



Era of Bioinformatics

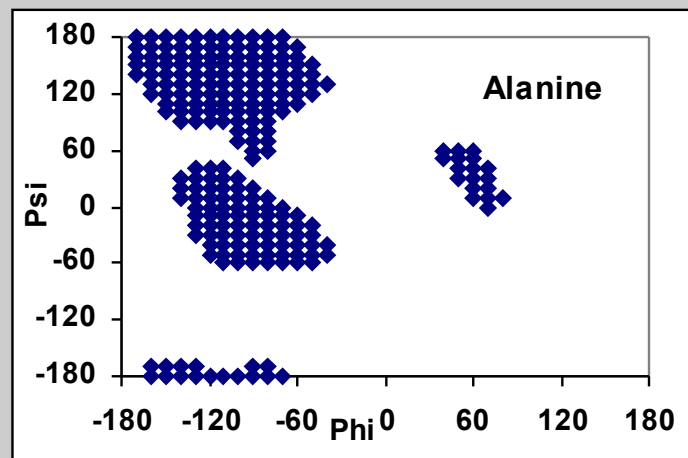
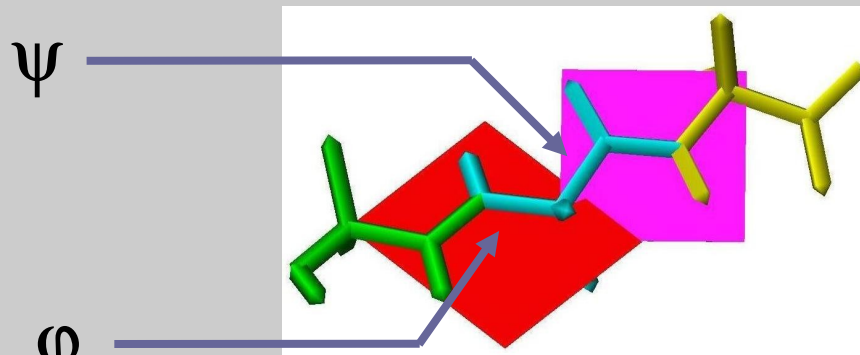
Homayoun Valafar

Department of Computer Science and Engineering, USC



Computational Complexity of Protein Folding

- For a protein of size N amino acids:
 - $df = 2 \times (N - 1)$
 - Each degree of freedom spans 0° - 360°
 - Possible conformations at 10° resolution: $36^{2(N-1)}$
 - $N = 100$, 10^6 struct / sec $\Rightarrow$ 4.4575E+291 millennia
 - NP class of problems.
 - $N=11 \Rightarrow$ 32 millennia
 - $N=11$, 50 angles $\Rightarrow$ 1 millennium





Impetus for Computational Protein Folding

- Origin of most diseases (if not all diseases) can be traced to one or a system of proteins.
- Structure elucidation takes about a year (average)
- Structure elucidation costs in average \$1M / Protein
- Computational protein folding significantly reduces both.
 - Cost to almost zero.
 - Time requirement of about a week (current state).
- Can study the entire proteome of an unknown organism in a matter of months!



Part II

Promise of Bioinformatics



Alternative Approach to Ab-Initio Structure Determination

- Protein folds are limited to only ~10,000 families.
- This observation provides an alternate approach to protein folding.
- Protein folding can be stated as a classification problem!
 - ANN, Bayesian analysis, Fuzzy logic, Cluster analysis & PCA.
 - SVD, Newton's method, Simplex, Gradient descent, SA, GA & DGO.
 - Convolution, DFT, Digital filter design & ICT.
 - Program development, updating of code, parallelizing programs.
- Requires a complete database of all folds.
- The main objective of the structural genomics initiative is the rapid completion of the family fold database.



NIH Initiative for Structural Genomics

- During the fall of 2000, NIGMS announced the following awardees for the pilot programs in the structural genomics.
 - Berkeley Structural Genomics Center
 - The Joint Center for Structural Genomics
 - The Midwest Center for Structural Genomics
 - New York Structural Genomics Research Consortium
 - Northeast Structural Genomics Consortium
 - The Southeast Collaboratory for Structural Genomics
 - TB Structural Genomics Consortium
 - Structural Genomics of Pathogenic Protozoa Consortium
 - Center for Eukaryotic Structural Genomics
- The objective is to develop high-throughput structure determination methods (200 structures per year).



