Sequence Alignment

Example:

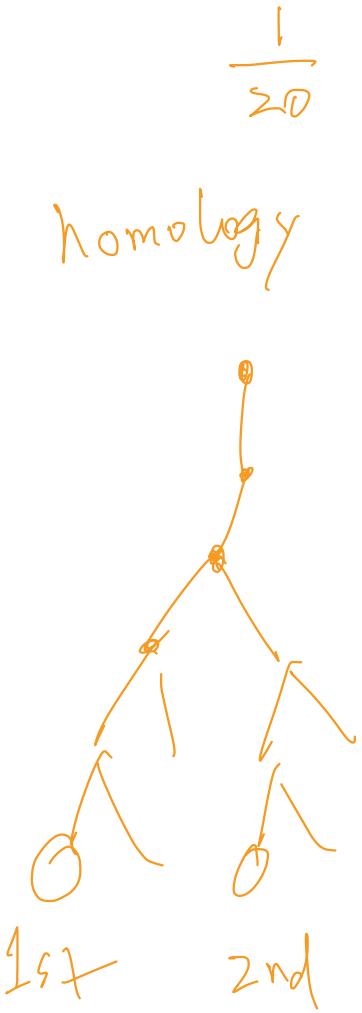
>AVP78042.1 spike protein [Bat SARS-like coronavirus]
MLFFLFLQFALVNSQCDLTGRTPLNPNYTNSSQRGVYYPDTIYRSDTLVLSQGYFLPFYSNVSWYYSLTT
NNAATKRTDNPILDFKDGIYFAATEHSNIVRGWIFGTTLDNTSQSLLIVNNATNVIIKVCNFDFCYDPYL
SGYYHNNKTWSIREFAVYSFYANCTFEYVSKSFMLNISGNGGLFNTLREFVFRNVDGHFKIYSKFTPVNL
NRGLPTGLSVLQPLVELPVSINITKFRTLLTIHRGDPMSNNGWTAFSAAYFVGYLKPRTFMLKYNENGTI
TDAVDCALDPLSETKCTLKSLSVQKGIYQTSNFRVQPTQSIVRFPNITNVCPFHKVFNATRFPSVYAWER
TKISDCIADYTVFYNSTSFSTFKCYGVSPSKLIDLCFTSVYADTFLIRFSEVRQVAPGQTGVIADYNYKL
PDDFTGCVIAWNTAKQDTGHYFYRSHRSTKLKPFERDLSSDENGVRTLSTYDFNPNVPLEYQATRVVVLS
FELLNAPATVCGPKLSTQLVKNQCVNFNFNGLKGTGVLTDSSKRFQSFQQFGKDASDFIDSVRDPQTLEI
LDITPCSFGGVSVITPGTNTSSEVAVLYQDVNCTDVPTTIHADQLTPAWRIYAIGTSVFQTQAGCLIGAE
HVNASYECDIPIGAGICASYHTASILRSTGQKAIVAYTMSLGAENSIAYANNSIAIPTNFSISVTTEVMP
VSMAKTSVDCTMYICGDSIECSNLLLQYGSFCTQLNRALSGIAIEQDKNTQEVFAQVKQIYKTPPIKDFG
GFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDISARDLICAQKFNGLTVLPPLLTDE
MIAAYTAALISGTATAGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQESL
TSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYV
TQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYIPSQEKNFTTA
PAICHEGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIINNTVYDPLQPELDSF
KEELDKYFKNHTSPDIDLGDISGINASVVNIQKEIDRLNEVARNLNESLIDLQELGKYEHYIKWPWYVWL
GFIAGLIAIVMVTILLCCMTSCCSCLKGCCSCGFCCKFDEDDSEPVLKGVKLHYT

>YP_009724390.1 surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHV
SGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPF
LGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPI
NLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYN
ENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV
YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD
YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF
PLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL
PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLT
PTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLG
AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGI
AVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC
LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIG
VTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDI
LSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLM
SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNT
FVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVA
KNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD
SEPVLKGVKLHYT

- How do we know these two proteins are similar?
- Many existing tools: such as Clustal Omega.

