CAP 5510/CGS 5166: Bioinformatics & Bioinformatic Tools

GIRI NARASIMHAN, SCIS, FIU

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

Integrated!

Entrez Portal to NCBI

- PubMed; Bookshelf
- DNA and Protein Sequence database
- Protein structure database
- Genome assemblies
- BLAST

Youtube videos: http://www.youtube.com/ncbinlm

- SNP
- TaxBrowser
- Population study data sets
- PubChem (small mols)
- GEO (Gene Expression Omnibus)
- OMIM (Mendelian Inheritance

Other critical databases

- PDB (http://www.wwpdb.org/)
- KEGG (http://www.genome.jp/kegg/)
- MetaCyc (http://metacyc.org)
- Reactome (http://www.reactome.org)
- ENCODE (http://encodeproject.org/ENCODE/ functional elements in human genome
- 1000 Genomes Project; Int'l HapMap Project
- Human Microbiome Project
- Human Epigenome Project
- Gene Ontology (GO)

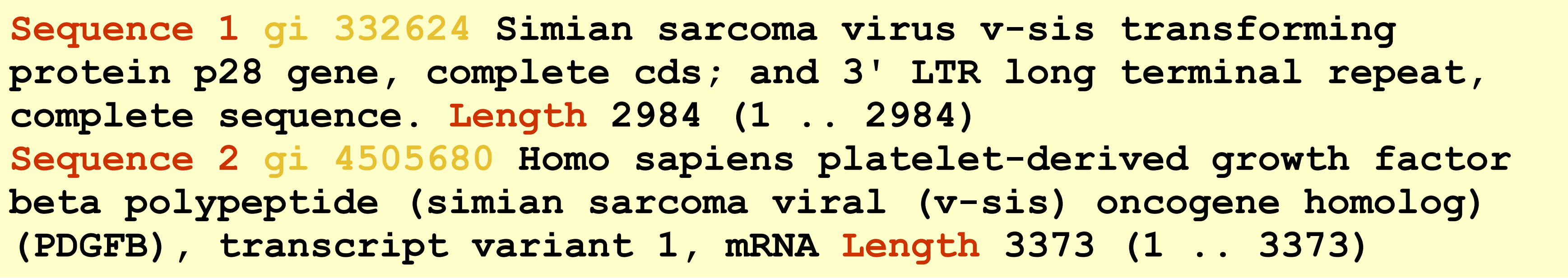
Sequence Alignment

1. Can show sequences are close

rpoA [Pseudomonas aeruginosa] with rpoA [Pseudomonas fluorescence]

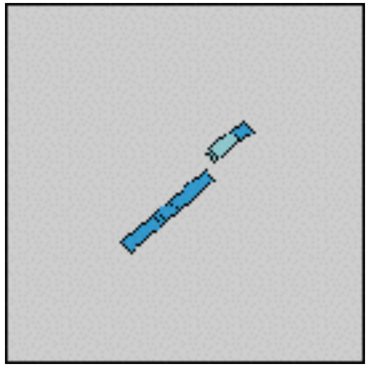
```
Query  1    MQISVNEFLTPRHIDVQVVSPTRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE  60
            MQ SVNEFLTPRHIDVQVVS TRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE
Sbjct  1    MQSSVNEFLTPRHIDVQVVSQTRAKITLEPLERGFGHTLGNALRRILLSSMPGCAVVEAE  60

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

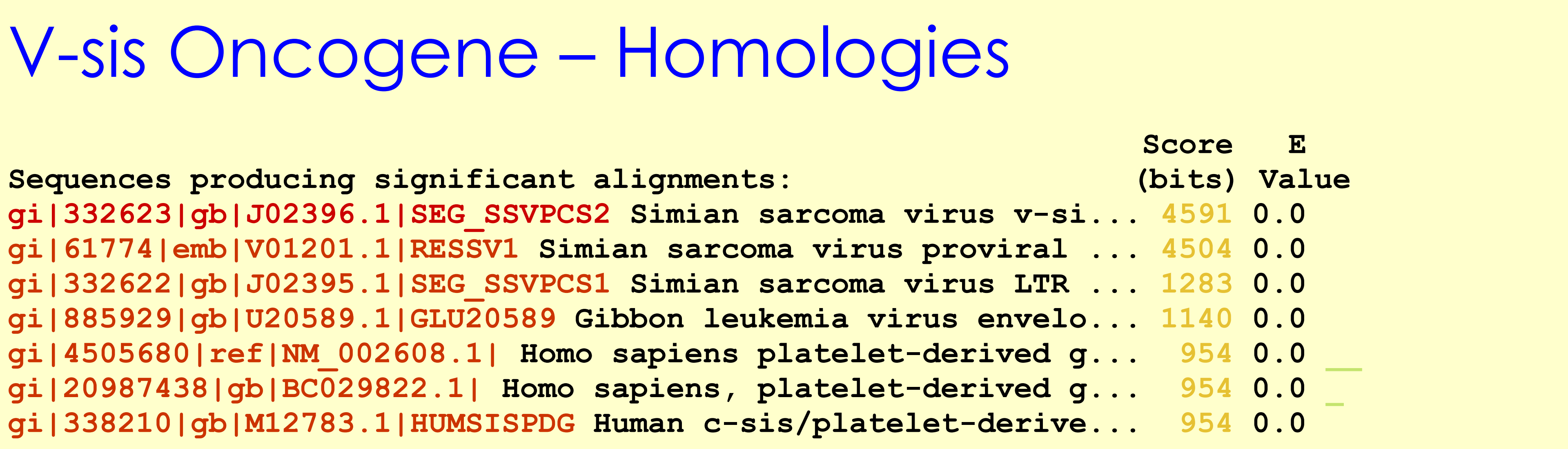


Sequence 1 gi 332624 Simian sarcoma virus v-sis transforming protein p28 gene, complete cds; and 3' LTR long terminal repeat, complete sequence. Length 2984 (1 .. 2984)

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