Influence of Bioinformatics in Computational Biology

- Traditionally, research in the field of structural biology is based on interest in function of a particular protein.
- Recent developments in bioinformatics have provided a nearly orthogonal path of research.
- Structure and function of an unknown protein may be predicted from the genome!
- Unimaginable advances can be made in the field of molecular biology and pharmaceutical endeavors.



Evolutionary Relationship

Homayoun Valafar

Department of Computer Science and Engineering, USC



Protein Sequence-Structure-Function Relationship

- Structure is necessary (not sufficient) for function
- Structure determination is very expensive
- Two identical sequences will produce the same structure
 - How about sequences that differ in only one amino acid?
 - How about sequences with 90% identity?
 - How far sequence similarity imposes/signifies structural similarity?
- Need to assess and quantify similarity between two sequences



Evolutionary Relation

- Evolution takes place at the DNA level while fitness is evaluated at the protein level.
- What is the likelihood of finding a particular amino acid in a protein sequence? Is it $1/20$ for all amino acids?
- Can any amino acid be substituted for any other amino acid with the same likelihood?
- Are all amino acids the same?
- Ref 1, 2, 3.
- What is the likelihood that two sequences are descendants of the same parent sequence?



Alignment Score S

- Total score S of an alignment is the sum of all s.
- Positive s or S is good.
- Negative s or S is not good.
- Example:
 - AIF and SIF? AIF and FIF? Which relationship is more likely?
 - AIF and FRD? AIF and SLL? Which pair are more likely relatives?
- Which is a better alignment:

BBAAACD
BBBAAAD

or

BBAAACD
BBBAAAD



Blosom Substitution Matrices

Blosom50

A	Ala
C	Cys
D	Asp
E	Glu
F	Phe
G	Gly
H	His
I	Ile
K	Lys
L	Leu
M	Met
N	Asn
P	Pro
Q	Gln
R	Arg
S	Ser
T	Thr
V	Val
W	Trp
Y	Tyr

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

$$s(x,y) = \log \left(\frac{p_{xy}}{p_x p_y} \right)$$

P_{xy} is the probability that x and y are evolutionarily related.

P_x is the probability of occurrence of x.

P_y is the probability of occurrence of y.



Alignment Example

- Align the following sequences:
 - HEAGAWGHEE
 - PAWHEAE
- Sometimes alteration of a sequence is not based on substitution.
 - Insertion or deletion of an amino acid.
 - How to deal with these?
 - Penalty for insertion is $-d$ ($d > 0$).
 - Penalty for extension of gap is $-e$ ($e > 0$ and normally less than $e < d$).
 - Gap-opening and gap-extension penalties



