

CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS08.html

Genomic Databases

- **Entrez Portal** at National Center for Biotechnology Information (**NCBI**) gives access to:
 - Nucleotide (*GenBank, EMBL, DDBJ*)
 - Protein (*PIR, SwissPROT, PRF*, and Protein Data Bank or *PDB*)
 - Genome
 - Structure
 - 3D Domains
 - Conserved Domains
 - Gene; UniGene; HomoloGene; SNP
 - GEO Profiles & Datasets
 - Cancer Chromosomes
 - PubMed Central; Journals; Books
 - OMIM
 - Database Neighbors and Interlinking

Sequence Alignment – Why?

>gi|12643549|sp|O18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)

```
MRNLPCLGTAGGSGGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNLGGVVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS
QYKRECPSIFAWEIRDRLLEQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPCTSDNDNGRQSEDCRRVCSPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSGFAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGRPAGVGLGSGQSANLGASCSSGSGYEVLSAYALPPPMASSAADSFFSAASSAS
ANVTPHHTIAQESPCSSASHFGVAHSSGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHV
APGKQQFFASCFYSPWV
```

>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)

```
MQNSHSGVNLGGVFNVRPLPDSTRQKIVELAHSGARPCDISRILQVSNNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLEQENVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL
QRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASN
TPSHIPISSSFSTSVYQPIQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQ
TSSYSCMLPTSPSVNGRSYDITYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
```

Drosophila Eyeless vs. Human Aniridia

```
Query: 57 HSGVNQLGGVFGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYETG 116
          HSGVNQLGGVFG GRPLPDSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYETG
Sbjct: 5   HSGVNQLGGVFGVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYETG 64

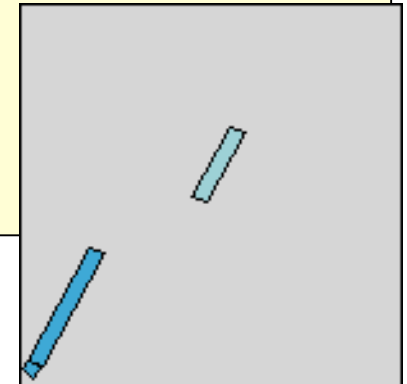
Query: 117 SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLEQENVCTNDNIPSVSSIN 176
          SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLLE E VCTNDNIPSVSSIN
Sbjct: 65  SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLESEGVCTNDNIPSVSSIN 124

Query: 177 RVLRLNLAQKEQ 188
          RVLRLNLA++K+Q
Sbjct: 125 RVLRLNLAASEKQQ 136
```

```
Query: 417 TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV 476
          +++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
Sbjct: 197 SDEAQMRLQLKRKLQRNRTSFTQEIQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQV 256

Query: 477 WFSNRRRAKWRREEKLRNQRR 496
          WFSNRRRAKWRREEKLRNQRR
Sbjct: 257 WFSNRRRAKWRREEKLRNQRR 276
```

E-Value = $2e^{-31}$



Implications of Sequence Alignment

- ❑ **Mutation** in DNA is a natural evolutionary process. Thus sequence similarity may indicate **common ancestry**.
- ❑ In biomolecular sequences (DNA, RNA, protein), high sequence similarity implies significant **structural and/or functional similarity**.

Discovery based on alignments

- ❑ **Early 1970s:** Simian sarcoma virus causes cancer in some species of monkeys.
- ❑ **1970s:** infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.
 - **Hypothesis:** Certain genes (oncogenes) in viruses encode cellular growth factors, which are proteins needed to stimulate growth of a cell colony. Thus uncontrolled quantities of growth factors produced by the infected cells cause cancer-like behavior.
- ❑ **1983:**
 - The oncogene from SSV called **v-sis** was isolated and sequenced.
 - The partial amino-acid sequence for platelet-derived growth factor (PDGF) was sequenced and published. It stimulates the proliferation of normal cells.
 - R.F. Doolittle was maintaining one of the earliest home-grown databases of published amino-acid sequences.
 - Sequence Alignment of v-sis and PDGF showed something surprising.

