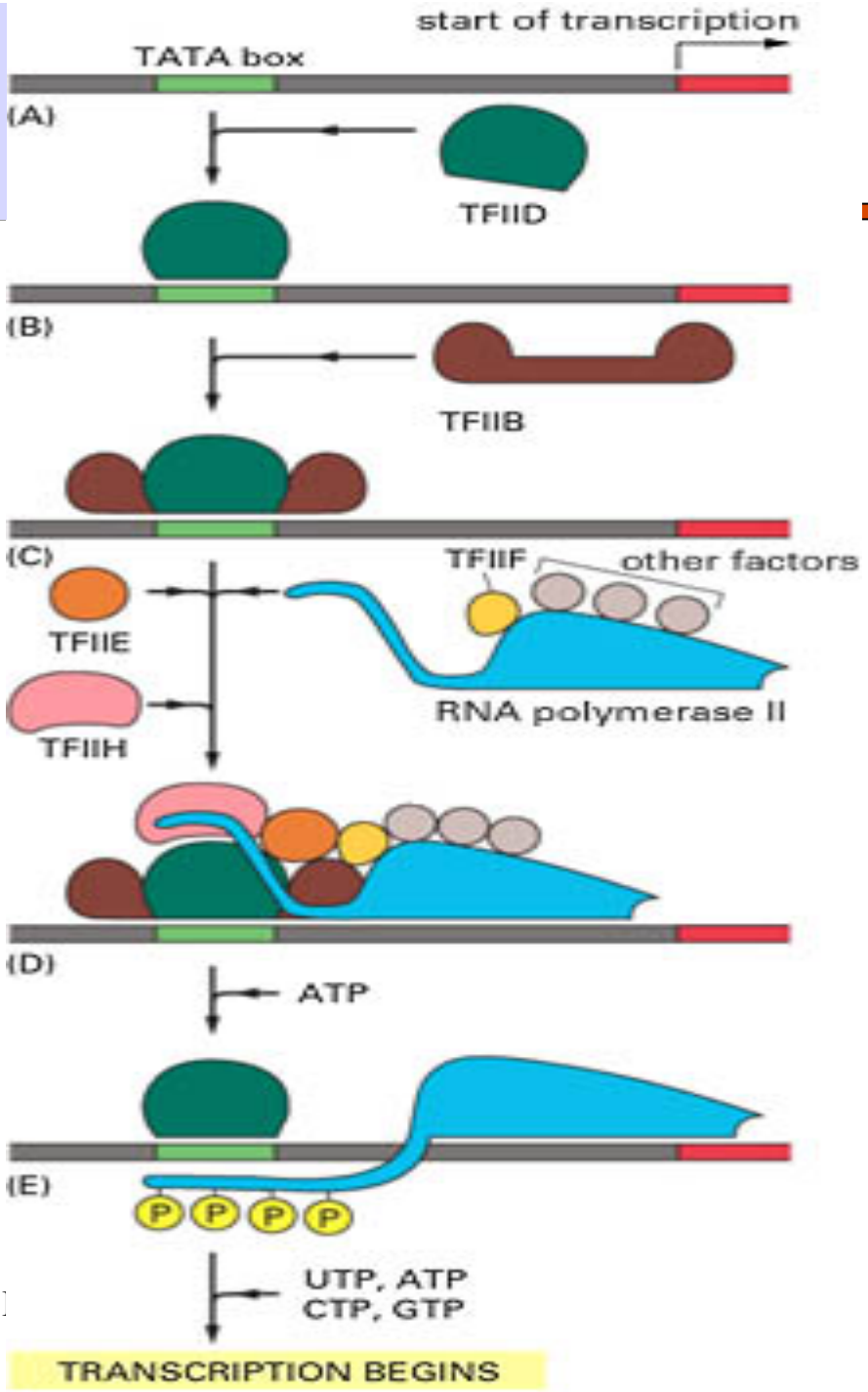
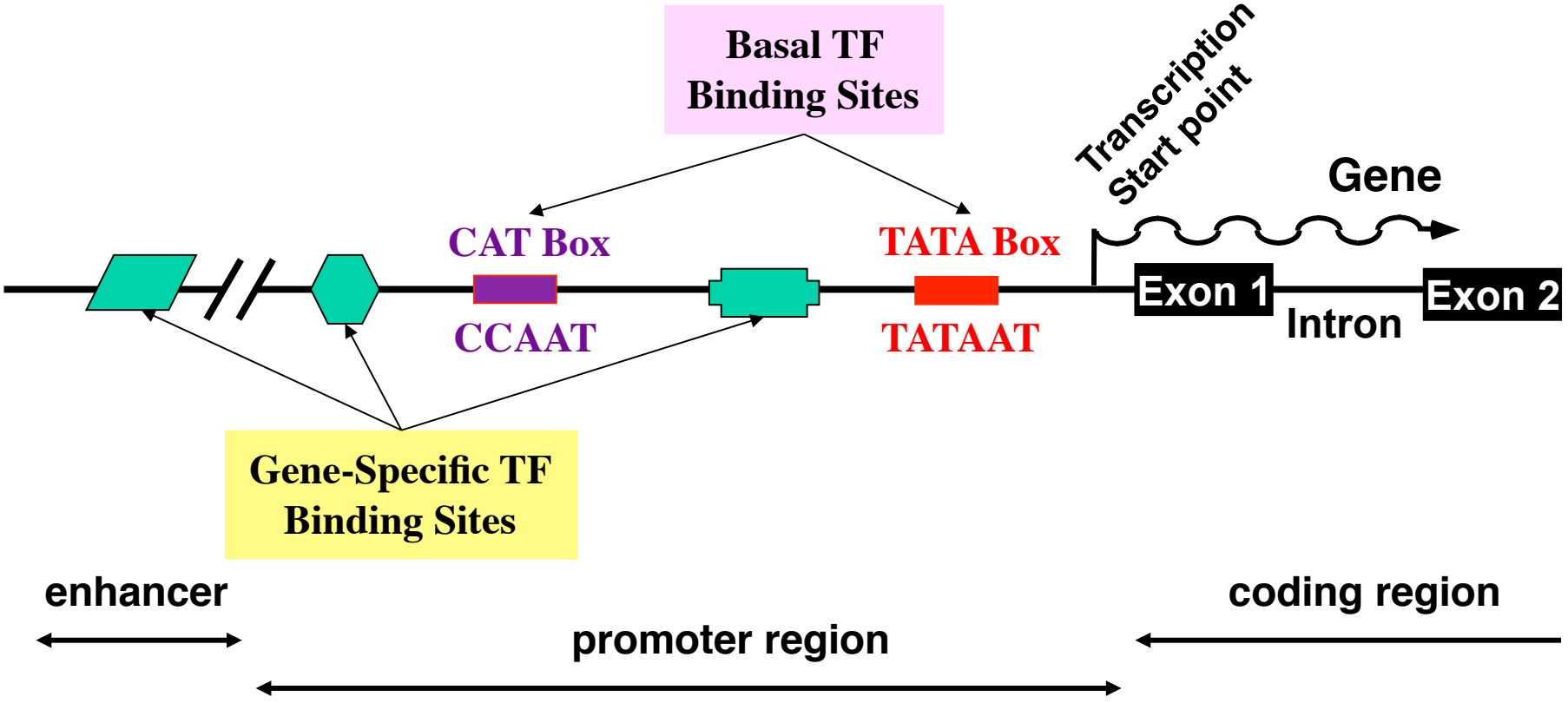
BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan
ECS 254A; Phone: x3748
giri@cs.fiu.edu
http://www.cs.fiu.edu/~giri/teach/BSC4934_Su10.html
July 2010

