CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

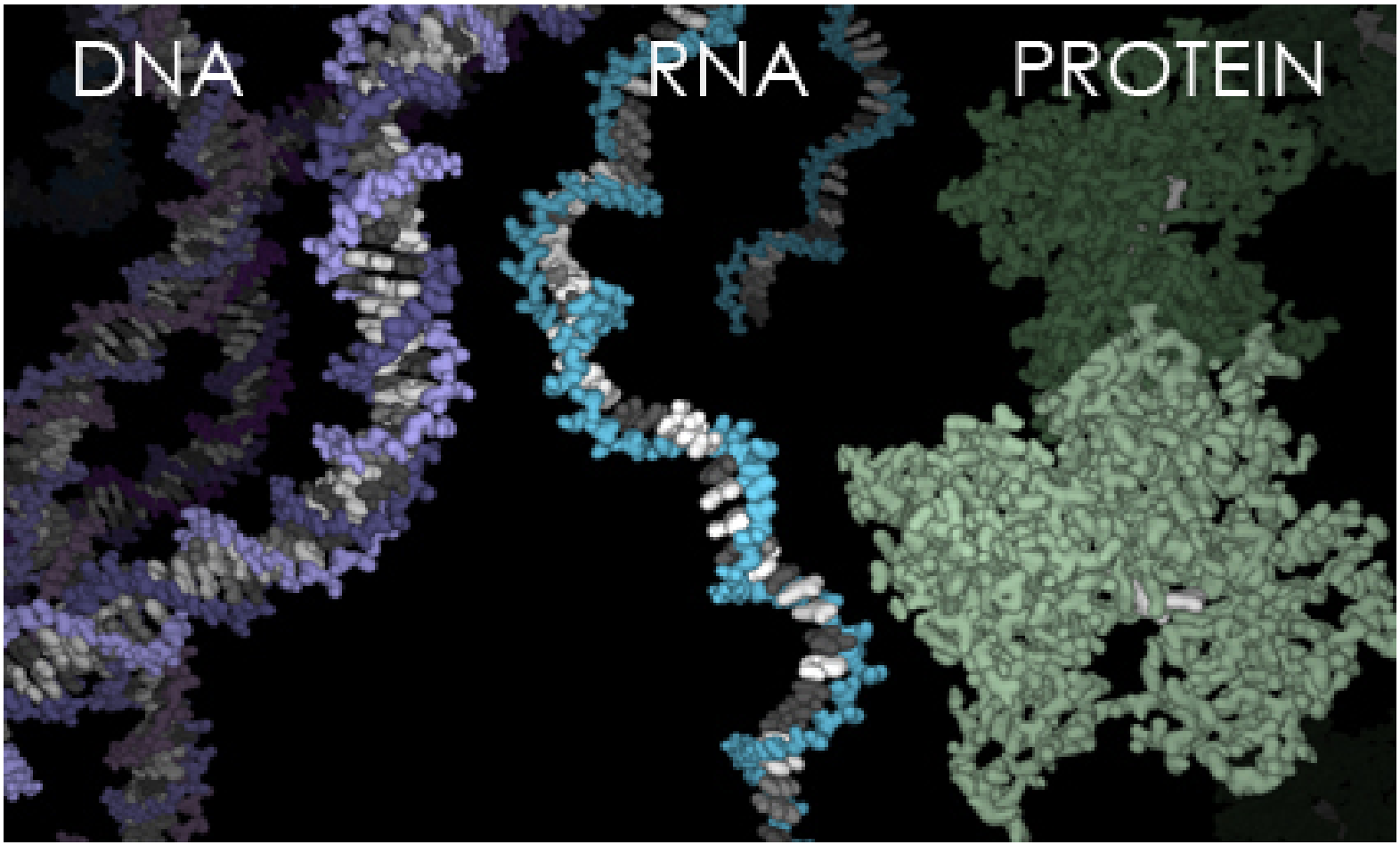
Giri Narasimhan

ECS 254A / EC 2474; Phone x3748; Email: giri@cis.fiu.edu
My Homepage: http://www.cs.fiu.edu/~giri
http://www.cs.fiu.edu/~giri/teach/BioinfS15.html
Office ECS 254 (and EC 2474); Phone: x-3748
Office Hours: By Appointment Only

Jan 19, 2015

Presentation Outline

The drama of Molecular Biology . . . the actors



http://exploringorigins.org/images/centralDogma.jpg

The Polymeric Players

Molecule	Unit Name	Unit Composition
DNA	Nucleotide or Base	A, C, G, T

The Polymeric Players

Molecule	Unit Name	Unit Composition
DNA	Nucleotide or Base	A, C, G, T
RNA	Nucleotide or Base	A, C, G, U

The Polymeric Players

Molecule	Unit Name	Unit Composition
DNA	Nucleotide or Base	A, C, G, T
RNA	Nucleotide or Base	A, C, G, U
Protein	Amino acid residue	amino acids represented by 20-letter alphabet missing {B, J, O, U, X, Z}

Typical DNA Sequence


```
   1  gggagaacac  ccggagaagg  aggaggaggc  gaagaaaagc  aacagaagcc  cagttgctgc
  61  tccaggtccc  tcggacagag  ctttttccat  gtggagactc  tctcaatgga  cgtgcccccc
 121  agtgcttctt  agacggactg  cggtctccta  aaggtcgacc  atggtggccg  ggacccgctg
 181  tcttctagtg  ttgctgcttc  cccaggtcct  cctgggcggc  gcggccggcc  tcattccaga
 241  gctgggccgc  aagaagttcg  ccgcggcatc  cagccgaccc  ttgtcccggc  cttcggaaga
 301  cgtcctcagc  gaatttgagt  tgaggctgct  cagcatgttt  ggcctgaagc  agagacccac
 361  ccccagcaag  gacgtcgtgg  tgcccccta  tatgctagat  ctgtaccgca  ggcactcagg
 421  ccagccagga  gcgcccgccc  cagaccaccg  gctggagagg  gcagccagcc  gcgccaacac
 481  cgtgcgcagc  ttccatcacg  aagaagccgt  ggaggaactt  ccagagatga  gtgggaaaac
 541  ggcccggcgc  ttcttcttca  atttaagttc  tgtccccagt  gacgagtttc  tcacatctgc
 601  agaactccag  atcttccggg  aacagataca  ggaagctttg  ggaaacagta  gtttccagca
 661  ccgaattaat  atttatgaaa  ttataaagcc  tgcagcagcc  aacttgaaat  ttcctgtgac
 721  cagactattg  gacaccaggt  tagtgaatca  gaacacaagt  cagtgggaga  gcttcgacgt
 781  caccccagct  gtgatgcggt  ggaccacaca  gggacacacc  aaccatgggt  ttgtggtgga
 841  agtggcccat  ttagaggaga  acccaggtgt  ctccaagaga  catgtgagga  ttagcaggtc
 901  tttgcaccaa  gatgaacaca  gctggtcaca  gataaggcca  ttgctagtga  cttttggaca
 961  tgatggaaaa  ggacatccgc  tccacaaacg  agaaaagcgt  caagccaaac  acaaacagcg
```