PDGF and v-sis

- ❑ One region of 31 amino acids had 26 exact matches
- ❑ Another region of 39 residues had 35 exact matches.
- ❑ **Conclusion:**
 - The previously harmless virus incorporates the normal growth-related gene (proto-oncogene) of its host into its genome.
 - The gene gets mutated in the virus, or moves closer to a strong enhancer, or moves away from a repressor.
 - This causes an uncontrolled amount of the product (the growth factor, for example) when the virus infects a cell.
- ❑ Several other oncogenes known to be similar to growth-regulating proteins in normal cells.

V-sis Oncogene - Homologies

Sequences producing significant alignments:		Score	E
		(bits)	Value
gi 332623 gb J02396.1 SEG_SSVPCS2	Simian sarcoma virus v-si...	4591	0.0
gi 61774 emb V01201.1 RESSV1	Simian sarcoma virus proviral ...	4504	0.0
gi 332622 gb J02395.1 SEG_SSVPCS1	Simian sarcoma virus LTR ...	1283	0.0
gi 885929 gb U20589.1 GLU20589	Gibbon leukemia virus envelo...	1140	0.0
gi 4505680 ref NM_002608.1	Homo sapiens platelet-derived g...	954	0.0
gi 20987438 gb BC029822.1	Homo sapiens, platelet-derived g...	954	0.0
gi 338210 gb M12783.1 HUMSISPDG	Human c-sis/platelet-derive...	954	0.0

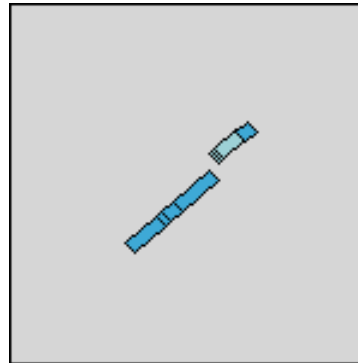
Sequence Alignment

```
>gi|4505680|ref|NM_002608.1| Homo sapiens platelet-derived growth
factor beta polypeptide (simian sarcoma viral (v-sis) oncogene homolog)
(PDGFB), transcript variant 1, mRNA Length = 3373 Score = 954 bits
(481), Expect = 0.0 Identities = 634/681 (93%), Gaps = 3/681 (0%)
Strand = Plus / Plus
Query: 1015 agggggacccattcctgaggagctctataagatgctgagtggccactcgattcgctcct 1074
      |||
Sbjct: 1084 agggggacccattcccgaggagctttatgagatgctgagtgaccactcgatccgctcct 1143
      > 21 E G D P I P E E L Y E M L S D H S I R S
Query: 1075 tcgatgacctccagcgcctgctgcagggagactccggaaaagaagatggggctgagctgg 1134
      | |||
Sbjct: 1144 ttgatgatctccaacgcctgctgcacggagaccccgagaggaagatggggccgagttgg 1203
      > 61 D L N M T R S H S G G E L E S L A R G R
```

Sequence Alignment

Sequence 1 [gi 332624](#) Simian sarcoma virus v-sis transforming protein p28 gene, complete cds; and 3' LTR long terminal repeat, complete sequence. **Length** 2984 (1 .. 2984)

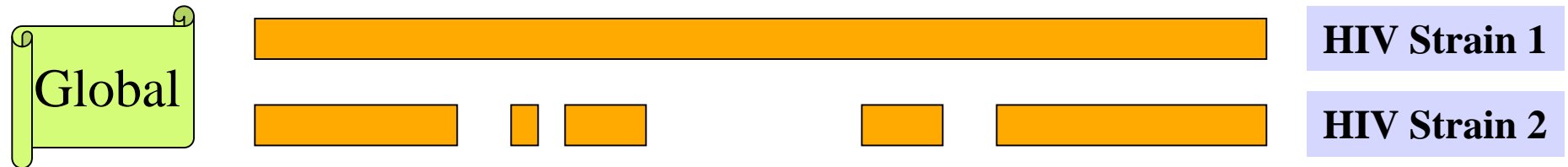
Sequence 2 [gi 4505680](#) Homo sapiens platelet-derived growth factor beta polypeptide (simian sarcoma viral (v-sis) oncogene homolog) (PDGFB), transcript variant 1, mRNA **Length** 3373 (1 .. 3373)