```
                                                                 Score    E
Sequences producing significant alignments:                      (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
```

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.

Can be the basis for discoveries … 2

- 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 agggggaccccattcctgaggagctctataagatgctgagtggccactcgattcgctcct 1074

            |||||||||||||||| |||||||| ||| |||||||||||| ||||||||| |||||||

Sbjct: 1084 agggggaccccattcccgaggagctttatgagatgctgagtgaccactcgatccgctcct 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

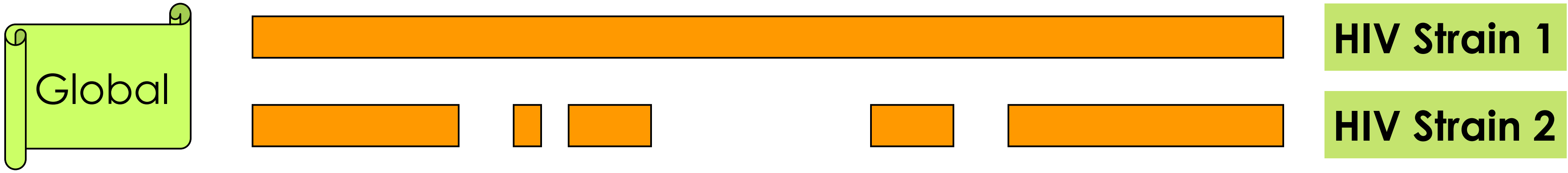
Implications of Sequence Alignment

- 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

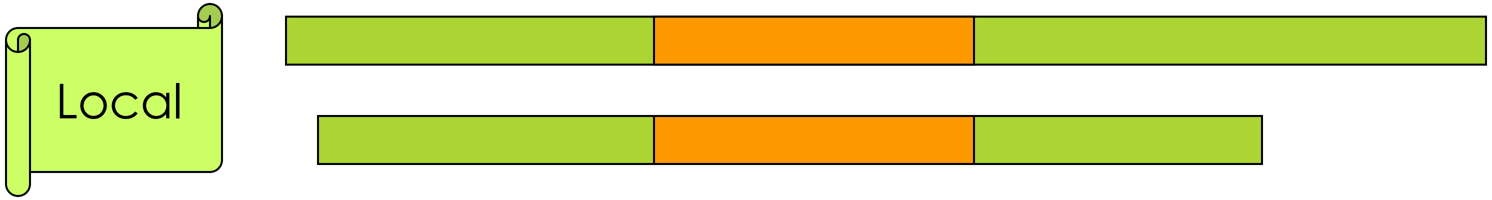


Global

HIV Strain 1

HIV Strain 2

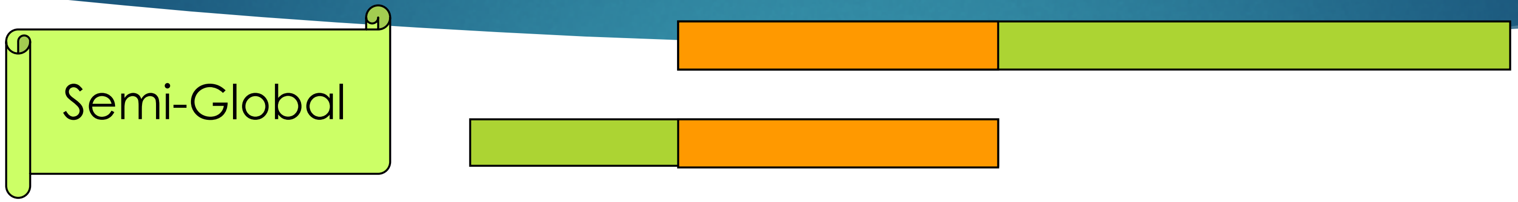
□ Global Alignment: similarity over entire length



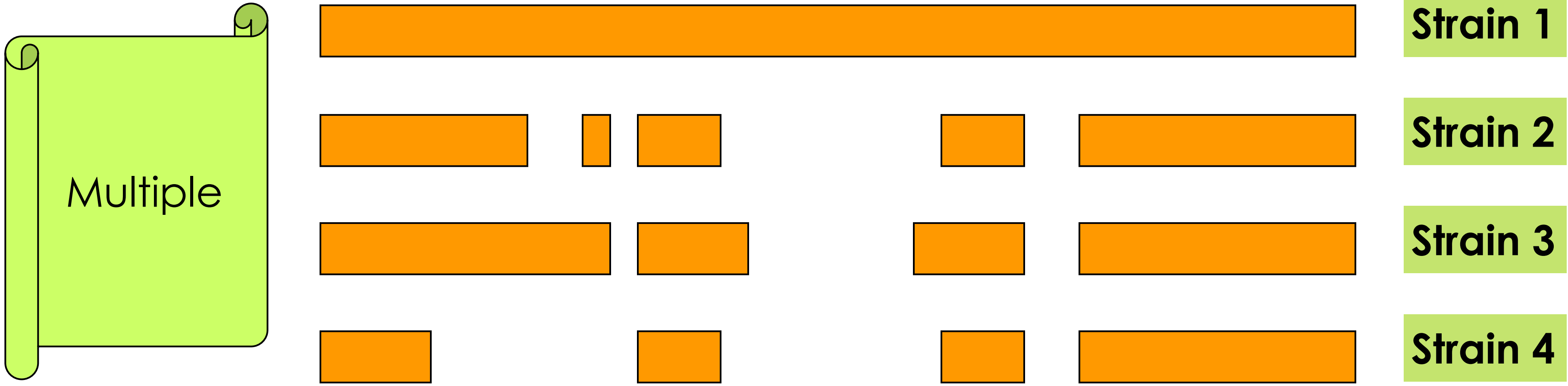
Local

□ 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



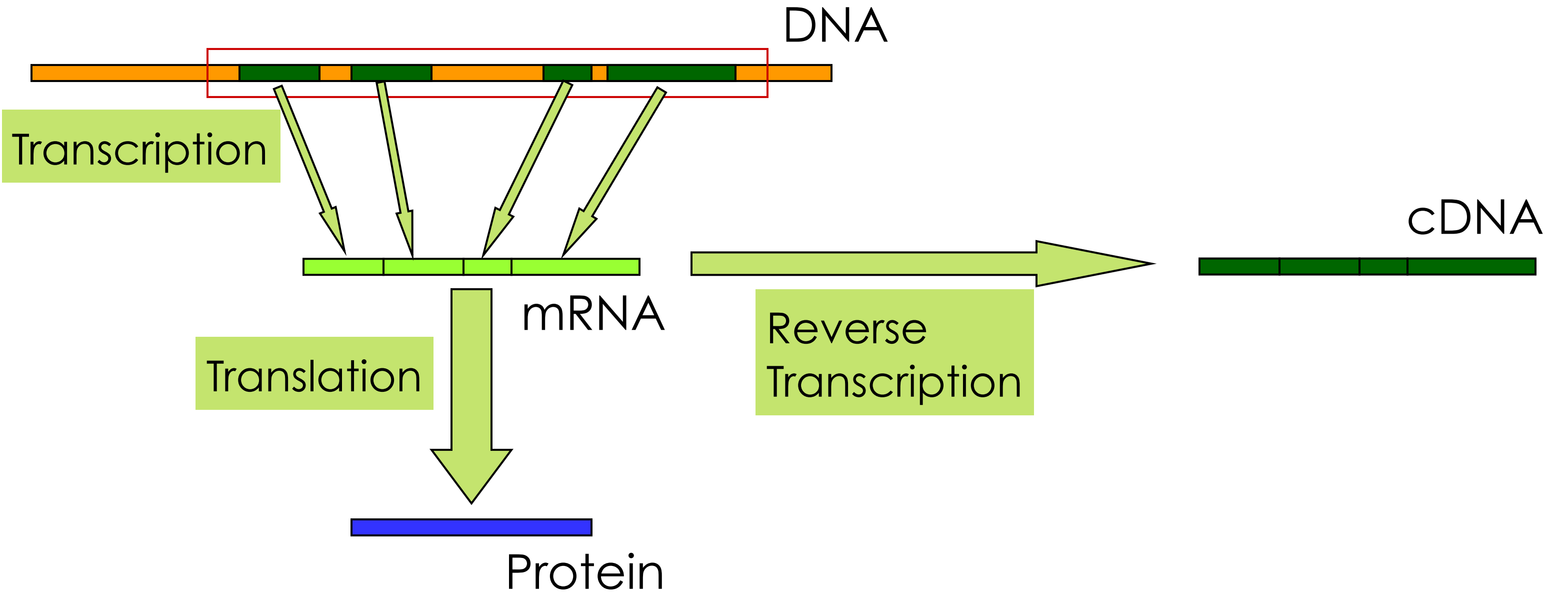
- □ Multiple Alignment: similarity between sets of sequences