Alignment Algorithms

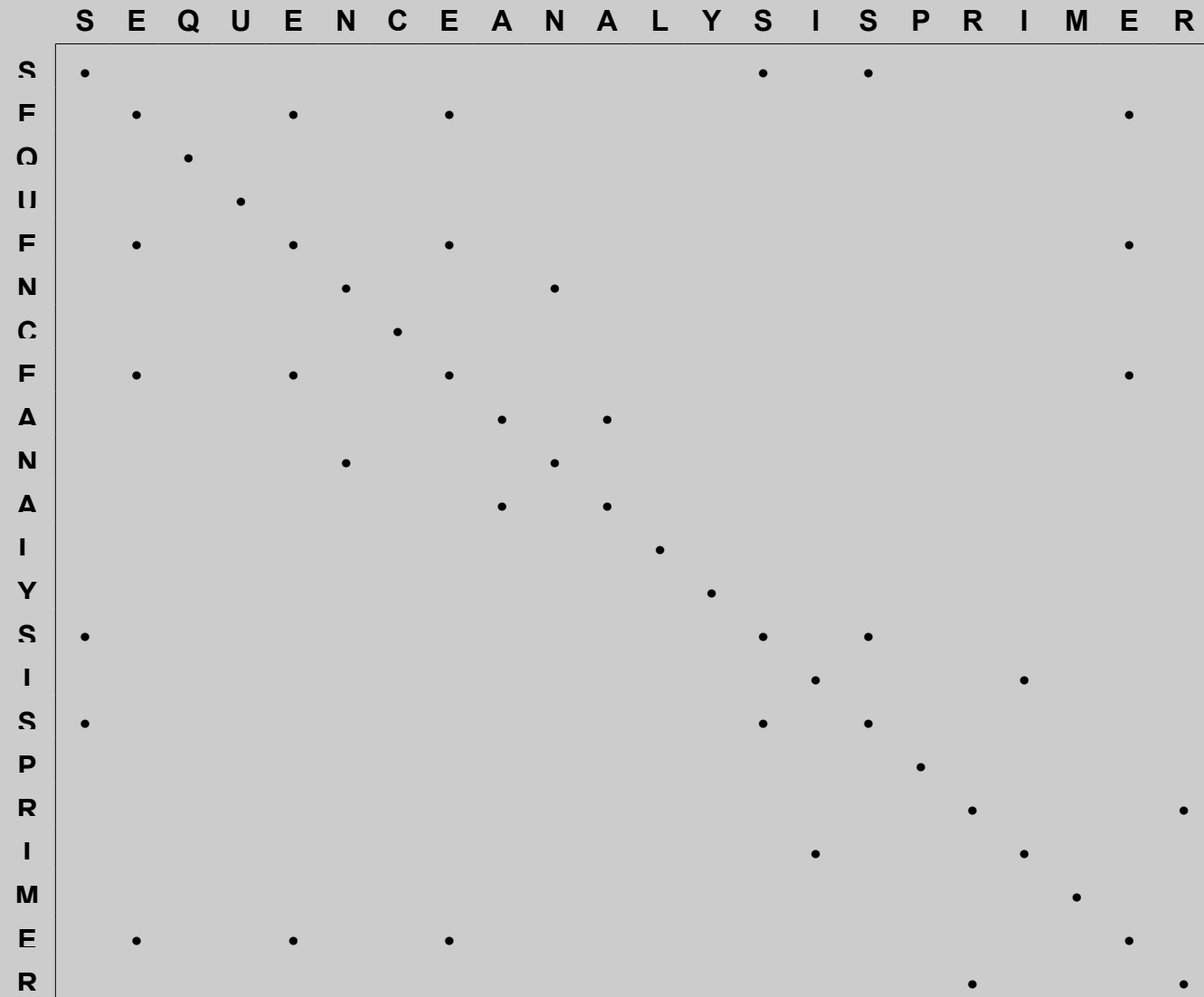
Homayoun Valafar