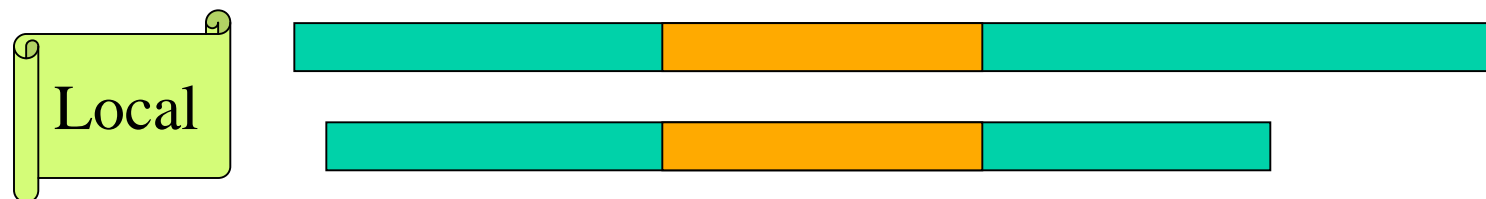
Similarity vs. Homology

- **Homologous** sequences share common ancestry.
- **Similar** sequences are "near" to each other by some appropriately defined measurable criteria.

Types of Sequence Alignments - 1



Global Alignment: similarity over entire length



Local Alignment: no overall similarity, but some segment(s) is/are similar

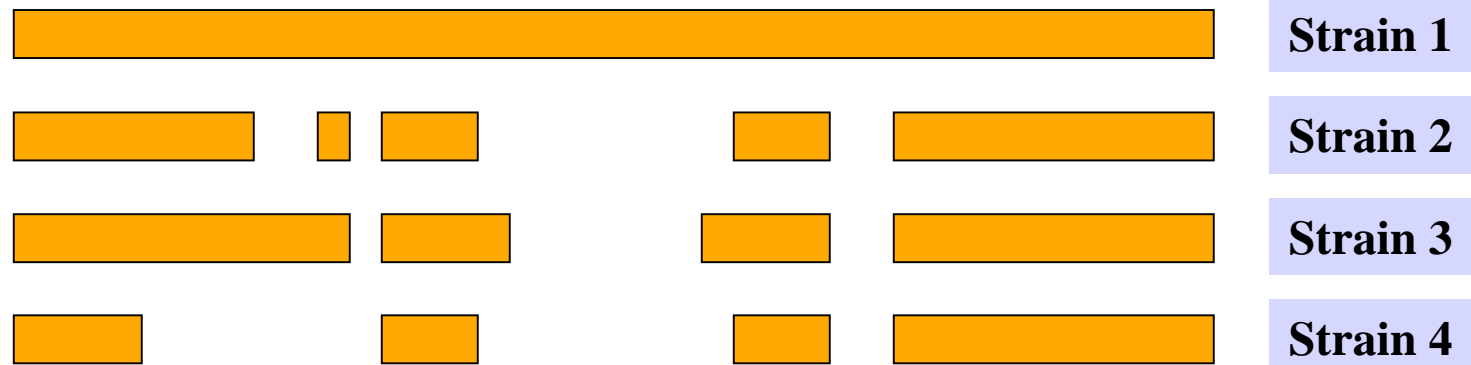
Types of Sequence Alignments - 2

Semi-Global



□ **Semi-global Alignment:** end segments may not be similar

Multiple



□ **Multiple Alignment:** similarity between sets of sequences

Sequence Alignment

Global:

- Needleman-Wunsch-Sellers (1970).