Building Blocks of DNA & RNA

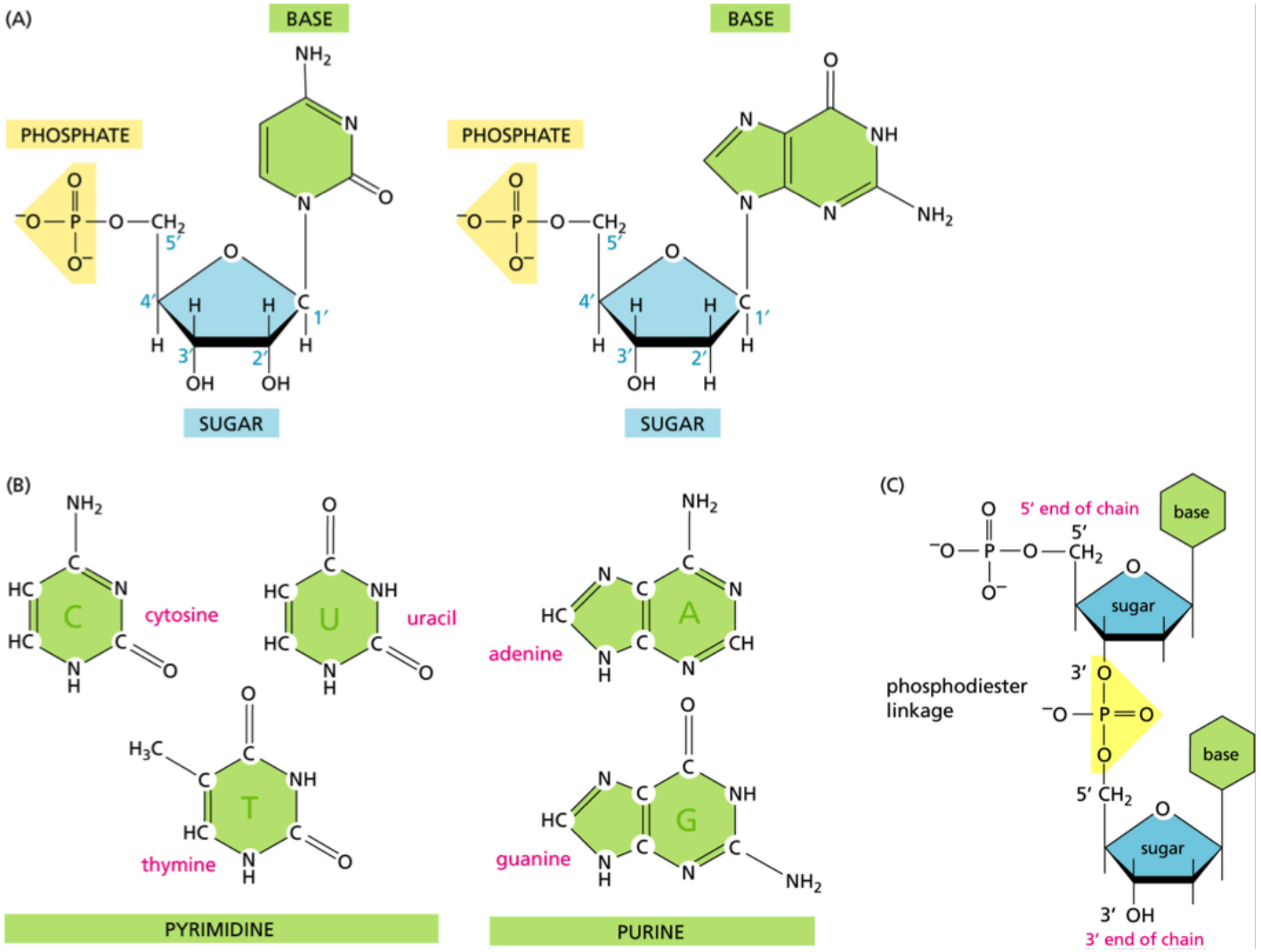


Fig 1.1, Zvelebil & Baum

DNA – Double Helix Structure

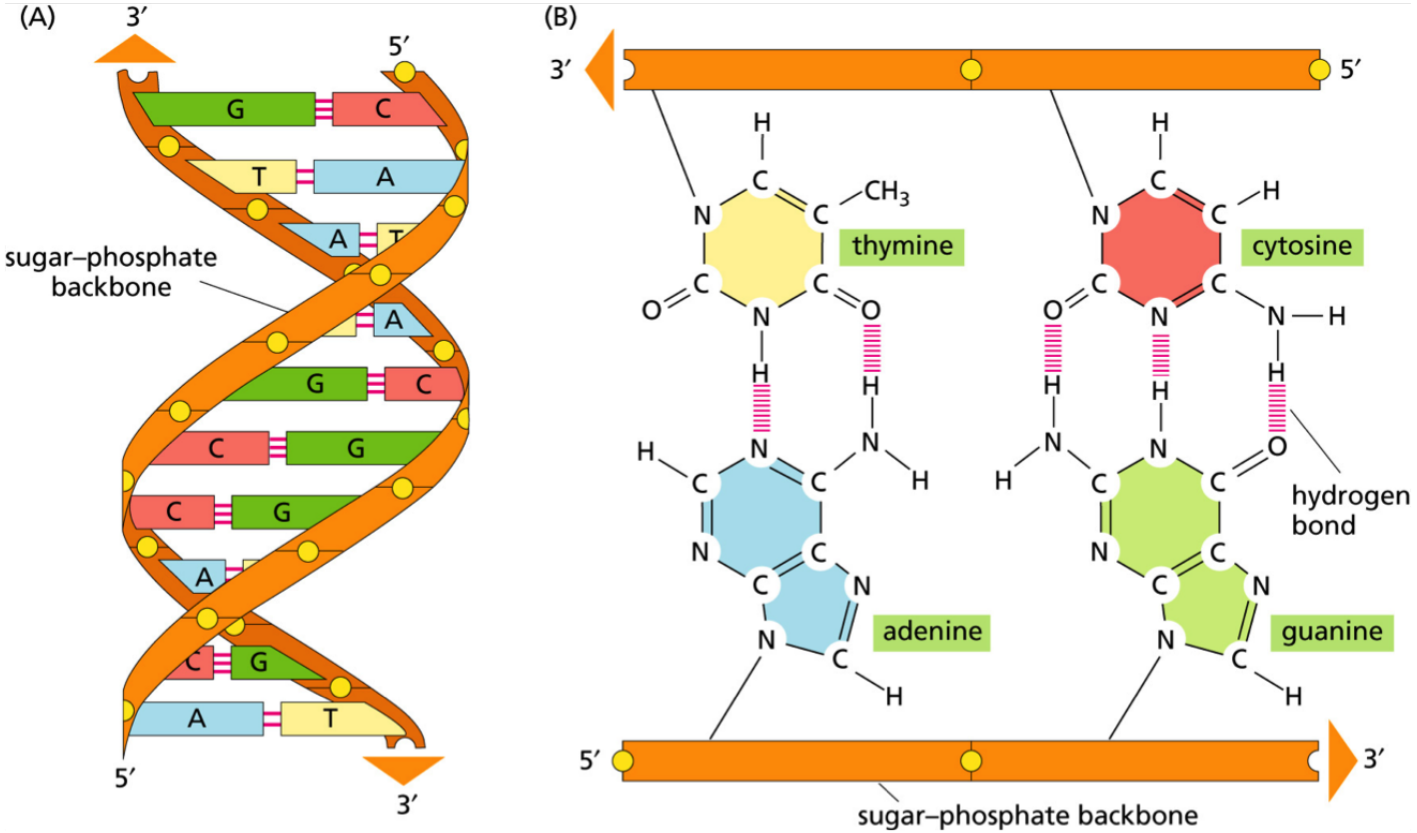
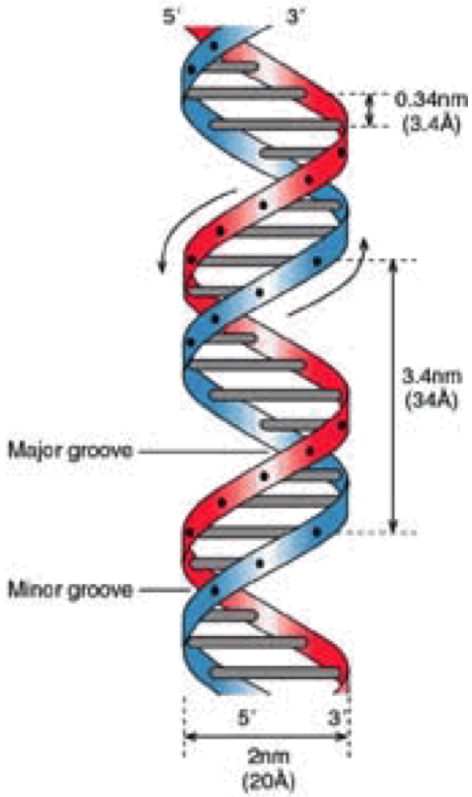
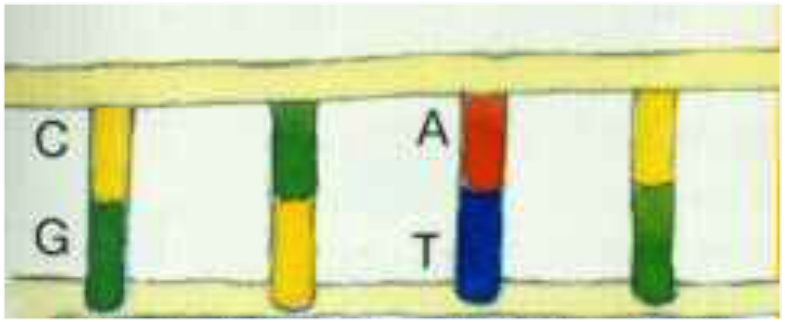


Fig 1.3, Zvelebil & Baum

DNA Molecule



Complementary Bases



From `http://www.cellsalive.com/cells/cell_model.htm`

RNA Molecule

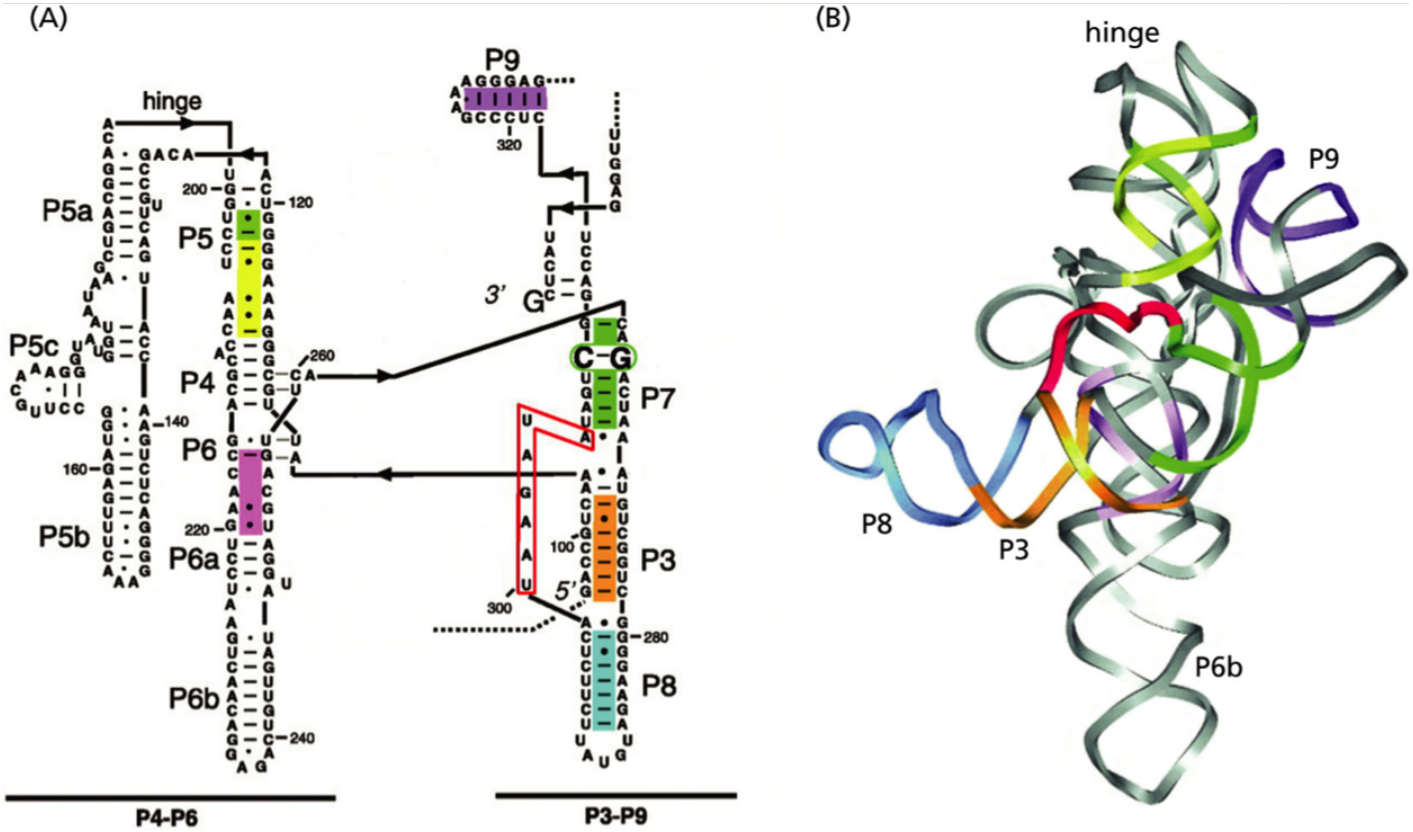


Fig 1.5, Zvelebil & Baum

Protein – The 20 Amino Acids

Letter Code	3 Letter Code	Amino Acid	Letter Code	3 Letter Code	Amino Acid
A	Ala	Alanine	M	Met	Methionine
C	Cys	Cysteine	N	Asn	Asparagine
D	Asp	Aspartic Acid	P	Pro	Proline
E	Glu	Glutamic Acid	Q	Gla	Glutamine
F	Phe	Phenylalanine	R	Arg	Arginine
G	Gly	Glycine	S	Ser	Serine
H	His	Histidine	T	Thr	Threonine
I	Ile	Isoleucine	V	Val	Valine
K	Lys	Lysine	W	Trp	Trypophan
L	Leu	Leucine	Y	Tyr	Tyrosine

Protein – Typical Sequence

>gi|23491729|dbj|BAC16799.1| P53 [Homo sapiens]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAA
PRVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN
TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVHVCACPGR
DRRTEEENLRKKGEPHHELPPGSTKRALSNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALEL
KDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

Protein molecules have a 3D structure

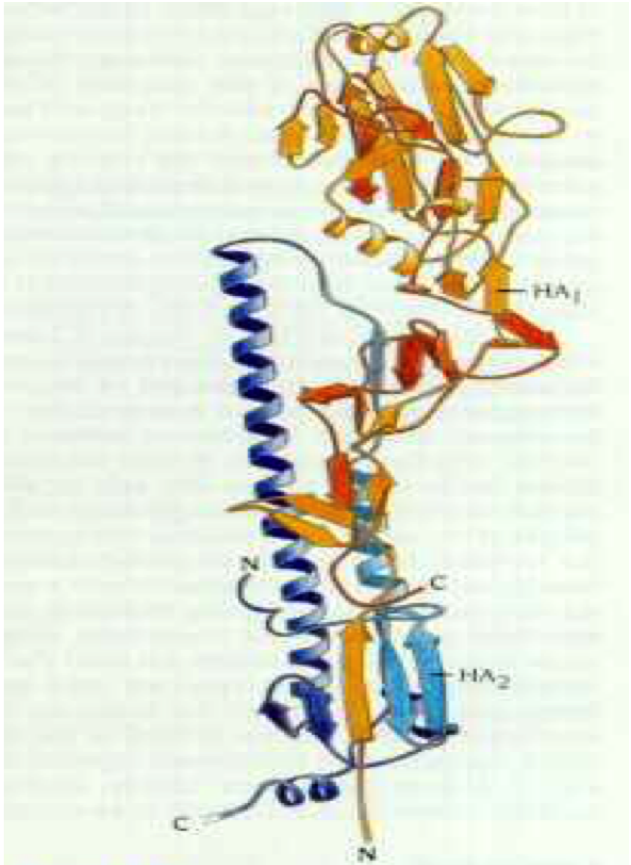


Figure 5.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA_1. (Courtesy of Don Wiley, Harvard University.)