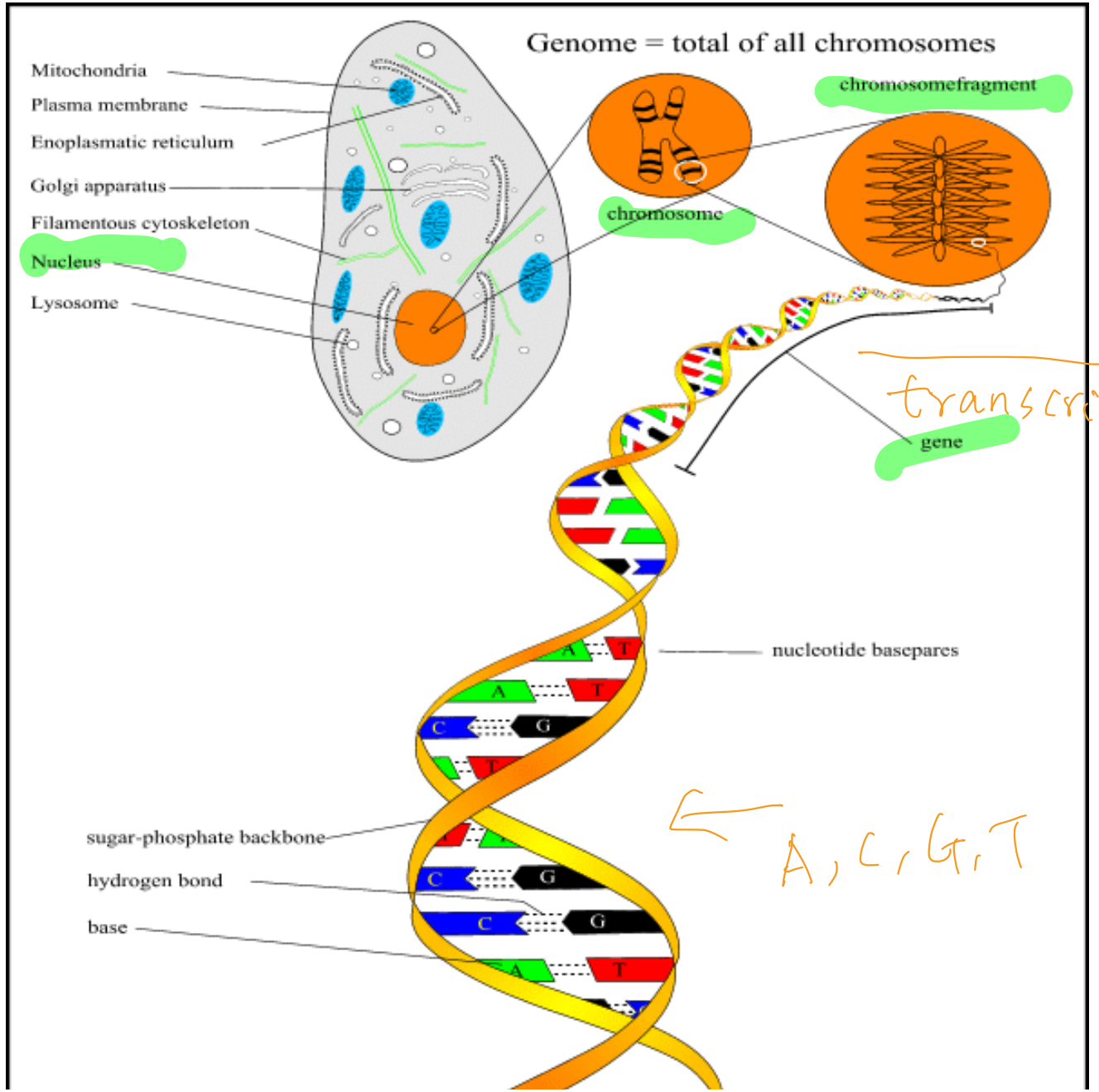
Sequence Alignment

```
AVP78042.1          VNNATNVIIKVCNFDFCYDPYLSGYYHN-NKTWSIREFAVYSFYANCTFEYVSKSFMLNI   177
YP_009724390.1      VNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDL   179
                    *******:****:*:** **:*. ***: **:*   ** ***   ********: *::::

AVP78042.1          SGNGGLFNTLREFVFRNVDGHFKIYSKFTPVNLNRGLPTGLSVLQPLVELPVSINITKFR   237
YP_009724390.1      EGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQ   239
                    .*: * *:.******:*:**:******.**:** *.** *:*.*:***:**:.****:*:

AVP78042.1          TLLTIHRGDP---MSNNGWTAFSAAYFVGYLKPRTFMLKYNENGTITDAVDCALDPLSET   294
YP_009724390.1      TLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSET   299
                    ***::**.      *..**** :***:****:****:**********************

AVP78042.1          KCTLKSLSVQKGIYQTSNFRVQPTQSIVRFPNITNVCPFHKVFNATRFPSVYAWERTKIS   354
YP_009724390.1      KCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRIS   359
                    ******::*:**************:**********:*** :******* *****:*.:**

AVP78042.1          DCIADYTVFYNSTSFSTFKCYGVSPSKLIDLCFTSVYADTFLIRFSEVRQVAPGQTGVIA   414
YP_009724390.1      NCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIA   419
                    :*:***:*:***:************:** *****.****:*:** .****:****** **

AVP78042.1          DYNYKLPDDFTGCVIAWNTAKQDT-----GHYFYRSHRSTKLKPFERDLSSDEN------   463
YP_009724390.1      DYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTP   479
                    ******************: : *:      :*:** .*.::********:*::
```