Local:

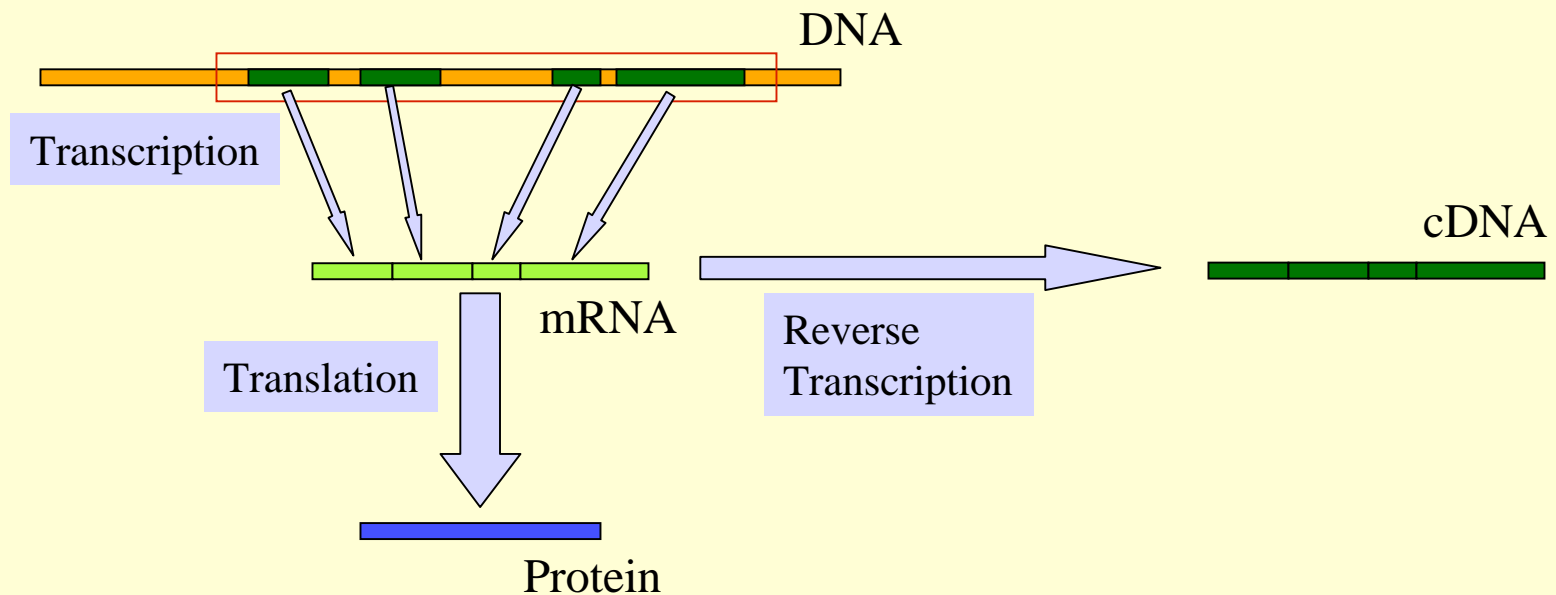
- Smith-Waterman (1981)

- Useful when commonality is small and global alignment is meaningless. Often unaligned portions "mask" short stretches of aligned portions. Example: comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.

Dynamic Programming (DP) based.

Why gaps?

- Example: Finding the gene site for a given (eukaryotic) cDNA requires "gaps".
- What is cDNA? cDNA = Copy DNA



How to score mismatches?

	A	C	D	E	F	G	H	→
A	4	0	-2	-1	-2	0	-2	
C	0	9	-3	-4	-2	-3	-3	
D	-2	-3	6	2	-3	-1	-1	
E	-1	-4	2	5	-3	-2	0	
F	-2	-2	-3	-3	6	-3	-	
G	0	-3	-1	-2	-3			
H	-2	-3	-1	0				

BLOSUM 62

BLAST & FASTA

- FASTA

 - [Lipman, Pearson '85, '88]

- Basic Local Alignment Search Tool

 - [Altschul, Gish, Miller, Myers, Lipman '90]

