



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΡΓΑΣΙΑ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

Ακαδημαϊκό έτος 2021-2022

ΑΥΓΕΡΙΝΟΣ ΧΡΗΣΤΟΣ - Π19020

ΒΙΤΑΚΗΣ ΑΘΑΝΑΣΙΟΣ - Π19247

ΠΑΝΑΓΙΩΤΟΠΟΥΛΟΣ ΔΗΜΗΤΡΙΟΣ - Π19130

ΑΛΕΞΑΝΔΡΗΣ ΙΩΑΝΝΗΣ - Π19006

Πίνακας Περιεχομένων

Θέμα 1	3
Άσκηση 7.2	3
Εκφώνηση.....	3
Υλοποίηση	4
Ερώτημα (1).....	4
Ερώτημα (2).....	4
Ερώτημα (3).....	5
Ερώτημα (4).....	6
Ερώτημα (5).....	7
Ερώτημα (6).....	8
Ερώτημα (7).....	12
Θέμα 2	13
Άσκηση 11.4	13
Εκφώνηση.....	13
Σχεδιασμός	13
Υλοποίηση	13
Αποτελέσματα	15
Τρόπος εκτέλεσης	15
Θέμα 3	16
Άσκηση 6.14	16
Εκφώνηση.....	16
Σχεδιασμός	16
Υλοποίηση	16
Νικηφόρα Στρατηγική	16
Αποτελέσματα	17
Τρόπος εκτέλεσης	18
Θέμα 4	19
Υλοποίηση	19
Προετοιμασία.....	19
Ερώτημα 1	20
α)	20
β).....	20
γ)	20

Ερώτημα 2	20
α)	20
β).....	21
γ)	21
δ).....	21
Ερώτημα 3	21
α)	21
β).....	22
Ερώτημα 4	23
α)	23
Ερώτημα 5	24
α)	24
β).....	25
γ)	26
Βιβλιογραφία	27


Θέμα 1

Άσκηση 7.2

Εκφώνηση

Πραγματοποιήστε φυλογενετικές αναλύσεις χρησιμοποιώντας το λογισμικό MEGA.

- (1) Μεταβείτε στη βάση δεδομένων των συντηρημένων δομικών επικρατειών (<http://www.ncbi.nlm.nih.gov/cdd>) στο NCBI.
- (2) Εισαγάγετε τον όρο λιποκαλίνες (lipocalins) ή άλλο όνομα οικογένειας της επιλογής σας. Εναλλακτικά, μπορείτε να ξεκινήσετε από την Ensembl, τη HomoloGene ή την Pfam.
- (3) Επιλέξτε τη μορφή αρχείου mFasta και στη συνέχεια κάντε κλικ στην επιλογή «Reformat». Το αποτέλεσμα είναι μια πολλαπλή στοίχιση αλληλουχιών. Αντιγράψτε το αποτέλεσμα σε έναν επεξεργαστή κειμένου (π.χ. Notepad++) και απλοποιήστε τα ονόματα των αλληλουχιών.
- (4) Εισαγάγετε το αρχείο (ή επικολλήστε τις αλληλουχίες) στο MEGA, όπως φαίνεται στην Εικόνα 7.9. Στοιχίστε τις αλληλουχίες και αποθηκεύστε τις σε μορφές αρχείων .mas και .meg.
- (5) Επιλέξτε Phylogeny > Construct/Test (Φυλογένεση > Κατασκευή/Δοκιμή) για να δημιουργήσετε δέντρα με τις μεθόδους ένωσης γειτόνων, μέγιστης πιθανοφάνειας ή άλλες.
- (6) Για κάθε δέντρο που δημιουργείτε, διαβάστε την αντίστοιχη λεζάντα. Δοκιμάστε τα εργαλεία δέντρων (π.χ. τοποθετήστε μια ρίζα, αναστρέψτε κόμβους, εμφανίστε ή αποκρύψτε τα μήκη των κλάδων, αλλάξτε μορφές απεικόνισης).
- (7) Πραγματοποιήστε bootstrapping. Προσδιορίστε τις συστάδες κλάδων που έχουν χαμηλά επίπεδα στήριξης. Γιατί συμβαίνει αυτό;


National Library of Medicine
National Center for Biotechnology Information

Log in

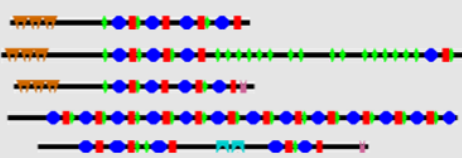
Conserved Domains

Conserved Domains

Advanced

Search

Help



CDD

The Conserved Domain Database is a resource for the annotation of functional units in proteins. Its collection of domain models includes a set curated by NCBI, which utilizes 3D structure to provide insights into sequence/structure/function relationships.

Using CDD

- [Quick Start Guide](#)
- [How To Guides](#)
- [Help](#)
- [FTP](#)
- [News](#)
- [Publications](#)

CDD Tools

- [Overview of CDD Resources](#)
- [CD-Search](#)
- [Batch CD-Search](#)
- [CDART \(domain architectures\)](#)
- [SPARCLE \(protein labeling engine\)](#)
- [BLAST](#)

Other Resources

- [Structure Group Home Page](#)
- [Entrez Structure \(Molecular Modeling Database\)](#)
- [Entrez Gene](#)
- [Entrez Protein](#)

Conserved Domains

Conserved Domains

lipocalins

Create alert

Advanced

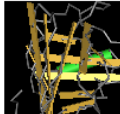
Search

Help

Summary ▾ 20 per page ▾ Sort by Default order ▾

Search results

Items: 1 to 20 of 57


☐


Lipocalin_FABP: lipocalin/cytosolic fatty acid-binding protein family

Lipocalins are diverse, mainly low molecular weight extracellular proteins that bind principally sma...

Accession: cl10502 ID: 415860

[View in Cn3D](#) [Protein](#) [Superfamily Members](#) [PubMed](#)

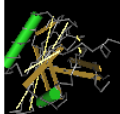
☐


lipocalin_LTBP1-like: Triatominae salivary lipocalins such as Rhodnius prolixus LTBP1 and Meccus pallidipennis triabin and similar proteins

This subfamily includes various insect proteins found in the saliva of Triatominae (kissing bugs), i...

Accession: cd19423 ID: 381198

[View in Cn3D](#) [Specific Protein](#) [Protein](#) [Superfamily](#) [Superfamily Members](#) [PubMed](#)


☐


Lipocalin: Lipocalin / cytosolic fatty-acid binding protein family

Lipocalins are transporters for small hydrophobic molecules, such as lipids, steroid hormones, bilin...

Accession: cl21528 ID: 419712

[View in Cn3D](#) [Protein](#) [Superfamily Members](#) [PubMed](#)

☐


Pallilysin: Pallilysin beta barrel domain

The *Treponema pallidum* protein, Tp0751 (also known as pallilysin), possesses adhesive properties and...

Accession: cl39980 ID: 423742

Send to: ▾

Filter your results:

- All (57)
- [NCBI-curated \(46\)](#)
- [families \(53\)](#)
- [superfamilies \(7\)](#)
- [imported \(7\)](#)

[Manage Filters](#)

Find related data

Database: Select

[Find items](#)

Search details

lipocalins[All Fields]

[Search](#) [See more...](#)

Recent activity

[See more...](#)

Ερώτημα (3)

Curated CD Hierarchy [?](#)

- cd00301 lipocalin_FABP
 - cd00742 FABP
 - cd19441 CRABP
 - cd19460 CRABP1
 - cd19461 CRABP2
 - cd19442 CRBP
 - cd19462 CRBP1
 - cd19463 CRBP2
 - cd19464 CRBP3
 - cd19465 CRBP4
 - cd19443 FABP3-like
 - cd19466 FABP3
 - cd19467 FABP4
 - cd19468 FABP5
 - cd19469 FABP8

Imported CD [?](#)