Central Dogma

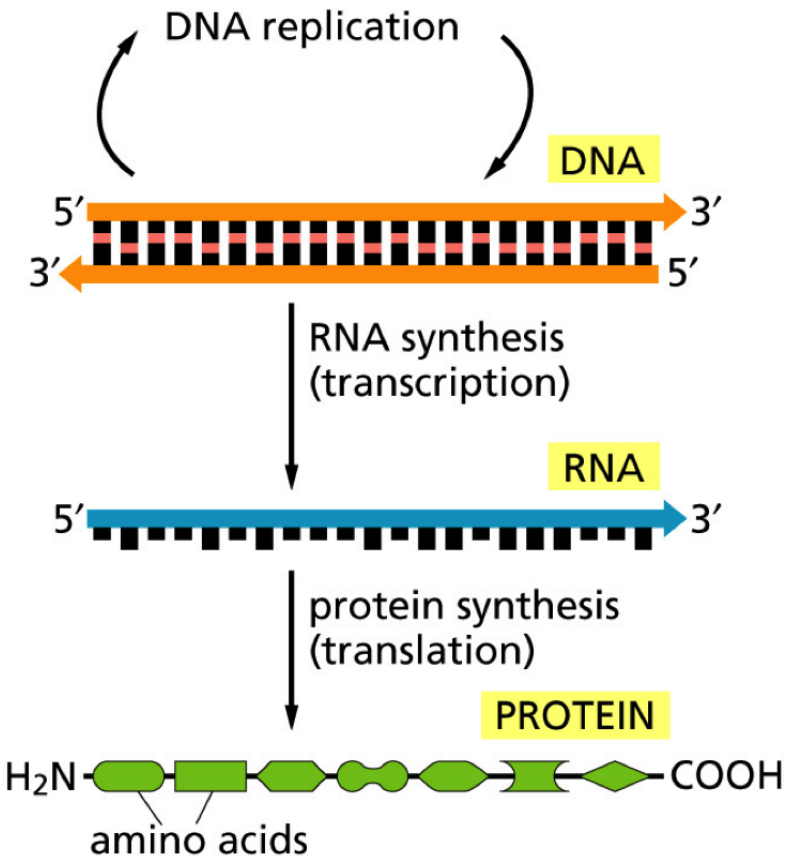


Fig 1.6, Zvelebil & Baum

DNA Replication

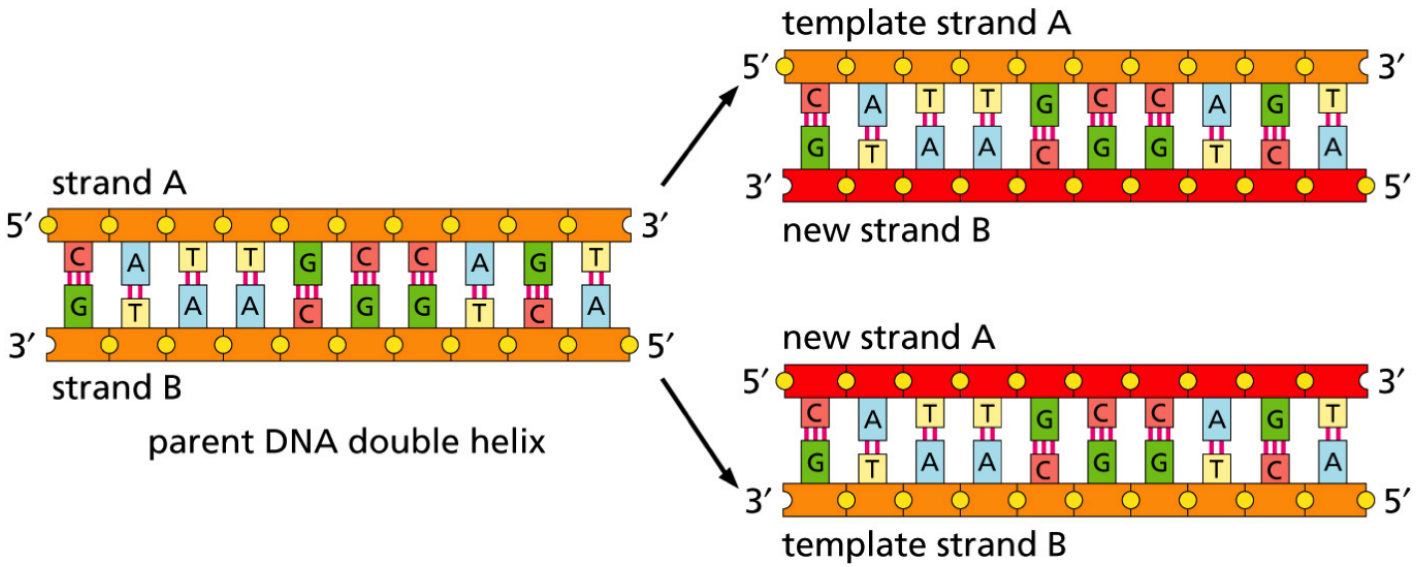
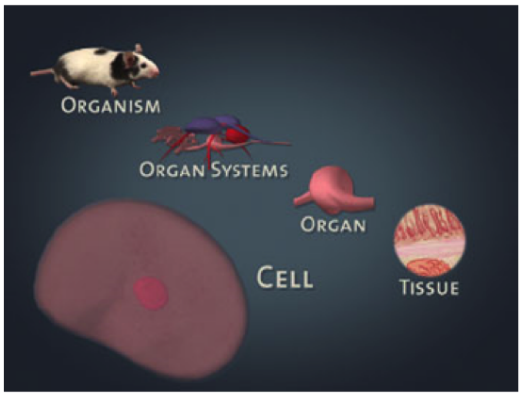
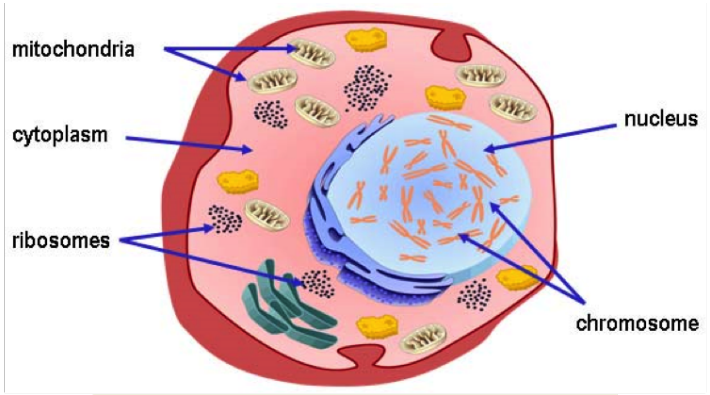


Fig 1.4, Zvelebil & Baum

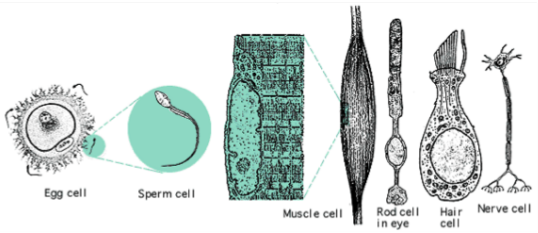
The Cell



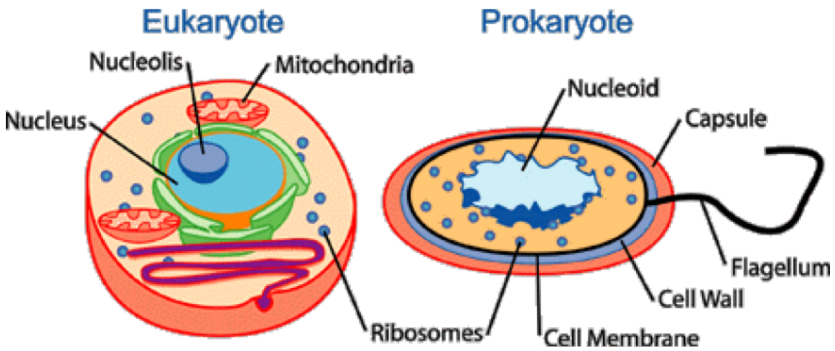
http://www.learner.org/channel/courses/essential/life/session1/closer1.html