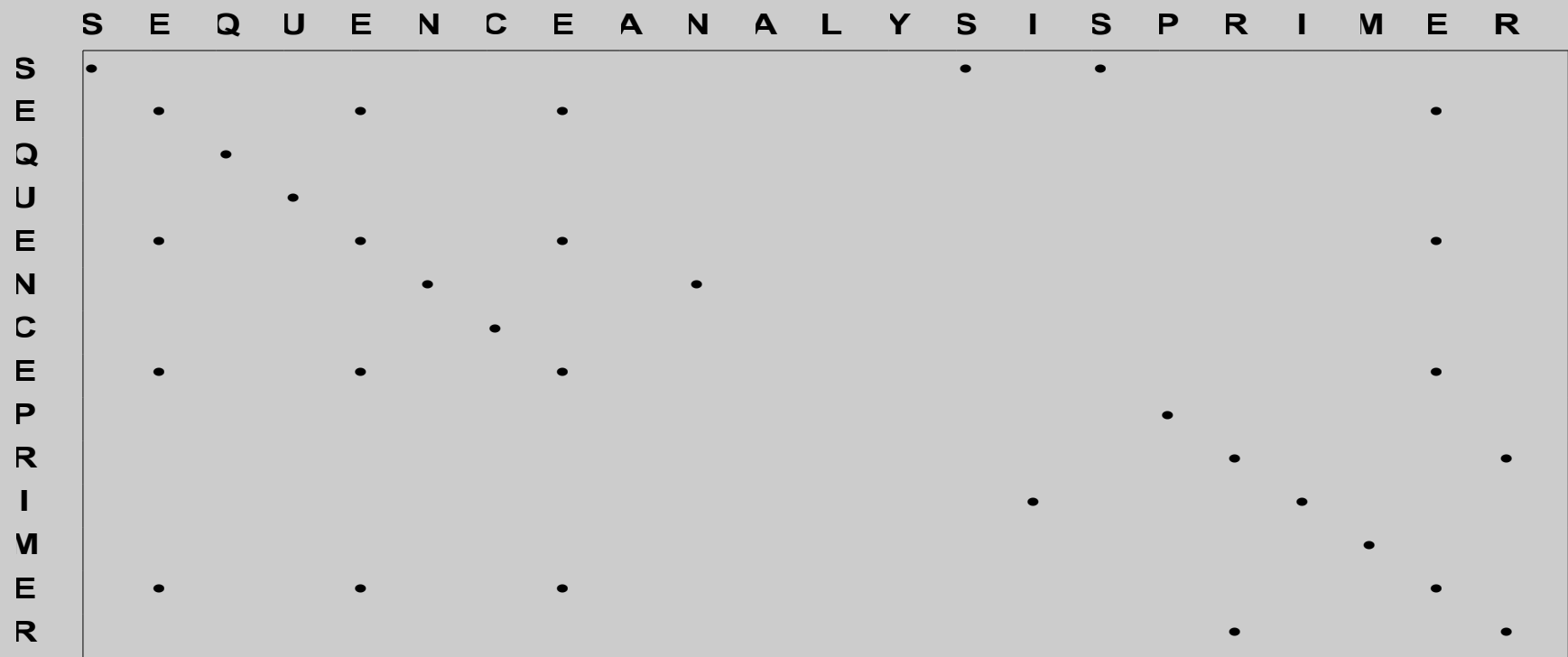
Department of Computer Science and Engineering, USC