Transcription Initiation



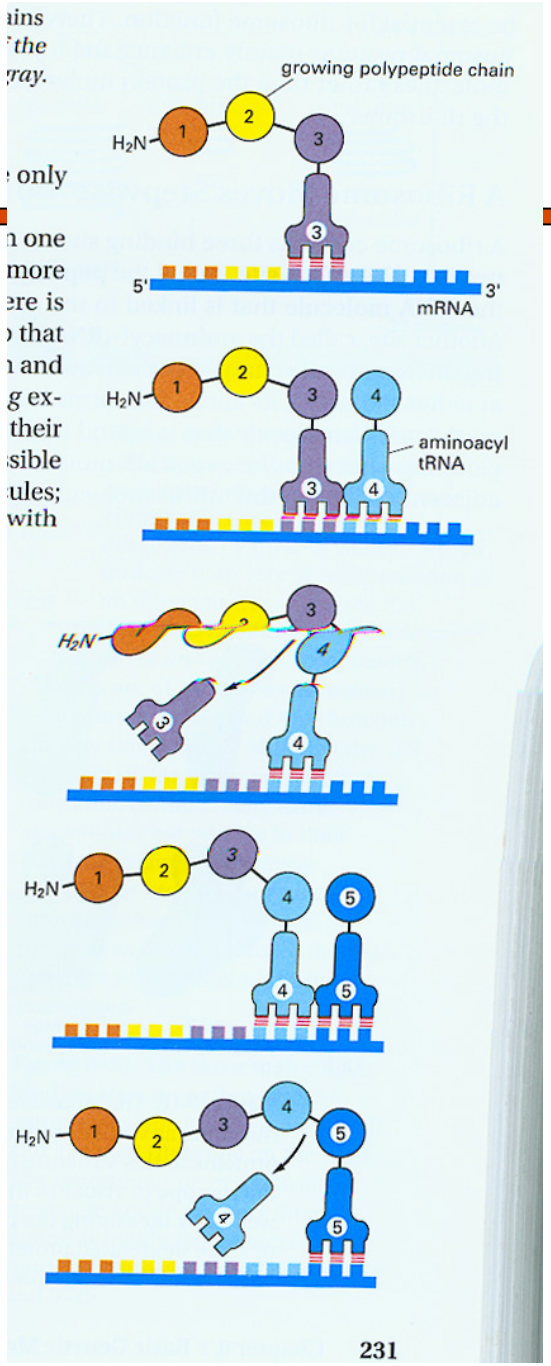
Transcription Regulation



Courtesy: Kalai Mathee

Protein Synthesis:

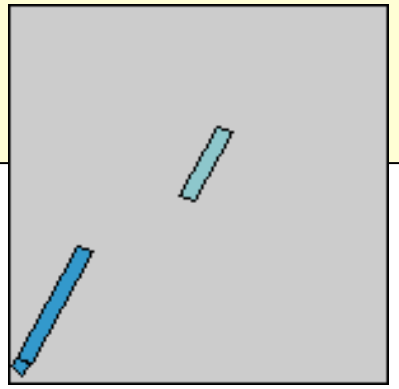
Incorporation of amino acid into protein



Drosophila Eyeless vs. Human Aniridia

```
Query: 57   HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 116
            HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 5    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64

Query: 117 SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSIN 176
            SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNDNIPSVSSIN
Sbjct: 65   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN 124

Query: 177 RVLRNLAAQKEQ 188
            RVLRNLA++K+Q
Sbjct: 125 RVLRNLASEKQQ 136
```

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

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



E-Value = 2e-31

Three major public DNA databases

- GenBank
  - NCBI (Natl Center for Biotechnology Information) **www.ncbi.nlm.nih.gov**
- EMBL
  - EBI (European Bioinformatics Inst)
- DDBJ
  - Japan's center

Entrez Portal @ NCBI

- PubMed
- DNA and Protein Sequence database
- Protein structure database
- Population study data sets
- Genome assemblies
- BLAST
- OMIM (Mendelian Inheritance in Man)
- TaxBrowser

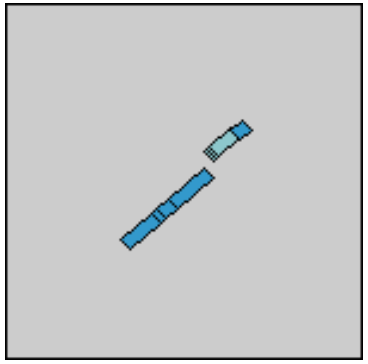
1. Can show sequences are close

rpoA [Pseudomonas aeruginosa] with rpoA [Pseudomonas fluorescence]

```
Query  1    MQISVNEFLTPRHIDVQVVSPTRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE  60
            MQ SVNEFLTPRHIDVQVVS TRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE
Sbjct  1    MQSSVNEFLTPRHIDVQVVSQTRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE  60

Query  61   IDGVLHEYSAIEGVQEDVIEILLNLKGLAIKLHGRDEVTLTLSKKGSGVVTAADIQLDHD  120
            IDGVLHEYSAIEGVQEDVIEILLNLKGLAIKLHGRDEVTLTL+KKGSGVVTAADIQLDHD
Sbjct  61   IDGVLHEYSAIEGVQEDVIEILLNLKGLAIKLHGRDEVTLTLAKKGSGVVTAADIQLDHD  120

Query  121  VEIVNPDHVIANLASNGALNMKLTVARGRGYEPADSRQSDEDESRSIGRLQLDSSFSPVR  180
            VEI+N DHVIANLA NGALNMKL VARGRGYEPAD+RQSDEDESRSIGRLQLD+SFSPVR
Sbjct  121  VEIINGDHVIANLADNGALNMKLKVARGRGYEPADARQSDEDESRSIGRLQLDASFSPVR  180

Query  181  RIAYVVENARVEQRTNLDKLVIDLETNGTLDPEEAIRRAATILQQQLAAFVDLKGDSEPV  240
            R++YVVENARVEQRTNLDKLV+DLETNGTLDPEEAIRRAATILQQQLAAFVDLKGDSEPV
Sbjct  181  RVSYVVENARVEQRTNLDKLVLDLETNGTLDPEEAIRRAATILQQQLAAFVDLKGDSEPV  240

Query  241  VIEQEDEIDPILLRPVDDLELTVRSANCLKAENIYYIGDLIQRTEVELLKTPNLGKKSLT  300
            V EQEDEIDPILLRPVDDLELTVRSANCLKAENIYYIGDLIQRTEVELLKTPNLGKKSLT
Sbjct  241  VEEQEDEIDPILLRPVDDLELTVRSANCLKAENIYYIGDLIQRTEVELLKTPNLGKKSLT  300

Query  301  EIKDVLASRGLSLGMRLDNWPPASLKKDDKATA  333
            EIKDVLASRGLSLGMRLDNWPPASLKKDDKATA
Sbjct  301  EIKDVLASRGLSLGMRLDNWPPASLKKDDKATA  333
```

