Sequence Alignment – Why?

```
>gi|12643549|sp|O18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)
MRNLPCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS
QYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQGDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSGYEVLSAYALPPPPMASSSAADSSFSAASSAS
ANVTPHHTIAQESCPSPCSSASHFGVAHSSGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHV
APGKQQFFASCFYSPWV
```

```
>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II protein)
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL
QRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASN
TPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQ
TSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
```

Drosophila Eyeless vs. Human Aniridia

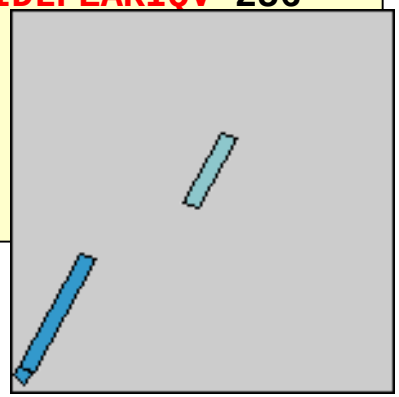
```
Query: 57   HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 116
            HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 5    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64

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

Why Sequence Analysis?

- 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 bound.
  - 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 (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

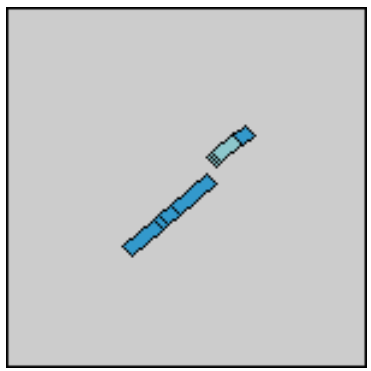
Sequence Alignment

>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
```

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)



