7.36 / 20.390 / 6.802

7.91 / 20.490 / 6.874 / HST.506

Lecture #3

C. Burge

Feb. 11, 2014

Global Alignment of Protein Sequences

(NW, SW, PAM, BLOSUM)

Topic 1 Info

- Overview slide has blue background - readings for upcoming lectures are listed at bottom of overview slide
- Review slides will have purple background
- Send your background/interests to TA for posting if reg’d for grad version
- PS1 is posted. BLAST tutorial may be helpful
- PS2 is posted. Look at the programming problem

Local Alignment (BLAST) and Statistics

- Sequencing
  - Conventional
  - 2nd generation
- Local Alignment:
  - a simple BLAST-like algorithm
  - Statistics of matching
  - Target frequencies and mismatch penalties for nucleotide alignments

Background for 2/7, 2/12 lectures: Z&B Ch. 4 & 5, **BLAST tutorial**

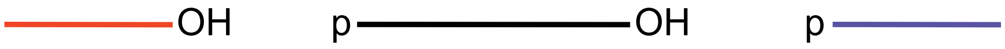
Questions: Chemistry / Library Prep

Dye terminator chemistry: dye is attached to base

How to put different adapters on the two ends?

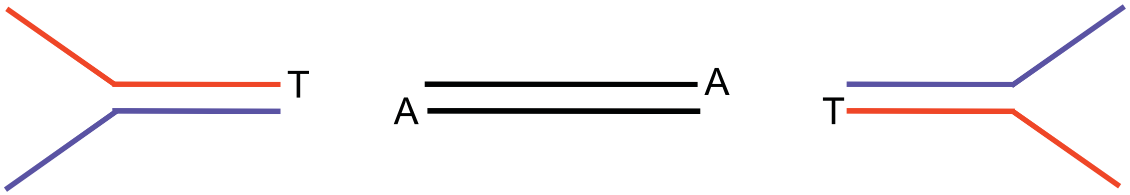
At least three ways:

1) RNA ligation



2) polyA tailing/polyTVN-ad2priming/circularization (PMID 19213877)

3) ligation of Y-shaped adapters



DNA Sequence Alignment I: Motivation

You are studying a recently discovered human non-coding RNA.

You search it against the mouse genome using BLASTN (N for nucleotide) and obtain the following alignment:

```
Q: 1   ttgacctagatgagatgtcgttcacttttactcaggtacagaaaa 45
       |||| |||||||||||| | |||||||||||| || |||||||||
S: 403 ttgatctagatgagatgccattcacttttactgagctacagaaaa 447
```

Is this alignment significant?
Is this likely to represent a homologous RNA?

How to find alignments?

DNA Sequence Alignment II

Identify high scoring segments whose score S exceeds a cutoff x using a **local alignment** algorithm (e.g., BLAST)

Scores follow an extreme value (aka Gumbel) distribution:

$$P(S > x) = 1 - \exp[-KMN\ e^{-\lambda x}]$$

For sequences/databases of length M, N where K, λ are positive parameters that depend on the score matrix and the composition of the sequences being compared

Conditions: expected score is negative, but positive scores possible

Alternate algorithm

Karlin & Altschul 1990

Computational Efficiency

Measure efficiency in cpu run time and memory

O() = "big-oh" notation (computational Order of problem)

Consider the number of individual computations required to run algorithm as a function of the number of 'units' in the problem (e.g., base pairs, amino acid residues)

Analyze the asymptotic worst-case running time or sometimes just do the experiment and measure run time

If problem scales as square of the number of units it is

$O(n^2)$ "order n-squared"

DNA Sequence Alignment III

How is λ related to the score matrix?

λ is the unique positive solution to the equation*:

$$\sum_{i,j} p_i r_j e^{\lambda s_{ij}} = 1$$

p_i = freq. of nt i in query, r_j = freq. of nt j in subject

s_{ij} = score for aligning an i,j pair

"Target frequencies"* : $q_{ij} = p_i r_j e^{\lambda s_{ij}}$

*Karlin & Altschul, 1990

DNA Sequence Alignment VI

Optimal mismatch penalty m for given target identity fraction r

$$\boxed{m = \ln(4(1-r)/3)/\ln(4r)}$$

Examples:

r	0.75	0.95	0.99
m	-1	-2	-3

r = expected fraction of identities in high-scoring BLAST hits

DNA Sequence Alignment VII

Meaning of mismatch penalty equation

$$\boxed{m = \ln(4(1-r)/3)/\ln(4r)}$$

Examples:

r	0.75	0.95	0.99
m	-1	-2	-3

So why is m = -3 better for finding matches with 99% identity?

Does it mean that you can only find 99% identical matches with a mismatch score of -3?

Answer: No. It's also possible to find 99% matches with m = -1 or -2.

But m changes the match length required to achieve statistical significance

λ is the unique positive solution to the equation

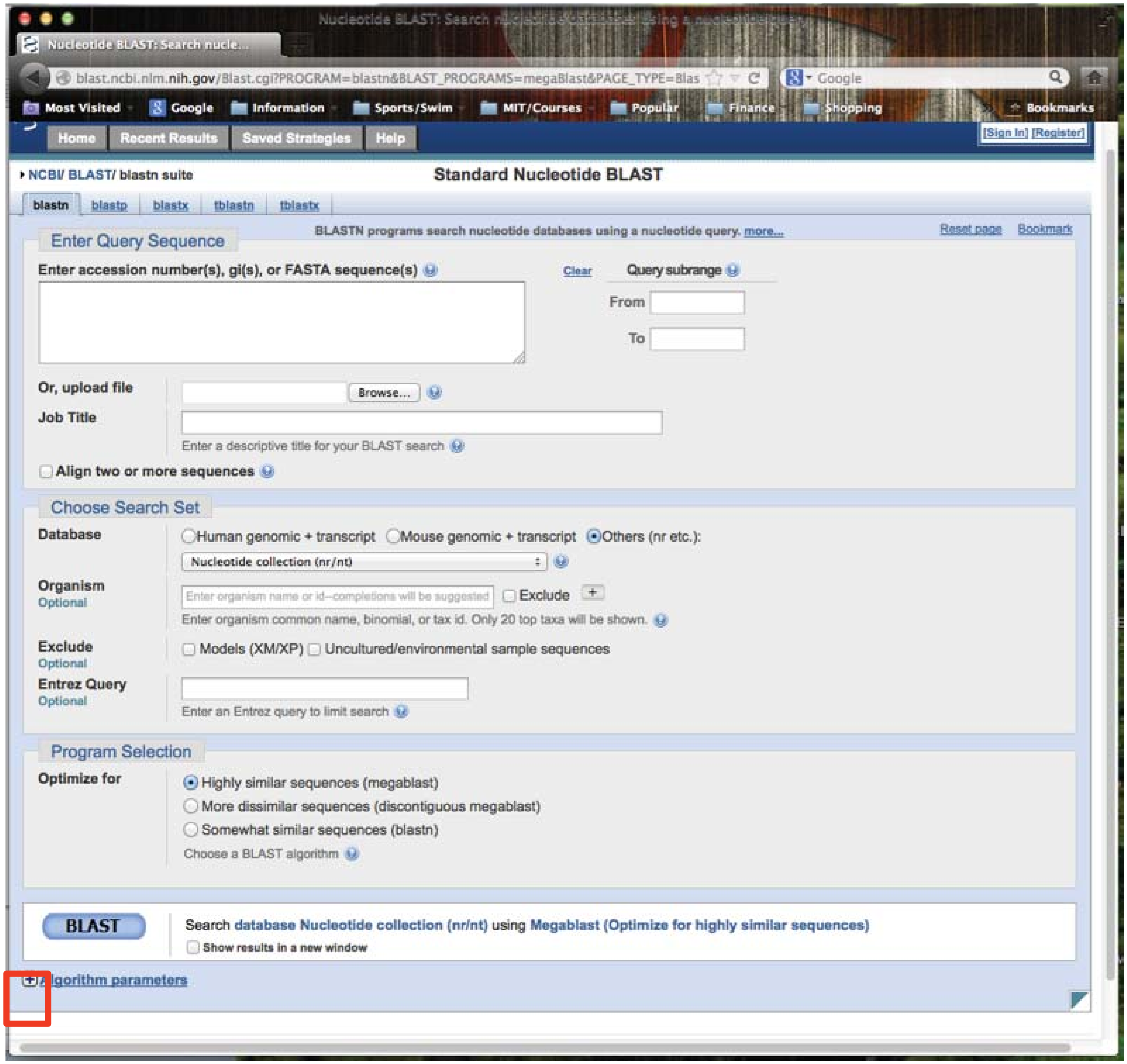
$\sum_{i,j} p_i p_j e^{\lambda s_{ij}} = 1$ p_i = frequency of nt i, s_{ij} = score for aligning an i,j pair

and $P(S > x) = 1 - \exp[-KMN\, e^{-\lambda x}]$

If we change the mismatch score from -1 to -3, λ will increase. Therefore, the score required to achieve a given level of significance will decrease, i.e. shorter hits will be significant.