2. Can show sequences have similar parts

```
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)
```



3. Can identify similar sequences from DB

V-sis Oncogene – Homologies

Sequences producing significant alignments:	Score (bits)	E 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

4. Can pinpoint mutations

870 GTGGCTGCTTCTTTGGTTGTGCTGTGGCTCCTTGGAAA

X

870 GTGGCTGCTTCTTTGGTTGTGCTGTAGCTCCTTGGAAA

5. Can be basis for discoveries

- 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.

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 aggggaccccattcctgaggagctctataagatgctgagtggccactcgattcgctcct 1074
           |||||||||||||||| |||||||| ||| |||||||||||| ||||||||| |||||||
Sbjct: 1084 aggggaccccattcccgaggagctttatgagatgctgagtgaccactcgatccgctcct 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 ttgatgatctccaacgcctgctgcacggagaccccggagaggaagatggggccgagttgg 1203
     > 61 D  L  N  M  T  R  S  H  S  G  G  E  L  E  S  L  A  R  G  R
```

6. Can help describe motifs, domains, and families of sequences

- Family alignment for the ITAM domain (Immunoreceptor tyrosine-based activation motif)

```
CD3D_MOUSE/1-2   EQLYQPLRDR EDTQ-YSRLG GN
Q90768/1-21      DQLYQPLGER NDGQ-YSQLA TA
CD3G_SHEEP/1-2   DQLYQPLKER EDDQ-YSHLR KK
P79951/1-21      NDLYQPLGQR SEDT-YSHLN SR
FCEG_CAVPO/1-2   DGIYTGLSTR NQET-YETLK HE
CD3Z_HUMAN/3-0   DGLYQGLSTA TKDT-YDALH MQ
C79A_BOVIN/1-2   ENLYEGLNLD DCSM-YEDIS RG
C79B_MOUSE/1-2   DHTYEGLNID QTAT-YEDIV TL
CD3H_MOUSE/1-2   NQLYNELNLG RREE-YDVLE KK
CD3Z_SHEEP/1-2   NPVYNELNVG RREE-YAVLD RR
CD3E_HUMAN/1-2   NPDYEPIRKG QRDL-YSGLN QR
CD3H_MOUSE/2-0   EGVYNALQKD KMAEAYSEIG TK
Consensus/60%    -.lYpsLspc pcsp.YspLs pp
```

Simple Modular Architecture Research Tool

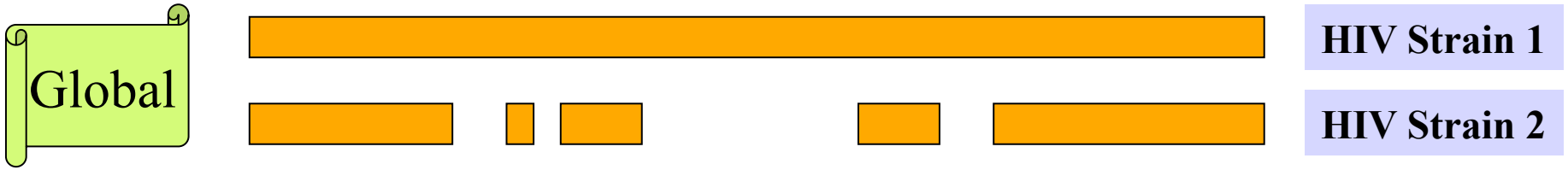
Implications of Sequence Alignment

- 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.

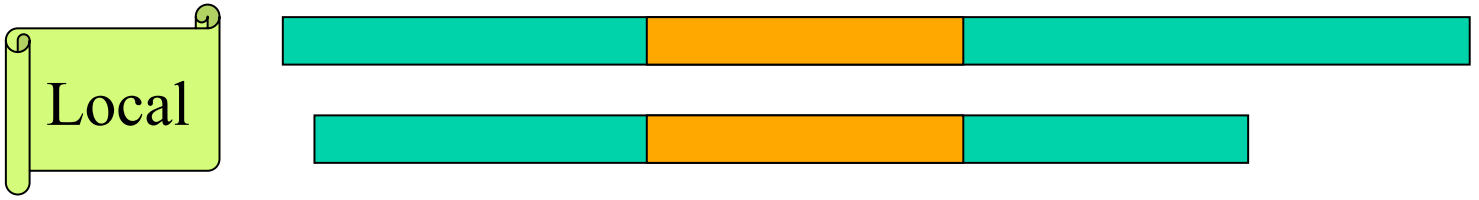
Similarity vs. Homology

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

Types of Sequence Alignments - 1

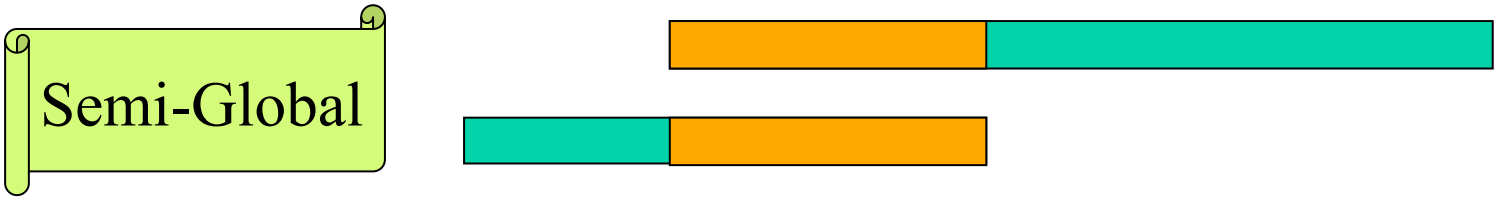


- Global Alignment: similarity over entire length

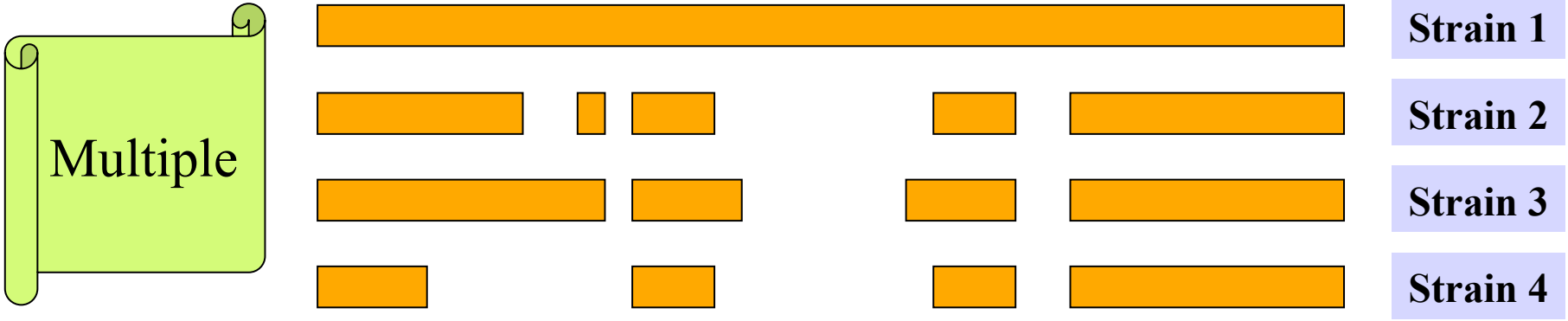


- Local Alignment: no overall similarity, but some segment(s) is/are similar

Types of Sequence Alignments - 2



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