BLAST Overview

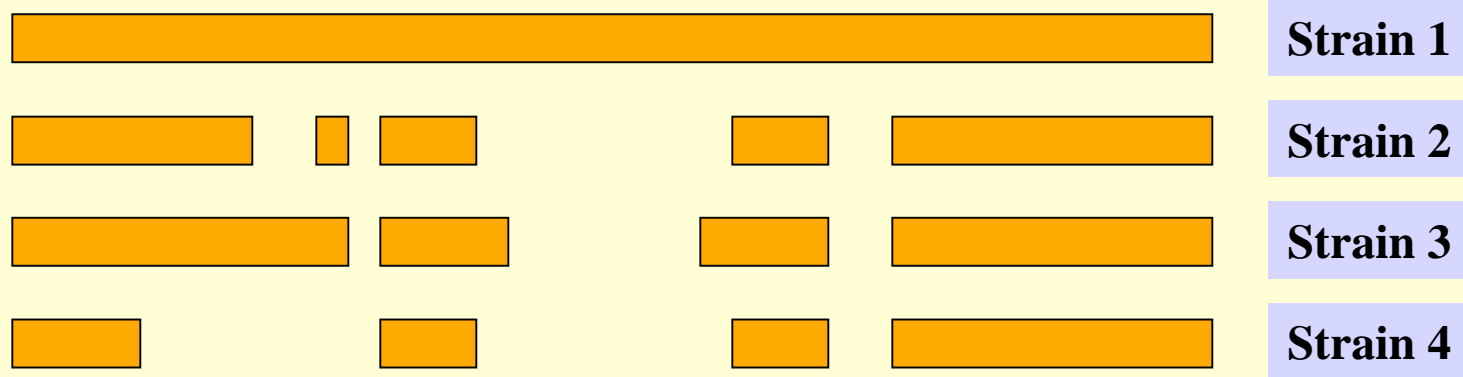
- ❑ Program(s) to search all sequence databases
- ❑ Tremendous Speed/Less Sensitive
- ❑ Statistical Significance reported
- ❑ WWWBLAST, QBLAST (send now, retrieve results later), Standalone BLAST, BLASTcl3 (Client version, TCP/IP connection to NCBI server), BLAST URLAPI (to access QBLAST, no local client)

BLAST Strategy & Improvements

- ❑ Lipman et al.: speeded up finding “runs” of “hot spots”.
- ❑ Eugene Myers '94: “Sublinear algorithm for approximate keyword matching”.
- ❑ Karlin, Altschul, Dembo '90, '91: “Statistical Significance of Matches”

Why Gaps?

□ Example: Aligning HIV sequences.



BLAST Variants

☐ Nucleotide BLAST

- **Standard blastn**
- **MEGABLAST** (Compare large sets, Near-exact searches)
- **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering)

☐ Protein BLAST

- **Standard blastp**
- **PSI-BLAST** (Position Specific Iterated BLAST)
- **PHI-BLAST** (Pattern Hit Initiated BLAST; reg expr. Or Motif search)
- **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering, PAM-30)

☐ Translating BLAST

- **Blastx**: Search nucleotide sequence in protein database (6 reading frames)
- **Tblastn**: Search protein sequence in nucleotide dB
- **Tblastx**: Search nucleotide seq (6 frames) in nucleotide DB (6 frames)

BLAST Cont'd

❑ RPS BLAST

- Compare protein sequence against Conserved Domain DB; Helps in predicting rough structure and function

❑ Pairwise BLAST

- blastp (2 Proteins), blastn (2 nucleotides), tblastn (protein-nucleotide w/ 6 frames), blastx (nucleotide-protein), tblastx (nucleotide w/6 frames-nucleotide w/ 6 frames)

❑ Specialized BLAST

- Human & Other finished/unfinished genomes
- *P. falciparum*: Search ESTs, STSs, GSSs, HTGs
- VecScreen: screen for contamination while sequencing
- IgBLAST: Immunoglobulin sequence database

BLAST Credits

- Stephen Altschul
- Jonathan Epstein
- David Lipman
- Tom Madden
- Scott McGinnis
- Jim Ostell
- Alex Schaffer
- Sergei Shavirin
- Heidi Sofia
- Jinghui Zhang

Databases used by BLAST

Protein

● nr (everything), swissprot, pdb, alu, individual genomes

Nucleotide

● nr, dbest, dbsts, htgs (unfinished genomic sequences), gss, pdb, vector, mito, alu, epd

Misc

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.