Sequence Alignment

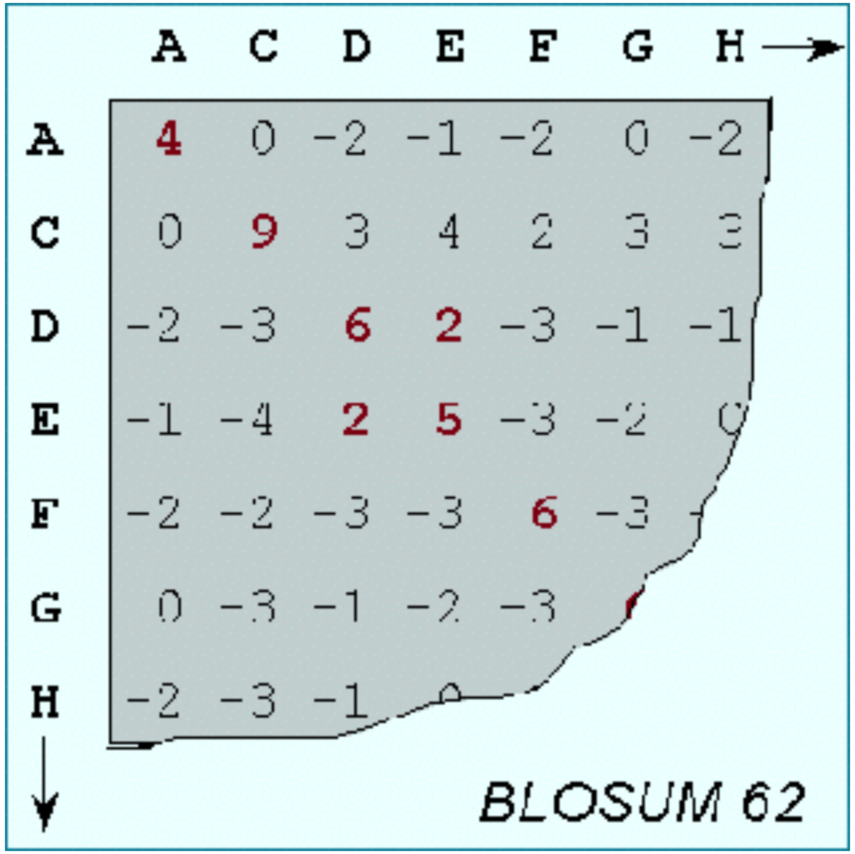
- Global:
  - Needleman-Wunsch-Sellers (1970).
- Local:
  - Smith-Waterman (1981)
  - Useful when commonality is small and global alignment is meaningless. Often unaligned portions "mask" short stretches of aligned portions. Example: comparing long stretches of anonymous DNA; aligning proteins that share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



	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	6	
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)