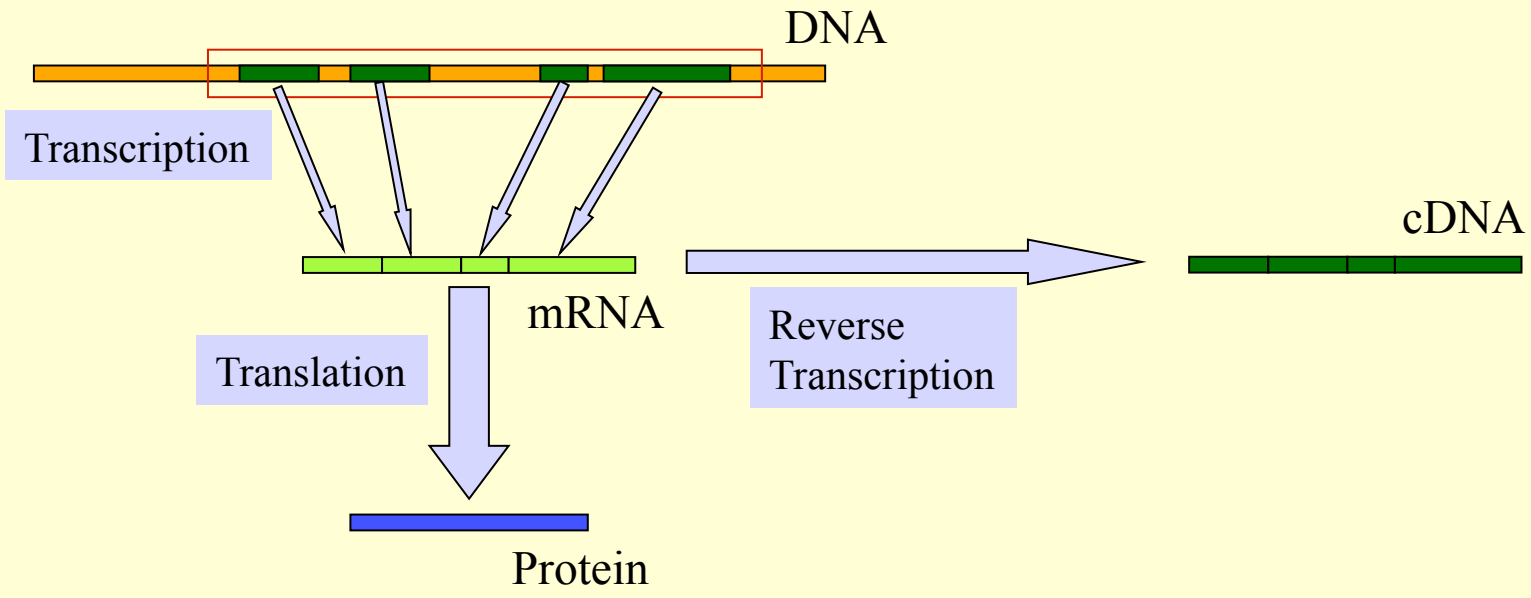
- Multiple Alignment: similarity between sets of sequences

Sequence Alignment

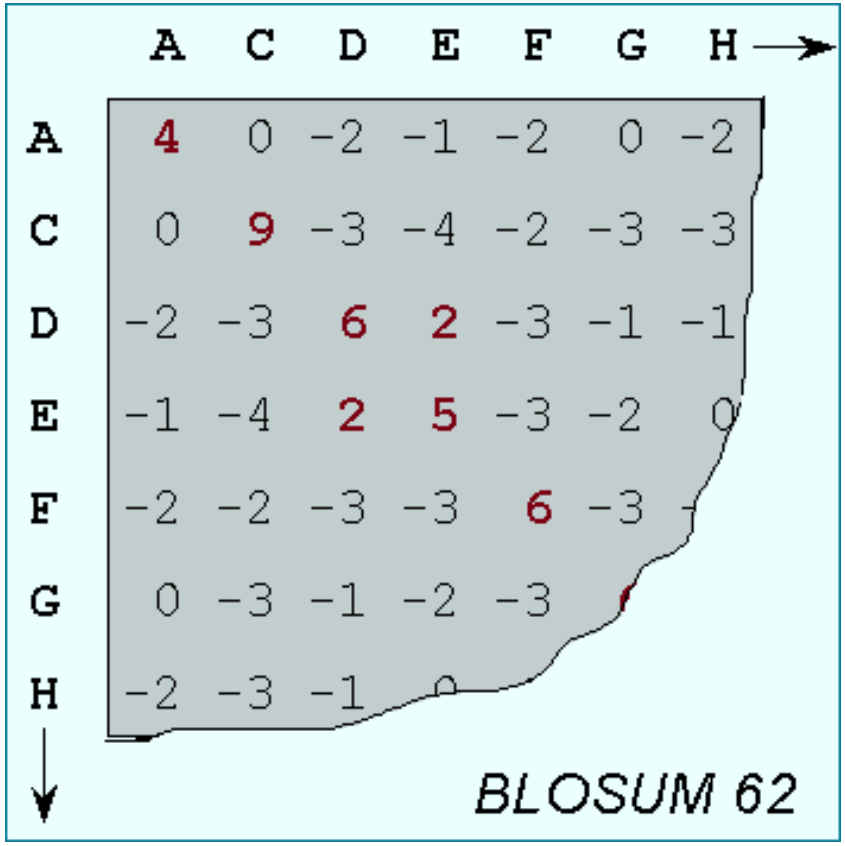
- Global:
  - Needleman-Wunsch-Sellers (1970).
- Local:
  - Smith-Waterman (1981)
  - Useful when commonality is small and global alignment is meaningless. Often unaligned portions "mask" short stretches of aligned portions. Example: comparing long stretches of anonymous DNA; aligning proteins that share 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?



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

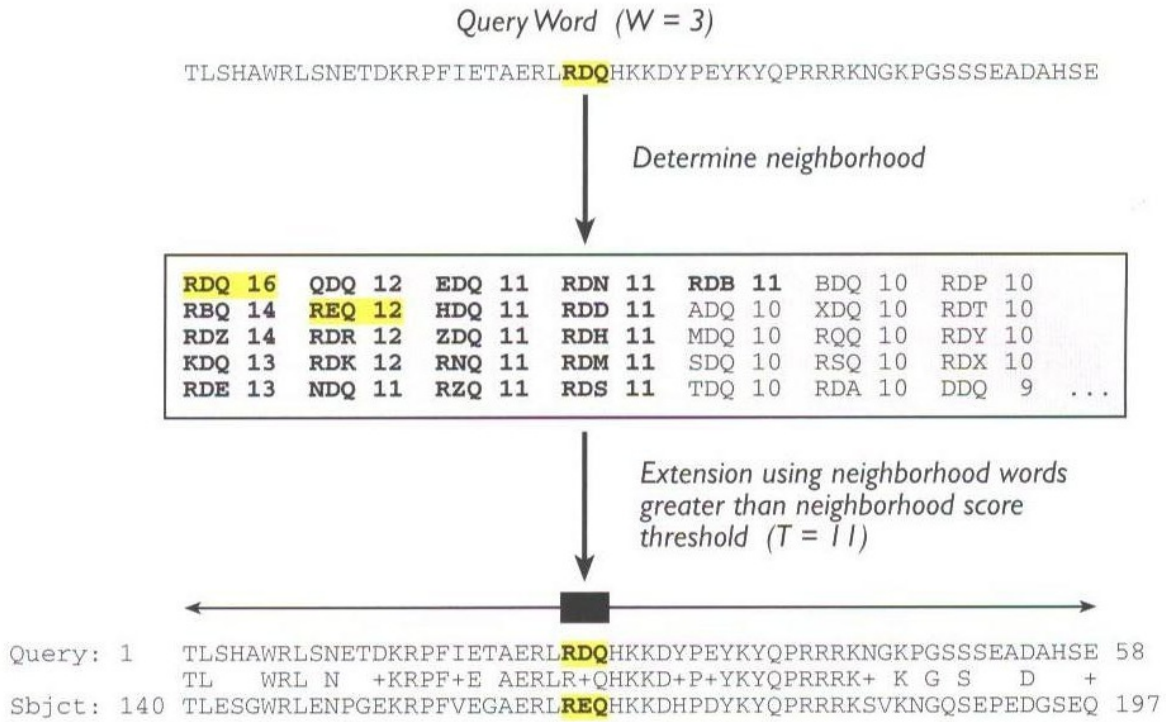


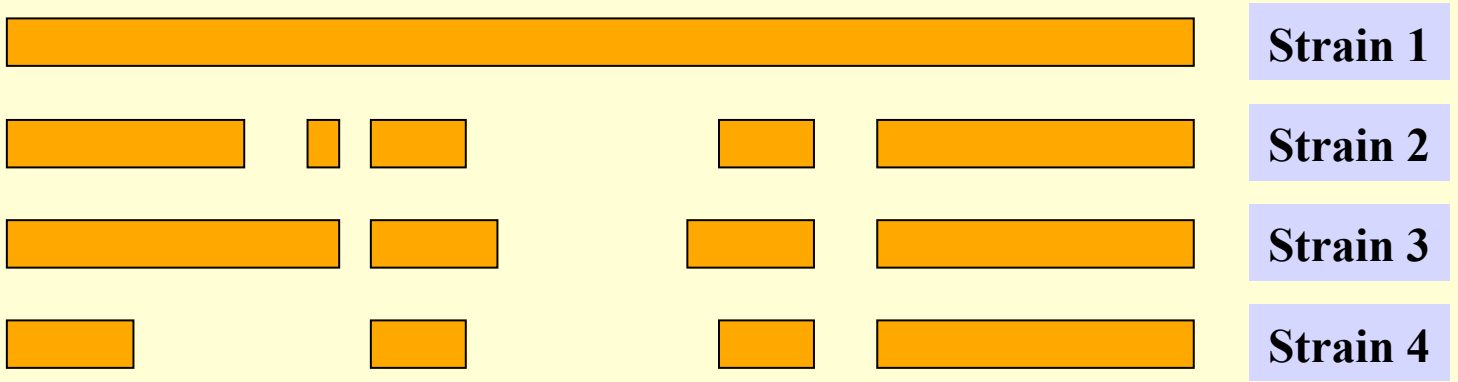
FIGURE 11.7 The initiation of a **BLAST** search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words "in the neighborhood" of the original query word. Any occurrences of these neighborhood words in sequences within the target database then are investigated. See text for details.

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: Immunoglobin 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**

BLAST Parameters and Output