- Too many identical positions to be random.
- Insertion/deletion (indel) needed for a proper comparison.

DNA



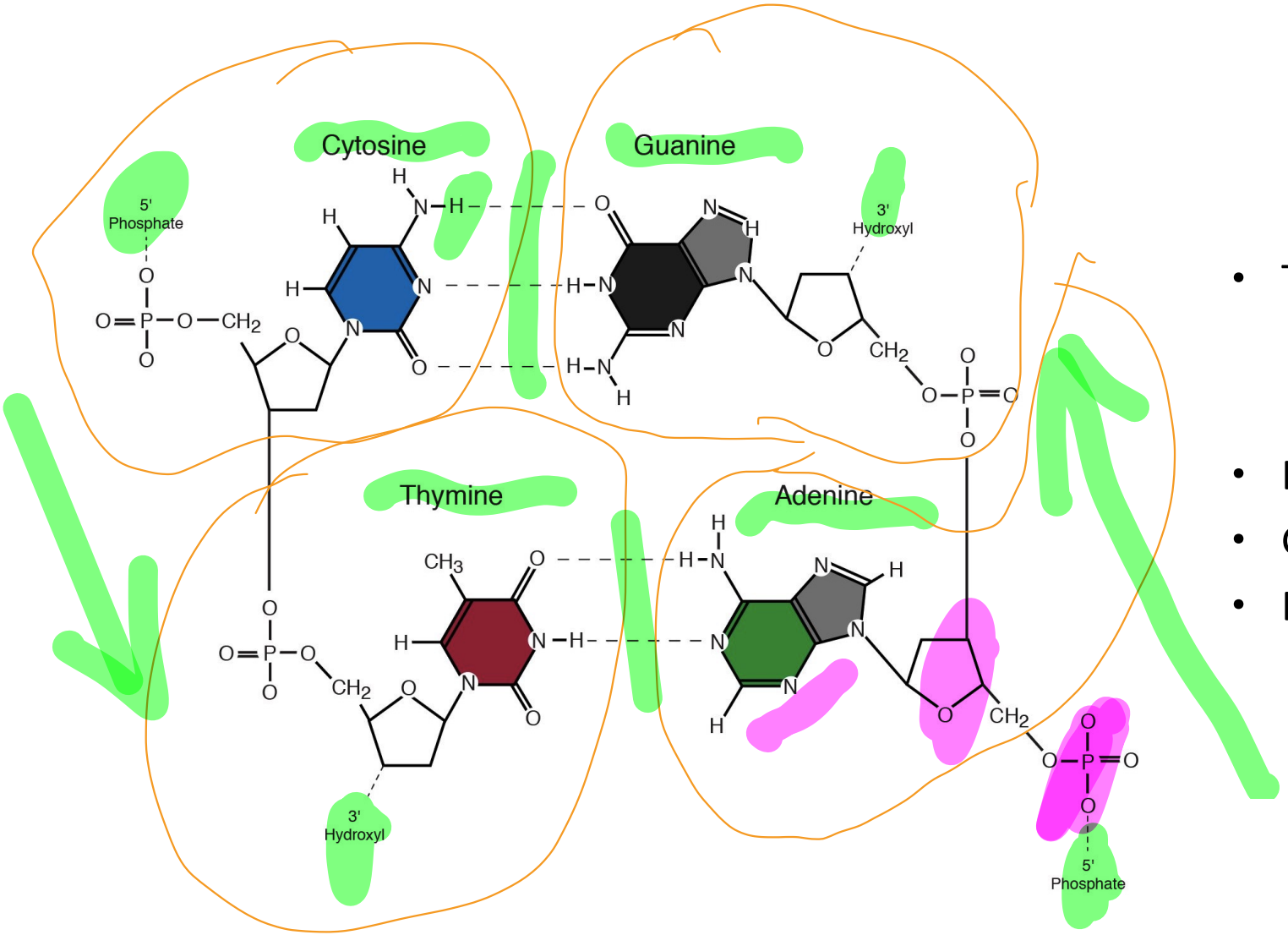
mRNA protein

transcription

translation

A, C, G, T

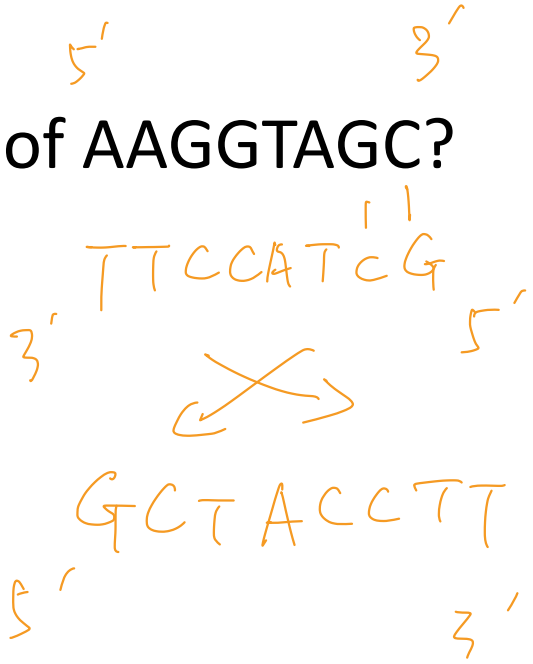
Nucleotide and Base Pairs



- Two classes of necleutide bases:
  - Purine: A and G
  - Pyrimidine: T and C
- Base pairs are due to hydrogen bonds.
- G-C bind stronger because of 3 H-bonds.
- DNA moleculre is oriented (5’ -> 3’).

Reverse Complement a DNA Sequence

- DNA is double-helical, with two complementary strands.
- Complementary bases:
  - Adenine (A) - Thymine (T)
  - Guanine (G) - Cytosine (C)
- Example: What is the reverse complement of AAGGTAGC?