http://www.biotechnologyonline.gov.au/popups/img_cellwithlabels.cfm



http://www.biology.eku.edu/RITCHISO/301notes1.htm

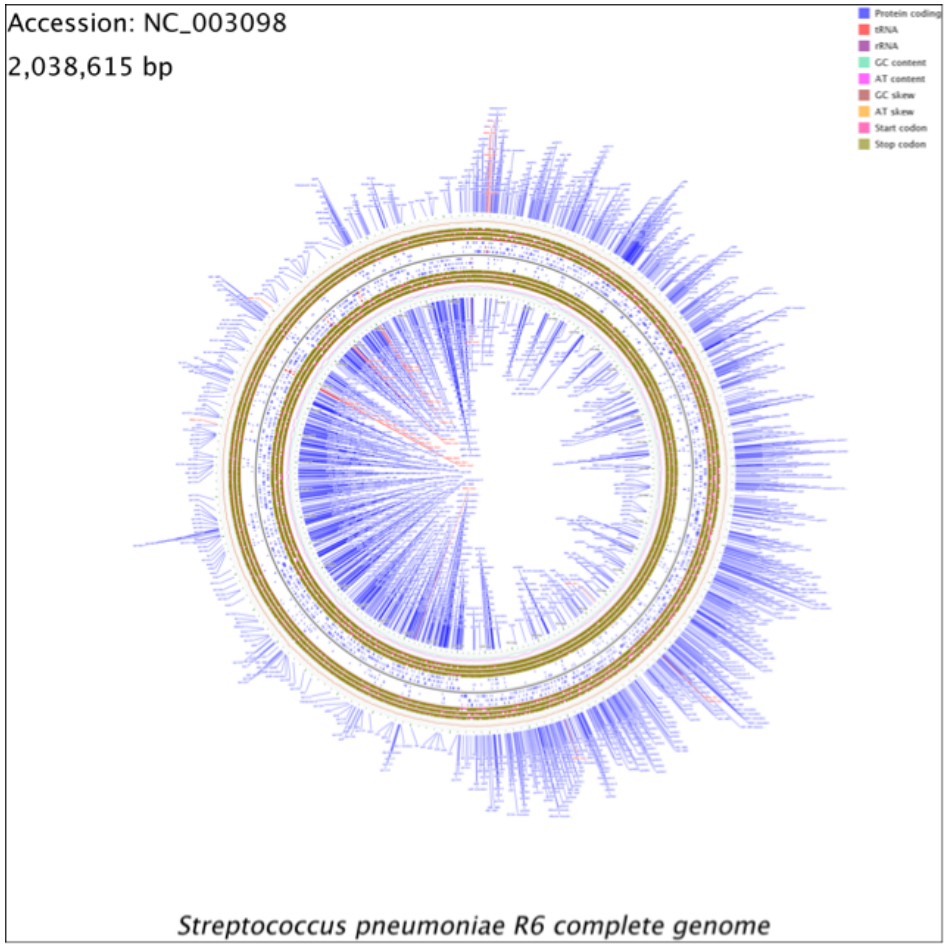


http://en.wikipedia.org/wiki/File:Celltypes.png

From

`http://www.cellsalive.com/cells/cell_model.htm`

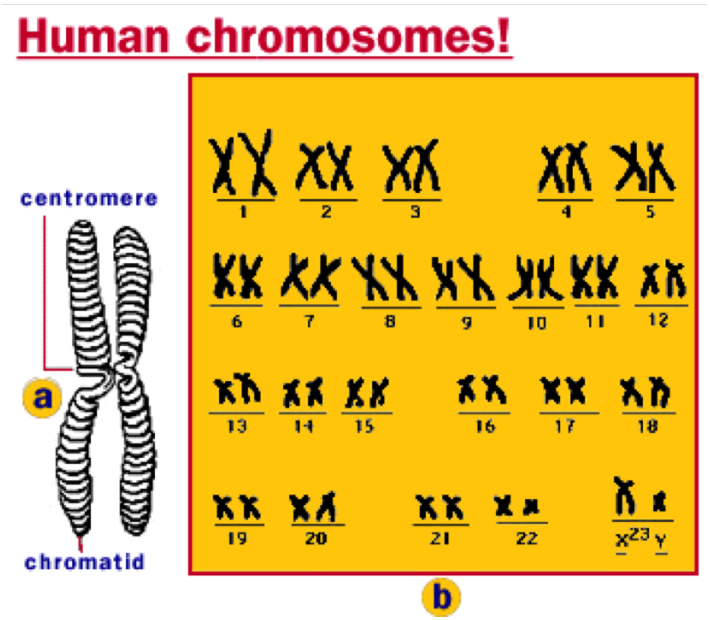
Bacterial Chromosomes

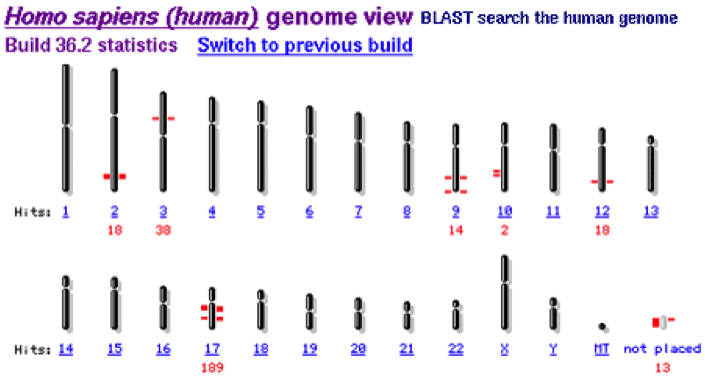


From

`http://www.cellsalive.com/cells/cell_model.htm`

Human Chromosomes





From

`http://www.cellsalive.com/cells/cell_model.htm`

Genes

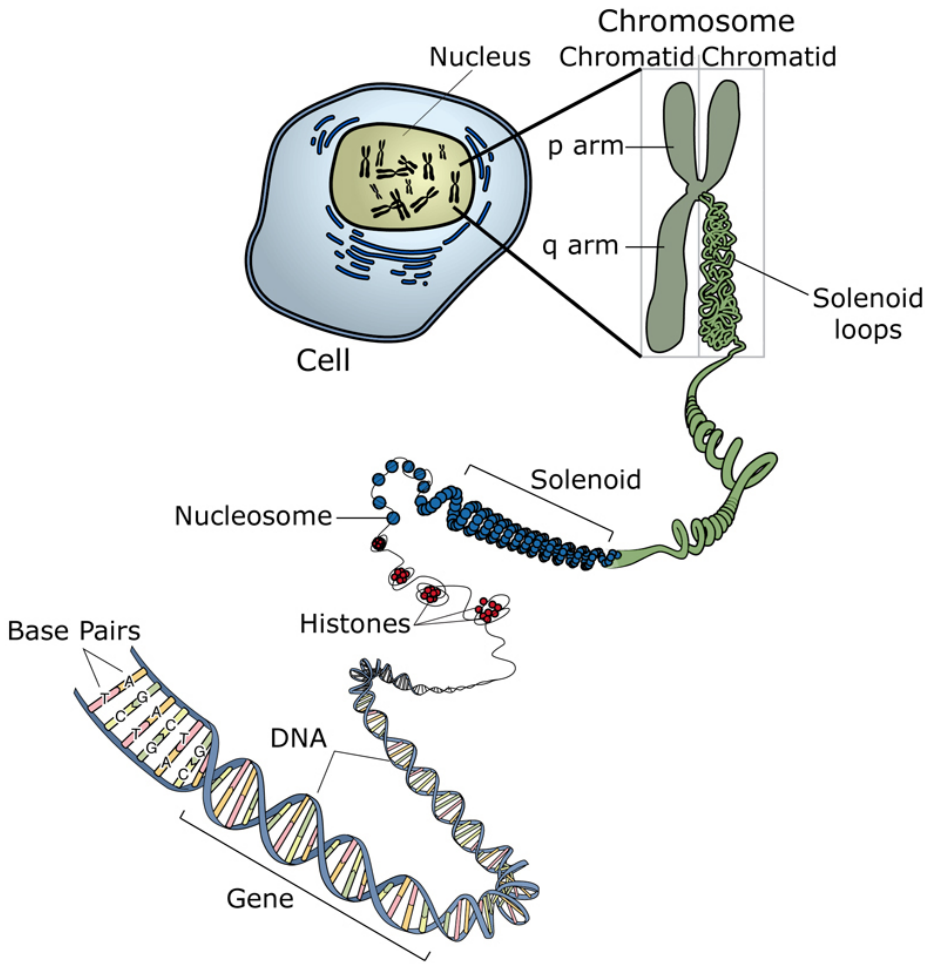
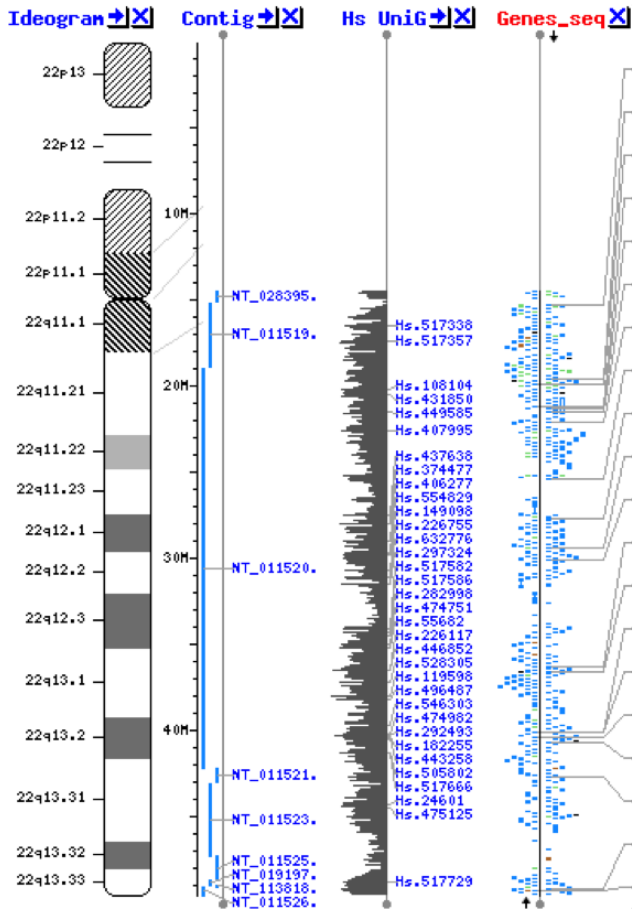


Image adapted from: National Human Genome Research Institute.

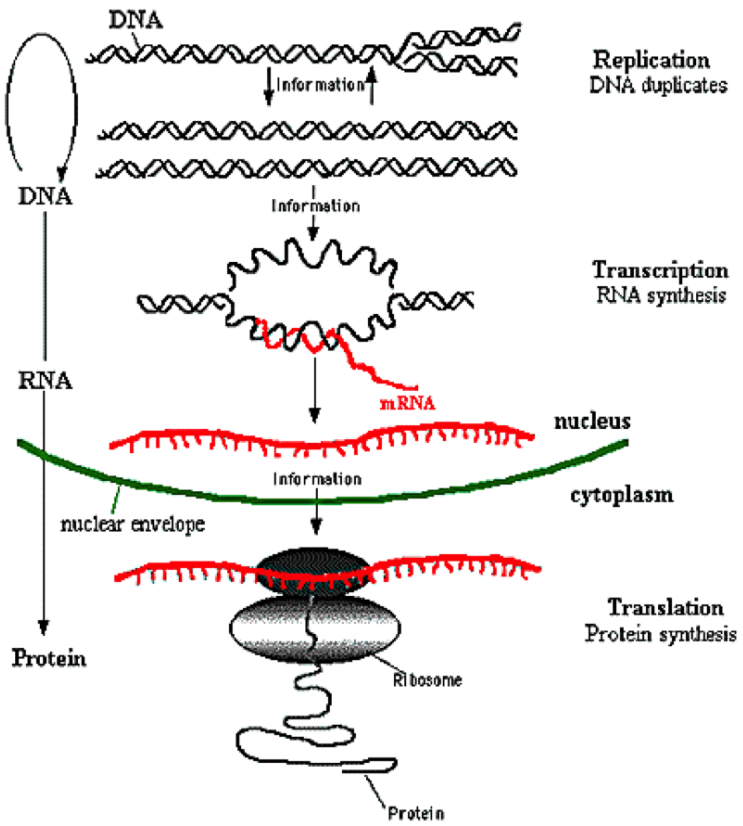
From `http://www2.le.ac.uk/departments/genetics/vgec/diagrams/36chromosomeunravel.jpg`

Human Chromosomes



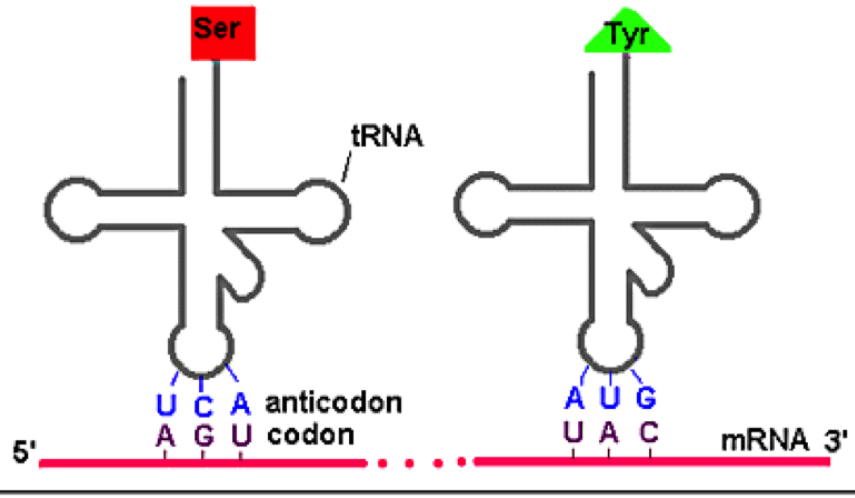
From `http://www.cellsalive.com/cells/cell_model.htm`