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

Scoring Matrix to Use

PAM 40	Short alignments with high similarity (70-90%)
PAM 160	Members of a protein family (50-60%)
PAM 250	Longer alignments (divergent sequences) (~30%)
BLOSUM90	Short alignments with high similarity (70-90%)
BLOSUM80	Members of a protein family (50-60%)
BLOSUM62	Finding all potential hits (30-40%)
BLOSUM30	Longer alignments (divergent sequences) (<30%)

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

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

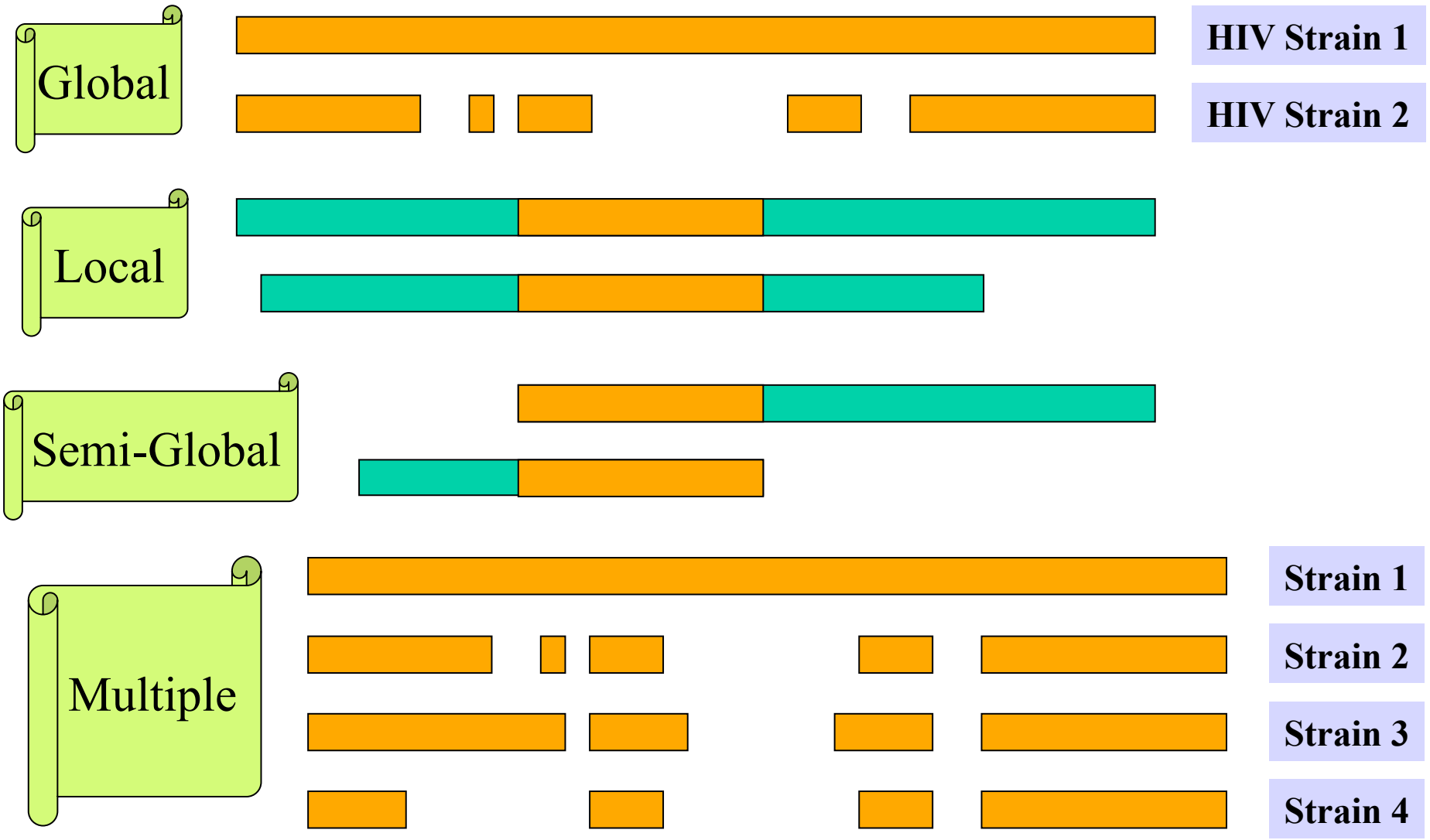
The parameters depend on the scoring system.

- **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 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
	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											

Given

δ[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

Match/Mismatch score

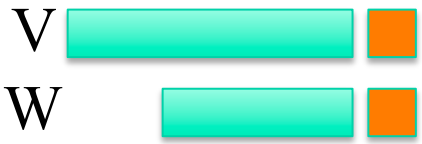
Recurrence Relation

S[I, J] = MAXIMUM {
S[I-1, J-1] + δ(V[I], W[J]),
S[I-1, J] + δ(V[I], —),
S[I, J-1] + δ(— , W[J]) }

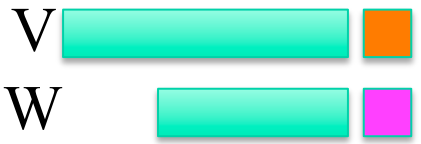
Gap Penalty

What happens with last character(s)?