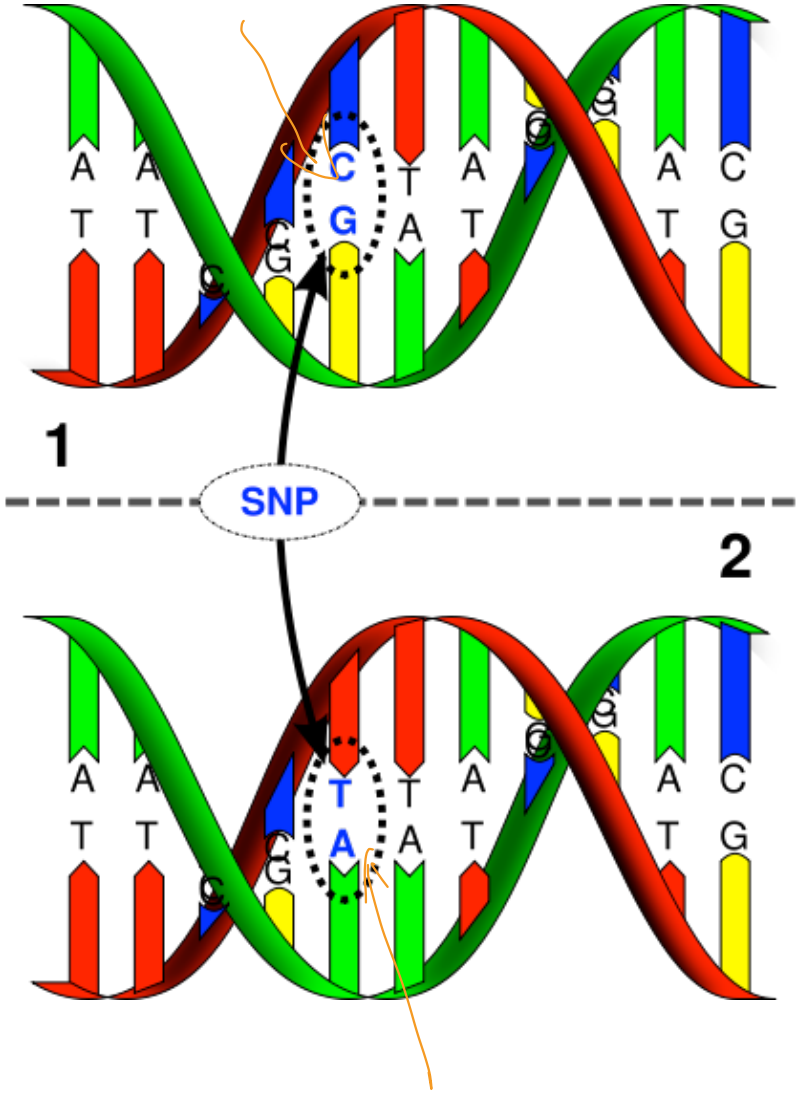


DNA Mutation

- DNA mutates with a small probability when inherited by the offspring.
  - For example, one base can be substituted by another.
  - This creates different alleles of the same gene.
  - An **allele** is a variant form of a gene at the same location of the genome among different individuals.
- Also, one only inherits half of each parent's genome.
- These together cause the differences between individuals of the same species.

Single Nucleotide Polymorphisms

SNP



- Single base variation between members of a species.
- For Human, 90% of all human genetic variation is caused by SNPs. SNPs occur every 100 to 300 bases along the 3-billion-base human genome.
- Major risk for genetic disease.

Compare DNA sequences

- The most often used distance on strings in computer science is Hamming distance.
  - **AGTTTAATCA**
  - **| | | | | | |**
