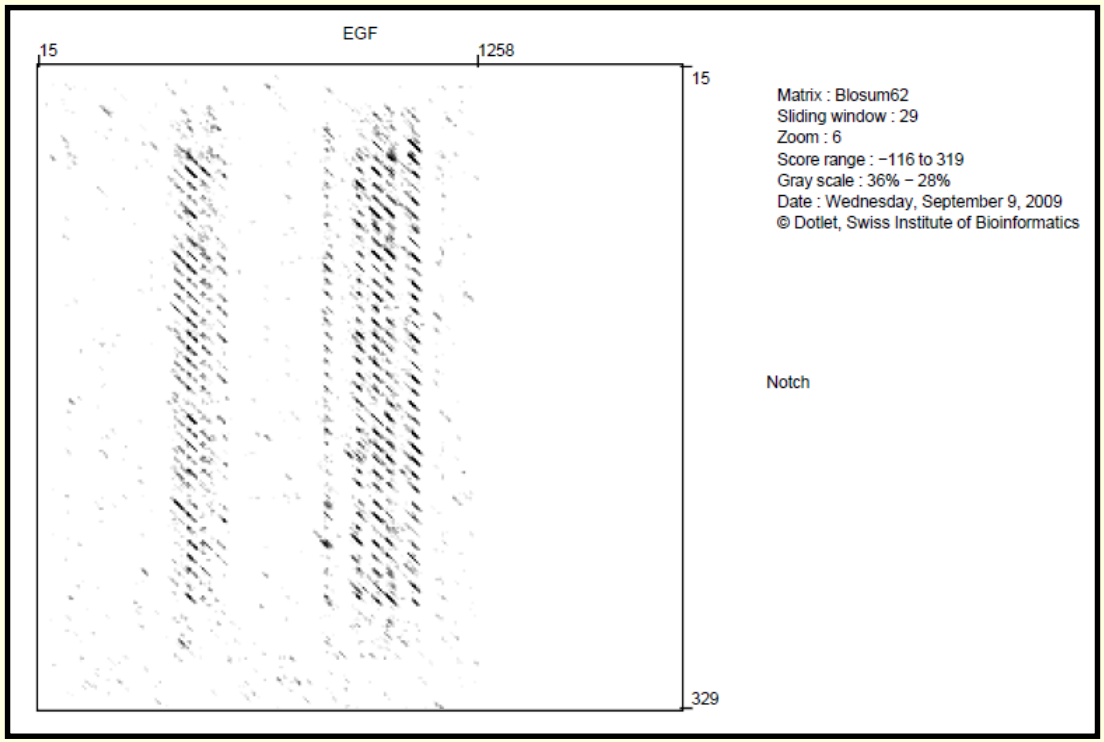
Practical Alignments

```
Statistics: (shuffled [500]) MLE statistics: Lambda= 0.2295;  K=0.1712
Threshold: E() < 1 score: 34
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.0 Oct 2008)
Parameters: BL50 matrix (15:-5), open/ext: -50/-1
 Scan time:  0.050

>>>AAA21284, 99 aa vs TMP.q2 library

>>P63120  HERV-K endogenous retrovirus type K,  protease                    (156 aa)
 Waterman-Eggert score: 83;  30.0 bits; E(1) <  1.4e-05
36.1% identity (66.7% similar) in 36 aa overlap (7-42:8-43)
Entrez Lookup  Re-search database  General re-search
         10        20        30        40
AAA212 QRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKW
       .::.  . : :.  :.:.:::::: ...   . : .:
P63120 NRPVCKAIIQGKQFEGLVDTGADVSIIALNQWPKNW
        10        20        30        40


 Waterman-Eggert score: 56;  21.1 bits; E(1) <  0.0069
36.8% identity (68.4% similar) in 19 aa overlap (76-94:83-101)
Entrez Lookup  Re-search database  General re-search
          80        90
AAA212 LVGPTPVNIIGRNMLTQIG
       ..   :.:. ::..: : :
P63120 MITSIPLNLWGRDLLQQWG
             90       100
```

Sequence Comparison

Practical Alignments

- ***Proteins with repeats***
  - Mouse Epidermal Growth Factor (EGF)
  - Drosophila Notch