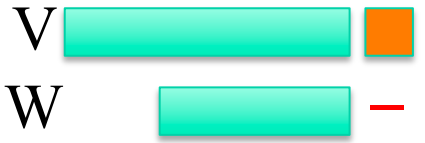
1. Last characters **MATCH**



2. Last characters **MISMATCH**

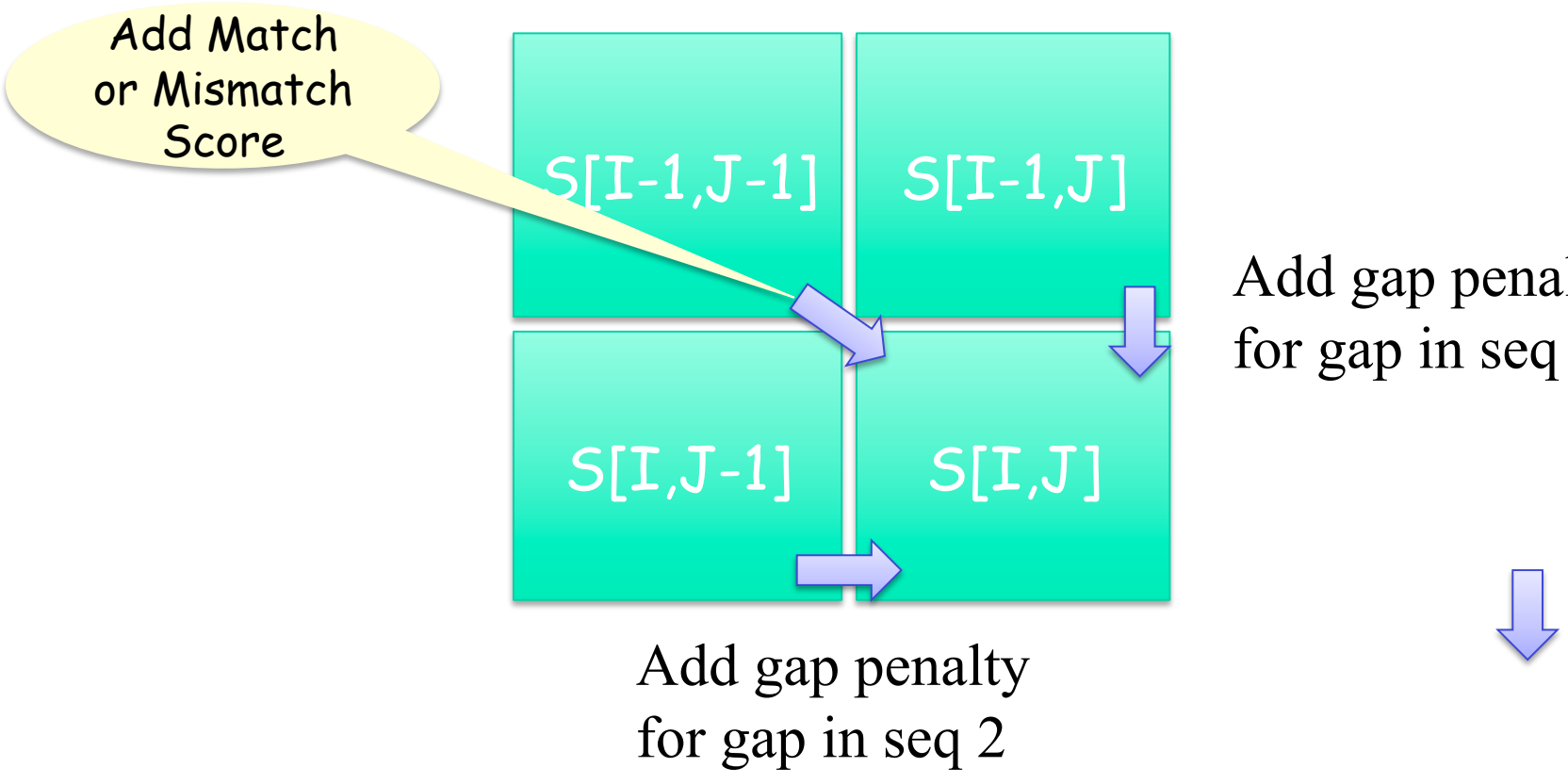


3. Last character of W aligned with GAP



4. Last character of V aligned with GAP

How to fill in the matrix?



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
	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											

Given

δ[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

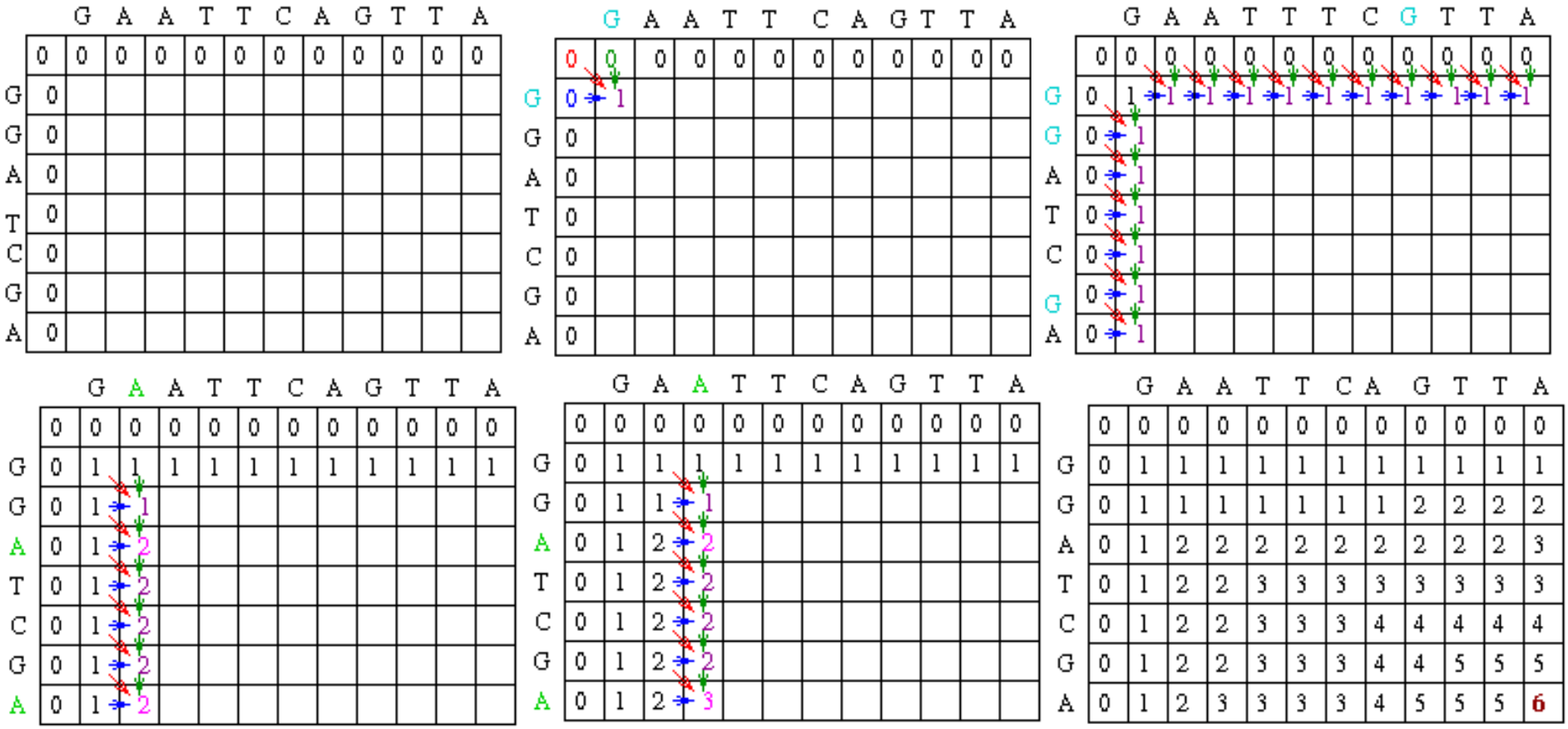
Recurrence Relation

S[I, J] = MAXIMUM {
S[I-1, J-1] + δ(V[I], W[J]),
S[I-1, J] + δ(V[I], —),
S[I, J-1] + δ(— , W[J]) }

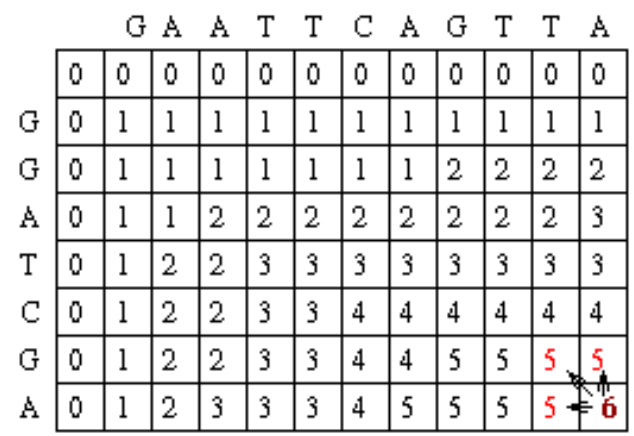
Global Alignment: An example

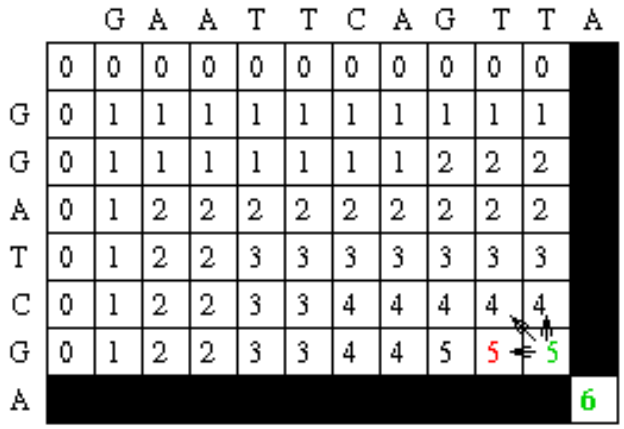
$$S[I, J] = MAXIMUM\{ S[I-1, J-1] + \delta(V[I], W[J]),\ S[I-1, J] + \delta(V[I], —),\ S[I, J-1] + \delta(— , W[J]) \}$$

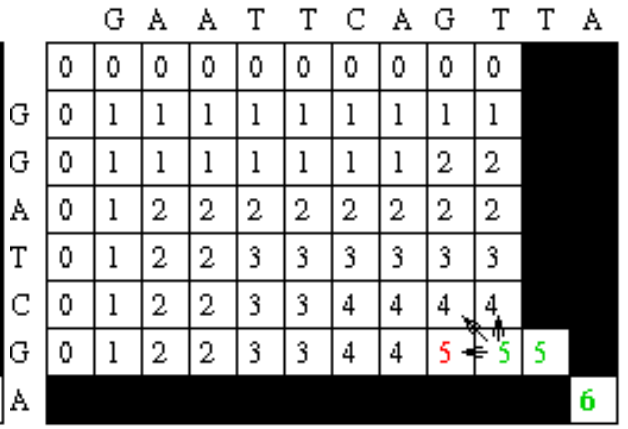
```
V: G A A T T C A G T T A
W: G G A T C G A
```

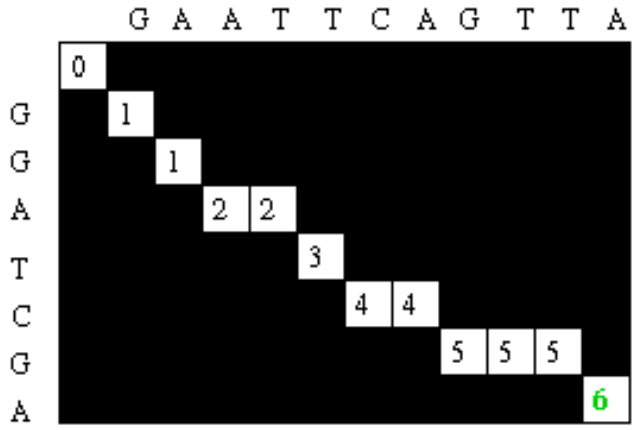


Traceback



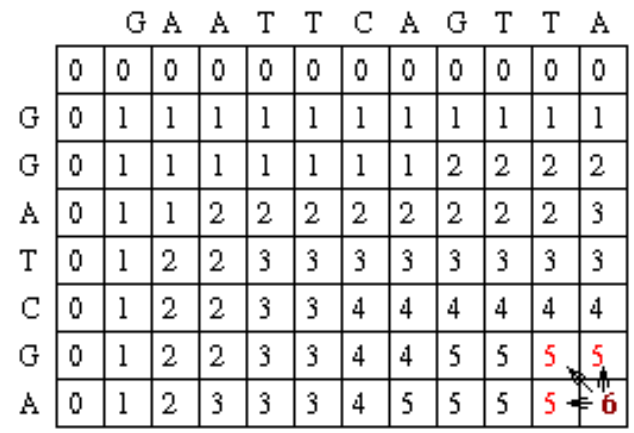


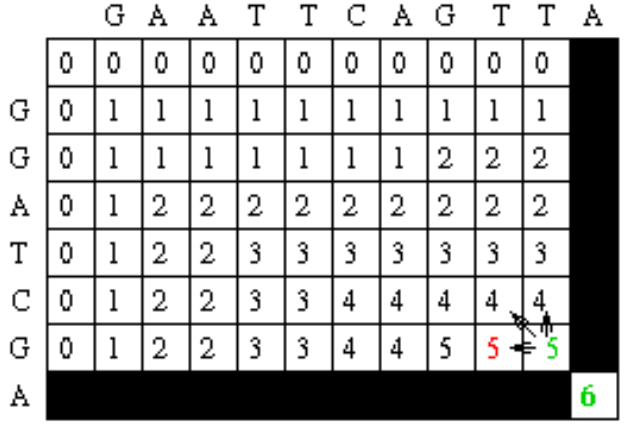


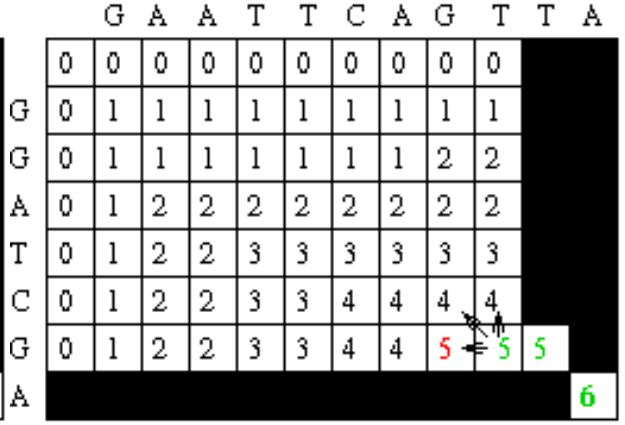


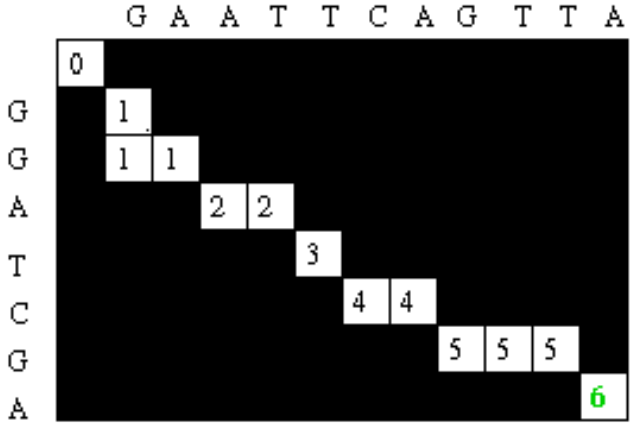