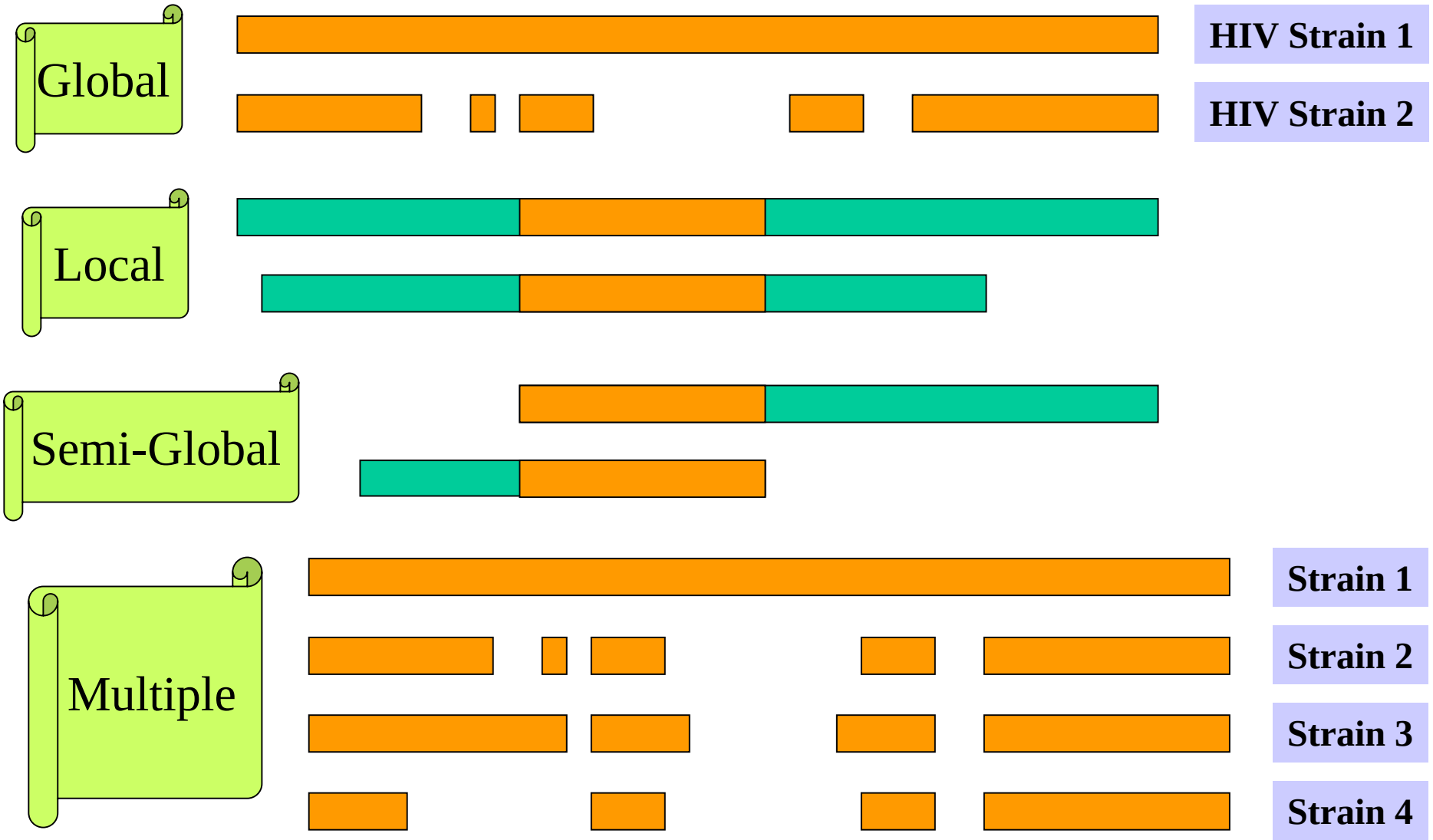
Genomic Databases

- 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

Similarity vs. Homology

- Homologous sequences share common ancestry.
- Similar sequences are "near" to each other by some criteria. Similarity can be measured using appropriate criteria.

Types of Sequence Alignments



Types of Sequence Alignments

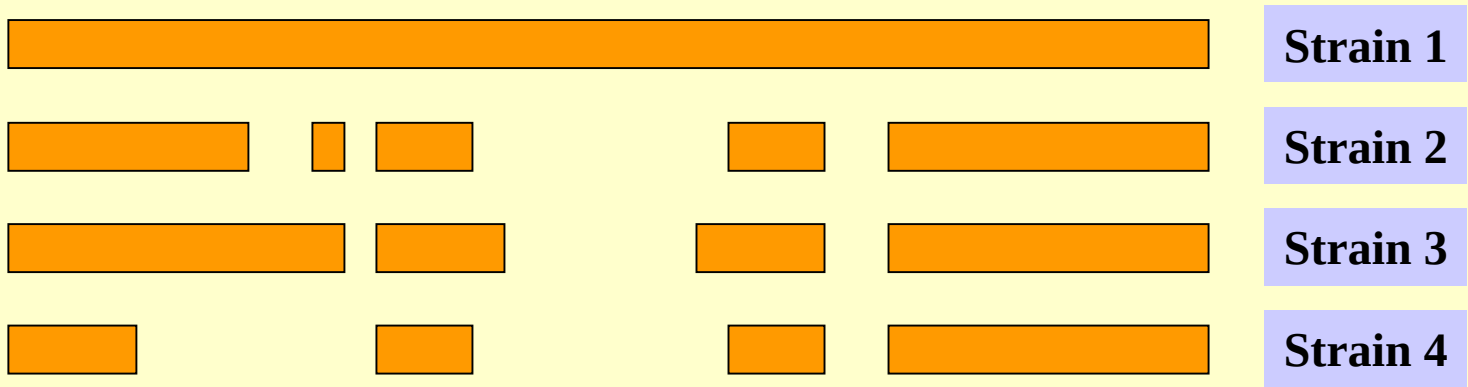
- Global Alignment: similarity over entire length
- Local Alignment: no overall similarity, but some segment(s) is/are similar
- Semi-global Alignment: end segments may not be similar
- Multiple Alignment: similarity between sets of sequences

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.