Central Dogma



The Central Dogma of Molecular Biology

RNA and Genetic Code

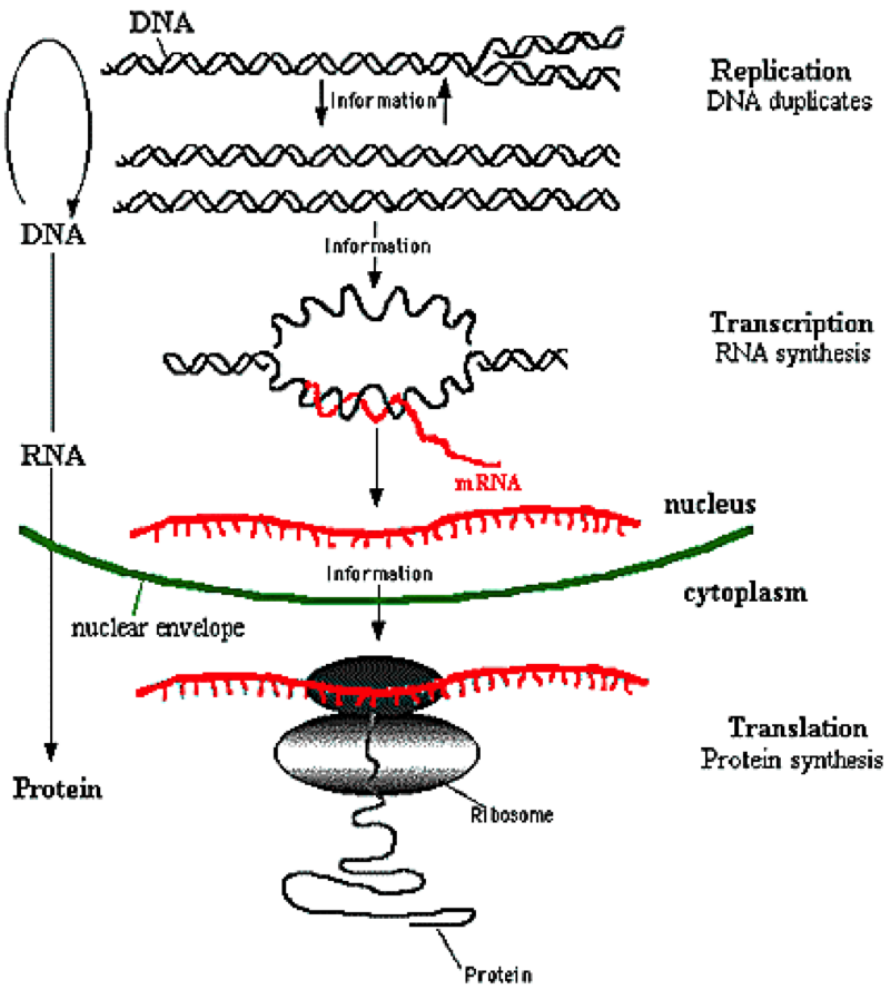


2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

The Genetic Code

Central Dogma



The Central Dogma of Molecular Biology

Transcription

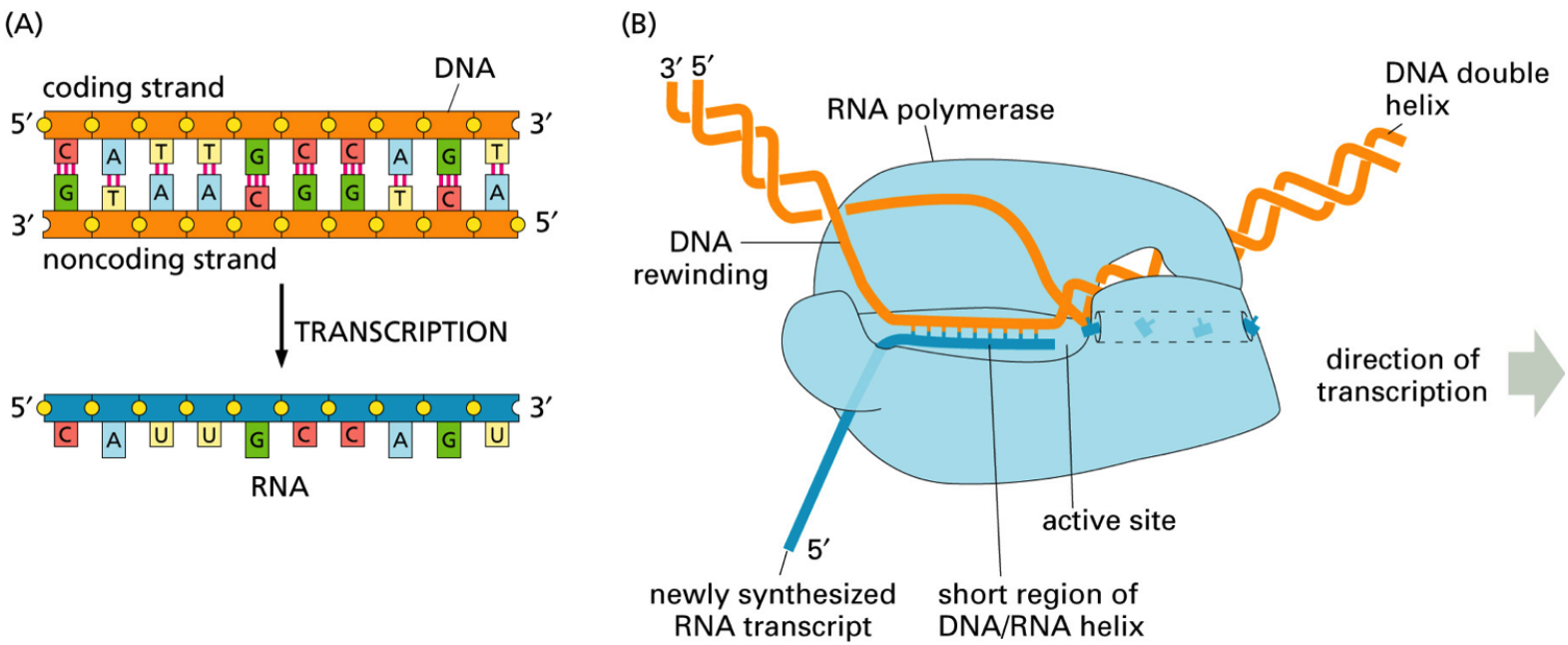
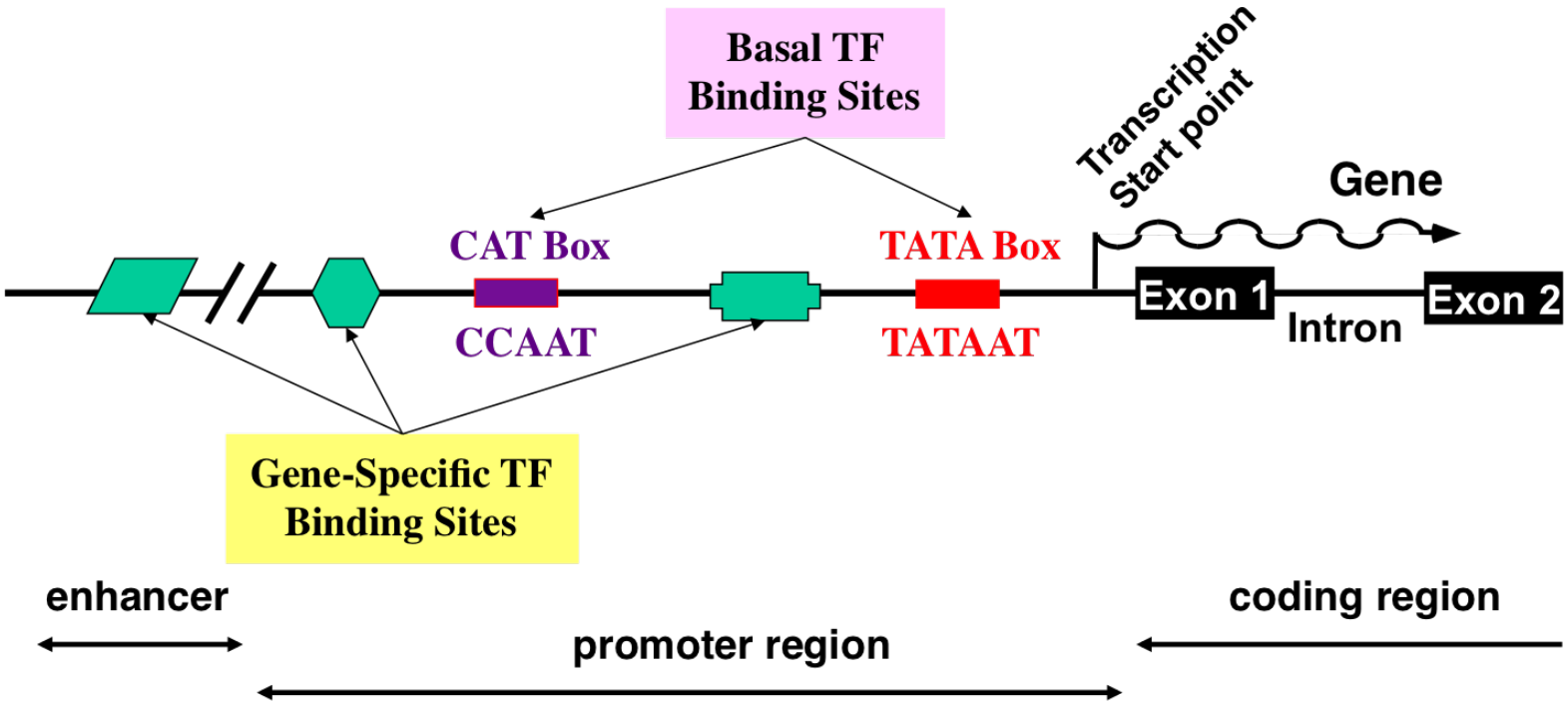


