Bioinformatics I

Tutorial Examples

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

1

Example 1

1. Transcribe the following DNA to RNA, then use the genetic code to translate it to a sequence of amino acids.

   TCATAATACGTTTTGTATTCGCCAGCG CTTCGGTGT

2

Solution 1...

- To transcribe the DNA, first substitute each DNA for it's counterpart (i.e., G for C, C for G, T for A and A for T):
- TCATAATACGTTTTGTATTCGCCAGCGCTTCGGTGT
- AGTATTATGCAAAACATAAGCGGTCGCGAAGCCACA
- Next, remember that the Thymine (T) bases become a Uracil (U). Hence our sequence becomes:
- AGUAUUAUGCAAAACAUAAGCGGUCGCGAAGCCACA
- Using the genetic code is also easy – just split the RNA sequence into triplets: :
- AGU AUU AUG CAA AAC AUA AGC GGU CGC GAA GCC ACA

3

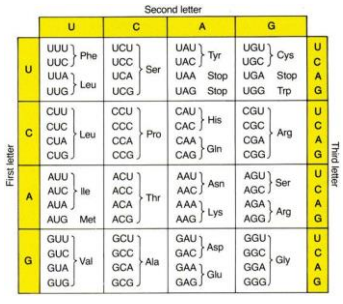
Solution 1...

- then look each triplet (codon) up in the genetic code table. So AGU becomes Serine, which we can write as Ser, or just S. AUU becomes Isoleucine (Ile), which we write as I. Carrying on in this way, we get:
- SIMQNISGREAT

- Homework: Write a Perl program that implements DNA translation to amino acit sequence

4

Genetic Code



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

5

Example 2

- Remove the first letter from the sequence given in example 1, and redo the translation. Explain what happened?

- New sequence

  CATAATACGTTTTGTATTCGCCAGCGCT TCGGTGT

6

Solution 2...

- To transcribe the DNA, first substitute each DNA for it's counterpart (i.e., G for C, C for G, T for A and A for T):
- CATAATACGTTTTGTATTCGCCAGCGCTTCGGTGT
- GTATTATGCAAAACATAAGCGGTCGCGAAGCCACA
- Next, remember that the Thymine (T) bases become a Uracil (U). Hence our sequence becomes:
- GUAUUAUGCAAAACAUAAGCGGUCGCGAAGCCACA
- Using the genetic code is also easy – just split the RNA sequence into triplets: :
- GUA UUA UGC AAA ACA UAA GCG GUC GCG AAG CCA CA

7

…Solution 2...

- Removing the first letter and splitting into codons again gives us:

  GUA UUA UGC AAA ACA UAA GCG GUC GCG AAG CCA CA

- GUA translates to Val (V), UUA translates to Leu (L), UGC translates to Cys (C), AAA translates to Lys (K), ACA translates to Thr (T), and UAA translates to STOP.
- This gives us the sequence:

  VLCKT STOP

- Continuing with the translation, we get:

  AVAKP

8

…Solution 2

- So, if the above DNA sequence from which the RNA was transcribe was actually a gene, it effective length would have been halved, in addition to all of the amino acids changing in the residue sequence it generated.
- Given that the protein structure is largely dictated by it's shape, and it's shape is largely dictated by the residue sequence, we see that it is not surprising that a random mutation such as a deletion will cause harm, or even death to an organism.

9

Example 3

- What is the Hamming distance and Levenshtein (or edit) distance between these two strings?

  BIOINFORMATICS_IS_THE
  BEST_FOR_STRUCTURE_PREDICTION

10

Solution 3

- To calculate the Hamming distance, just count the number of pairs of letters in the alignment which are not the same ignoring indels.

B	I	O	I	N	F	O	R	M	A	T	I	C	S	_	I	S	_	T	H	E	_	_	_	_	_	_	_	_
	*	*	*						*		*	*	*		*	*			*	*								
B	E	S	T	_	F	O	R	_	S	T	R	U	C	T	U	R	E	_	P	R	E	D	I	C	T	I	O	N

- 11
- To calculate the Levenshtein distance, just count the number of pairs of letters in the alignment which are not the same including indels.