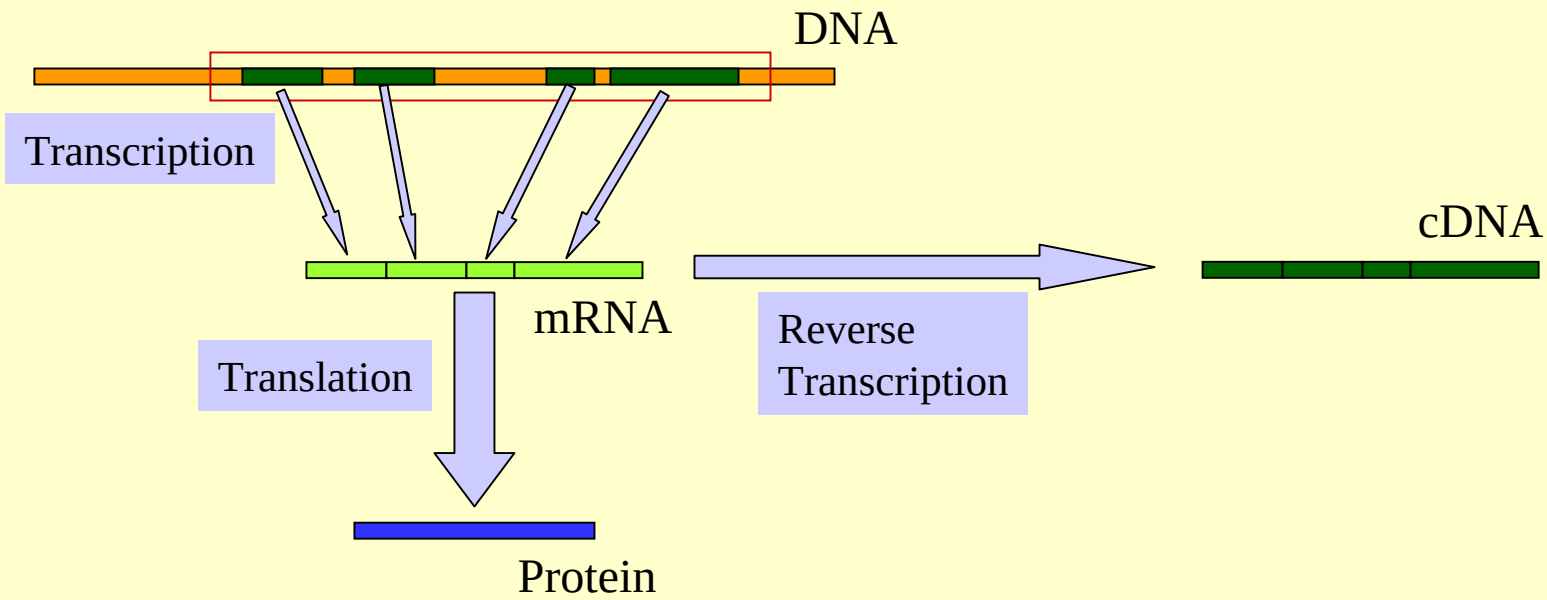
Why Gaps?

- Example: Aligning HIV sequences.

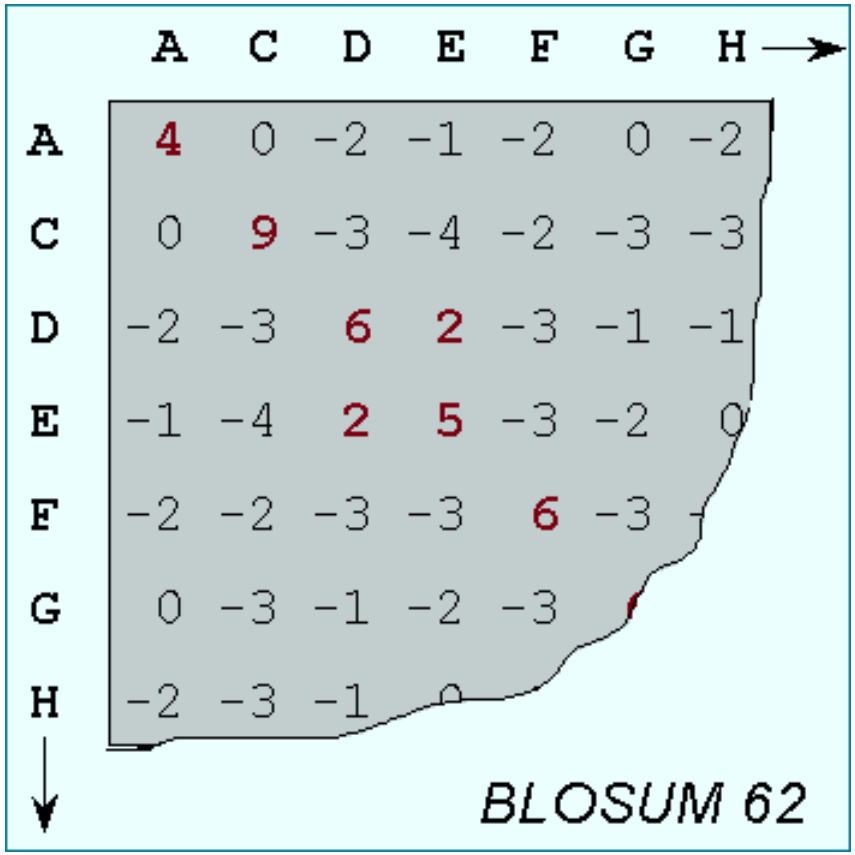


Why gaps?

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



How to score mismatches?



General Bioinformatics Resources

- GenBank: (Portal) PubMed at NCBI, NIH
  - Try Lambda Cro (73101), Ecoli Sigma-70 (1SIG), Ecoli Sigma factor (1072030), Bacteriorhodopsin (14194473), 1baza vs. 1myka (P-22 Arc repressors)
- BLAST
- SwissPROT
- InterPro

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"

BLAST Variants

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

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.