COG5514

pfam08768



Sequence Alignment

Reformat

Format: mFasta

Row Display: up to 10

Color Bits: 2.0 bit

Type Selection: the most d



```

>gi|609411913|pdb|3WJB|B
--mwshpqfeknqlqqlqnpgeppvhpFVAPLSYLLGTWRGQGEGEYPTIPSFYRGEIIRFSH-SGKPVIAYTQKTWKL
ESgA-PLLAESGYFRprPDGS--IEVVIACSTGLVEVQKGTYNv--dEqSIKLKS---DLV---GNASKVKEISREFELv
DGK---LSYVVRLSTTTNPLQLKAILDKl-----
>gi|504799881|ref|WP_014986983.1|
-----mvesqpiaphpdIAPLAALLGTWRGNHGHEYPTIQPFDYLEEVRFGH-LGRPFLTYRQRTRAA
DD-GrPMHAETGYLR--CPRPdrVELILAHPTGITEICEGALTvddgAlhLEFDS---TSIgrsSTAKLVLTALGRTFQV-
KGDt--IDYTVRMAAVGEPLQHHLAATLIrae-----
>gi|218551765|sp|A1T297.1|Y449_MYCVP
-----madvpalhpdVAALAPLLGTWVGEGSGEYPTIEPFGYTEEITFGH-VGKPFLTYAQRTRAA
DD-GrPLHAETGYLR--ASAPdrIEWILAHPTGITEIQEGQLTadgdGlrMELVS---SSIgrsGSAKEVTDVGRSIEL-
RGDt--LTYTLRMAAVGQPLQHHLsAVLRrvr-----
>gi|333487064|gb|AEF36456.1|
-----mELAPLLGTWSGRGRGVYPTIASFDYLEEVTFSH-VGKPFLVYGQKTKSA
AD-GIPLHAETGYLR--VPQPgrIEWVLAHPSGITEIEVGSYRvtadGieLEMSA---PTIglPTAKEVTALSRRYRL-
ARDe--LSYTLDMGAVGEPANHLTAALRrtg-----
>gi|218551734|sp|B2HLY1.1|Y1995_MYCMM
-----mpadlhpdlDALAPLLGTWAGQGSGEYPTIEPFEYLEEIVVFSH-VGKPFLVYAQKTRAV
AD-GaPLHAETGYLR--VPKPgqVELVLAHPSGITEIEVGTSasggVieMEMVT---TAIgmtPTAKEVTALSRsFRM-
VGDe--LSYRLRMGAVGLPLQHHLGARLRrks-----
>gi|81413567|sp|Q73W27.1|Y2833_MYCPA
-----mptdlhpdlAALAPLLGTWTRGSGKYPTIQPFDYLEEVTFSH-VGKPFLAYAQKTRAA
AD-GkPLHAETGYLR--VPQPgrLELVLAHPSGITEIEVGSYAvttggLieMRMST---TSIglTSAKEVTALARWFRI-
DGDe--LSYSVQMGAVGQPLQDHLAAVLHrqr-----
>gi|88193107|pdb|2FR2|A
-----mtrdlapaLQALSPLLGSWAGRGAGKYPTIRPFEYLEEIVVFAH-VGKPFLTYTQQTRAV
AD-GkPLHSETGYLR--VCRPgCVELVLAHPSGITEIEVGTSvtgdVieLELSTradGSiglaPTAKEVTALDRSYRI-
DGDe--LSYSLQMRVAVGQPLQDHLAAVLHrqrshhhhhh
>gi|81537071|sp|Q9CCB8.1|Y1006_MYCLE
-----mpsdlcpdLQALAPLLGSWVGRGMGKYPTIQPFYELEEIVVFSH-LDRPFLTYTQKTRAI
TD-GkPLHAETGYLR--VPQPghIELVLAHSDIAEIEVGTSvtgdLieVELVT---TTIglvPTAKQVTALGRFFRI-

```

Μεταφορά σε επεξεργαστή κειμένου και απλοποίηση των ονομάτων αλληλουχιών.

```
>3WJB
--mwshpqfeknqlqqlqnpgesppvhp fVAPLSYLLGTWRGQGEYPTIPSFYRGEEIRFSH-SGKPVIAYTQKTWKL
ESgA-PLLAESGYFRprPDGS--IEVVIACSTGLVEVQKGTYNv--dEqSIKLKS---DLV---GNASKVKEISREFELv
DGK---LSYVVRLSTTTNPLQPLLKAILDkl-----
>WP_014986983.1
-----mvesqppiaphpdIAPLAALLGTWRGNHGGEYPTIQPFDYLEEVRFGH-LGRPFLTYRQRTRAA
DD-GrPMHAETGYLR--CPRPdrVELILAHPTGITEICEGALTvddgAlhLEFDS---TSigrsSTAKLVTALGRTFQV-
KGDt--IDYTVRMAAVGEPLQHHLAATLirae-----
>A1T297.1|Y449_MYCVP
-----madsvpalhpdVAALAPLLGTWVGESEYPTIEPFGYTEETFGH-VGKPFLTYAQRTAA
DD-GrPLHAETGYLR--ASAPdrIEWILAHPTGITEIQEQQLTadgdGlrMELVS---SSigrsGSAKEVTDVGRSIEL-
RGDt--LTYTLRMAAVGQPLQHHLSAVLRrvr-----
>AEF36456.1
-----mELAPLLGTWSGRGRGVYPTIASFDYLEEVTFSH-VGKPFLVYGQTKSA
AD-GlPLHAETGYLR--VPQPgrIEWVLAHPSGITEIEVGSYRvtadGieLEMSA---PTIglapTAKEVTALSRRYRL-
ARDe--LSYTLDMGAVGEPAQNHLTAALRrtg-----
>B2HLY1.1|Y1995_MYCMM
-----mpadlhpdlDALAPLLGTWAGQGSGEYPTIEPFEYLEEVVFSH-VGKPFLVYAQKTRAV
AD-GaPLHAETGYLR--VPKPgcVELVLAHPSGITEIEVGTYsasggVieMEMVT---TAIgmtPTAKEVTALSRSFRM-
VGDe--LSYRLRMGAVGLPLQHHLGARLRrks-----
>Q73W27.1|Y2833_MYCPA
-----mptdlhpdlAALAPLLGTWTGRGSGKYPTIQPFDYLEEVTFSH-VGKPFLAYAQKTRAA
AD-GkPLHAETGYLR--VPQPgrLELVLAHPSGITEIEVGSYAvtgglieMRMST---TSigltPSAKEVTALARWFRI-
DGDe--LSYSVQMGA VGQPLQDHLLAAVLHrqr-----
>2FR2
-----mtrdlapaLQALSPLLGSWAGRGAGKYPTIRPFEYLEEVVFAH-VGKPFLTYTQOTRAV
AD-GkPLHSETGYLR--VCRPgcVELVLAHPSGITEIEVGTYsvtgdvieLELSTradGSiglapTAKEVTALDRSYRI-
DGDe--LSYSLQMRAGVQPLQDHLLAAVLHrqrshhhhhh
>Q9CCB8.1|Y1006_MYCLE
-----mpsdlcpdlQALAPLLGSWVGRGMKYPTIQPFYEYLEEVVFSH-LDRPFLTYTQKTRAI
AD-GkPLHSETGYLR--VPQPgrLELVLAHPSGITEIEVGSYAvtgglieMRMST---TSigltPSAKEVTALARWFRI-
DGDe--LSYSLQMRAGVQPLQDHLLAAVLHrqrshhhhhh
```

Ερώτημα (4)

M11: Alignment Explorer (lipocalin.fas)

Data Edit Search Alignment Web Sequencer Display Help

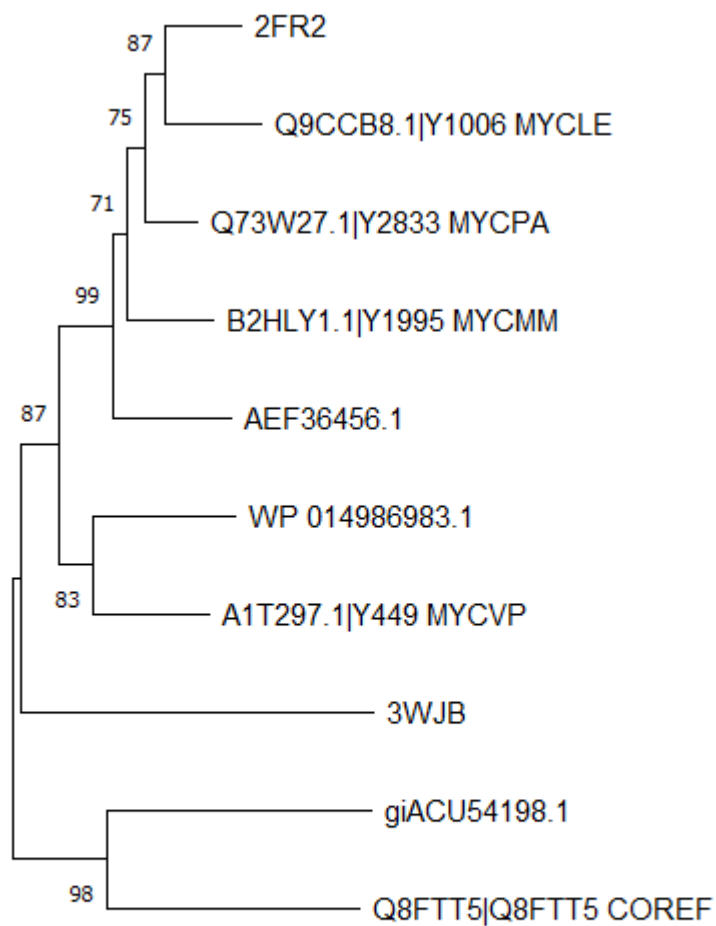
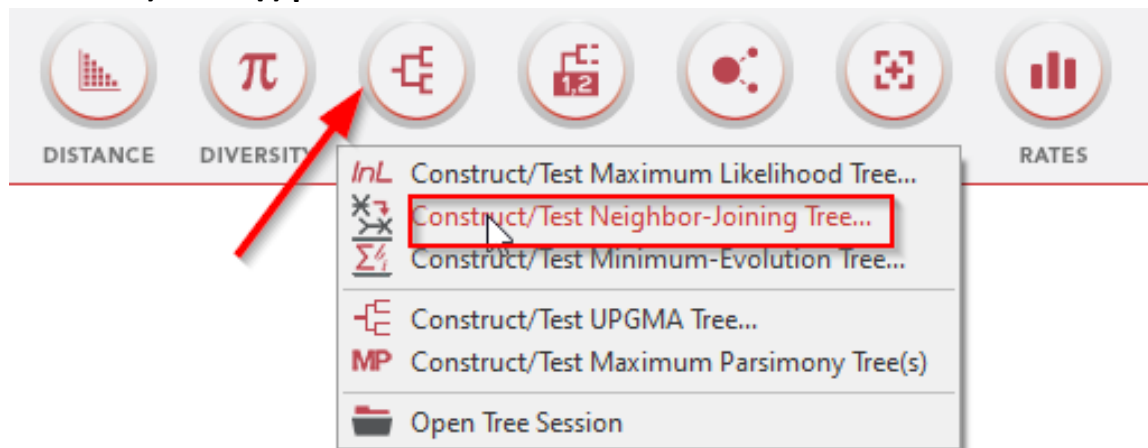
Protein Sequences