B	I	O	I	N	F	O	R	M	A	T	I	C	S	_	I	S	_	T	H	E	_	_	_	_	_	_	_	_
	*	*	*	*				*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
B	E	S	T	_	F	O	R	_	S	T	R	U	C	T	U	R	E	_	P	R	E	D	I	C	T	I	O	N

- 24

11

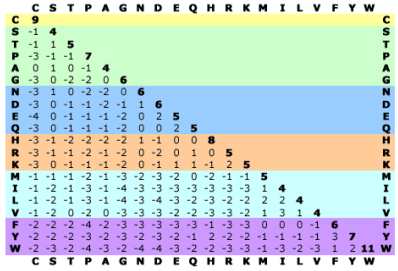
Example 4

- Using the BLOSUM62 substitution matrix, what is the best alignment of these two sequences? (Slide one over the other, and score –1 for end gaps, i.e., letters hanging over either ends).

  FYGNYK
  DGSFNW

12

BLOSUM62 Substitution Matrix



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Solution 4…

- To work out the best alignment,
  - write down all the ways to overlap these sequences and work out the BLOSUM scores for each alignment,
    - remembering to take off 1 for every gap (-).
- It is possible to use a heuristic approach, and have a look to see if there are any obviously good overlaps.
  - If we score these first, then it may become obvious that all the others will not give us a good score.

…Solution 4

	score		score
FYGNYK----- -----DGSFNW	-11	-FYGNYK DGSFNW-	-2
FYGNYK---- ----DGSFNW	-13	--FYGNYK DGSFNW--	-7
FYGNYK--- ---DGSFNW	-8	---FYGNYK DGSFNW---	-4
FYGNYK-- --DGSFNW	-10	----FYGNYK DGSFNW----	-9
FYGNYK- -DGSFNW	5	-----FYGNYK DGSFNW-----	-9
FYGNYK DGSFNW	-14		

All possible overlaps are given in the two boxes with their scores.

The best overlap is therefore the only one scoring a positive number (5)

FYGNYK -
- DGSFNW

Example 5

- Draw a dotplot for these two sequences:

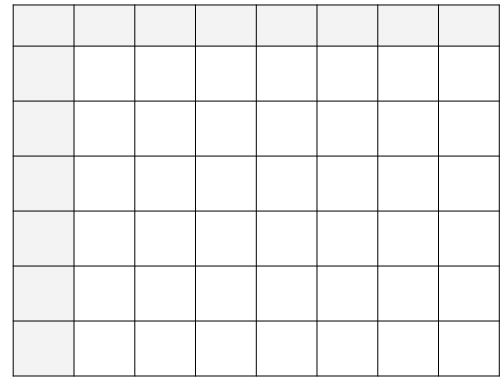
  DILVDEQ

  IVQDEQ
- Then find a likely global alignment for these two sequences.
- Show on the dotplot how you produced this alignment

Solution 5…

- To produce the dotplot,
  - draw a matrix with the first sequence going along the top,
  - and the second sequence going down the left hand side.
- Then mark with a dot all the entries in the matrix
  - where the letter along the top and the letter down the side are equal.
- This gives you:

…Solution 5…



…Solution 5…

	D	I	L	V	D	E	Q
I							
V							
Q							
D							
E							
Q							

19

…Solution 5…

	D	I	L	V	D	E	Q
I		●					
V							
Q							
D							
E							
Q							

20

…Solution 5…

	D	I	L	V	D	E	Q
I		●					
V				●			
Q							
D							
E							
Q							

21

…Solution 5…

	D	I	L	V	D	E	Q
I		●					
V				●			
Q							●
D							
E							
Q							

22

…Solution 5…

	D	I	L	V	D	E	Q
I		●					
V				●			
Q							●
D	●				●		
E							
Q							

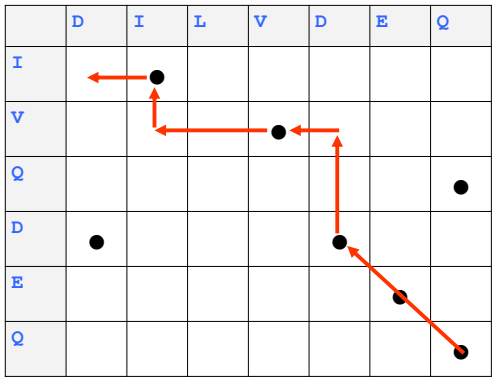
23

…Solution 5…

	D	I	L	V	D	E	Q
I		●					
V				●			
Q							●
D	●				●		
E						●	
Q							

24

…Solution 5…



25

…Solution 5…

D I L V - D E Q

- I - V Q D E Q

26

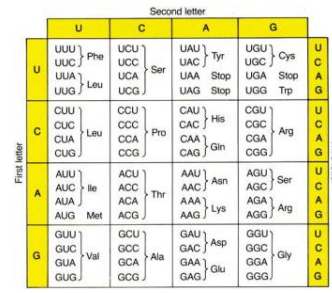
Example 6

An amino acid sequence is given as **DIK**.