So why would you ever want to use m = -1?

Google: blastn



NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

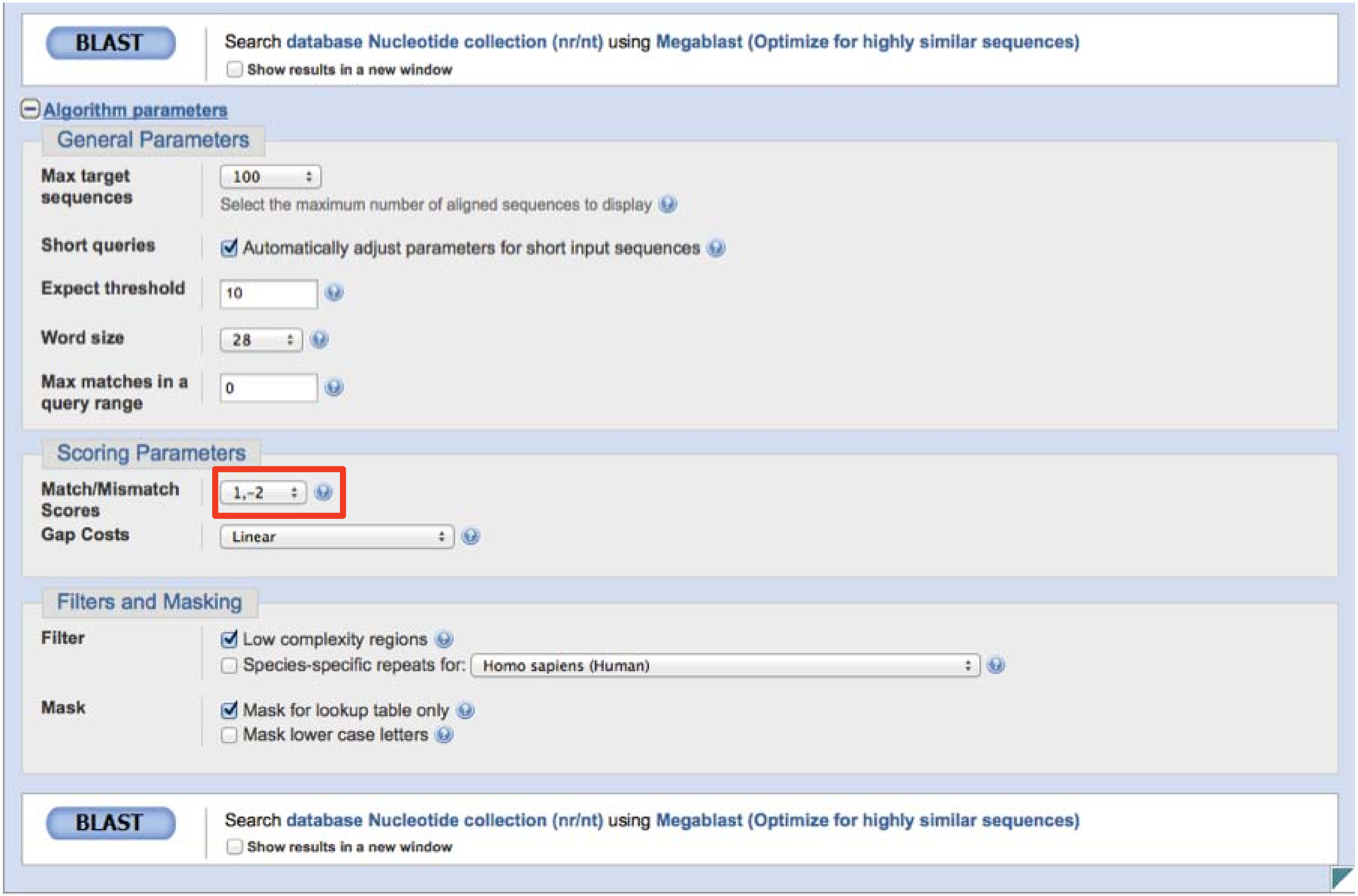
Choose a BLAST algorithm

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

+ Algorithm parameters

Courtesy of National Library of Medicine. In the public domain.



Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

Algorithm parameters

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 28

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

Filters and Masking

Filter: ☑ Low complexity regions

□ Species-specific repeats for: Homo sapiens (Human)

Mask: ☑ Mask for lookup table only

□ Mask lower case letters

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

Courtesy of National Library of Medicine. In the public domain.

DNA Sequence Alignment VIII

Translating searches:

- translate in all possible reading frames
- search peptides against protein database (BLASTP)

```
ttgacctagatgagatgtcgttcacttttactgagctacagaaaa

ttg|acc|tag|atg|aga|tgt|cgt|tca|ctt|tta|ctg|agc|tac|aga|aaa
 L   T   x   M   R   C   R   S   L   L   L   S   Y   R   K

t|tga|cct|aga|tga|gat|gtc|gtt|cac|ttt|tac|tga|gct|aca|gaa|aa
   x   P   R   x   D   V   V   H   F   Y   x   S   T   E

tt|gac|cta|gat|gag|atg|tcg|ttc|act|ttt|act|gag|cta|cag|aaa|a
    D   L   D   E   M   S   F   T   F   T   E   L   Q   K
```

Also consider reading frames on complementary DNA strand

DNA Sequence Alignment IX

Common flavors of BLAST:

Program	Query	Database
BLASTP	aa	aa
BLASTN	nt	nt
BLASTX	nt (⇒ aa)	aa
TBLASTN	aa	nt (⇒ aa)
TBLASTX	nt (⇒ aa)	nt (⇒ aa)
PsiBLAST	aa (aa msa) aa	

msa = multiple sequence alignment

Which would be best for searching ESTs against a genome?

Global Alignment of Protein Sequences (NW, SW, PAM, BLOSUM)

- Global sequence alignment (Needleman-Wunch-Sellers)
- Gapped local sequence alignment (Smith-Waterman)
- Substitution matrices for protein comparison

Background for today: Z&B Chapters 4,5 (esp. pp. 119-125)

Why align protein sequences?