Species/Abbrv	Sequence
1. 3WJB	- mwshpqfeknqlqqlqnpgesppvhp fVAPLSYLLGTWRGQGEYPTIPSFYRGEEIRFSH
2. WP_014986983.1	-----mvesqppiaphpdIAPLAALLGTWRGNHGGEYPTIQPFDYLEEVRFGH
3. A1T297.1 Y449_MYCVP	-----madsvpalhpdVAALAPLLGTWVGESEYPTIEPFGYTEETFGH
4. AEF36456.1	-----mELAPLLGTWSGRGRGVYPTIASFDYLEEVTFSH
5. B2HLY1.1 Y1995_MYCMM	-----mpadlhpdlDALAPLLGTWAGQGSGEYPTIEPFEYLEEVVFSH
6. Q73W27.1 Y2833_MYCPA	-----mptdlhpdlAALAPLLGTWTGRGSGKYPTIQPFDYLEEVTFSH
7. 2FR2	-----mtrdlapaLQALSPLLGSWAGRGAGKYPTIRPFEYLEEVVFAH
8. Q9CCB8.1 Y1006_MYCLE	-----mpsdlcpdlQALAPLLGSWVGRGMKYPTIQPFYEYLEEVVFSH
9. giACU54198.1	-----mtireMLEGTWTGSGIGSYPEVAEFSYQERLRFES
10. Q8FTT5 Q8FTT5 COREF	mgarvermlfdarsapiplrstgmdihpniAPYGF LTGTWTGKGHGFYPTIEDFSYEETLNFST

lipocalin.mas

Ερώτημα (5)

Μέθοδος ένωσης γειτόνων



0.10

Ερώτημα (6)

Μέθοδος γένεσης γειτόνων

Λεζάντα

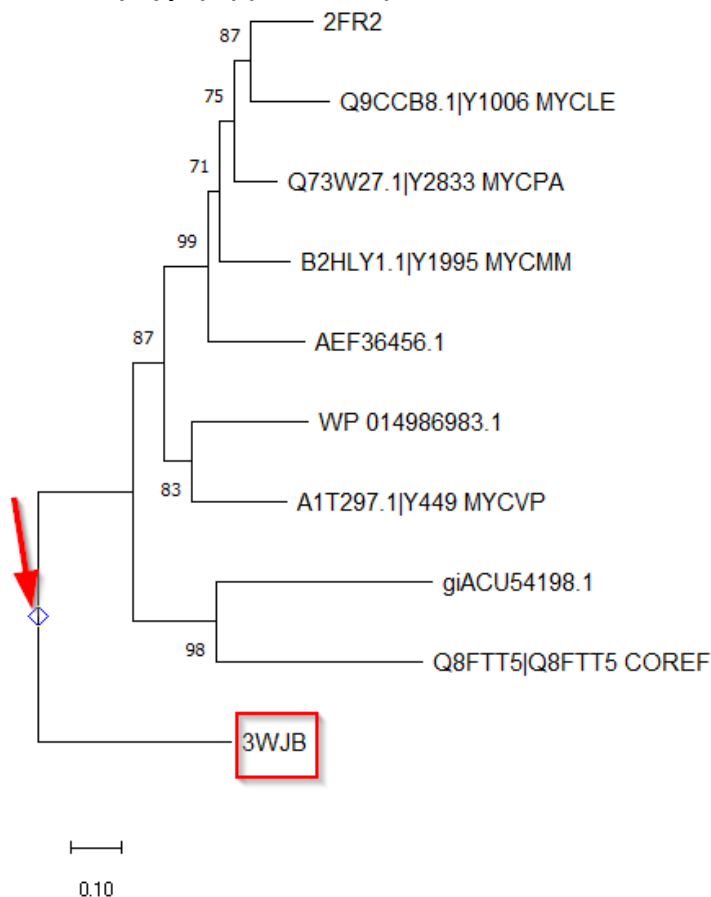
Evolutionary relationships of taxa

The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [3] and are in the units of the number of amino acid substitutions per site. This analysis involved 10 amino acid sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 200 positions in the final dataset. Evolutionary analyses were conducted in MEGA11 [4]

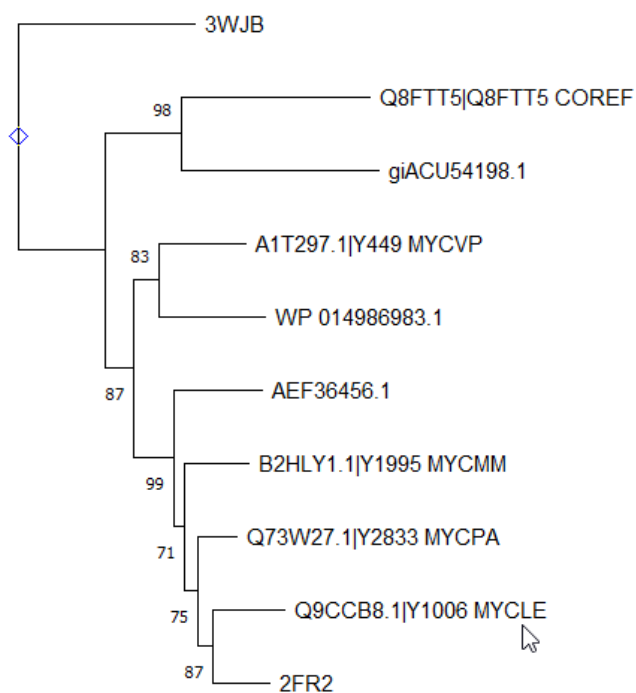
1. Saitou N. and Nei M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.
2. Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.
3. Zuckerkandl E. and Pauling L. (1965). Evolutionary divergence and convergence in proteins. Edited in *Evolving Genes and Proteins* by V. Bryson and H.J. Vogel, pp. 97-166. Academic Press, New York.
4. Tamura K., Stecher G., and Kumar S. (2021). MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* <https://doi.org/10.1093/molbev/msab120>.

Disclaimer: Although utmost care has been taken to ensure the correctness of the caption, the caption text is provided "as is" without any warranty of any kind. Authors advise the user to carefully check the caption prior to its use for any purpose and report any errors or problems to the authors immediately (www.megasoftware.net). In no event shall the authors and their employers be liable for any damages, including but not limited to special, consequential, or other damages. Authors specifically disclaim all other warranties expressed or implied, including but not limited to the determination of suitability of this caption text for a specific purpose, use, or application.