Sequence Comparison

Practical Alignments

- *EGF vs Notch*
- *First of ~ 50 significant alignments*

```
Statistics: (shuffled [500]) MLE statistics: Lambda= 0.0970;  K=0.0007773
Threshold: E() < 1 score: 81
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.0 Oct 2008)
Parameters: BL50 matrix (15:-5), open/ext: -12/-2
 Scan time:  8.160

>>>sp|P07207|NOTCH_DROME, 2703 aa vs TMP.q2 library

>>sp|P01132|EGF_MOUSE Pro-epidermal growth factor OS=Mus musculus GN=Egf PE=  (1217 aa)
 Waterman-Eggert score: 387;  64.5 bits; E(1) <  1.3e-13
22.2% identity (45.5% similar) in 734 aa overlap (445-1060:314-1021)
Entrez Lookup  Re-search database  General re-search
          450       460       470       480       490       500       510
sp|P07 DCSEDIDEC---DQGSPCEHNGICVNTPGSYRCNCSQGFTGPR----CETNINECESHPCQNEGSCLDDPGTFRCVCMPG
       : ..: :     ..: ::.  :.:   : :.   :.:.:  :    :: ..::: ..      .: . ::...:.:  :
sp|P01 DAAKDPDPELLKQRGRPCRF-GLCERDPKSHSSACAEGYTLSRDRKYCE-DVNECATQNHGCTLGCENTPGSYHCTCPTG
          320       330       340       350       360       370       380       390

          520       530       540       550       560       570       580       590
sp|P07 FT----GTQCEIDIDECQSNPCLNDGTCHDKINGFKCSCALGFTGARCQINIDDCQSQPCRN-RGICHDSIAG-YSCECP
       :.    : ::. ..  : .:    .  :    .: .: :  : . .:   .   :.:    .   ::     : . :.:
sp|P01 FVLLPDGKQCH-ELVSCPGNVSKCSHGCVLTSDGPRCICPAGSVLGRDGKTCTGCSSPDNGGCSQICLPLRPGSWECDCF
             400       410       420       430       440       450       460       470

            600       610       620                       630           640
sp|P07 PGYTGTSCEININDCDSNPCHRGKCIDDV-------NSFKCLC------------DPG----YTGYICQKQINECESNPC
       ::    : . .     .:       .:.       ...: :             ::     : .    : :.. . .
sp|P01 PGYDLQSDRKSCAASGPQPLLLFANSQDIRHMHFDGTDYKVLLSRQMGMVFALDYDPVESKIYFAQTALKWIERANMDGS
             480       490       500       510       520       530       540       550

               650       660       670               680           690       700
sp|P07 Q-----------FDGHCQDRVGSYYCQCQAGTS--------GKNCEVNVNECHSNP----CNNGATCIDGINSYKCQCVP
       :          ..:   : .:      ..: :        ::. .. ..:  : :     .  :  .   .     .
sp|P01 QRERLITEGVDTLEGLALDWIGRRIYWTDSGKSVVGGSDLSGKHHRIIIQERISRPRGIAVHPRARRLFWTDVGMSPRIE
             560       570       580       590       600       610       620       630

         710       720       730         740       750             760            770
sp|P07 GFTGQHCEKNVDECISSPCANNGVCIDQVNG--YKCECPRGFYDAHCLSD------VDECASNP---CVNEGR--CEDGI
       . . :  .. :    :.   .:. :: ..   : :.  :.  .   :.       ... ...:    : : .    :
sp|P01 SASLQGSDR-VLIASSNLLEPSGITIDYLTDTLYWCDTKRSVIEMANLDGSKRRRLIQNDVGHPFSLAVFEDHLWVSDWA
              640       650       660       670       680       690       700
```

```
           780       790                       800           810       820
sp|P07 NEFICHCPPGYTGKRCELD--------------IDECSSNPC--QHGGT---CYDKLNAFSCQCMPGYT----GQKC---
         . .       .: .:.                . . ...::  ..::    : ..:..  : :  :..    :. :
sp|P01 IPSVIRVNKRTGQNRVRLQGSMLKPSSLVVVHPLAKPGADPCLYRNGGCEHICQESLGTARCLCREGFVKAWDGKMCLPQ
     710       720       730       740       750       760       770       780

                830       840       850          860       870       880       890
sp|P07 --------ETNIDDCVTNPCGNGGTCIDKVNGYKC---VCKVPFTGRDCESKMDPCASNRCKNEAKCTPSSNFLDFSCTC
               .....  ::.  ..  . .   .: .    : ..  .: . :   : :. . : ..:.:. ...  . .:
sp|P01 DYPILSGENADLSKEVTSLSNSTQAEVPDDDGTESSTLVAEIMVSGMNYE---DDCGPGGCGSHARCVSDGETAECQCLK
     790       800       810       820       830          840       850       860

         900       910         920         930       940         950          960       970
sp|P07 KLGYTGRYCDEDIDECSLS-SPCRNGAS-CLNVPGSYRCLCTKGYEGRDCA-INTDDC--ASFPCQNGGTCLDGIGDYSC
       ..  :  :. ::::: :. : : . .: :.:. :.: : :..::::   . .. :.:  ..  : ....: .  : :.:
sp|P01 GFARDGNLCS-DIDECVLARSDCPSTSSRCINTEGGYVCRCSEGYEGDGISCFDIDECQRGAHNCAENAACTNTEGGYNC
        870         880       890       900         910         920       930       940

               980       990      1000      1010      1020      1030        1040
sp|P07 LCVD--GFDGKHCETDINECLSQPCQNGATCSQYVNSYTCTCPLGFSGINCQTNDEDCTESSCLNGGSC--IDGINGYNC
       :.   .  :. :  .    :   ..:   ..  :::   :: ...:                  ::::: :  :.....:.:
sp|P01 TCAGRPSSPGRSCPDSTAPSLLG--EDGHHLDR--NSYP-GCPSSYDGY-------------CLNGGVCMHIESLDSYTC
         950       960         970         980                       990      1000

      1050      1060
sp|P07 SCLAGYSGANCQYK
       .:. ::::  :: .
sp|P01 NCVIGYSGDRCQTR
     1010      1020
```

Schedule Week 3
4 – 6 September

- ***Wednesday - Alignments /Scoring Systems***
- ***Friday –Searching for homologs***

- ***9 Sep Monday***
  - Searching for homologs (Ch 5.3-5.5, Ch 4.6-4.7)
- ***11 Sep Wednesday***
  - Searching for homologs
- ***13 Sep Friday***
  - Sequence motifs (Ch 4.8-4.10)

Database Searching

Sequence database searching

- *Null hypothesis: a pair sequences are unrelated*

If the pair of sequences have a more significant match than could reasonably be expected from unrelated (not homologous) sequences
Then they must be related (homologous)