- Functional predictions based on identifying homologous proteins or protein domains

Assumes

Sequence similarity → Similarity in function (and/or structure)

implies

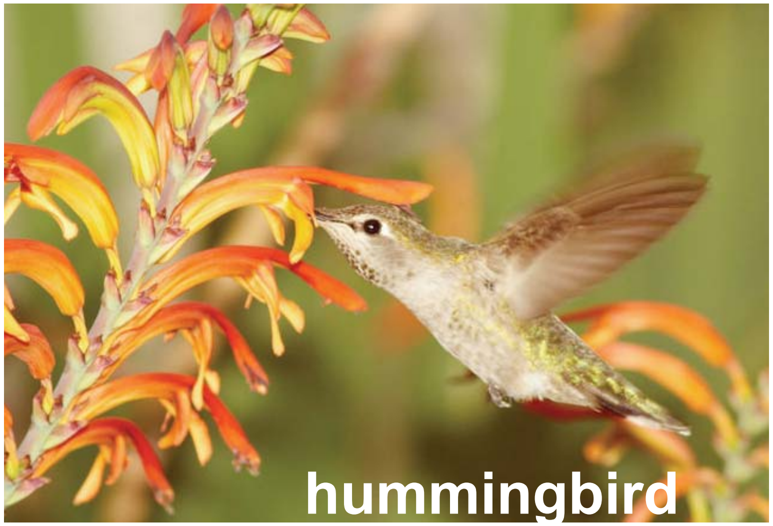
- almost always true for similarity > 30%
- 20-30% similarity is “the twilight zone”

BUT: Function carried out at level of folded protein, i.e. 3-D structure
Sequence conservation occurs at level of 1-D sequence

Converse is not true

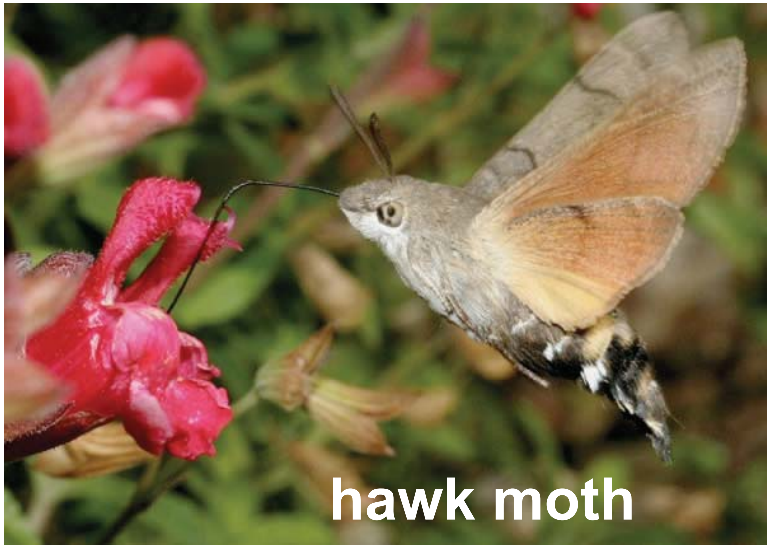
Structural similarity ↛ Sequence similarity
(or even homology)

Convergent Evolution



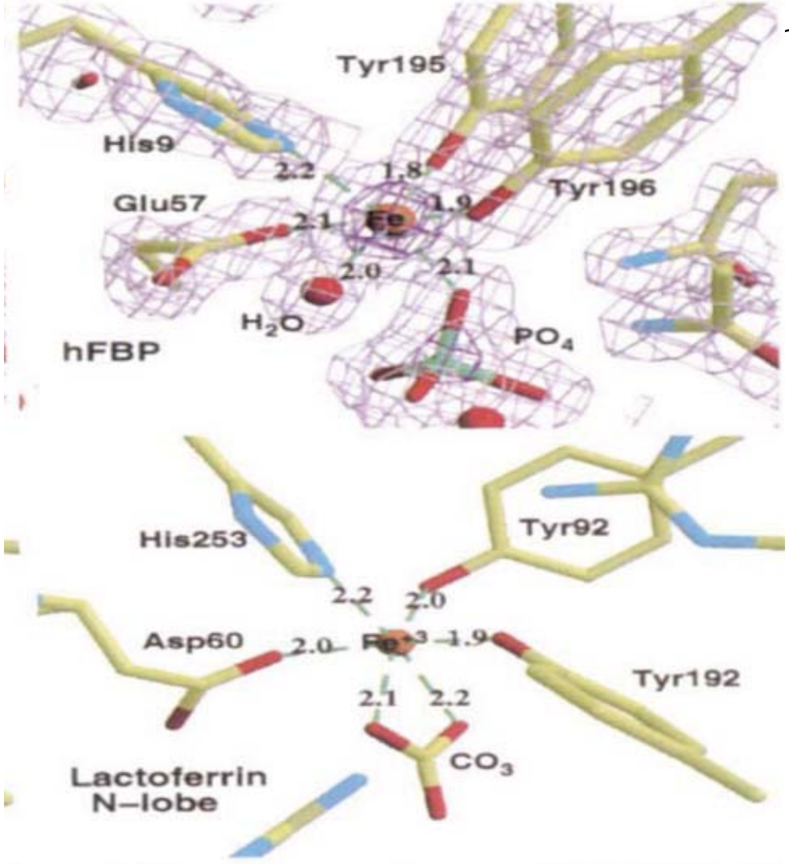
Courtesy of Matthew Field. License: CC-BY.

Last common ancestor lived > 500 Mya and lacked wings (and probably legs and eyes)



Same idea for proteins - can result in similar structures with no significant similarity in sequence

Convergent Evolution of Fe3+-binding Proteins



Haemophilus Fe3+-binding protein (hFBP)

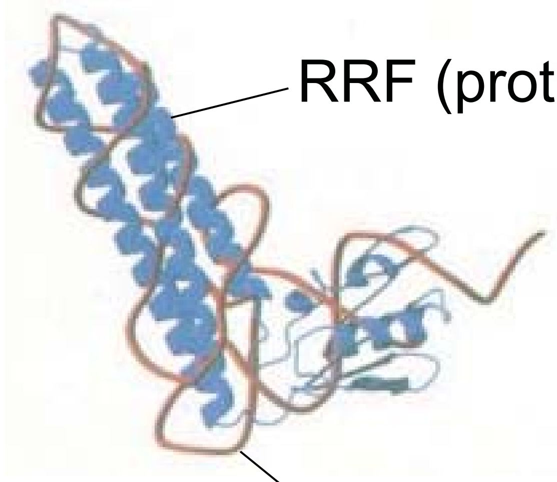
Eukaryotic lactoferrin

Last common ancestor occurred > 2Bya and bound anions

Courtesy of Nature Publishing Group. Used with permission.
Source: Bruns, Christopher M., Andrew J. Nowalk, et al. "Structure of Haemophilus Influenzae Fe+3-Binding Protein Reveals Convergent Evolutionwithin a Superfamily." *Nature Structural & Molecular Biology* 4, no. 11 (1997): 919-24.

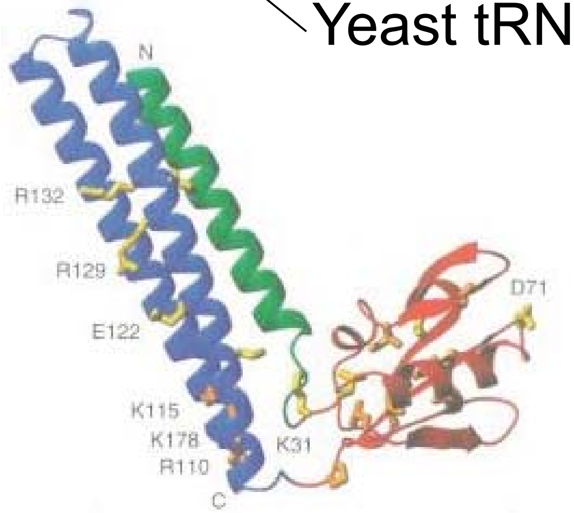
Bruns et al. Nature Struct. Biol. 1997

Convergent Evolution of a Protein and an RNA



RRF (protein)

Yeast tRNA^{Phe}



Source: Selmer, Maria, Salam Al-Karadaghi, et al. "Crystal Structure of Thermotoga Maritima Ribosome Recycling Factor: A tRNA Mimic." *Science* 286, no. 5448 (1999): 2349-52.