Τοποθέτηση ρίζας (στο 3WJB)

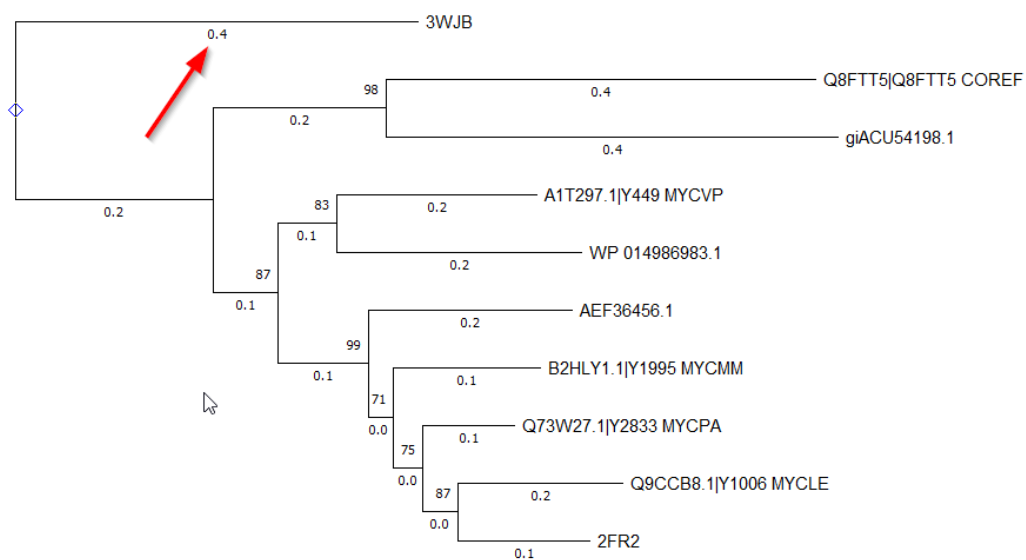


Αναστροφή κόμβων



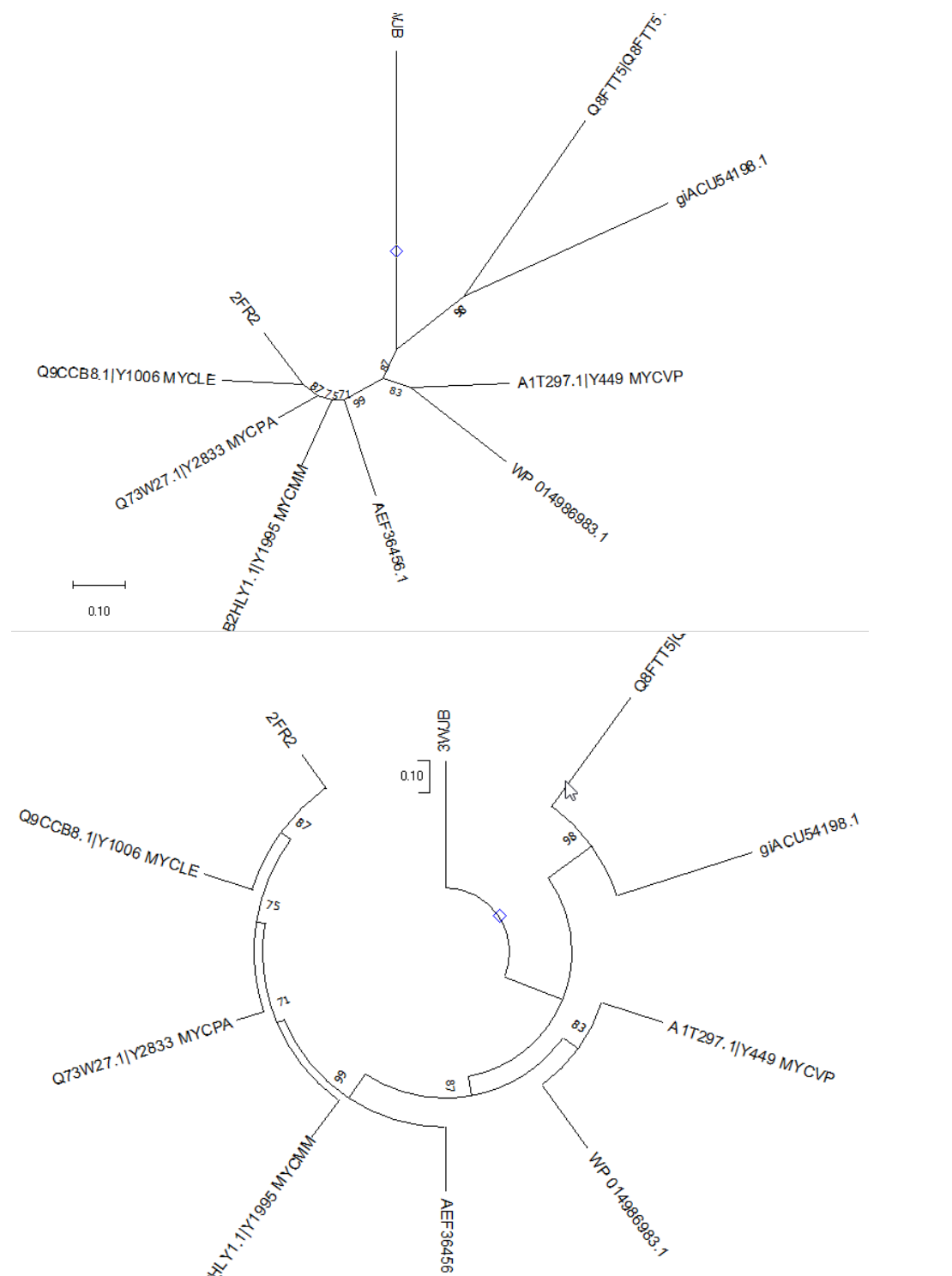
0.10

Εμφάνιση μηκών κλαδών



0.10

Αλλαγή μορφών απεικόνισης(radiation,circle)



Μέθοδος μέγιστης πιθανοφάνειας

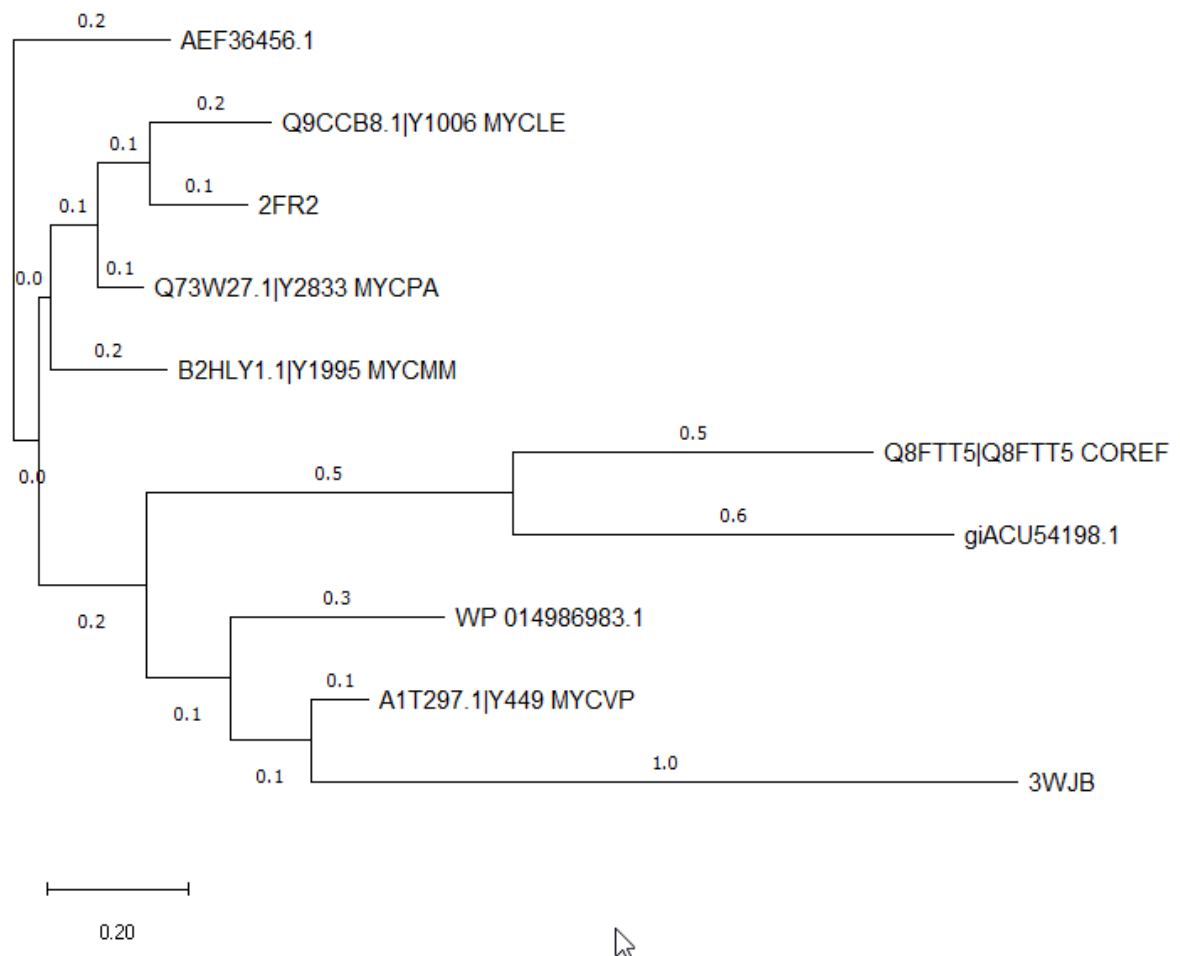
Λεζάντα

Evolutionary analysis by Maximum Likelihood method

The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model [1]. The tree with the highest log likelihood (-2721.70) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (next to the branches). This analysis involved 10 amino acid sequences. There were a total of 200 positions in the final dataset. Evolutionary analyses were conducted in MEGA11 [2]