Determine the possible DNA sequences which results in the synthesis of the given amino acid sequence. (Use the genetic code table)

27

Genetic Code

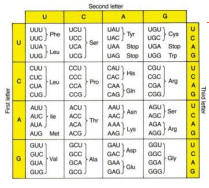


A=Ala=Alanine
C=Cys=Cysteine
D=Asp=Aspartic acid
E=Glu=Glutamic acid
F=Phe=Phenylalanine
G=Gly=Glycine
H=His=Histidine
I=Ile=Isoleucine
K=Lys=Lysine
L=Leu=Leucine
M=Met=Methionine
N=Asn=Asparagine
P=Pro=Proline
Q=Gln=Glutamine
R=Arg=Arginine
S=Ser=Serine
T=Thr=Threonine
V=Val=Valine
W=Trp=Tryptophan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

28

Solution 6…

- To determine possible DNA sequences, we need to apply phases of central dogma of MB in reverse order.



– Looking at the genetic code table, Amino acids in DIK sequence can be translated from the following RNA triplets.

Abbreviation	Amino Acid	Possible triplets
D	Aspartic Acid	GAU, GAC
I	Isoleucine	AUU, AUC, AUA
K	Lysine	AAA, AAG

29

…Solution 6…

- We have 12 possibilities to obtain the DIK sequence.
  - We can visualize them in the following table.

*											
GAU						GAC					
AUU		AUC		AUA		AUU		AUC		AUA	
AAA	AAG	AAA	AAG	AAA	AAG	AAA	AAG	AAA	AAG	AAA	AAG

- We, then apply reverse transcription to find possible DNA sequences by backsubstituting each RNA with its DNA counterpart.

30

…Solution 6

- 12 Possible sequences are (RNA → DNA defines reverse transcription operation):

01-) GAU AUU AAA => CTATAATTT
02-) GAU AUU AAG => CTATAATTC
03-) GAU AUC AAA => CTATAGTTT
04-) GAU AUC AAG => CTATAGTTC
05-) GAU AUA AAA => CTATATTTT
06-) GAU AUA AAG => CTATATTTC
07-) GAC AUU AAA => CTGTAATTT
08-) GAC AUU AAG => CTGTAATTC
09-) GAC AUC AAA => CTGTAGTTT
10-) GAC AUC AAG => CTGTAGTTC
11-) GAC AUA AAA => CTGTATTTT
12-) GAC AUA AAG => CTGTATTTC

31

Example 7…

What is the compositional complexity of these residue sequences?

KKKKTRAITERMMMM and TRAITER

- There are over 30 different definitions of complexity in modern science .
- Biopolymers (nucleic acids and proteins) are represented in the form of sequences of symbols from finite alphabets.

32

…Example 7…

- Term compositional complexity is related to the concept of algorithmic complexity in a sense that
  - repetitive sequences over a given finite alphabet A are considered simple
  - nonrepetitive sequences over a given finite alphabet A are considered complex
- Random (i.e. patternless) sequences are considered maximally complex

33

…Example 7

- The numerical value of compositional complexity of a string of symbols depends on both the choice of alphabet and the frequencies with which specific letters are used.
- Applications of compositional complexity to sequence analysis include
  - functionally or structurally relevant segmenting of nucleotide and protein sequences,
  - genome sequence annotation,
  - finding new functionally relevant properties through studies of large collections of functionally equivalent sequence data.

34

Solution 7…

- Formula for compositional complexity (for protein sequences) is the following:

$$K = \frac{1}{L} log_{20} \left(\frac{L!}{\prod_{i=1}^{20} n_i!} \right)$$

- Note that L is the sequence length and the n_i's are the number of occurrences of the letters of the alphabet that can occur in the sequence.
- As our sequence is a residue sequence, there can only be twenty different letters in the sequence.

35

…Solution 7…

We'll work out the complexity of the longer sequence first.

To calculate the compositional complexity using this formula, we need to work out the values we will be putting into it.

Firstly, we need length, L, of the sequence, which is 15.

Next, we need the number of occurrences of each letter in the sequences.

There are 4 Ks, 2 Ts, 2 Rs, 1 A, 1 I, 1 E and 4 Ms.

So we can write:

$n_K = 4$, $n_T = 2$, $n_R = 2$, $n_A = 1$, $n_I = 1$, $n_E = 1$ and $n_M = 4$

36

…Solution 7…

Now we need to multiply together all the factorials of these numbers.