  - **AGTATAACGA**
- This makes some sense on comparing DNA sequences in some cases. But there are other mutations
  - Substitution ACAGT → ACGGT
  - Insertion/deletion (indel) ACAGT → ACGT
- Other DNA rearrangements can also happen. But substitutions and indel are the two mutations we concern the most for this course.

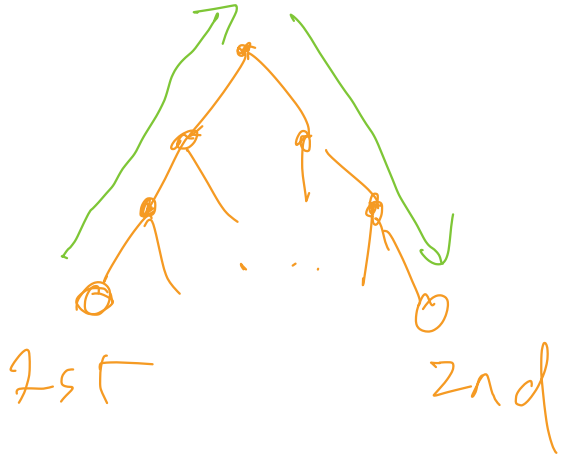
Edit Distance

- Let's focus on substitution and indel only. How "far" away are two sequences from each other?
- E.g. CGATA and GGATTA

- **Edit distance**: the minimum number of edit operations needed to convert one to another. Here edit operations include substitutions and indels.

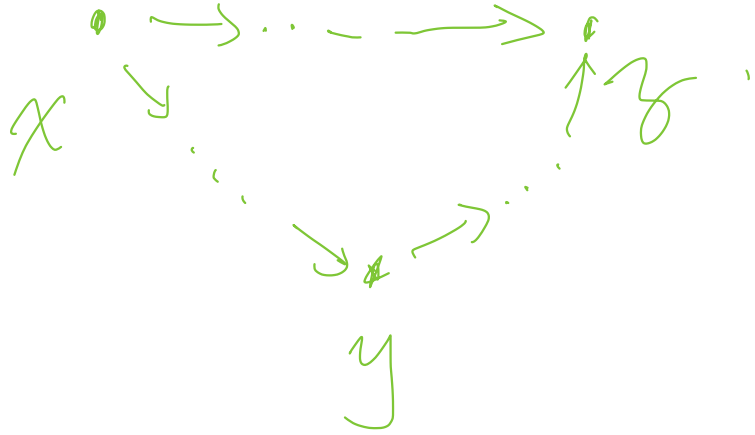
```
d(ATGCATTTA,ATGTACTTTC)
   ATGCATTTA
   ATGTACTTTC
```

3



Edit distance is a distance metric

- Identity: d(x,y)=0 iff x=y
- Symmetry: d(x,y) = d(y,x)
- Triangular Inequality: d(x,z) <= d(x,y) + d(y,z)