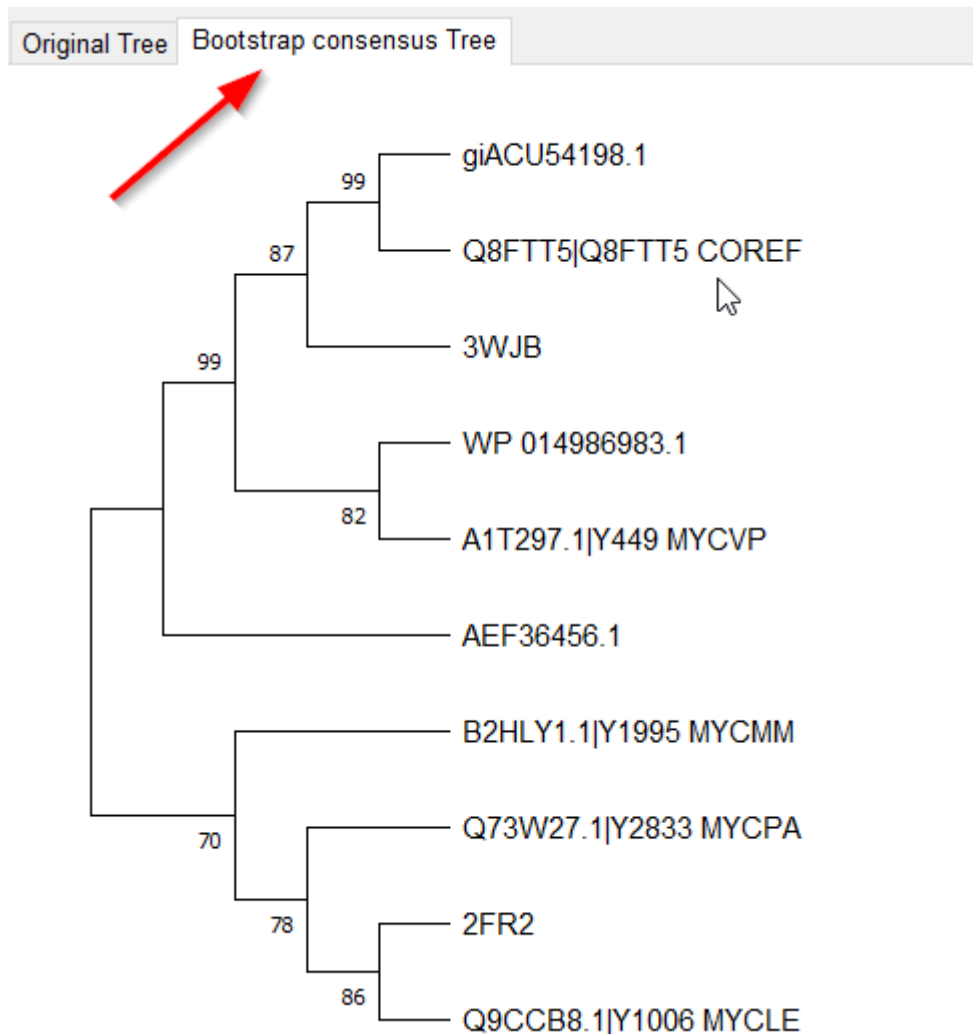
1. Jones D.T., Taylor W.R., and Thornton J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8: 275-282.
2. Tamura K., Stecher G., and Kumar S. (2021). MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* <https://doi.org/10.1093/molbev/mab120>

Δέντρο(με εφαρμογή δόκιμων εργαλείων δέντρων)



Ερώτημα (7)

Bootstrapping (εκτίμηση της εμπιστοσύνης-ακεραιότητας των κλάδων σε ένα φυλογενετικό δέντρο.)



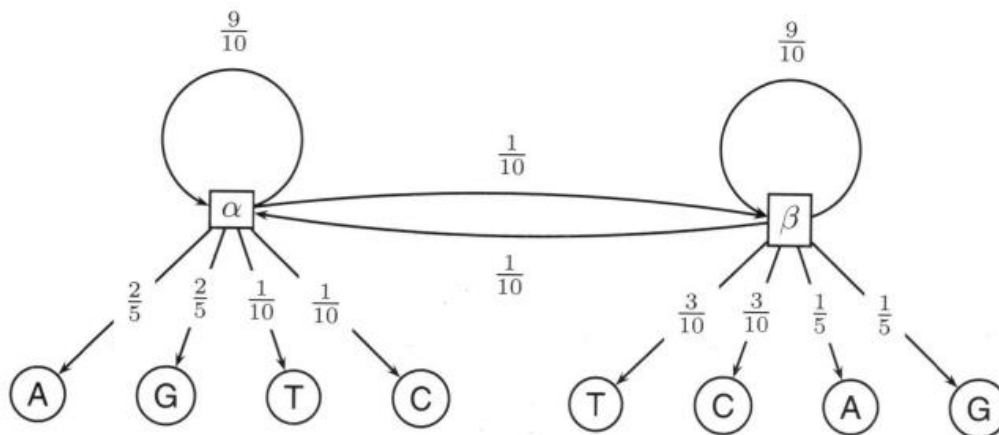
Παρατηρούμε ότι δεν υπάρχουν χαμηλά επίπεδα στήριξης στο συγκεκριμένο φυλογενετικό δέντρο καθώς εκτιμάται ότι τα χαμηλά επίπεδα στήριξης κυμαίνονται στα επίπεδα 50%-70%.Επίσης παρατηρούμε ότι δεν απορρίπτεται κανένα επίπεδο στήριξης καθώς δεν υπάρχουν τιμές κάτω από 50%. Αυτό σημαίνει ότι στο συγκεκριμένο δέντρο έχουμε καλή ευθυγράμμιση και σαφή υποστήριξη για τη συγκεκριμένη σχεδίαση των κλάδων.

Θέμα 2

Άσκηση 11.4

Εκφώνηση

Στο σχήμα 11.7 φαίνεται ένα HMM με δύο καταστάσεις α και β . Όταν το HMM βρίσκεται στην κατάσταση α , έχει μεγαλύτερη πιθανότητα να εκπέμψει πουρίνες (A και G). Όταν βρίσκεται στην κατάσταση β έχει μεγαλύτερη πιθανότητα να εκπέμψει πυριμιδίνες (C και T). Αποκωδικοποιήστε την πιο πιθανή ακολουθία των καταστάσεων (α/β) για την αλληλουχία GGCT. Χρησιμοποιήστε λογαριθμικές βαθμολογίες αντί για κανονικές βαθμολογίες πιθανοτήτων.



Σχήμα 11.7 Το HMM που περιγράφεται στο Πρόβλημα 11.4.

Σχεδιασμός

Για την επίλυση της άσκησης χρησιμοποιούμε τη Python 3.10 και τη βιβλιοθήκη numpy.

Υλοποίηση

Για να αποκωδικοποιήσουμε την ακολουθία των καταστάσεων (α/β) για την αλληλουχία GGCT θα χρησιμοποιήσουμε τον αλγόριθμο του Viterbi ο οποίος στηρίζεται στις αρχές του δυναμικού προγραμματισμού.

Αρχικά θεωρούμε ότι η πιθανότητα για την επιλογή αρχικής κατάστασης είναι $1/2$ για την μετάβαση στην κατάσταση α και αντίστοιχα $1/2$ για την β . Έτσι κατασκευάζουμε τους πίνακες μεταβολής καταστάσεων και εκπομπής συμβόλων όπως βλέπουμε παρακάτω.

	α	β
INITIALS	0.5	0.5

Αρχικές πιθανότητες

	α	β
α	0.9	0.1
β	0.1	0.9

Πίνακας μεταβάσεων

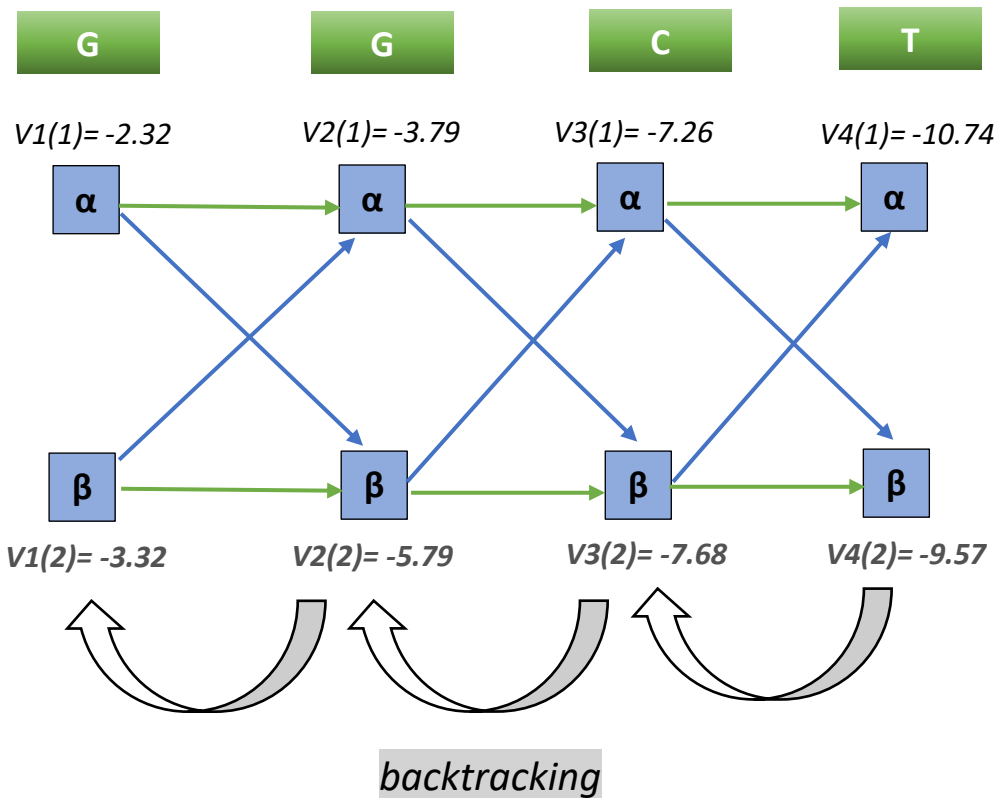
	A	G	T	C
α	0.4	0.4	0.1	0.1
β	0.2	0.2	0.3	0.3