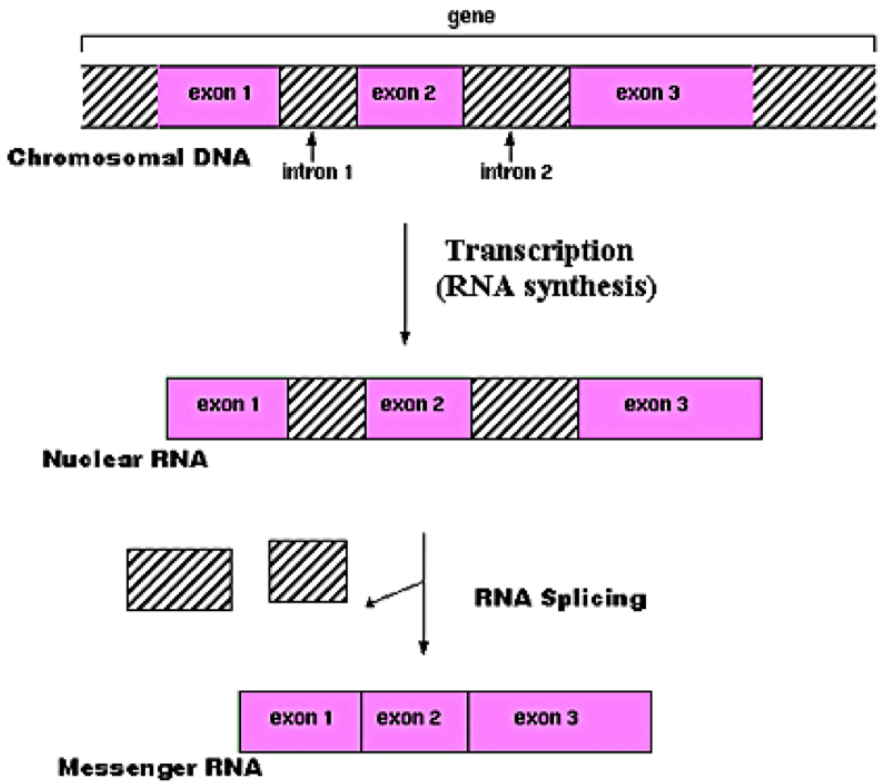
Fig 1.7, Zvelebil & Baum

Transcription



Courtesy: Dr. Kalai Mathee

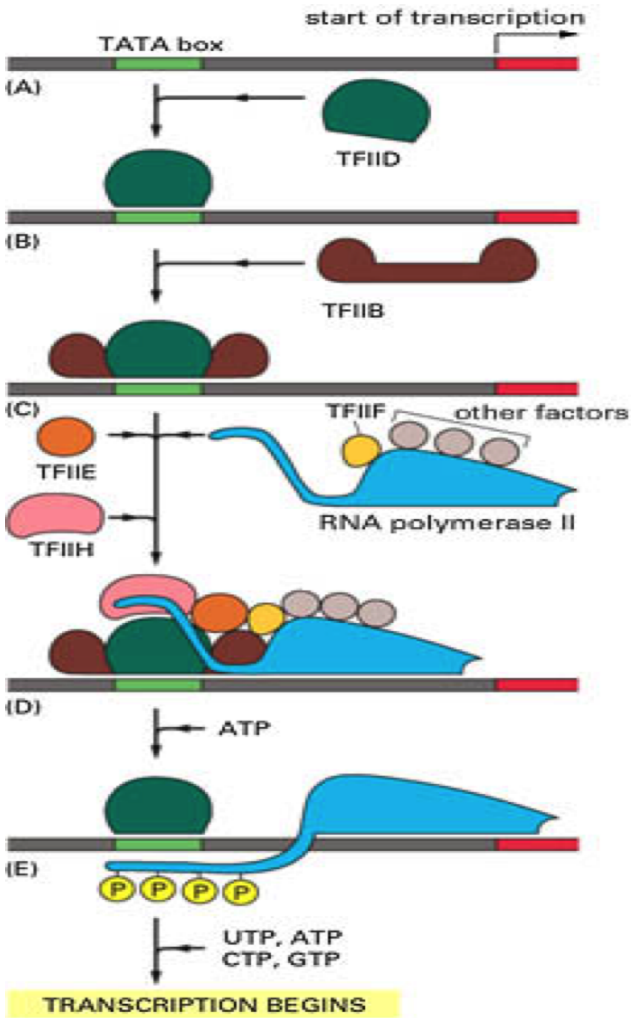
Transcription



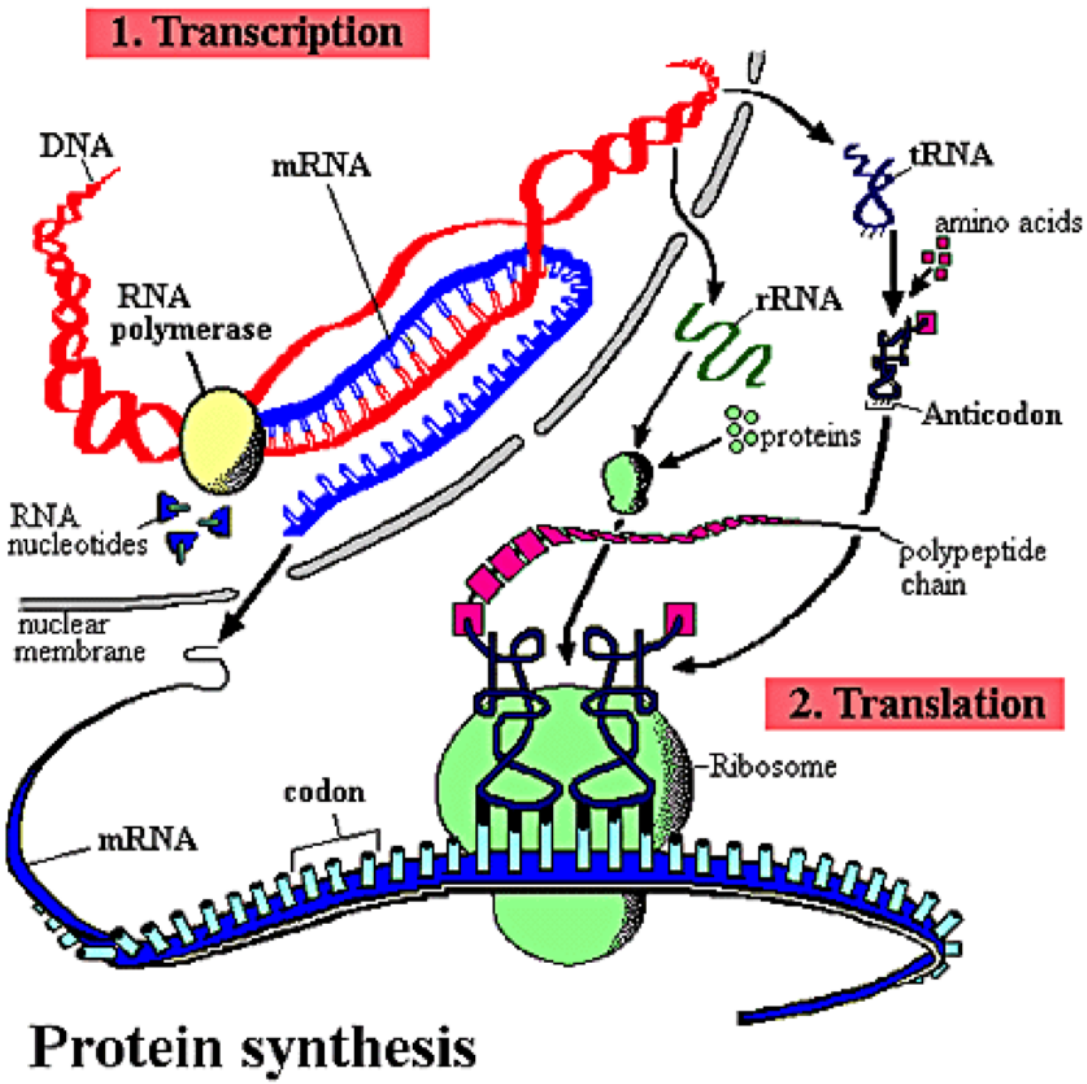
RNA synthesis and processing

Fig 1.6, Zvelebil & Baum

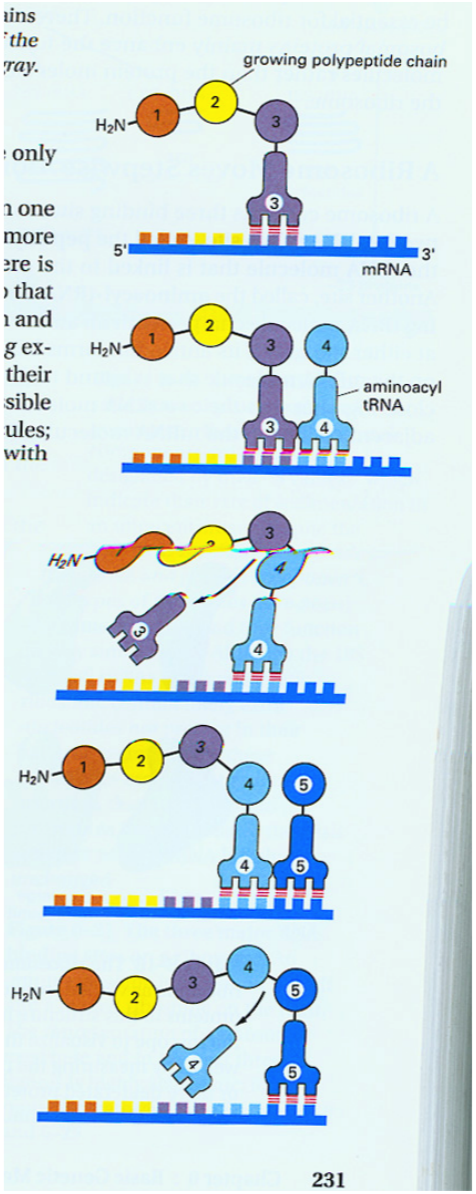
Transcription



Translation



Translation



Presentation Outline

3 Major Public Databases

- GenBank
  - National Center for Biotechnology Information (NCBI)
- EMBL European Mol Biol Laboratory
  - European Bioinformatics Institute (EBI)
- DDBJ: DNA Data Bank of Japan
  - National Institute of Genetics (NIG)
- All 3 have been completely integrated!

Entrez Portal @ NCBI

- PubMed; Bookshelf
- DNA and Protein Sequence database
- Protein Structure database
- Genome Assemblies
- BLAST
- dbSNP
- Taxonomy Browser
- Population study data sets
- PubChem
- GEO (Gene Expression Omnibus)
- OMIM (Mendelian Inheritance in Man)

Other Important Databases

- PDB `http://www.wwpdb.org/`
- KEGG `http://www.genome.jp/kegg/`
- MetaCyc `http://metacyc.org`
- ENCODE `http://encodeproject.org/ENCODE/` (functional elements in human genome)
- 1000 Genomes Project
- International HapMap Project
- Human Microbiome Project
- Human Epigenome Project
- Gene Ontology (GO)
- Human Connectome Project

Presentation Outline

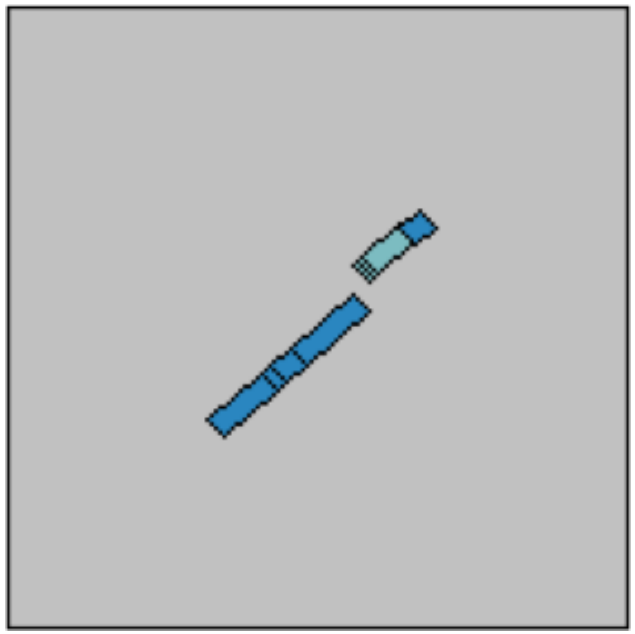
1. Can show sequences are close

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

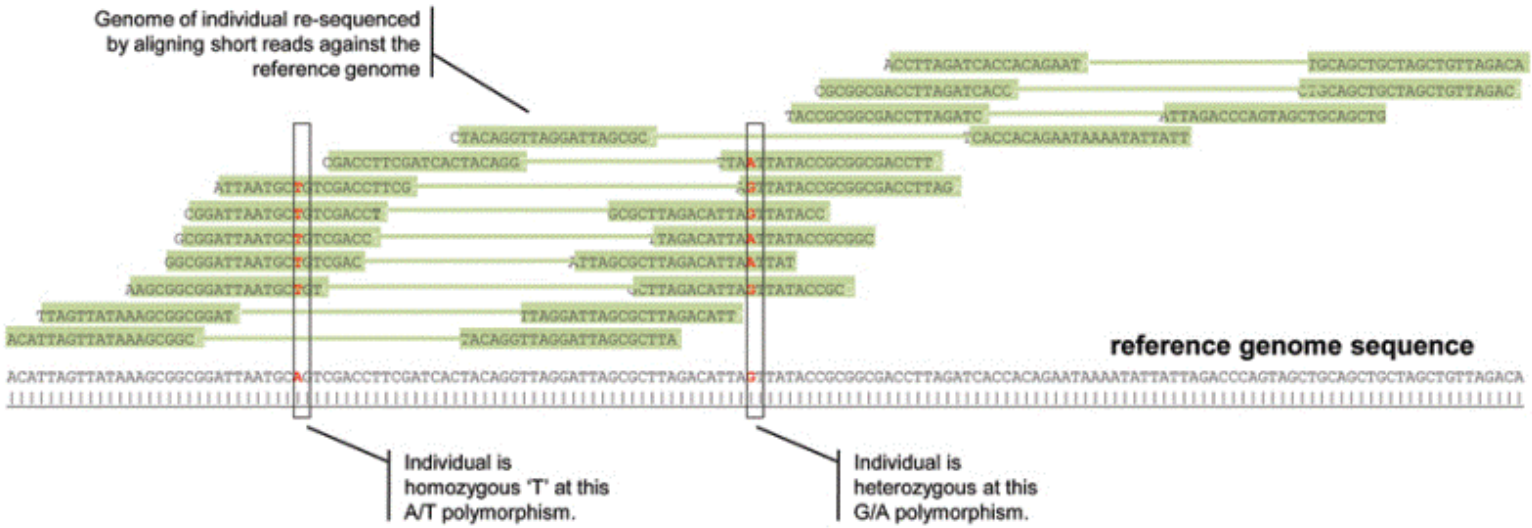
V-sis Oncogene – Homologies

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 help in sequence assembly



6. Can be basis for discovery

- Early 1970s: SSV causes cancer in some species of monkeys.