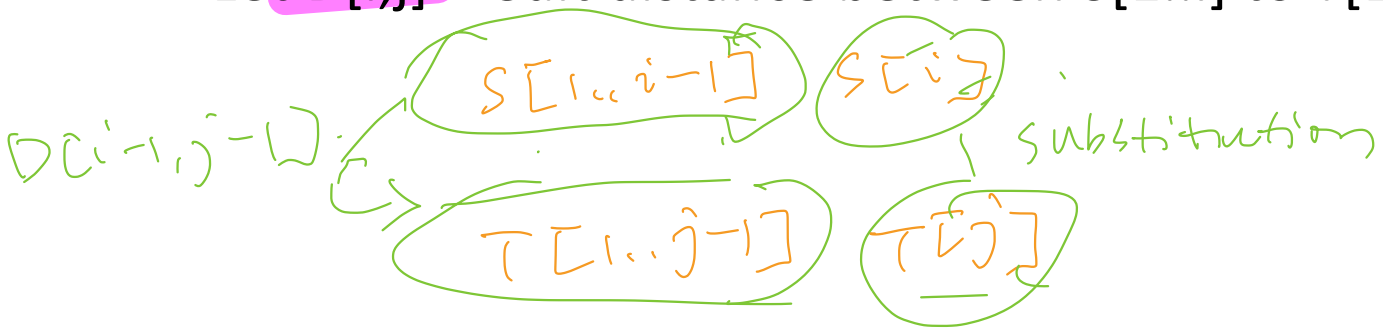


Preparation for the Algorithm

- For convenience of the proof, we treat each occurrence of the same letter different.
- E.g. ATAA -> ATA can be done by either deleting the 2nd or 3rd letter A from the first string. These are different editing paths.
- This does not affect our definition of edit distance, but makes our later proof more precise.

Dynamic Programming Algorithm for Edit distance

- Let D[i,j] = edit distance between S[1..i] to T[1..j].



- Consider the edit operations associated with S[i] and T[j] in the optimal edit operations. One of the following cases will happen (why?):
  1. S[i] is deleted
  2. T[j] is inserted
  3. S[i] becomes T[j]

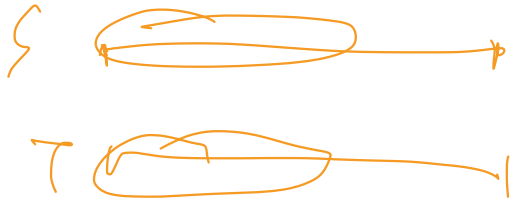
Recurrence Relation

- $D[i,j] = \min \begin{cases} D[i-1,j]+1 \\ D[i,j-1]+1 \\ D[i-1,j-1]+\delta(S[i],T[j]) \end{cases}$
- Here $\delta(S[i],T[j]) = 0$ if S[i]=T[j] and 1 if not.

Dynamic Programming Algorithm for Edit distance

- D[0,0] = 0.
- D[0, i] = i for i=1..|S|
- D[i, 0] = i for i=1..|T|
- for i from 1..|S|
- for j from 1..|T|
- D[i,j] = min {D[i-1,j]+1, D[i,j-1]+1, D[i-1, j-1]+d(S[i], T[j])}.
- Return D[|S|, |T|]

A Note about Dynamic Programming



- Define “subproblems”
- Develop recurrence relation to compute subproblems
- Initialization (base cases)
- Determine the computation order for solving the subproblems.
  - Usually bottom-up (smaller to larger)
- Find the solution of the original input

Longest Common Subsequence

- The second way to evaluate the similarity of two sequences is through LCS.
- A subsequence is obtained by deleting some of the letters from the supersequence and concatenating the remaining letters together.
- What is the LCS of the following two sequences?
  - ATGCATTTA
  - ATGTACTTTC
- LCS can be computed with dynamic programming as well. (Exercise)

Alignment

- The third way to compare to sequences is through sequence alignment.
- Align the two sequences by inserting spaces, so that they are the most **similar** column-wisely.
  - ATGCA-TTTA
    ||| | || |
    ATGTACTT-A
- What does “similar” mean? Usually we need a “scoring function” or a “score function”.
- Let’s define the alignment score to be **the total of column scores**. And each column is assigned by a constant score depending on matching conditions.
- E.g. Match = 1, mismatch =-1, indel =-1. This is sometimes called the “score scheme”.

Two Example Alignments

- AATGCGA-TTTT
  || | |||
  G-TG--ACTTTC

- AATG-CGATTTT
  || | ||
  G-TGAC-TTTC-

- +1 = match
- -1 = mismatch
- -1 = indel

- Which of the two alignments better?