Πίνακας εκπομπών

Στη συνέχεια μετατρέπουμε όλες τις τιμές των πινάκων πιθανοτήτων σε λογαριθμικές τιμές αντί για κανονικές πιθανότητες για να μην υπάρχει underflow στους πολλαπλασιασμούς. Έτσι, κατά την εφαρμογή του αλγορίθμου Viterbi είναι αποδοτικότερο να υπολογίσουμε το άθροισμα των πιθανοτήτων αντί για το γινόμενο τους.

Αλγόριθμος Viterbi

Η υλοποίηση του αλγορίθμου Viterbi σε αυτήν την άσκηση έγινε ως εξής: Ξεκινάμε υπολογίζοντας τις πιθανότητες για τις καταστάσεις α και β με βάση τις αρχικές πιθανότητες των καταστάσεων μας και συνεχίζουμε για κάθε σύμβολο της ακολουθίας τον υπολογισμό της πιθανότητας να έχει προκύψει από τη κατάσταση α ή β , συν τη πιθανότητα να πάμε από την προηγούμενη κατάσταση στη τωρινή, συν τη συνολική πιθανότητα της προηγούμενης κατάστασης. Παράλληλα σε κάθε βήμα του αλγορίθμου σημειώνουμε την κατάσταση από την οποία προήλθε για την κατάσταση στην οποία καταλήγουμε κάθε φορά ώστε να μας βοηθήσει στην οπισθοδρόμηση. Αφού υπολογίσουμε όλες τις διαδρομές για όλες τις καταστάσεις και βρούμε την μέγιστη πιθανότητα στο τέλος ανάμεσα στις δύο καταστάσεις ξεκινάμε την οπισθοδρόμηση η οποία μας φανερώνει και τη πιο πιθανή ακολουθία καταστάσεων(α/β). Η παραπάνω διαδικασία απεικονίζεται στο παρακάτω διάγραμμα.



ΣΗΜΕΙΩΣΗ

Τα πράσινα βελάκια υποδεικνύουν ποια διαδρομή επικρατεί για κάθε κατάσταση σε κάθε βήμα. Τέλος χρησιμοποιούμε την μέθοδο οπισθοδρόμησης (με τα γκρι βελάκια) μέχρι και το πρώτο σύμβολο της ακολουθίας για να βρούμε το καλύτερο μονοπάτι το οποίο θα είναι και η πιο πιθανή ακολουθία καταστάσεων.

Αποτελέσματα

```
a: -2.321928 (previous state:None)    -3.795859 (previous state:a)    -7.269790 (previous state:a)    -10.743722 (previous state:a)
b: -3.321928 (previous state:None)    -5.795859 (previous state:b)    -7.684828 (previous state:b)    -9.573797 (previous state:b)
```

The best path for the sequence GGCT is: b b b b with highest probability of -9.573796658442287

Εδώ βλέπουμε ότι για την ακολουθία GGCT η καλύτερη ακολουθία καταστάσεων που δίνει μέγιστη πιθανότητα (λογαριθμική) -9.573797 είναι η ββββ

Τρόπος εκτέλεσης

- Ανοίγουμε ένα παράθυρο cmd ή terminal όπου βρίσκεται στο path όπου υπάρχει το αρχείο 11_4.py.
- Εκτελούμε την εντολή `python 11_4.py`.

Θέμα 3

Άσκηση 6.14

Εκφώνηση

Δύο παίκτες παίζουν το εξής παιχνίδι με δύο αλληλουχίες που έχουν μήκος n και m νουκλεοτίδια αντίστοιχα. Σε κάθε γύρο του παιχνιδιού, ένας παίκτης μπορεί να αφαιρέσει δύο νουκλεοτίδια από τη μία αλληλουχία (είτε την πρώτη είτε τη δεύτερη) και ένα νουκλεοτίδιο από την άλλη. Ο παίκτης που δεν μπορεί να κάνει κίνηση κερδίζει. Ποιος θα κερδίσει; Περιγράψτε τη νικηφόρα στρατηγική για όλες τις τιμές των n και m .

Σχεδιασμός

Για την επίλυση της άσκησης χρησιμοποιούμε τη Python 3.10 και τις βιβλιοθήκη Bio και Random. Τα ισοένζυμα τα οποία θα χρησιμοποιήσουμε είναι το brain με 39253 νουκλεοτίδια και το liver με 49971 νουκλεοτίδια τα οποία βρίσκονται στα αρχεία brain.fasta και liver.fasta.

Υλοποίηση

Το συγκεκριμένο παιχνίδι παίζεται με 2 παίκτες και 2 αλληλουχίες νουκλεοτιδίων. Οι δυο παίκτες παίζουν μεταξύ τους. Αρχικά παίζει ο πρώτος παίκτης ο οποίος πρέπει να επιλέξει μια ακολουθία από την οποία θα διαγράψει 2 στοιχεία και από την άλλη 1 στοιχείο. Υστέρα παίζει ο δεύτερος παίκτης ο οποίος πρέπει να κάνει αντίστοιχα μια επιλογή. Η διαδικασία επαναλαμβάνεται έως ότου κάποιος παίκτης δεν μπορεί να κάνει άλλη κίνηση, έτσι θα είναι και ο νικητής του παιχνιδιού.

Νικηφόρα Στρατηγική

Έστω, λοιπόν, m η ακολουθία 1 και n η ακολουθία 2. Ο πρώτος παίκτης θα βρίσκεται σε ευνοϊκή θέση όταν παραλαμβάνει τις ακολουθίες ως εξής:

- $m = 3x$ and $n \geq m$ για $x \in \mathbb{N}^0$ και το αντίθετο.
- $m = n = 3k + 1$ για $x \in \mathbb{N}^0$.

Για παράδειγμα αν είναι η σειρά του πρώτου παίκτη και οι ακολουθίες είναι της μορφής:

- $[m,n]$ = Για $[0,0]$ η $[0,3]$ η $[0,2]$ κερδίζει αφού είναι τελικές καταστάσεις ενώ για $[3,3]$ η $[3,4]$ η $[3,6]$ η $[6,7]$ μεγιστοποιεί τις πιθανότητες να κερδίσει.
- $[m,n]$ = Για $[1,1]$ κερδίζει ενώ για $[4,4]$ η $[7,7]$ η $[10,10]$ εξίσου μεγιστοποιεί τις πιθανότητες να κερδίσει.

Σύμφωνα με τα παραπάνω συμπεραίνουμε ότι για να κερδίσει ένας παίκτης εφόσον είναι η σειρά του, θα πρέπει να επιλέξει την κίνηση η οποία φέρνει τον αντίπαλο στην χειρότερη δυνατή θέση.

Αποτελέσματα

Πρώτες επαναλήψεις

```
Initial lengths = liver length: 39253      brain length: 49971

      plays: First Player
liver length: 39251      brain length: 49970

      plays: Second Player
liver length: 39250      brain length: 49968

      plays: First Player
liver length: 39248      brain length: 49967

      ...
      ...
      ...|
      ...

      plays: First Player
liver length: 9252      brain length: 19969
```

Τελευταίες επαναλήψεις

```
liver length: 4      brain length: 10746
      plays: First Player
liver length: 3      brain length: 10744
      plays: Second Player
liver length: 1      brain length: 10743
      plays: First Player
liver length: 0      brain length: 10741
      ----- Second Player wins! -----
```

Όπως βλέπουμε ο δεύτερος παίκτης δεν μπορεί να κάνει κάποια κίνηση καθώς η πρώτη ακολουθία έχει μήκος 0, οπότε είναι και ο νικητής.

Τρόπος εκτέλεσης

- Ανοίγουμε ένα παράθυρο cmd ή terminal όπου βρίσκεται στο path όπου υπάρχει το αρχείο 6_14.py.
- Εκτελούμε την εντολή `python 6_14.py`.