Types of Sequence Alignments

Global



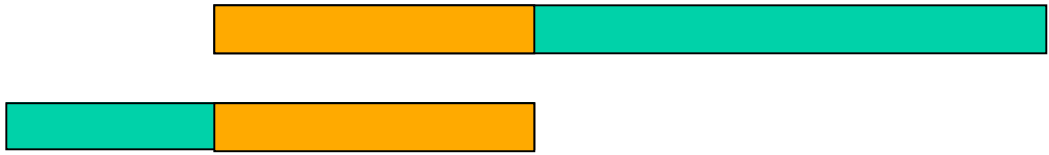
HIV Strain 1

HIV Strain 2

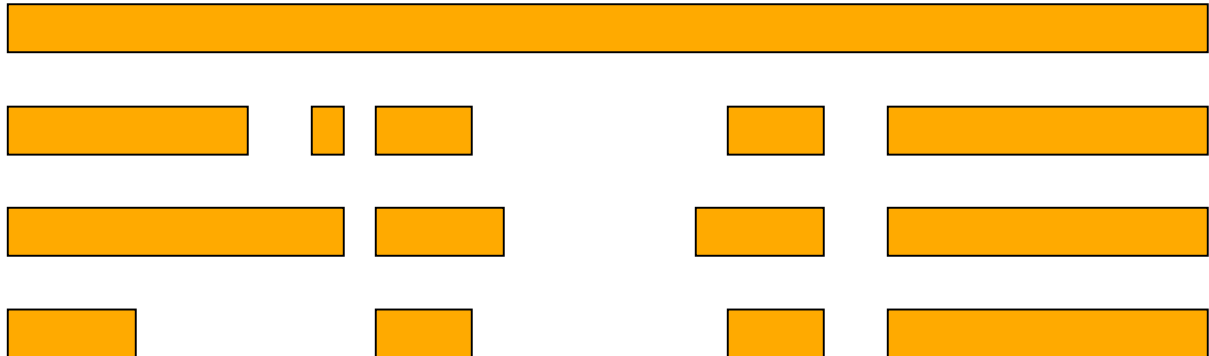
Local



Semi-Global



Multiple



Strain 1

Strain 2

Strain 3

Strain 4

Global Alignment: An example

V: G A A T T C A G T T A
W: G G A T C G A

	G	A	A	T	T	C	A	G	T	T	A
G	0										
G	0										
A	0										
T	0										
C	0										
G	0										
A	0										

Given

$\delta[I, J]$ = Score of Matching
the I^{th} character of sequence V &
the J^{th} character of sequence W

Compute

$S[I, J]$ = Score of Matching
First I characters of sequence V &
First J characters of sequence W

Recurrence Relation

$S[I, J] = \text{MAXIMUM} \{$
 $S[I-1, J-1] + \delta(V[I], W[J]),$
 $S[I-1, J] + \delta(V[I], -),$
 $S[I, J-1] + \delta(-, W[J]) \}$

Global Alignment: An example

$$S[I, J] = \text{MAXIMUM} \{$$

$$S[I-1, J-1] + \delta(V[I], W[J]),$$

$$S[I-1, J] + \delta(V[I], -),$$

$$S[I, J-1] + \delta(-, W[J]) \}$$

V: G A A T T C A G T T A
W: G G A T C G A

	G	A	A	T	T	C	A	G	T	T	A
0	0	0	0	0	0	0	0	0	0	0	0
G	0										
G	0										
A	0										
T	0										
C	0										
G	0										
A	0										

	G	A	A	T	T	C	A	G	T	T	A
0	0	0	0	0	0	0	0	0	0	0	0
G	0	1									
G	0										
A	0										
T	0										
C	0										
G	0										
A	0										