6. Can be basis for discovery

- Early 1970s: SSV causes cancer in some species of monkeys.
- 1970s: infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.

6. Can be basis for discovery

- Early 1970s: SSV causes cancer in some species of monkeys.
- 1970s: infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.
  - Hypothesis: Oncogenes in viruses encode cellular growth factors (proteins to stimulate growth); Uncontrolled quantities of growth factors produced by infected cells cause cancer-like behavior.

6. Can be basis for discovery

- Early 1970s: SSV causes cancer in some species of monkeys.
- 1970s: infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.
  - Hypothesis: Oncogenes in viruses encode cellular growth factors (proteins to stimulate growth); Uncontrolled quantities of growth factors produced by infected cells cause cancer-like behavior.
- 1983:

6. Can be basis for discovery

- Early 1970s: SSV causes cancer in some species of monkeys.
- 1970s: infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.
  - Hypothesis: Oncogenes in viruses encode cellular growth factors (proteins to stimulate growth); Uncontrolled quantities of growth factors produced by infected cells cause cancer-like behavior.
- 1983:
  - Oncogene from SSV called v-sis isolated & sequenced.
  - Partial sequence for platelet-derived growth factor (PDGF) sequenced & published; PDGF stimulates proliferation of cells.

6. Can be basis for discovery

- Early 1970s: SSV causes cancer in some species of monkeys.
- 1970s: infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.
  - Hypothesis: Oncogenes in viruses encode cellular growth factors (proteins to stimulate growth); Uncontrolled quantities of growth factors produced by infected cells cause cancer-like behavior.
- 1983:
  - Oncogene from SSV called v-sis isolated & sequenced.
  - Partial sequence for platelet-derived growth factor (PDGF) sequenced & published; PDGF stimulates proliferation of cells.
  - R.F. Doolittle was maintaining one of the earliest home-grown databases of published aa sequences.

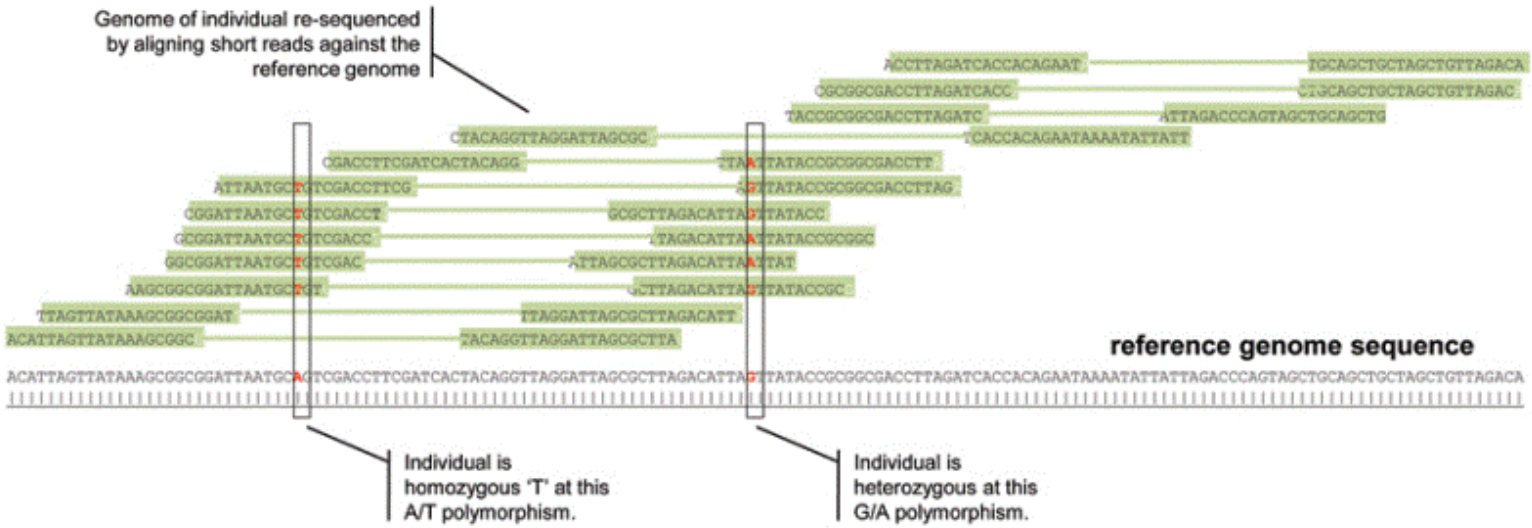
6. Can be basis for discovery

- Early 1970s: SSV causes cancer in some species of monkeys.
- 1970s: infection by certain viruses cause some cells in culture (in vitro) to grow without bounds.
  - Hypothesis: Oncogenes in viruses encode cellular growth factors (proteins to stimulate growth); Uncontrolled quantities of growth factors produced by infected cells cause cancer-like behavior.
- 1983:
  - Oncogene from SSV called v-sis isolated & sequenced.
  - Partial sequence for platelet-derived growth factor (PDGF) sequenced & published; PDGF stimulates proliferation of cells.
  - R.F. Doolittle was maintaining one of the earliest home-grown databases of published aa sequences.
  - Sequence Alignment of v-sis and PDGF had surprises.

PDGF and v-SIS

- Alignment was good.
- Two regions of alignment
  - region of 31 aa with 26 matches
  - region of 39 with 35 matches
- Conclusion:
  - Previously harmless virus incorporates growth-related gene (proto-oncogene) of its host into its genome.
  - Gene gets mutated in the virus, or moves closer to a strong enhancer, or moves away from a repressor.
  - When virus infects a cell, it causes uncontrolled amount of growth factor.
- Several other oncogenes known to be similar to growth-regulating proteins in normal cells.

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



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.

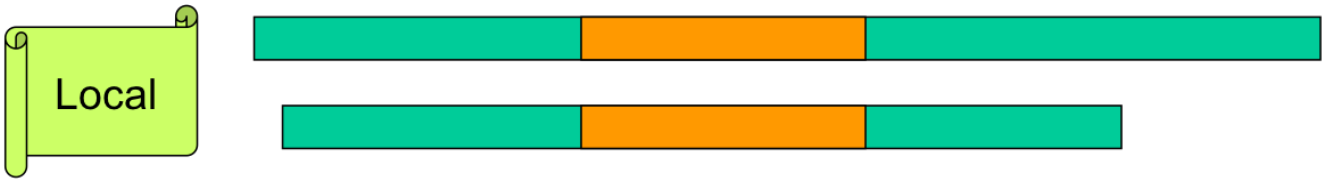
Similarity vs. Homology

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

Types of Sequence Alignment ... 1

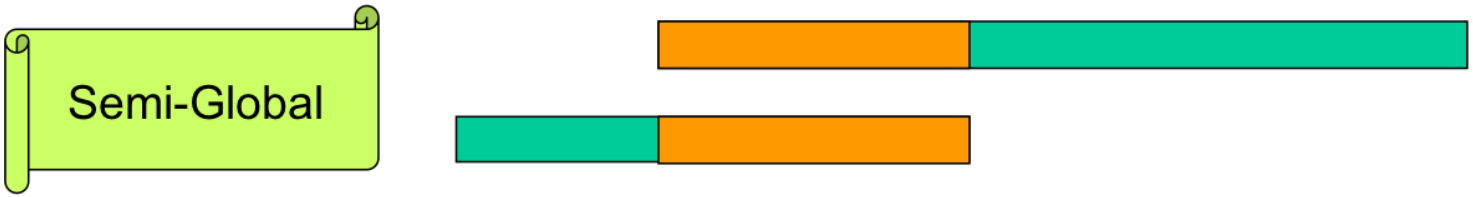
- Global Alignment: similarity over entire length

Types of Sequence Alignment ... 2



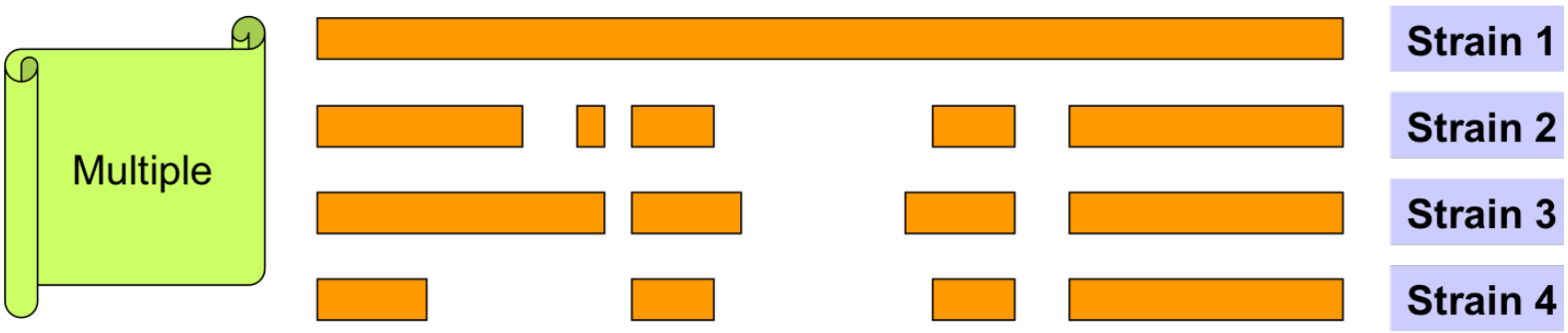
- Local Alignment: no overall similarity, but some segment(s) is/are similar