Θέμα 4

Υλοποίηση

Αναζητήστε τον κωδικό **PDB-code: 7NEH**, στην πρωτεϊνική βάση δεδομένων : <https://www.rcsb.org/> Πρόκειται για την τρισδιάστατη δομή του συμπλόκου ενός αντισώματος με μια πρωτεΐνη ακίδα.

Προετοιμασία:

- Παρατηρήστε τα στοιχεία που δίνει η PDB για αυτήν τη δομή

7NEH

Crystal structure of the receptor binding domain of SARS-CoV-2 Spike glycoprotein in complex with COVOX-269 Fab

PDB DOI: [10.2210/pdb7NEH/pdb](https://doi.org/10.2210/pdb7NEH/pdb)

Classification: **VIRAL PROTEIN/IMMUNE SYSTEM**

Organism(s): *Homo sapiens*, *Severe acute respiratory syndrome coronavirus 2*

Expression System: *Homo sapiens*

Mutation(s): Yes ⓘ

Deposited: 2021-02-04 Released: 2021-03-03

Deposition Author(s): Zhou, D., Ren, J., Stuart, D.

Funding Organization(s): Medical Research Council (MRC, United Kingdom), CAMS Innovation Fund for Medical Sciences (CIFMS)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

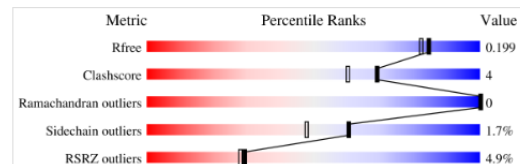
Resolution: 1.77 Å

R-Value Free: 0.198

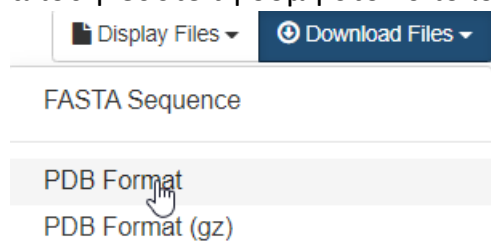
R-Value Work: 0.188

R-Value Observed: 0.189

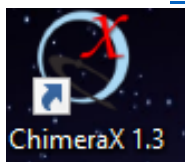
wwPDB Validation ⓘ



- Χρησιμοποιώντας πάνω δεξιά την επιλογή «Download Files» → PDB format αποθηκεύστε τη δομή στον υπολογιστή σας (είναι αρχείο plain text)



- Εγκαταστήστε στον υπολογιστή σας το λογισμικό Chimera-X από την ιστοσελίδα <https://www.rbvi.ucsf.edu/chimerax/download.html>



- Δείτε τις διαφάνειες του μαθήματος 19.05.2021 για λεπτομέρειες σχετικά με τη χρήση του λογισμικού Chimera-X και σε ό,τι αφορά τις εντολές συμπληρωματικό υλικό – σε περίπτωση που θέλετε να ενημερωθείτε περαιτέρω- μπορείτε να βρείτε στις αντίστοιχες σελίδες του Chimera.

Ερώτημα 1:

α) Δείτε τα στοιχεία που παρουσιάζονται στην πρωτεϊνική βάση δεδομένων και προσδιορίστε τη μέθοδο με την οποία έχει προσδιορισθεί η δομή του συμπλόκου;

- Η μέθοδος με την οποία έχει προσδιορισθεί η δομή του συμπλόκου είναι η : **X-RAY DIFFRACTION**

β) Ποιο το resolution (διακριτική ικανότητα) στο οποίο προσδιορίστηκε η δομή;

- Η διακριτική ικανότητα της δομής είναι **1.77Å**

γ) Παραθέστε το Ψηφιακό αναγνωριστικό (Digital Object Identifier, DOI) της σχετικής επιστημονικής δημοσίευσης



- Το DOI της δημοσίευσης είναι το [10.2210/pdb7NEH/pdb](https://doi.org/10.2210/pdb7NEH/pdb)

Ερώτημα 2:

α) Πόσες διακριτές πρωτεϊνικές αλυσίδες (molecular entities, macromolecules) περιλαμβάνει η εν λόγω δομή;

- Περιλαμβάνει **3** διακριτές πρωτεϊνικές αλυσίδες.

Macromolecules

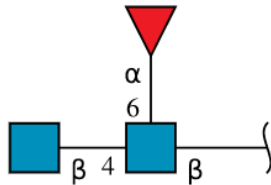
Entity ID: 1					
Molecule	Chains	Sequence Length	Organism	Details	Image
COVOX-269 Fab heavy chain	A [auth H]	222	Homo sapiens	Mutation(s): 0 ⓘ	
Entity ID: 2					
Molecule	Chains	Sequence Length	Organism	Details	Image
COVOX-269 fab light chain	B [auth L]	215	Homo sapiens	Mutation(s): 0 ⓘ	
2b4ke antibody	C [auth E]	302	αλυσίδα κοινής αλυσίδας 2b4ke	Gene names: 2 ⓘ Μητρώου(α): 1 ⓘ	
Entity ID: 3					

β) Για κάθε μια από αυτές σημειώστε το πλήθος των αμινοξέων (sequence length)

- Το πλήθος των αμινοξέων για την COVOX-269 Fab heavy chain είναι **222**.
- Το πλήθος των αμινοξέων για την COVOX-269 fab light chain είναι **215**.
- Το πλήθος των αμινοξέων για την Spike glycoprotein είναι **205**.

γ) Πόσους ολιγοσακχαρίτες περιλαμβάνει η δομή του συμπλόκου;

- Περιέχει **1** ολιγοσακχαρίτη.

Oligosaccharides			
Entity ID: 4			
Molecule	Chains	Chain Length	2D Diagram ⓘ
2-acetamido-2-deoxy-beta-D-glucopyranose-(1-4)-[alpha-L-fucopyranose-(1-6)]2-acetamido-2-deoxy-beta-D-glucopyranose	D [auth A]	3	

δ) Η δομή του συμπλόκου έχει ένα άτομο χλωρίου (Cl⁻). Παραθέστε την αλυσίδα την οποία ανήκει.

- Ανήκει στην αλυσίδα **CA**.

Ερώτημα 3:

α) Με χρήση του λογισμικού Chimera-X «διαβάστε» το αρχείο 7neh.pdb για να απεικονίσετε τη δομή. Παραθέστε με τη μορφή πίνακα τα στοιχεία που εμφανίζονται στο Log αρχείο και δείχνουν τις επί μέρους αλυσίδες της γλυκοπρωτεΐνης και του αντισώματος (heavy and light chain) καθώς και των επιπλέον στοιχείων (non-standard residues) που εμφανίζονται στο αρχείο.

Chain information for 7neh.pdb #1		
Chain	Description	UniProt
E	spike glycoprotein	SPIKE_SARS2
H	covox-269 fab heavy chain	
L	covox-269 fab light chain	

Non-standard residues in 7neh.pdb #1	
CL	chloride ion
EDO	1,2-ethanediol (ethylene glycol)
FUC	α-L-fucopyranose (α-L-fucose; 6-deoxy-α-L-galactopyranose; L-fucose; fucose)
NAG	2-acetamido-2-deoxy-β-D-glucopyranose (N-acetyl-β-D-glucosamine; 2-acetamido-2-deoxy-β-D-glucose; 2-acetamido-2-deoxy-D-glucose; 2-acetamido-2-deoxy-glucose; N-acetyl-D-glucosamine)
NO3	nitrate ion
PEG	di(hydroxyethyl)ether
SO4	sulfate ion

β) Επιλέξτε την αλυσίδα που αντιστοιχεί στην πρωτεΐνη ακίδα είτε μέσω του log αρχείου είτε χρησιμοποιώντας τη γραμμή εντολών στο κάτω μέρος της οθόνης

Command: [select /E](#)

Ή εναλλακτικά

Command: [select /E:332-527](#)

Στη συνέχεια χρησιμοποιώντας τη γραμμή εργαλείων :

[Actions](#) → [colour](#) → [all options](#)

Επιλέξτε μόνο το [Cartoons](#) και χρωματίστε την αλυσίδα με το χρώμα της αρεσκείας σας.