T. maritima ribosome recycling factor (RRF)

Unlikely to have ever had a common molecular ancestor

Selmer et al. Science 286. 2349 -. 1999

Types of Alignments

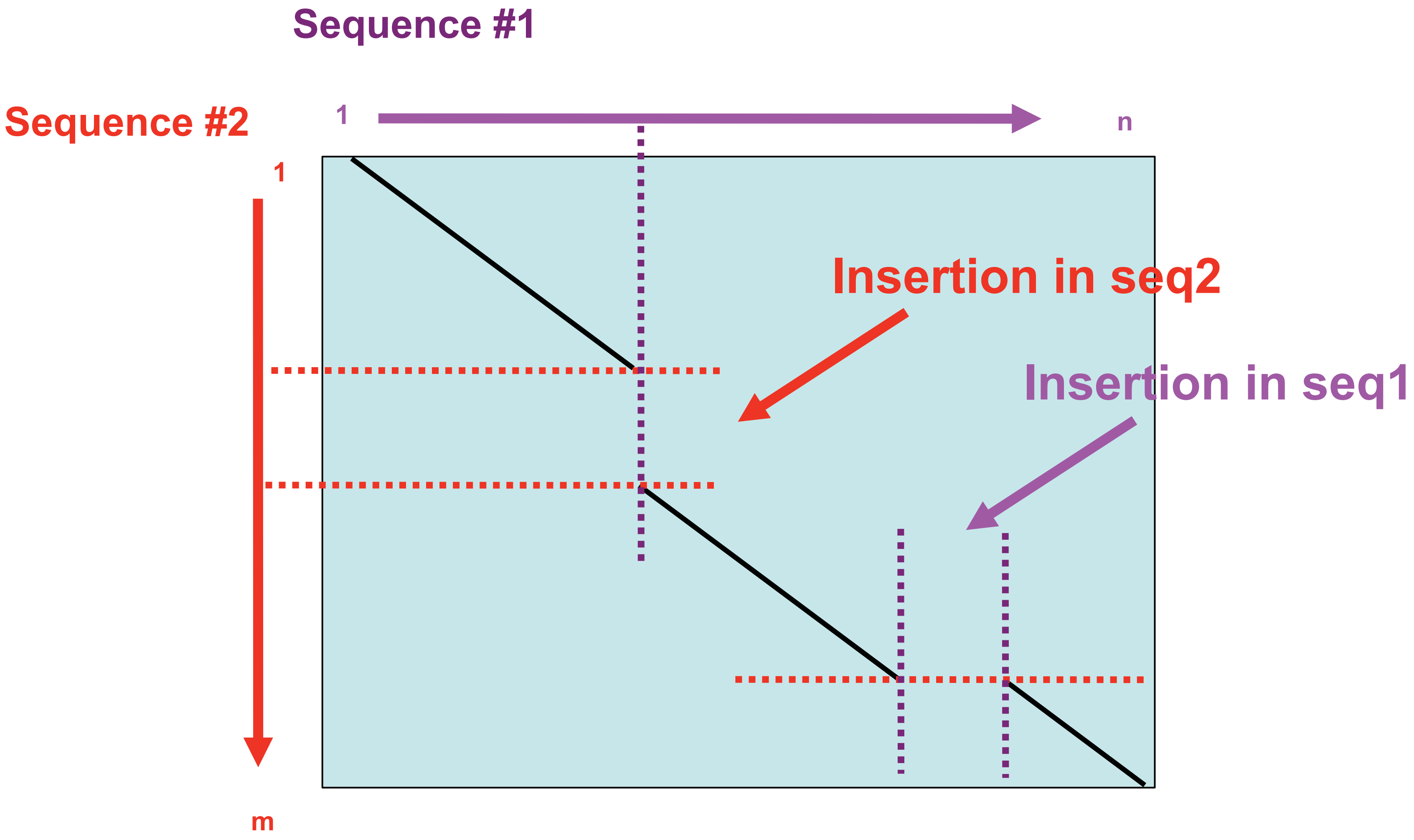
Scope:

- Local
- Global
- Semiglobal

Scoring system:

- Ungapped
- Gapped
  - linear
  - affine

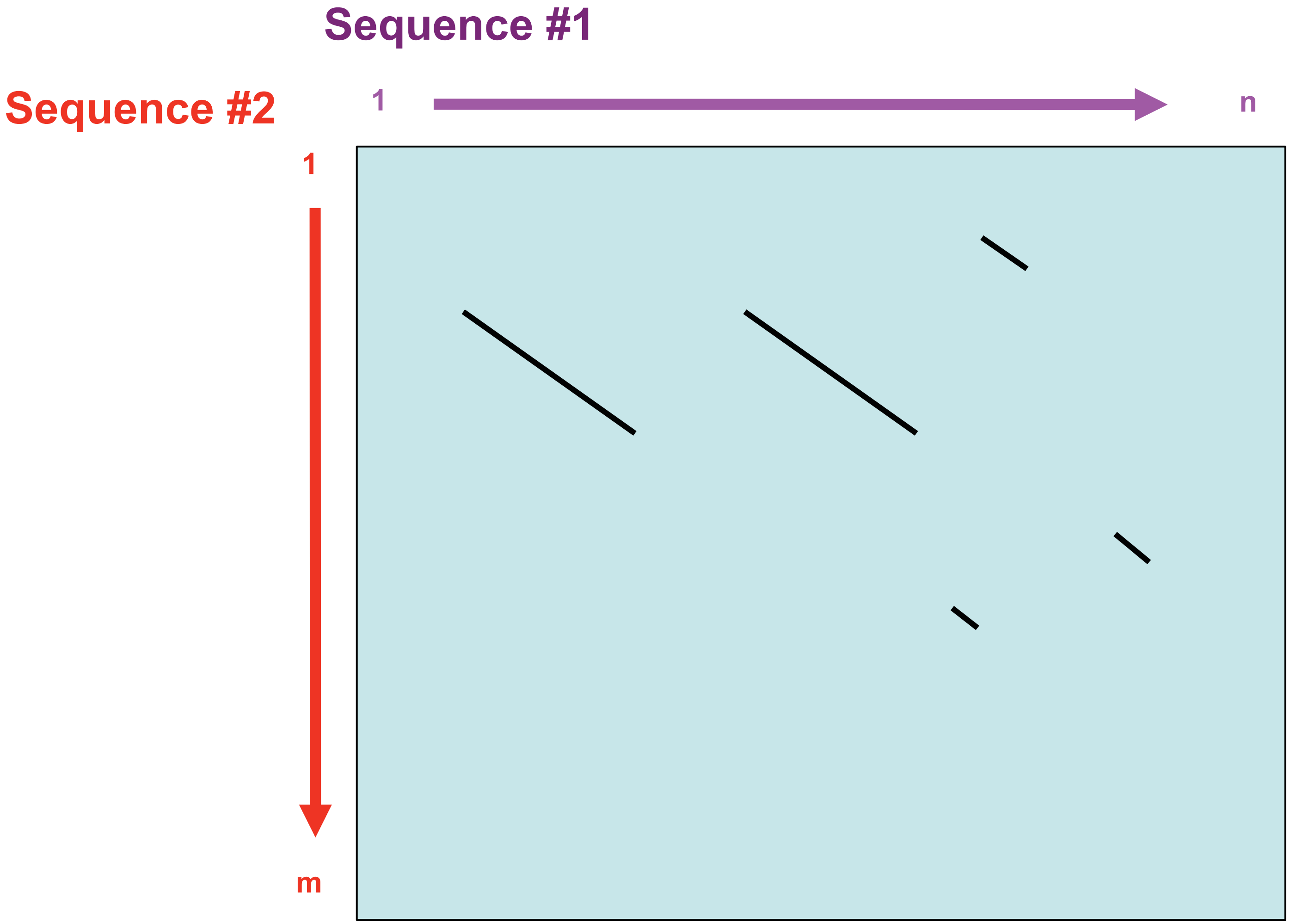
Dot Matrix Alignment Example



What type of alignment would be most appropriate for this pair of sequences?

Global

Dot Matrix Alignment Example 2



What type of alignment would be most appropriate for this pair of sequences?

Local

Gaps (aka "Indels")

```
AKHFRGCVS
AKKF--CVG
```

- Linear Gap Penalty
  - $\gamma(n) = nA$, n= no. of gaps, A = gap penalty

- "Affine" gap penalty

$$W_n = G + n\gamma,$$

n = no. of gaps, γ = gap extension penalty, and G = gap opening penalty

Or:

$$W_n = G + (n-1)\gamma$$

with alternative definition of gap opening penalty

Obtain optimal global alignment using *Dynamic Programming:*

First write one sequence across the top, and one down along the side

	Gap	V	D	S	C	Y
Gap	0	1 gap	2 gaps	→		
V	1 gap					
E	2 gaps					
S	↓					
L						
C						
Y						

Note – linear gap penalty: γ(n)=nA, where A=gap penalty
a negative number

Dynamic Programming:

Initialize the alignment matrix

		i =0	1	2	3	4	5
j =		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	S_{ij}				
2	E	-16					
3	S	-24					
4	L	-32					
5	C	-40					
6	Y	-48					

Sij = score of optimal alignment ending at position **i** in seq 1 and **j** in seq 2. Requires that we know **S(i-1, j-1), S(i, j-1), S(i-1, j)…**

Recursive: Solution to larger problem is built up from solutions to smaller problems

Store **Sij** and how we arrived at **Sij** in a matrix

Often called ‘dynamic programming’ or more generally ‘recursive optimization’

What is the gap penalty in this example?