```
V: G A A T T C A G T T A
   |   |   |  |   |     |
W: G G A - T C - G - - A
```

Alternative Traceback










```
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	0
G	0	×1	←1	← 1	← 1	← 1	← 1	← 1	×1	← 1	← 1	← 1
G	0	×1	↑1	↑1	↑1	↑1	↑1	↑1	×2	← 2	← 2	← 2
A	0	↑1	↑1	×2	← 2	← 2	← 2	×2	↑2	↑2	↑2	×3
T	0	↑1	← 2	↑2	×3	×3	← 3	← 3	← 3	×3	×3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	×4	← 4	← 4	← 4	← 4	← 4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	×5	← 5	← 5	← 5
A	0	↑1	↑2	×3	↑3	↑3	↑4	×5	↑5	↑5	↑5	×6

Improved Traceback

		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	← 1
G	0	×1	↑1	↑1	↑1	↑1	↑1	↑1	×2	← 2	← 2	← 2
A	0	↑1	↑1	×2	← 2	← 2	← 2	×2	↑2	↑2	↑2	×3
T	0	↑1	← 2	↑2	×3	×3	← 3	← 3	← 3	×3	×3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	×4	← 4	← 4	← 4	← 4	← 4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	×5	← 5	← 5	← 5
A	0	↑1	↑2	×3	↑3	↑3	↑4	×5	↑5	↑5	↑5	×6

Improved Traceback

		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	← 1
G	0	×1	↑1	↑1	↑1	↑1	↑1	↑1	×2	← 2	← 2	← 2
A	0	↑1	↑1	×2	← 2	← 2	← 2	×2	↑2	↑2	↑2	×3
T	0	↑1	← 2	↑2	×3	×3	← 3	← 3	← 3	×3	×3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	×4	← 4	← 4	← 4	← 4	← 4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	×5	← 5	← 5	← 5
A	0	↑1	↑2	×3	↑3	↑3	↑4	×5	↑5	↑5	↑5	×6

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

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 = $\Omega(a)$