Types of Sequence Alignment ... 3



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

Types of Sequence Alignment ... 4

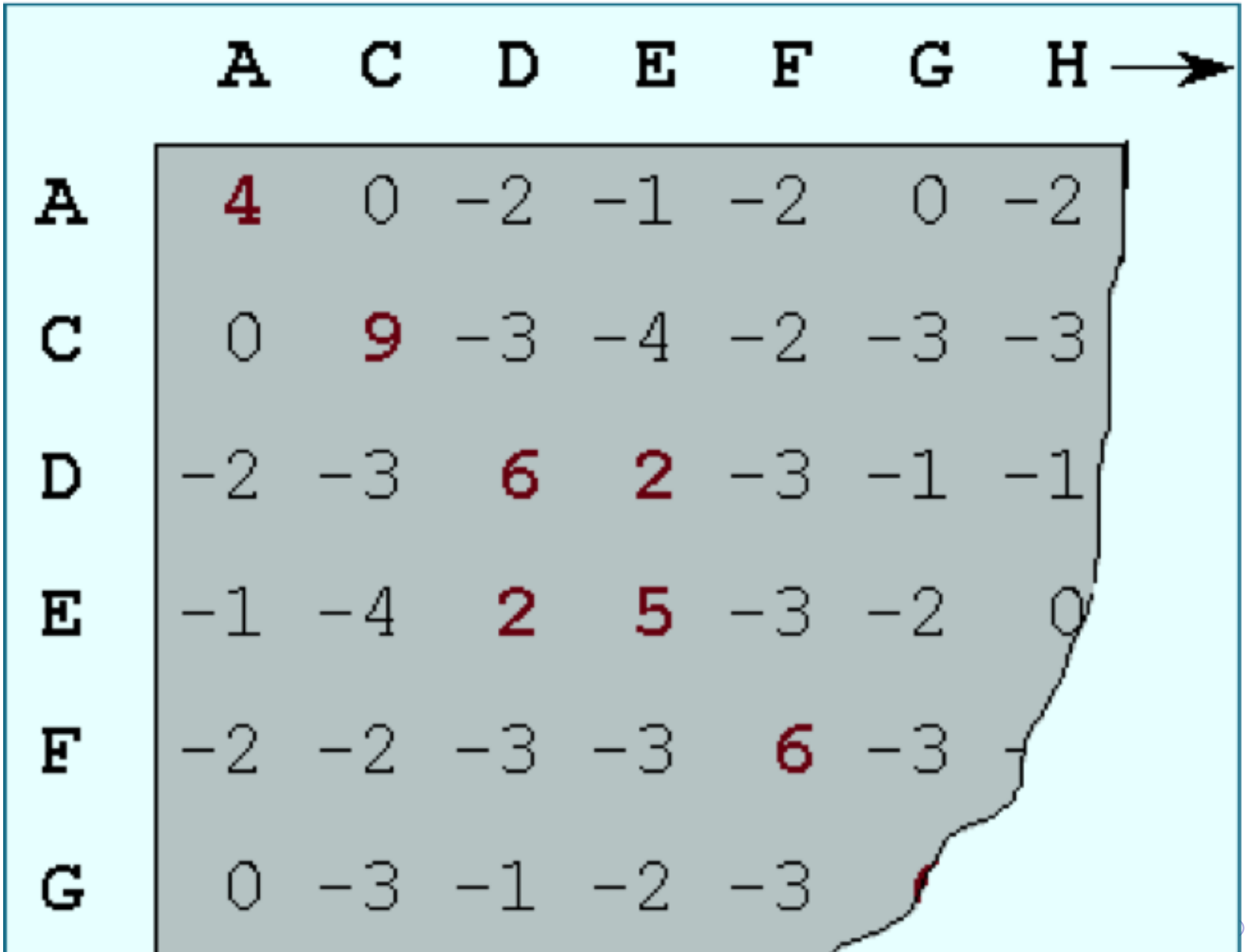


- Multiple Alignment: similarity between sets of sequences

Sequence Alignment Algorithms

- Global alignments: Needleman-Wunsch-Sellers 1970
- Local alignments: Smith-Waterman 1981
- Both use Dynamic Programming

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		

Revolution in Sequence Alignment Algorithms

- FASTA: Lipman, Pearson '85, '88
- Basic Local Alignment Search Tool (BLAST): Altschul, Gish, Miller, Myers, Lipman '90

Revolution in Sequence Alignment Algorithms

- FASTA: Lipman, Pearson '85, '88
- Basic Local Alignment Search Tool (BLAST): Altschul, Gish, Miller, Myers, Lipman '90
- Both programs:
  - search entire databases
  - tremendous speed and sensitivity
  - report statistical significance

BLAST

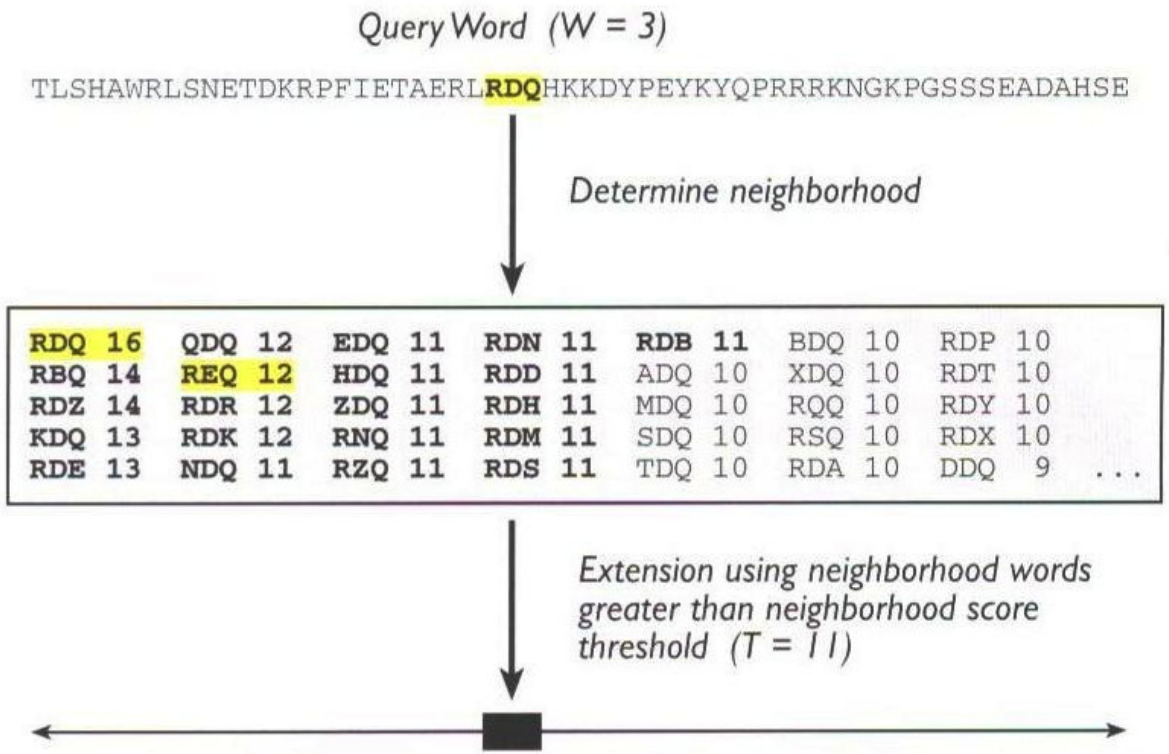


FIGURE 11.7 The initiation of a **BLAST** search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words "in the neighborhood" of the original query word. Any occurrences of these neighborhood words 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

BLAST Variants

- Nucleotide BLAST
  - blastn
  - MEGABIAST
  - Short Sequences (higher E-value threshold, smaller word size, no low-complexity filtering)
- Protein BLAST
  - blastp
  - PSI-BIAST
  - PHI-BLAST
  - 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 Parameters

- Type of query: nucleotide / protein

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w
- Gap penalties, p_1, p_2

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w
- Gap penalties, p_1, p_2
- Threshold scores, S, T

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w
- Gap penalties, p_1, p_2
- Threshold scores, S, T
- E-value cutoff, E

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w
- Gap penalties, p_1, p_2
- Threshold scores, S, T
- E-value cutoff, E
  - E-value, E, is the expected number of sequences that would have an alignment score greater than the current score, S

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w
- Gap penalties, p_1, p_2
- Threshold scores, S, T
- E-value cutoff, E
  - E-value, E, is the expected number of sequences that would have an alignment score greater than the current score, S
- Number of hits to display, H

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w
- Gap penalties, p_1, p_2
- Threshold scores, S, T
- E-value cutoff, E
  - E-value, E, is the expected number of sequences that would have an alignment score greater than the current score, S
- Number of hits to display, H
- Database to search, D

BLAST Parameters

- Type of query: nucleotide / protein
- Word size, w
- Gap penalties, p_1, p_2
- Threshold scores, S, T
- E-value cutoff, E
  - E-value, E, is the expected number of sequences that would have an alignment score greater than the current score, S
- Number of hits to display, H
- Database to search, D
- Scoring Matrix, M

BLAST Database and Scoring Matrix

- Databases:

BLAST Database and Scoring Matrix

- Databases:
  - Protein: NR)non-redudant, SwissPROT/UniPROT, pdb, custom

BLAST Database and Scoring Matrix

- Databases:
  - Protein: NR)non-redudant, SwissPROT/UniPROT, pdb, custom
  - Nucleotide: NR, dbest, dbsts, htgs, gss, pdb, vector, ...

BLAST Database and Scoring Matrix

- Databases:
  - Protein: NR)non-redudant, SwissPROT/UniPROT, pdb, custom
  - Nucleotide: NR, dbest, dbsts, htgs, gss, pdb, vector, ...
- Scoring Matrices

BLAST Database and Scoring Matrix

- Databases:
  - Protein: NR)non-redudant, SwissPROT/UniPROT, pdb, custom
  - Nucleotide: NR, dbest, dbsts, htgs, gss, pdb, vector, ...
- Scoring Matrices
  - PAM Matrices: PAM 40, 160, 250

BLAST Database and Scoring Matrix

- Databases:
  - Protein: NR)non-redudant, SwissPROT/UniPROT, pdb, custom
  - Nucleotide: NR, dbest, dbsts, htgs, gss, pdb, vector, ...
- Scoring Matrices
  - PAM Matrices: PAM 40, 160, 250 going fmor short alignments with high similarity (70-90 %) to members of a protein family (50-60 %), to longer alignments with divergent homologous sequences (less than 30 %)
  - BLOSUM Matrices: BLSOUM90, 80, 62, 30

BLAST Database and Scoring Matrix

- Databases:
  - Protein: NR)non-redudant, SwissPROT/UniPROT, pdb, custom
  - Nucleotide: NR, dbest, dbsts, htgs, gss, pdb, vector, ...
- Scoring Matrices
  - PAM Matrices: PAM 40, 160, 250 going fmor short alignments with high similarity (70-90 %) to members of a protein family (50-60 %), to longer alignments with divergent homologous sequences (less than 30 %)
  - BLOSUM Matrices: BLSOUM90, 80, 62, 30 going fmor short alignments with high similarity (70-90 %) to members of a protein family (50-60 %), to weak homologs (30-40 %), to longer alignments with divergent homologous sequences (less than 30 %) vector, ...

BLAST Database and Scoring Matrix

- Databases:
  - Protein: NR)non-redudant, SwissPROT/UniPROT, pdb, custom
  - Nucleotide: NR, dbest, dbsts, htgs, gss, pdb, vector, ...
- Scoring Matrices
  - PAM Matrices: PAM 40, 160, 250 going fmor short alignments with high similarity (70-90 %) to members of a protein family (50-60 %), to longer alignments with divergent homologous sequences (less than 30 %)
  - BLOSUM Matrices: BLSOUM90, 80, 62, 30 going fmor short alignments with high similarity (70-90 %) to members of a protein family (50-60 %), to weak homologs (30-40 %), to longer alignments with divergent homologous sequences (less than 30 %) vector, ...

Rules of Thumb

- Homology is often characterized by significant similarity over entire sequence or strong similarity in key places.

Rules of Thumb

- Homology is often characterized by significant similarity over entire sequence or strong similarity in key places.
- Matches that are $> 50\%$ identical in a 20-40 aa region occur frequently by chance

Rules of Thumb

- Homology is often characterized by significant similarity over entire sequence or strong similarity in key places.
- 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.

Rules of Thumb

- Homology is often characterized by significant similarity over entire sequence or strong similarity in key places.
- 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.
- Homology is transitive.

Rules of Thumb

- Homology is often characterized by significant similarity over entire sequence or strong similarity in key places.
- 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.
- Homology is transitive. A homologous to B & B to $C \Rightarrow A$ homologous to C.

Rules of Thumb

- Homology is often characterized by significant similarity over entire sequence or strong similarity in key places.
- 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.
- Homology is transitive. 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.

Rules of Thumb

- Homology is often characterized by significant similarity over entire sequence or strong similarity in key places.
- 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.
- Homology is transitive. 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 ... 2

- Results of searches using different scoring systems may be compared directly using normalized scores.

Rules of Thumb ... 2

- 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.

Rules of Thumb ... 2

- 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,

$$S' > \log(N/E)$$

  where E-value $= E$ and $N =$ size of search space.