Dynamic Programming: Recursion

Sequence 1

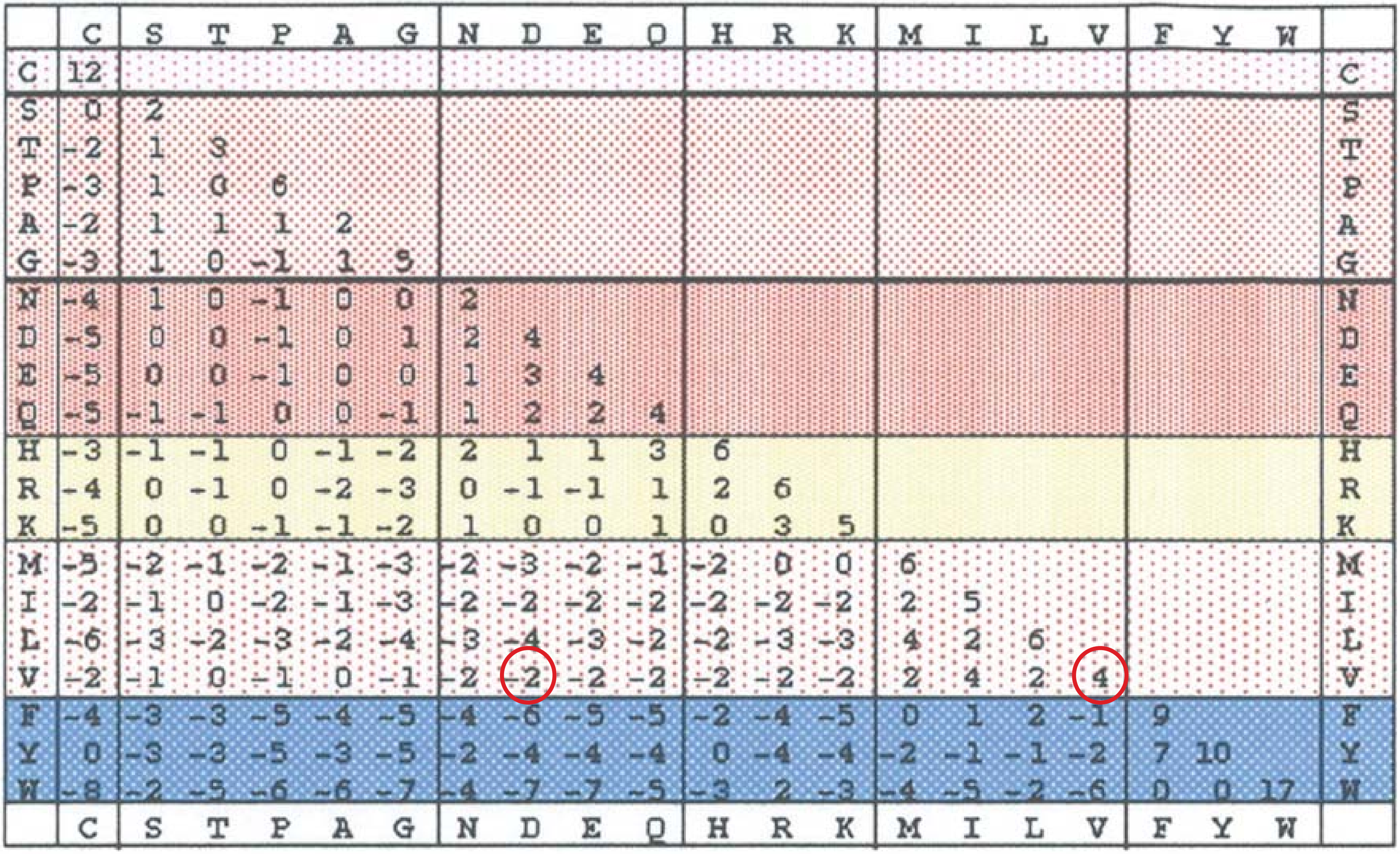
Sequence 2		i =0	1	2	3	4	5
j =		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	S_{ij}				
2	E	-16					
3	S	-24					
4	L	-32					
5	C	-40					
6	Y	-48					

Global alignments: Needleman-Wunsch-Sellers

$$S_{ij} = \text{max of: } \begin{cases} S_{i-1,\, j-1} + \sigma(x_i, y_j) \text{ (diagonal)} \\ S_{i-1,\, j} + A \text{ (from left to right)} \\ S_{i,\, j-1} + A \text{ (from top to bottom)} \end{cases}$$

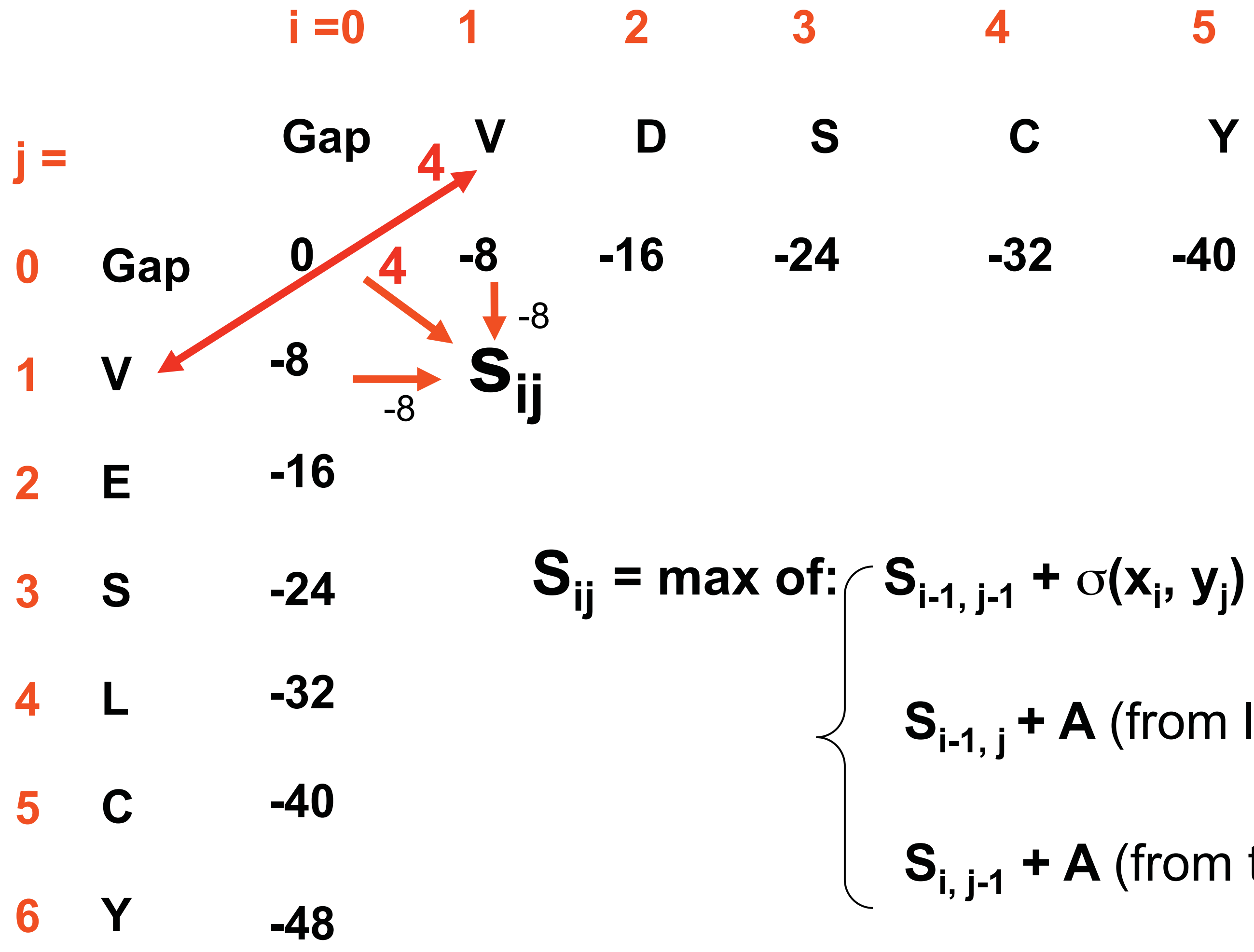
Computational complexity? O(mn) with linear gap penalty