0! = 1, so we don't need to worry about the letters which aren't there, as we will just be multiplying by 1.

Hence, we need to calculate:

4! * 2! * 2! * 1! * 1! *1! *4!

= 24 * 2 * 2 * 1 *1 * 1 * 24

= 2304

…Solution 7…

We now divide $L!$ by this number:

15!/2304 = 567567000,

and take log to the base 20 of this big number: $\log_{20}(567567000) = 6.729$.

To do this calculation with your calculator, you may need to remember that:

$\log_x(y) = \ln(y)/\ln(x)$, where ln(y) is the natural log of y, and your calculator should handle this.

We finish by dividing our value by the length of the sequence, 15.

So finally, our answer is:

6.729/15 = 0.449.

…Solution 7

Do the same calculations for TRAITER:

L=7

$n_T = 2$, $n_R = 2$, $n_A = 1$, $n_I = 1$, $n_E = 1$

$$K = \frac{1}{L} log_{20}\left(\frac{L!}{\prod_{i=1}^{20} n_i!}\right)$$

1/7 ($\log_{20}$(7!/(2!*2!)))

= 1/7 (log20(7!/4))

= 1/7 (log20(1260))

= 1/7 (2.383)

= 0.340.

Hence we see that the second sequence is less complex than the first (0.340 < 0.449).

Example 8

a. Construct the genetic distance matrix for these four sequences in an alignment.

(1) H Y Y – A U G W V M L L
(2) H A Y A A U G W U M L M
(3) H U – – A G W W U M A V
(4) A V Y – V V A W W L – A

Use each pair as they appear in the alignment given.

b. Use this matrix to infer a phylogenetic relationship between these genes (there is no algorithm here, just use your eye, and draw a phylogenetic tree).

Solution 8.a…

- The genetic distance between two sequences in an alignment is calculated by first determining the number of aligned pairs of letters where
  - neither is a gap, i.e., a dash, and
  - the two letters are different.
- We then divide this by the number of pairs of aligned letters where neither is a gap.
- Considering sequences (1) and (2), there are 11 such pairs
  - In 3 pairs, the letters are different:
    - (Y,A) (V,U) and (L,M).
  - Hence sequences (1) and (2) have a genetic distance of
    - 3/11 = 0.27.

…Solution 8.a

- Performing similar calculations with (1) and (3); (1) and (4); (2) and (3); (3) and (4) enables us to put these values into the following matrix:

	(1)	(2)	(3)	(4)
(1)				
(2)	0.27			
(3)	0.6	0.5		
(4)	0.8	0.8	0.89	

Solution 8.b…

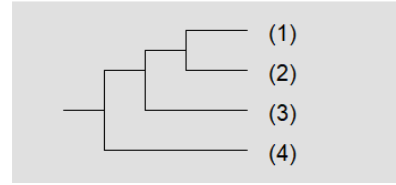
- Looking at the scores in the matrix:

	(1)	(2)	(3)	(4)
(1)				
(2)	0.27			
(3)	0.6	0.5		
(4)	0.8	0.8	0.89	

  - it seems that the sequences (1) and (2) are closely related genetically.
  - sequence (3) is more closely related to (1) and (2) than (4) is related to (1) and (2)
    - Hence, in phylogenetic tree, (3) would follow more closely to (1) and (2) than (4).

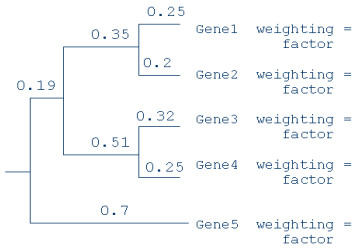
…Solution 8.b

- The tree would therefore be drawn as follows:



Example 9

a. Fill in the weighting factors for Genes 1 to 5 in the given phylogenetic tree below.



b. What order does the above guide tree dictate to the CLUSTAL algorithm for adding sequences/alignments/MSAs to its MSA?

Solution 9.a…

- To work out the weighting factor for a gene:
  - follow the path through the phylogenetic tree from the far left hand side all the way to the gene you are interested in.
  - Whenever you get to a number, add this to a running total, but not before you have divided it by the number of genes you could still get to along the path you are on.
    - For example, for Gene1, starting at the left, we first get to 0.19.
    - At this stage, we could follow the tree to any of Gene1, Gene2, Gene3 or Gene4.
      - Hence we divide 0.19 by 4 (0.19/4 = 0.0475) and this starts our running total.