	G	A	A	T	T	T	C	G	T	T	A
0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1									
A	0	1									
T	0	1									
C	0	1									
G	0	1									
A	0	1									

	G	A	A	T	T	C	A	G	T	T	A
0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1								
A	0	1	2								
T	0	1	2								
C	0	1	2								
G	0	1	2								
A	0	1	2								

	G	A	A	T	T	C	A	G	T	T	A
0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1							
A	0	1	2	2							
T	0	1	2	2							
C	0	1	2	2							
G	0	1	2	2							
A	0	1	2	3							

	G	A	A	T	T	C	A	G	T	T	A
0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5
A	0	1	2	3	3	3	3	4	5	5	6

Traceback

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	1	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A	0	1	2	3	3	3	4	5	5	5	6

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	1	1	1	1	1	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											6

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	1	1	1	1	1	1	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											6

	G	A	A	T	T	C	A	G	T	T	A
G	0										
G		1									
A			1								
T				2	2						
C					3						
G						4	4				
A							5	5	5		
A											6

V: G A A T T C A G T T A
 | | | | |
 W: G G A - T C - G - - A

Alternative Traceback

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A	0	1	2	3	3	3	4	5	5	5	6

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											6

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											6

	G	A	A	T	T	C	A	G	T	T	A
G	0										
G		1									
A		1	1								
T				2	2						
C						3					
G							4	4			
A									5	5	5
											6

V: G - A A T T C A G T T A
 | | | | | | | |
 W: G G - A - T C - G - - A

V: G A A T T C A G T T A
 | | | | | | | |
 W: G G A - T C - G - - A

Previous

Improved Traceback

	G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0
G	0	x1	←1	←1	←1	←1	←1	←1	x1	←1	←1
G	0	x1	↑1	↑1	↑1	↑1	↑1	↑1	x2	←2	←2
A	0	↑1	↑1	x2	←2	←2	←2	x2	↑2	↑2	↑2
T	0	↑1	←2	↑2	x3	x3	←3	←3	←3	x3	x3
C	0	↑1	↑2	↑2	↑3	↑3	x4	←4	←4	←4	←4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	x5	←5	←5
A	0	↑1	↑2	x3	↑3	↑3	↑4	x5	↑5	↑5	↑5

Improved Traceback

	G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0
G	0	x1	←1	←1	←1	←1	←1	←1	x1	←1	←1
G	0	x1	↑1	↑1	↑1	↑1	↑1	↑1	x2	←2	←2
A	0	↑1	↑1	x2	←2	←2	←2	x2	↑2	↑2	↑2
T	0	↑1	←2	↑2	x3	x3	←3	←3	←3	x3	x3
C	0	↑1	↑2	↑2	↑3	↑3	x4	←4	←4	←4	←4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	x5	←5	←5
A	0	↑1	↑2	x3	↑3	↑3	↑4	x5	↑5	↑5	↑5

Improved Traceback

	G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0
G	0	x1	←1	←1	←1	←1	←1	←1	x1	←1	←1
G	0	x1	↑1	↑1	↑1	↑1	↑1	↑1	x2	←2	←2
A	0	↑1	↑1	x2	←2	←2	←2	x2	↑2	↑2	↑2
T	0	↑1	←2	↑2	x3	x3	←3	←3	←3	x3	x3
C	0	↑1	↑2	↑2	↑3	↑3	x4	←4	←4	←4	←4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	x5	←5	←5
A	0	↑1	↑2	x3	↑3	↑3	↑4	x5	↑5	↑5	↑5

V: G A - A T T C A G T T A

| | | | | | |

W: G - G A - T C - G - - A

Subproblems

- Optimally align $V[1..I]$ and $W[1..J]$ for every possible values of I and J .
 - Having optimally aligned
 - $V[1..I-1]$ and $W[1..J-1]$
 - $V[1..I]$ and $W[1..J-1]$
 - $V[1..I-1]$ and $W[1, J]$
- it is possible to optimally align $V[1..I]$ and $W[1..J]$

- $O(mn)$,
where m = length of V ,
and n = length of W .