PAM250 Scoring Matrix



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

Dynamic Programming: filling in matrix



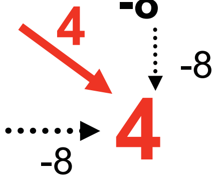
j =		i =0	1	2	3	4	5
		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	S_{ij}				
2	E	-16					
3	S	-24					
4	L	-32					
5	C	-40					
6	Y	-48					

$$S_{ij} = \max \text{ of: } \begin{cases} S_{i-1, j-1} + \sigma(x_i, y_j) \text{ (diagonal)} \\ S_{i-1, j} + A \text{ (from left to right)} \\ S_{i, j-1} + A \text{ (from top to bottom)} \end{cases}$$

Sequence 1

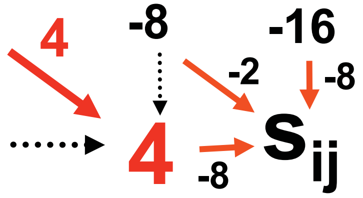
Sequence 2

		i =0	1	2	3	4	5
j =		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4				
2	E	-16					
3	S	-24					
4	L	-32					
5	C	-40					
6	Y	-48					

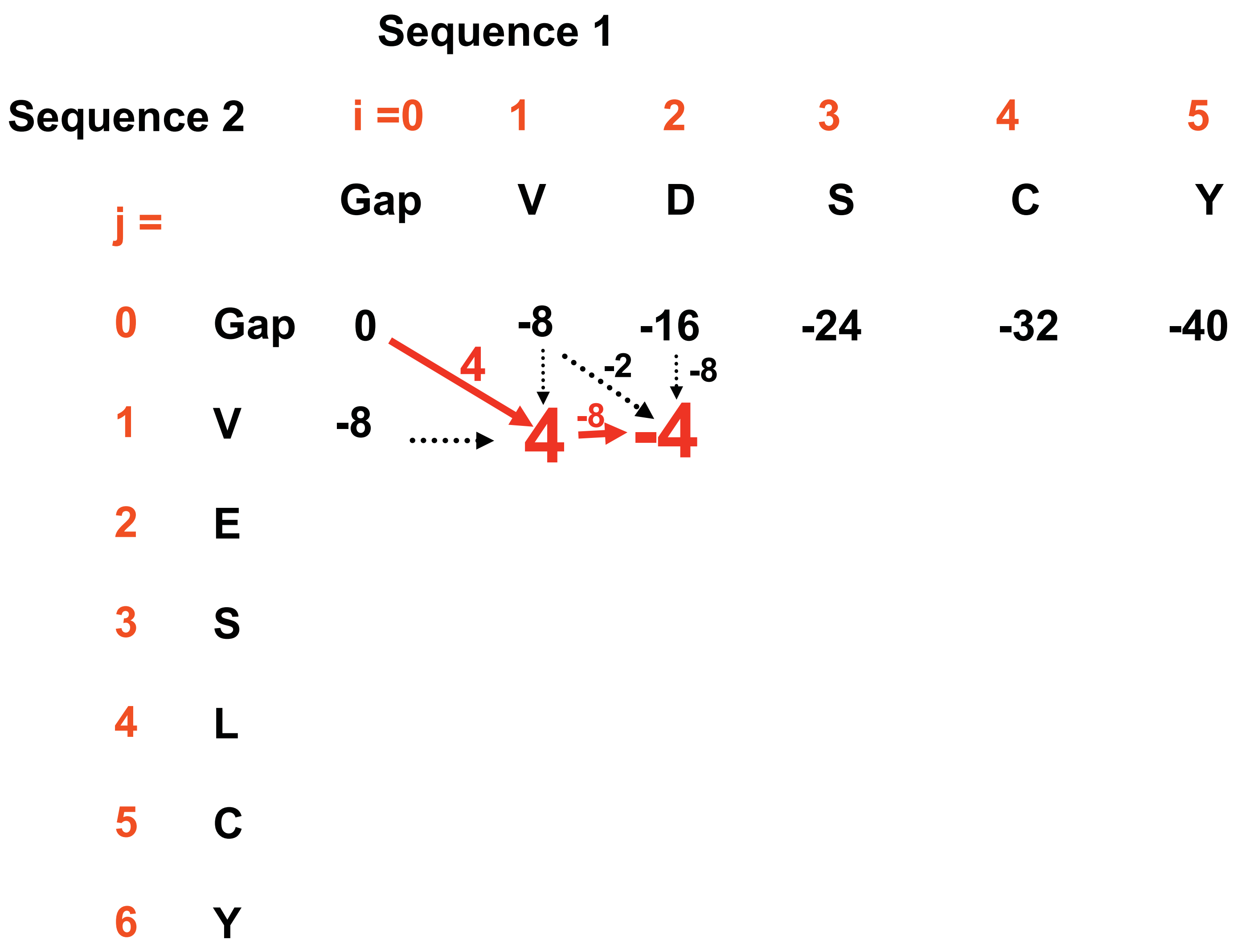


Sequence 1

Sequence 2		i =0	1	2	3	4	5
j =		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4	S_{ij}			
2	E	-16					
3	S	-24					
4	L	-32					
5	C	-40					
6	Y	-48					