…Solution 9.a…

  - Carrying on our journey to Gene1, we next come to the number 0.35.
  - At this stage, we could still go to either Gene1 or Gene2,
    - so we divide 0.35 by 2 and add it to our total:
      - 0.0475 + (0.35/2) = 0.2225.
  - Finally, we get to the number 0.25.
  - By this stage, we've arrived at Gene1,
    - so there are no other possibilities.
  - Hence, we add 0.25 to our running total to give us the final weighting factor for Gene1:
    - 0.2225 + 0.25 = 0.4725.

…Solution 9.a

- Following the same routine, we get the following weighting factors:

Gene1 = 0.19/4 + 0.35/2 + 0.25 = 0.4725
Gene2 = 0.19/4 + 0.35/2 + 0.2 = 0.4225
Gene3 = 0.19/4 + 0.51/2 + 0.32 = 0.6225
Gene4 = 0.19/4 + 0.51/2 + 0.25 = 0.5525
Gene5 = 0.7 = 0.7

Solution 9.b…

- To determine the ordering into CLUSTAL:
  - we move from the right hand side of the tree to the left hand side.
  - Whenever the tree joins two paths, this indicates that we should perform an alignment of
    - a pair of sequences,
    - a pair of aligned sequences,
    - a pair of MSAs.
  - The join also indicates which two things are to be aligned.

49

…Solution 9.b…

- So, starting from the right hand side, we move past the first set of numbers.
- At this stage, two pairs of paths get joined.
  - This indicates that we should align the sequences for Gene1 and Gene2 and separately align the sequences for Gene3 and Gene4.
    - It doesn't matter in which order we do the two alignments,
    - but we might as well start with Gene1 and Gene2
      - call this alignment A1,
    - followed by aligning Gene3 and Gene4
      - call this alignment A2.

50

…Solution 9.b…

- Moving on from right to left, we go past the second lot of numbers, and come to another join.
  - This indicates that we're now going to try to align the sequences for Genes 1 to 4.
    - As we've already aligned them in pairs, we actually need to align the alignments A1 and A2 at this stage
      - call this multiple sequence alignment MS1.
- Finally, we reach a join where the path from Gene5 meets the paths from Genes 1 to 4.
  - Hence this indicates that we perform an alignment of MS1 with the single sequence for Gene5 at the final stage.

51

…Solution 9.b

- Therefore, we perform the alignments in the following order:

(i)	(Gene1+Gene2)	= A1
(ii)	(Gene3+Gene4)	= A2
(iii)	(A1+A2)	= MS1
(iv)	(MS1+Gene5)	= MSA

- The final alignment gives us the MSA we were looking for.

52

Example 10

- Consider four species characterized by homologous sequences ATCC, ATGC, TTCG, and TCGG.
- Taking the number of differences as the measure of dissimilarity between each pair of species, use a simple clustering procedure to derive phylogenetic tree.

53

Solution 10…

- First form the distance matrix:

	ATCC	ATGC	TTCG	TCGG
ATCC				
ATGC				
TTCG				
TCGG				

54

…Solution 10...

- The distance matrix:

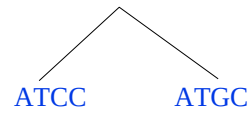
	ATCC	ATGC	TTCG	TCGG
ATCC	0	1	2	4
ATGC	1	0	3	3
TTCG	2	3	0	2
TCGG	4	3	2	0

- Smallest nonzero distance is 1

55

…Solution 10…

- Smallest nonzero distance is 1 (between ATCC and ATGC)
- Therefore first cluster is {ATCC and ATGC}
- The tree will contain the following fragment:



56

…Solution 10…

- Reduced distance matrix is:

	{ATCC,ATGC}	TTCG	TCGG
{ATCC,ATGC}	0	2.5	3.5
TTCG		0	2
TCGG			0

57

…Solution 10…

- Reduced distance matrix is:

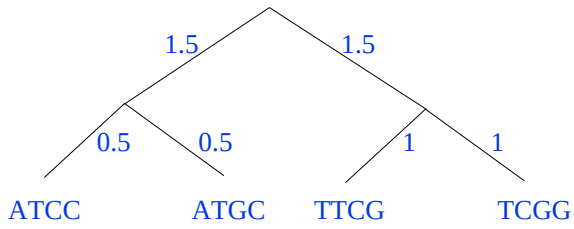
	{ATCC,ATGC}	TTCG	TCGG
{ATCC,ATGC}	0	(2+3)/2	4+3)/2
TTCG		0	2
TCGG			0

- Next cluster is: {TTCG, TCGG}

58

…Solution 10

- Linking the clusters gives the following tree:



59