Generalizations of Similarity Function

- ❑ Mismatch Penalty = α
- ❑ Spaces (Insertions/Deletions, InDels) = β
- ❑ Affine Gap Penalties:
(Gap open, Gap extension) = (γ, δ)
- ❑ Weighted Mismatch = $\Phi(a, b)$
- ⊖ Weighted Matches = $\Omega(a)$

Alternative Scoring Schemes

	G	A	A	T	T	C	A	G	T	T	A	
0	0	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
G	-2	x 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
G	-3	↑ -1	x -1	← -3	← -4	← -5	← -6	← -7	x -5	← -7	← -8	← -9
A	-4	↑ -2	x 0	x 0	← -2	← -3	← -4	← -5	← -6	← -7	← -8	x -7
T	-5	↑ -3	↑ -2	↑ -2	x 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7
C	-6	↑ -4	↑ -3	↑ -3	↑ -1	x -1	x 0	← -2	← -3	← -4	← -5	← -6
G	-7	↑ -5	↑ -4	↑ -4	↑ -2	↑ -3	↑ -2	x -2	x -1	← -3	← -4	← -5
A	-8	↑ -6	↑ -5	↑ -5	↑ -3	↑ -4	↑ -3	x -1	↑ -3	x -3	x -5	x -3

Match +1
Mismatch -2
Gap (-2, -1)

V: G A A T T C A G T T A
| | | | | |
W: G G A T - C - G - - A

Local Sequence Alignment

- **Example:** comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.
- **Smith-Waterman** Algorithm

Recurrence Relations (Global vs Local Alignments)

□ $S[I, J] = \text{MAXIMUM} \{$
 $S[I-1, J-1] + \delta(V[I], W[J]),$
 $S[I-1, J] + \delta(V[I], -),$
 $S[I, J-1] + \delta(-, W[J]) \}$

Global
Alignment

□ $S[I, J] = \text{MAXIMUM} \{ 0,$
 $S[I-1, J-1] + \delta(V[I], W[J]),$
 $S[I-1, J] + \delta(V[I], -),$
 $S[I, J-1] + \delta(-, W[J]) \}$

Local
Alignment

Local Alignment: Example

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	×1	0	0	0	0	0	0	0	0	0	0
G	0	×1	←0	0	0	0	0	0	×1	0	0	0
A	0	0	×2	×1	0	0	0	×1	0	0	0	×1
T	0	0	↑0	×1	×2	←1	0	0	0	×1	×1	0
C	0	0	0	0	↑0	×0	×2	0	0	0	0	0
G	0	0	0	0	0	0	0	0	×1	0	0	0
A	0	0	×1	×1	0	0	0	×1	0	0	0	×1

Match +1
Mismatch -1
Gap (-1, -1)

V: - G A A T T C A G T T A
 | | | |
 W: G G - A T - C - G - - A

Properties of Smith-Waterman Algorithm

- How to find all regions of "high similarity"?
 - Find **all** entries above a threshold score and traceback.
- What if: Matches = 1 & Mismatches/spaces = 0?
 - Longest Common Subsequence Problem
- What if: Matches = 1 & Mismatches/spaces = $-\alpha$?
 - Longest Common Substring Problem
- What if the average entry is positive?
 - Global Alignment

How to score mismatches?

	A	C	D	E	F	G	H	→
A	4	0	-2	-1	-2	0	-2	
C	0	9	-3	-4	-2	-3	-3	
D	-2	-3	6	2	-3	-1	-1	
E	-1	-4	2	5	-3	-2	0	
F	-2	-2	-3	-3	6	-3	-1	
G	0	-3	-1	-2	-3	6	-1	
H	-2	-3	-1	0	-1	-1	6	

BLOSUM 62

BLOSUM n Substitution Matrices

□ For each amino acid pair a, b

● For each BLOCK

- Align all proteins in the BLOCK
- Eliminate proteins that are more than $n\%$ identical
- Count $F(a), F(b), F(a,b)$
- Compute **Log-odds Ratio**

$$\log\left(\frac{F(a,b)}{F(a)F(b)}\right)$$