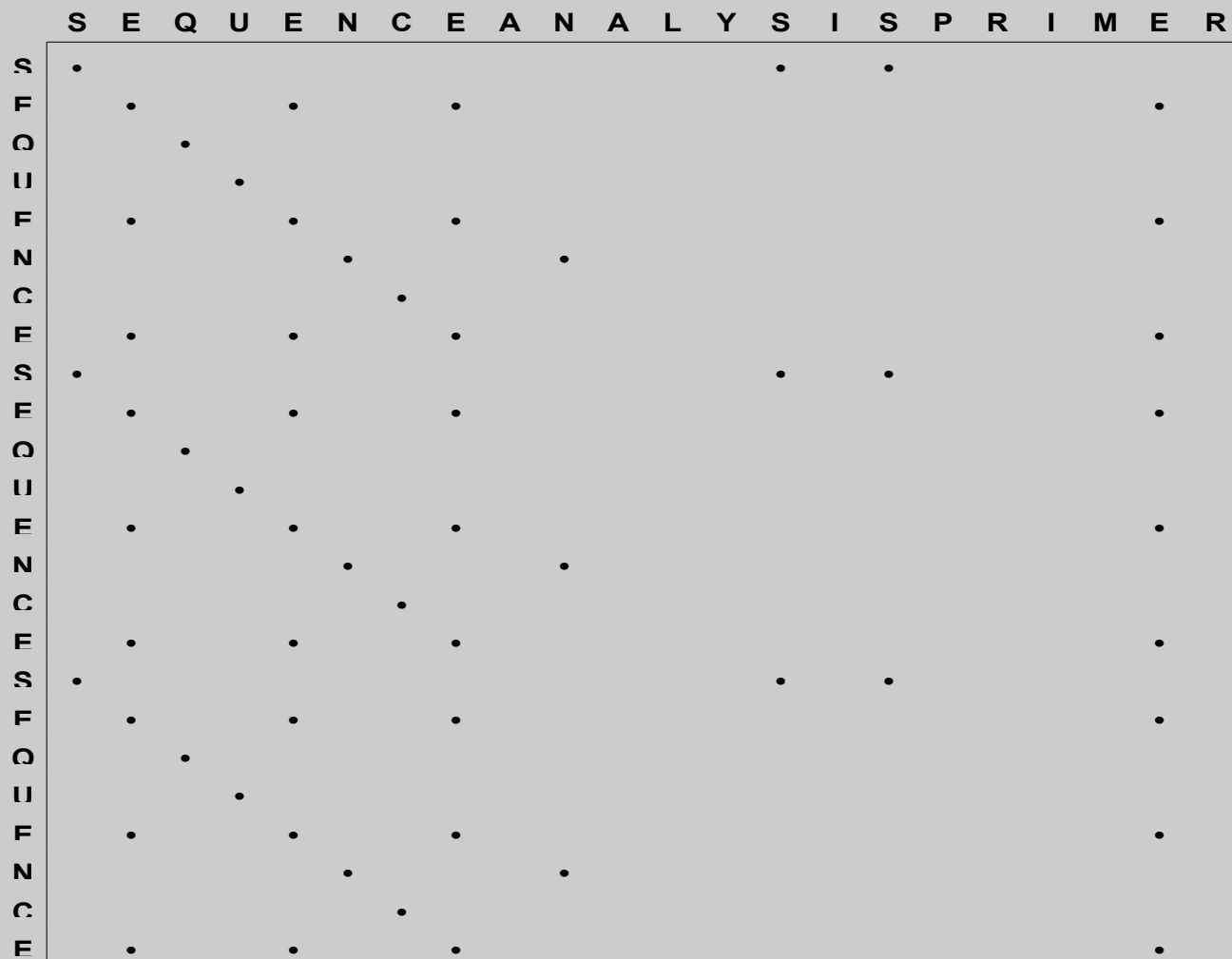


Dot Matrix

- Put one sequence on top.
- Put one sequence on side.
- Put a dot on every grid with matching letters.
- Patterns will imerge.
- Advantages:
 - Very simple and requires no a-priori knowledge of anything.
- Disadvantages:
 - Does not take into account a-priori knowledge.
 - Does not allow global alignment.
 - Requires human intervention.









Needleman-Wunsch Algorithm

- Produces optimal global alignment of two sequences
- First sequence X with size m and elements x_i
- Second sequence Y with size n and elements y_j
- Create a matrix/table $F(i,j)$ of size $(m+1) \times (n+1)$
- Each index corresponds to i -th character of X and j -th character of Y
- X spans the columns of F and Y spans the rows of F
- Each $F(i,j)$ contains the best score of alignment up to location i in sequence X and j in sequence Y
- Horizontal move is a gap in Y , vertical move is a gap in X and diagonal move is matching of x_i to y_j



Alignment Example

- Align the following sequences:
 - HEAGAWGHEE
 - PAWHEAE
 - Gap penalty of -8, extension penalty of -8.



The Score Matrix F

- Using the following rules, complete the F matrix in three steps

- 1) Complete the first row
- 2) Complete the first column
- 3) Complete the internal cells

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{array} \right\}$$

		H	E	A	G	A	W	G	H	E
	0									
P										
A										
W										
H										
E										
A										
E										

↓ j

→ i



Step 1 – Complete first row

Horizontal transition on the $F(i,j)$ matrix signifies a “GAP” in the Y sequence

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{array} \right\}$$

		H	E	A	G	A	W	G	H	E
	0	-8								
P										
A										
W										
H										
E										
A										
E										



Step 1 – Complete first row

Subsequent horizontal transitions on the $F(i,j)$ matrix signify “Gap Extensions” in the Y sequence

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{array} \right\}$$

		H	E	A	G	A	W	G	H	E
	0	-8	-16							
P										
A										
W										
H										
E										
A										
E										



Step 1 – Complete first row

Complete the $F(i,0)$

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{array} \right\}$$

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P										
A										
W										
H										
E										
A										
E										



Step 2 – Complete first column

Vertical transition on the $F(i,j)$ matrix signifies a “GAP” in the X sequence

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{array} \right\}$$

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8									
A										
W										
H										
E										
A										
E										



Step 2 – Complete first column

Subsequent vertical transitions on the $F(i,j)$ matrix signify “Gap Extensions” in the Y sequence

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{array} \right\}$$

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8									
A	-16									
W										
H										
E										
A										
E										



Step 2 – Complete first column

Complete $F(0,j)$

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{array} \right\}$$

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8									
A	-16									
W	-24									
H	-32									
E	-40									
A	-48									
E	-56									



Step 3 – Complete internal elements

- For each cell (i,j) three scores can be computed:

- Vertical move from F(i,j-1)
- Horizontal move from F(i-1,j)
- Diagonal move from F(i-1,j-1)

$$F(i,j) = \max \left\{ \begin{array}{l} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{array} \right\}$$

- Select and record the max score and direction

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8									
A	-16									
W	-24									
H	-32									
E	-40									
A	-48									
E	-56									

↓ j

→ i



Step 3 – Complete internal elements

$$F(1,1) = \max \begin{pmatrix} F(0,0) + s(x_1, y_1) \\ F(0,1) - 8 \\ F(1,0) - 8 \end{pmatrix} = \max \begin{pmatrix} 0 + s(H, P) \\ -8 - 8 \\ -8 - 8 \end{pmatrix} = \max \begin{pmatrix} 0 + (-2) \\ -16 \\ -16 \end{pmatrix}$$

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8	-2								
A	-16									
W	-24									
H	-32									
E	-40									
A	-48									
E	-56									



Blosom Substitution Matrices

Blosom50

A	Ala
C	Cys
D	Asp
E	Glu
F	Phe
G	Gly
H	His
I	Ile
K	Lys
L	Leu
M	Met
N	Asn
P	Pro
Q	Gln
R	Arg
S	Ser
T	Thr
V	Val
W	Trp
Y	Tyr

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

$$s(x,y) = \log \left(\frac{p_{xy}}{p_x p_y} \right)$$

P_{xy} is the probability that x and y are evolutionarily related.

P_x is the probability of occurrence of x.

P_y is the probability of occurrence of y.



Step 3 – Complete internal elements

- Trace back your transition from the bottom right corner to the top left corner by referring back to the transition matrix

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9



Step 3 – Complete internal elements

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9



Interpret Alignment

- Horizontal transition represents a gap in the vertical sequence
- Vertical transition represents a gap in the horizontal sequence
- Diagonal transition represents a match in the corresponding characters of the two sequences

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9

H E A G A W G H _ E
- - P - A W H E A E



Needleman-Wunsch Algorithm

- Very useful for global alignment of sequences:

VLSEGEWQLVLHVVAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED 60

VLSEGEWQLVLHVVAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED

VLSEGEWQLVLHVVAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED 60

- Global alignment implies close evolutionary relation.
- What if two sequences are distantly related?
 - A large middle section of a protein is deleted.
- Need to perform local alignment.
 - Smith Waterman Algorithm.



Smith-Waterman Algorithm

- Find the best local alignment of the following sequences:
 - HEAGAWGHEE
 - PAWHEAE
 - Gap penalty of -8, extension penalty of -8.
- Start from the largest score and trace back

$$F(i,j) = \max \begin{pmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{pmatrix}$$

		H	E	A	G	A	W	G	H	E
	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0
W	0	0	0	0	2	0	20	12	0	0
H	0	10	2	0	0	0	12	18	22	14
E	0	2	16	8	0	0	4	10	18	28
A	0	0	8	21	13	5	0	4	10	20
E	0	0	6	13	18	12	4	0	4	16



Sequence Alignment

Homayoun Valafar

Department of Computer Science and Engineering, USC



Basic Local Alignment Search Tool (BLAST)

- Exercise: Perform BLAST search on the following sequences:
- 1I92:A NA+/H+ EXCHANGE REGULATORY CO-FACTOR mutated by 0.5 45 out of 91.

```
CAAATGCTTCCTTGTCTTTGTTGGTGTATAAAGGTCCTAATGTTATTGCTTTTCATTGT  
GTTATTTCTAAATGGTATCTTGGTCAATATATTGAAGATGTTGATAAACATTTTCCTGCT  
ATGTCTGCTTCTATTATTGCTGGTTATGATTGTTTTGAAGTTAATAATAAAAATGTTGAA  
AAACTACTCATCCTGAAGAAGTTTCTTTTATTCTTGCTGCTCGTAATAATAAACGTATG  
CTTCTTTGGGATCCTGAACAAGCTGCTCGTCTT
```

- 1SF0

```
AHHHHHHGSK MIKVKVIGRN IEKEIEWREG MKVRDILRAV GFNTESAIK VNGKVVLEDD  
  
EVKDGDFVEV IPVVS GG
```