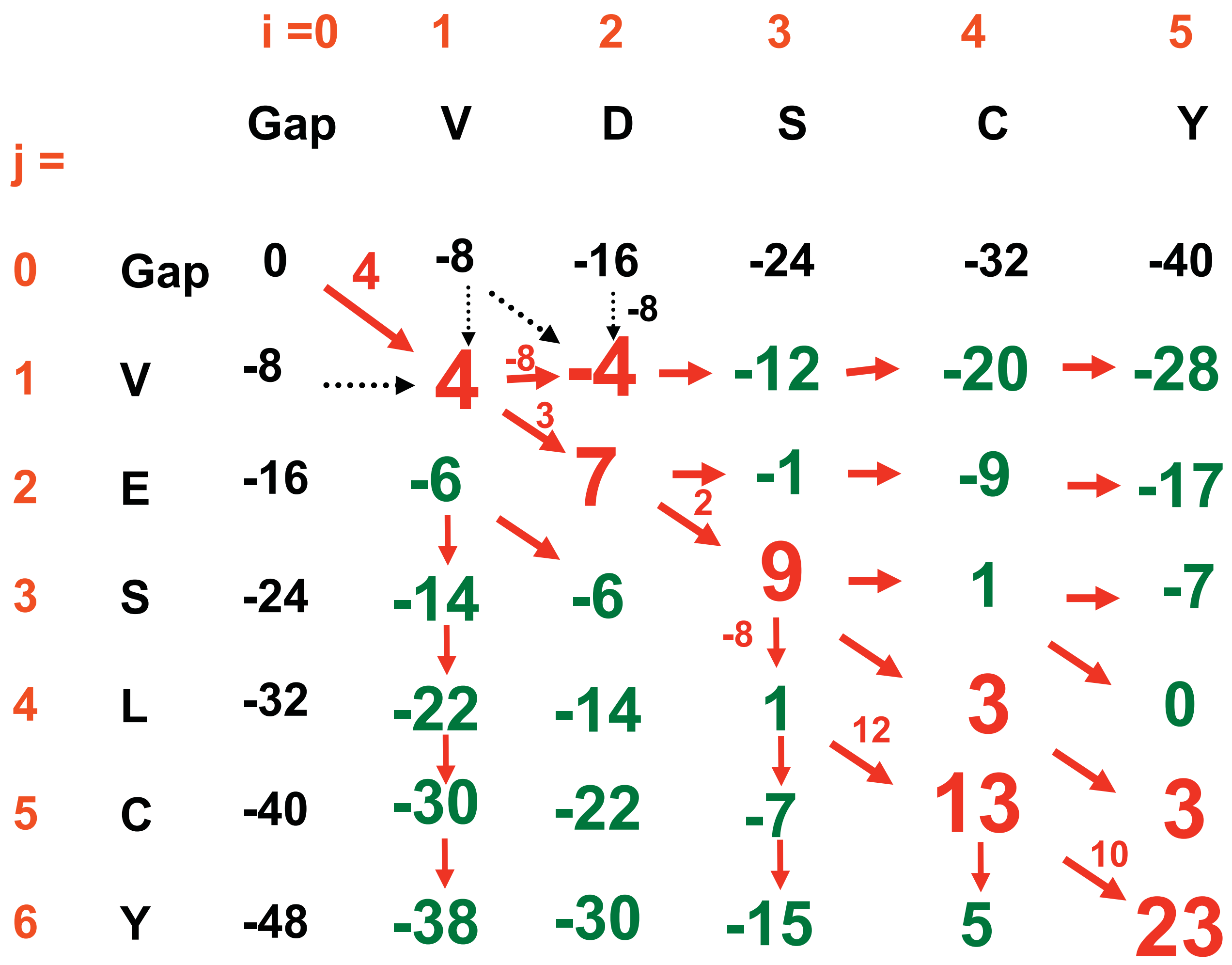


$$S_{ij} = \text{max of: } \begin{cases} S_{i-1,\, j-1} + \sigma(x_i, y_j) \text{ (diagonal)} \\ S_{i-1,\, j} + A \text{ (from left to right)} \\ S_{i,\, j-1} + A \text{ (from top to bottom)} \end{cases}$$



	Sequence 2	Sequence 1 i =0	1	2	3	4	5
j =		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4	-4			
2	E						
3	S						
4	L						
5	C						
6	Y						

Completed Dynamic Programming Matrix

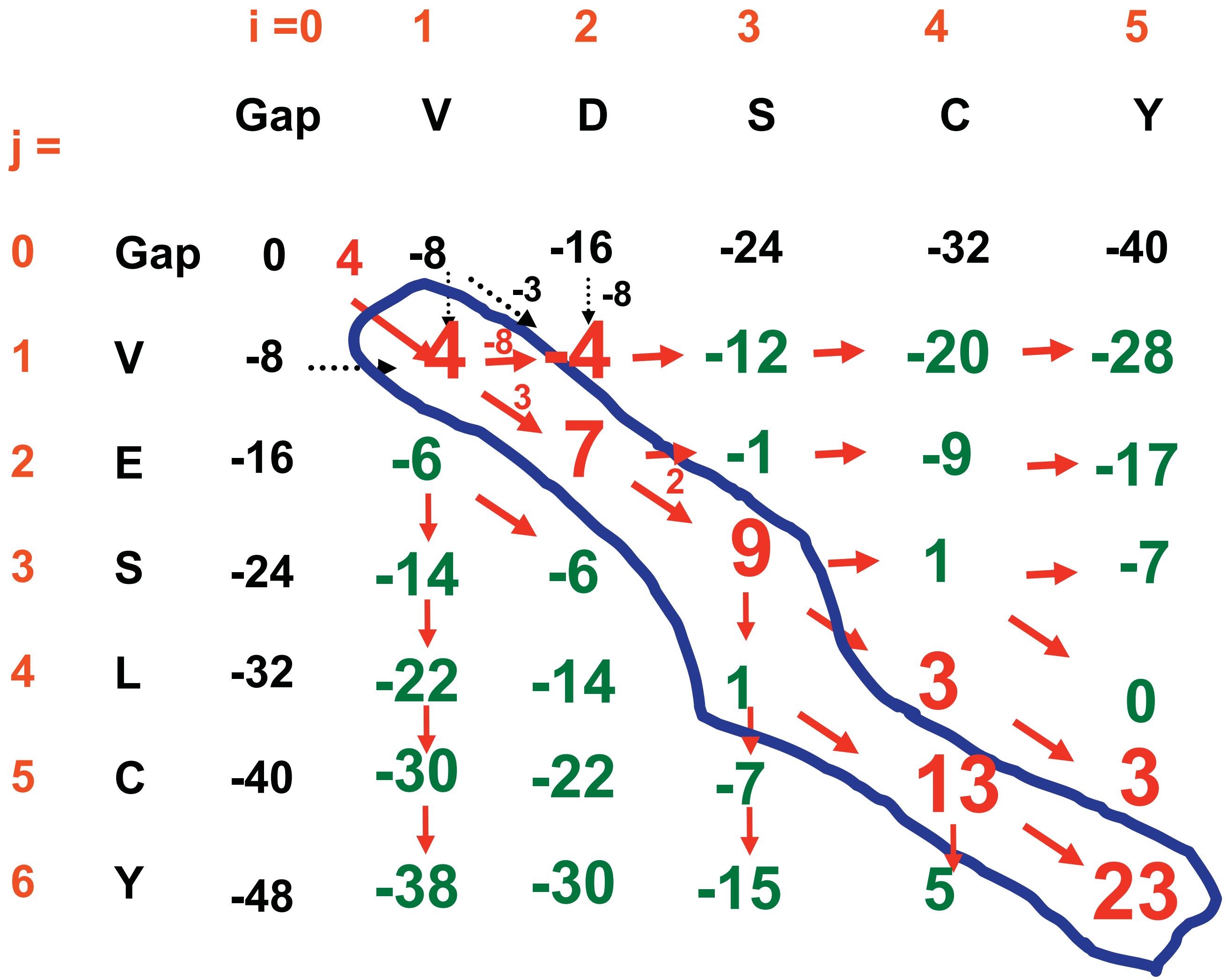


j =		i =0	1	2	3	4	5
		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4	-4	-12	-20	-28
2	E	-16	-6	7	-1	-9	-17
3	S	-24	-14	-6	9	1	-7
4	L	-32	-22	-14	1	3	0
5	C	-40	-30	-22	-7	13	3
6	Y	-48	-38	-30	-15	5	23

Keep track of scores AND how we got them → “traceback matrix”

The Traceback:

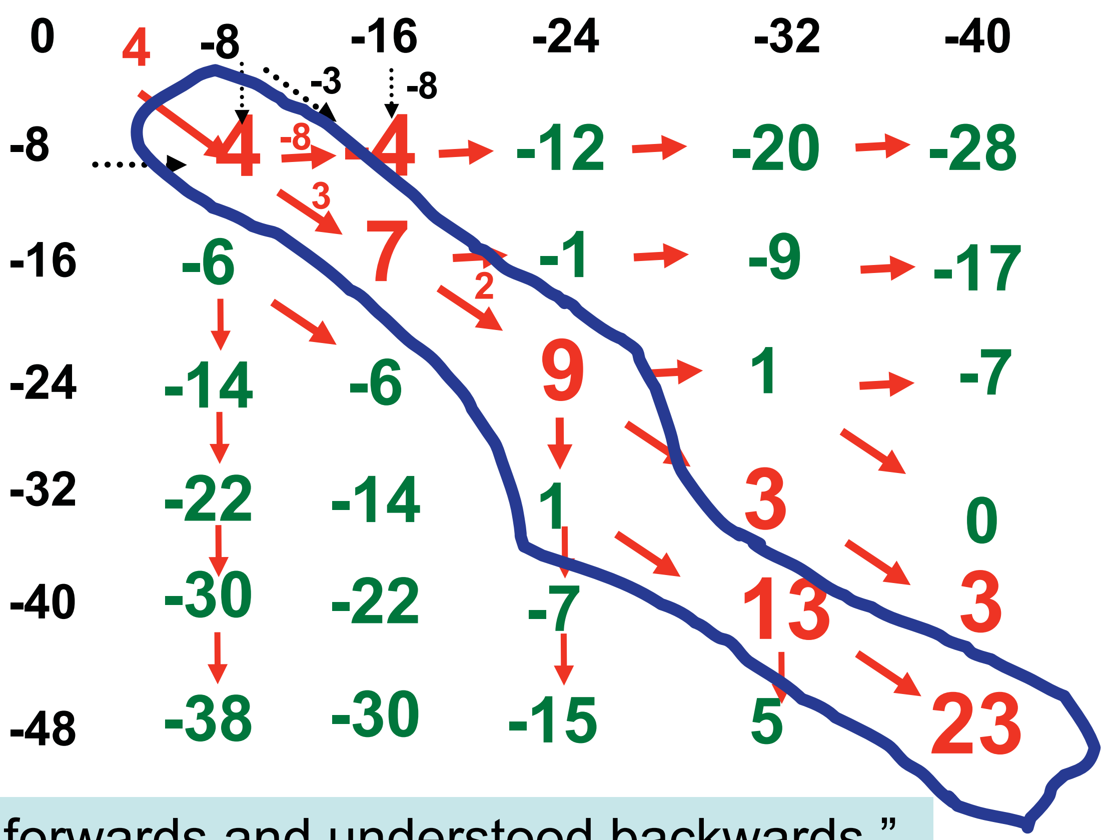
After the alignment square is finished, start at the lower right and work backwards following the arrows to see how you got there…



		i =0	1	2	3	4	5
j =		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4	-4	-12	-20	-28
2	E	-16	-6	7	-1	-9	-17
3	S	-24	-14	-6	9	1	-7
4	L	-32	-22	-14	1	3	0
5	C	-40	-30	-22	-7	13	3
6	Y	-48	-38	-30	-15	5	23

The Traceback gives the alignment:

```
V D S – C Y
V E S L C Y
```



j =		i =0	1	2	3	4	5
		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4	-4	-12	-20	-28
2	E	-16	-6	7	-1	-9	-17
3	S	-24	-14	-6	9	1	-7
4	L	-32	-22	-14	1	3	0
5	C	-40	-30	-22	-7	13	3
6	Y	-48	-38	-30	-15	5	23

> "Life must be lived forwards and understood backwards."
>
> - Søren Kierkegaard

Semiglobal Alignment

Allow sequences to overhang at either end without penalty -usually gives better alignments of homologous sequences of different lengths

Same algorithm as before except

• initialize edges of DP matrix $S_{i,0}$ and $S_{0,j}$ to 0

• instead of requiring traceback to begin at $S_{m,n}$, allow it to begin at highest score in bottom row or rightmost column

Gapped Local Alignment

Temple Smith and Michael Waterman, 1981 – modified Needleman-Wunsch-Sellers

Local alignment is the best scoring alignment of a substring in sequence x to a substring in sequence y.

Key idea is not to force the alignment to extend to the ends of the sequences

Photograph of scientists removed due to copyright restrictions.

Smith-Waterman Local Alignment

Again, use dynamic programming

Same basic scheme as before except

- **similarity matrix MUST include negative values for mismatches**

and

- **when the value calculated for a position in the scoring matrix is negative, the value is set to zero - this terminates the alignment**

Smith-Waterman:

Write one sequence across the top, and one down along the side

		i =0	1	2	3	4	5
j =		Gap	V	D	S	C	Y
0	Gap	0	0	0	0	0	0
1	V	0	S_{ij}				
2	E	0					
3	S	0					
4	L	0					
5	C	0					
6	Y	0					

Local alignments: Smith-Waterman

$$S_{ij} = \text{max of: } \begin{cases} S_{i-1,\, j-1} + \sigma(x_i, y_j) \text{ (diagonal)} \\ S_{i-1,\, j} - A \text{ (from left to right)} \\ S_{i,\, j-1} - A \text{ (from top to bottom)} \\ 0 \end{cases}$$

Need a metric of similarity between amino acid pairs

Simplest metric – identity matrix

	A	C	D	E	F	G	H	I	K
A	1	0	0	0	0	0	0	0	0
C		1	0	0	0	0	0	0	0
D			1	0	0	0	0	0	0
E				1	0	0	0	0	0
F					1	0	0	0	0
G						1	0	0	0
H							1	0	0
I								1	0
K									1

OK for nucleic acids,
but for proteins can
do substantially better

What properties should an
amino acid similarity matrix
have?

Refer to
Z&B pp. 119-125

Scoring system should favor matching identical or related amino acids and penalize for poor matches and for gaps

Need to know how often a particular amino acid pair is found in related proteins compared with its occurence by chance, and also how often gaps (insertions/deletions) are found in related proteins relative to dissimilar amino acid pairs

Scores and Evolution

Any alignment scoring system brings with it an implicit evolutionary model

Amino Acid Substitution Matrices

Margaret Dayhoff, 1978, PAM Matrices

Explicit evolutionary model
Assumes symmetry: A → B = B → A
Assumes amino acid substitutions observed over short periods of time can be extrapolated to long periods of time

71 groups of protein sequences, 85% similar
1572 amino acid changes.

Functional proteins → mutations "accepted" by natural selection

PAM1 matrix means 1% divergence between proteins - i.e. 1 amino acid change per 100 residues. Some texts re-state this as the probability of each amino acid changing into another is ~ 1% and probability of not changing is ~99%

Construction of a Dayhoff Matrix: PAM1

Step 1: *Measure pairwise substitution frequencies* for each amino acid within families of related proteins that can be confidently aligned

```
…. GDSFHYFVSHG…. .
…. GDSFHYYVSFG…. .
…. GDSYHYFVSFG…. .
…. GDSFHYFVSFG…. .
…. GDSFHFFVSFG…. .
```

900 Phe (F) remained F

100 Phe (F) → 80 Tyr (Y), 3 Trp (W), 2 His (H)….

Gives n_{ab}, i.e. $n_{YF}=80$

$n_{WF}=3$

n indicates raw count of events

….in evolution

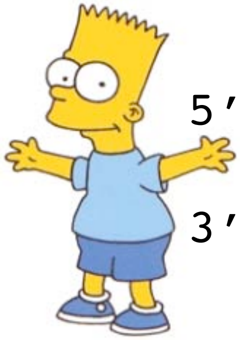
DNA Sequence Evolution

Generation ***n-1*** (grandparent)

```
5’ TGGCATGCACCCTGTAAGTCAATATAAATGGCTACGCCTAGCCCATGCGA 3’
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3’ ACCGTACGTGGGACATTCAGTTATATTTACCGATGCGGATCGGGTACGCT 5’
```

Generation ***n*** (parent)

```
5’ TGGCATGCACCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCATGCGA 3’
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3’ ACCGTACGTGGGACATTCAGTTATATTTACCGATACGGATCGGGTACGCT 5’
```



Generation ***n+1*** (child)

```
5’ TGGCATGCACCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCGTGCGA 3’
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3’ ACCGTACGTGGGACATTCAGTTATATTTACCGATACGGATCGGGCACGCT 5’
```

Markov Model (aka Markov Chain)

Stochastic Process:

- a random process or
- a sequence of Random Variables

Classical Definition

A discrete stochastic process $X_1, X_2, X_3, \ldots$ which has the Markov property:

$$P(X_{n+1} = j \mid X_1=x_1, X_2=x_2, \ldots X_n=x_n) = P(X_{n+1} = j \mid X_n=x_n\,)$$

(for all x_i, all j, all n)

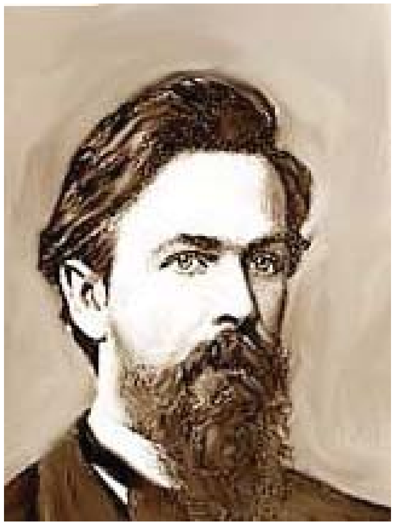


Image is in the public domain.

In words:

A random process which has the property that the future (next state) is conditionally independent of the past given the present (current state)

Andrey Markov, a Russian mathematician (1856 - 1922)