Ακυρώστε την επιλογή χρησιμοποιώντας τη γραμμή εργαλείων :

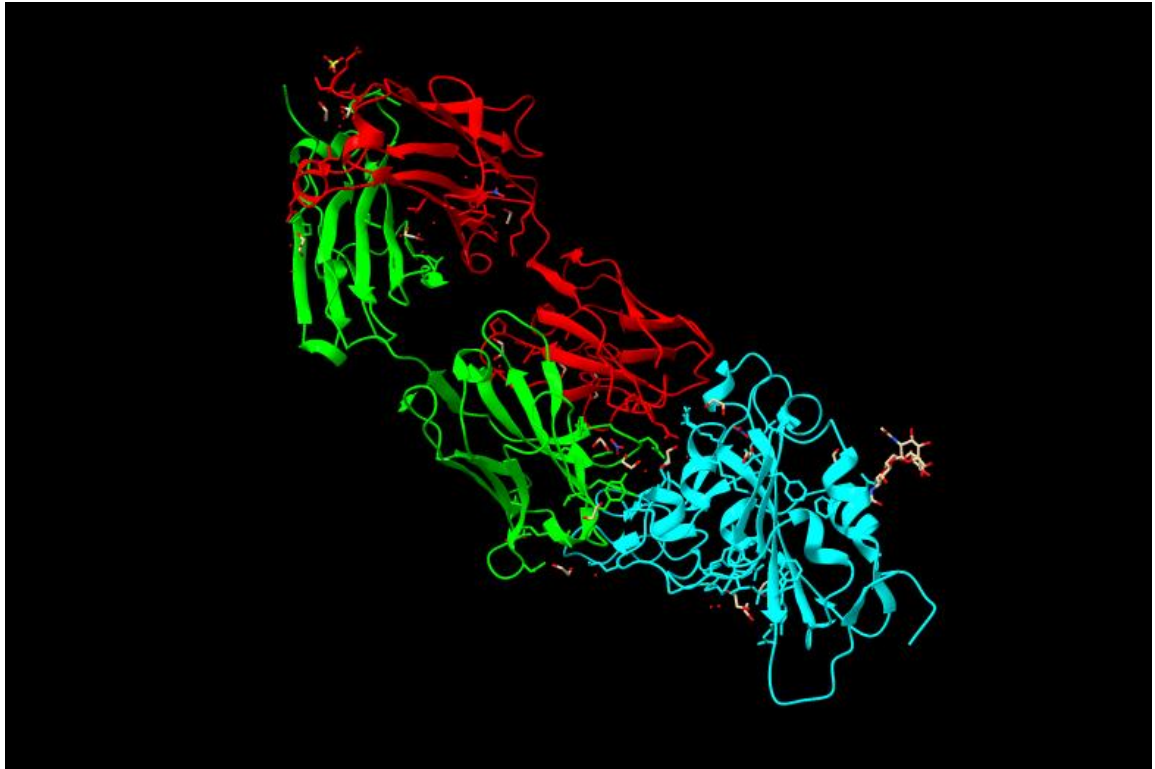
[Select](#) → [clear](#)

Επαναλάβετε για τις υπόλοιπες πρωτεϊνικές αλυσίδες (βλ. ερώτημα 2^α)

Αποθηκεύστε την εικόνα που δημιουργήσατε

[File](#) → [save](#) → (επιλέξτε το format)

Παρουσιάστε την εικόνα που δημιουργήσατε- Απάντηση για το (3β)



Ερώτημα 4:

α) Σε συνέχεια του ερωτήματος 3: Επιλέξτε όπως και πριν μια μια τις αλυσίδες π.χ.

Command: `select /E`

Στη συνέχεια χρησιμοποιώντας τη γραμμή εργαλείων που βρίσκεται το ίδιο το παράθυρο των γραφικών:

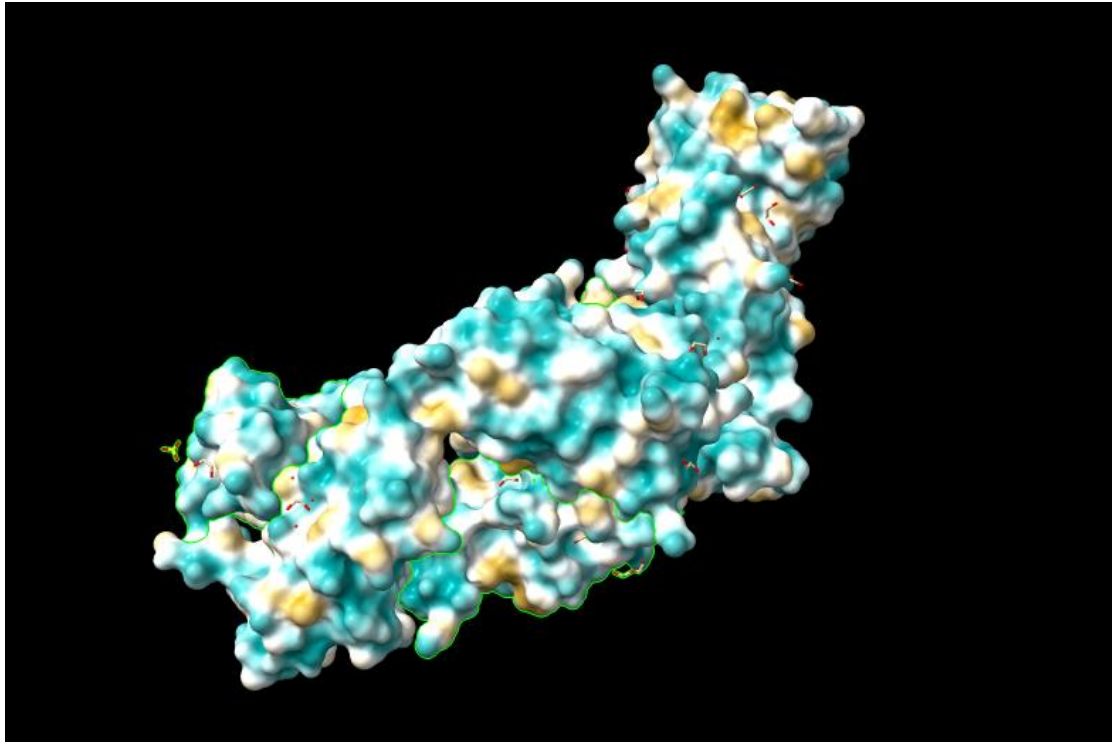
Molecule Display → `hydrophobic`

Εμφανίστε την επιφάνεια πρωτεΐνης για κάθε μία αλυσίδα της πρωτεΐνης (επαναλάβετε δηλαδή εκτός από την E και για τις υπόλοιπες)

Αποθηκεύστε την εικόνα που δημιουργήσατε

File → `save` → (επιλέξτε το `format`)

Παρουσιάστε την εικόνα που δημιουργήσατε- Απάντηση για το (3β)



Ερώτημα 5:

α) Σε συνέχεια του ερωτήματος 3: Επιλέξτε όπως και πριν μια μια τις αλυσίδες π.χ.

Command: `select /E`

Στη συνέχεια χρησιμοποιώντας τη γραμμή εργαλείων

Tools → Sequence → Show Sequence viewer

Εκεί με διαφορετικό χρώμα φαίνονται τα δευτεροταγή στοιχεία της πρωτεΐνης. Επιλέξτε μόνο τους β-κλώνους (β-strands) με το ποντίκι σας ως εξής:

Επιλέξτε με το ποντίκι μια ζώνη όπως υποδεικνύεται και κρατώντας πατημένο το shift προσθέστε επιπλέον ζώνες ώστε να επιλέξετε όλες τις περιοχές που έχουν την ίδια απόχρωση και αντιστοιχούν σε β-stands.

Στη συνέχεια χρησιμοποιώντας τη γραμμή εργαλείων :

Actions → colour → all options

Επιλέξτε μόνο το **Cartoons** και χρωματίστε την αλυσίδα με το χρώμα της αρεσκείας σας.

Επαναλάβετε για τις α-έλικες επιλέγοντας τις περιοχές με το άλλο χρώμα.

Παρουσιάστε την εικόνα που δημιουργήσατε- Απάντηση για το (5α)



Τα σκουρα μπλε είναι οι β-κλώνοι και τα κιτρινα οι α-κλώνοι.

β) Σε συνέχεια του ερωτήματος 3: Επιλέξτε το σάκχαρο που είναι προσδεδεμένο στην πρωτεΐνη ακίδα.

Command: select :NAG

Στη συνέχεια χρησιμοποιώντας τη γραμμή εργαλείων που βρίσκεται το ίδιο το παράθυρο των γραφικών:

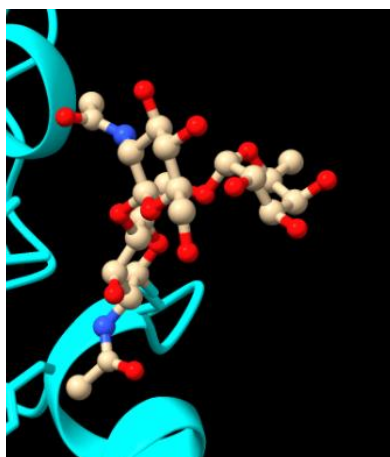
Molecule Display → Ball and stick

Θα παρατηρήσετε ότι ένα μέρος του σακχάρου δεν έχει εφαρμόσει την εντολή. Αν περάσετε τον cursor πάνω από αυτό θα δείτε ότι το «όνομα» του σακχάρου δεν είναι NAG αλλά FUC είναι δηλαδή ένα άλλο είδος σακχάρου συνδεδεμένο με το πρώτο. Συνεπώς για να το φτιάξετε όλο με την ίδια αναπαράσταση :

Command: select :FUC

Molecule Display → Ball and stick

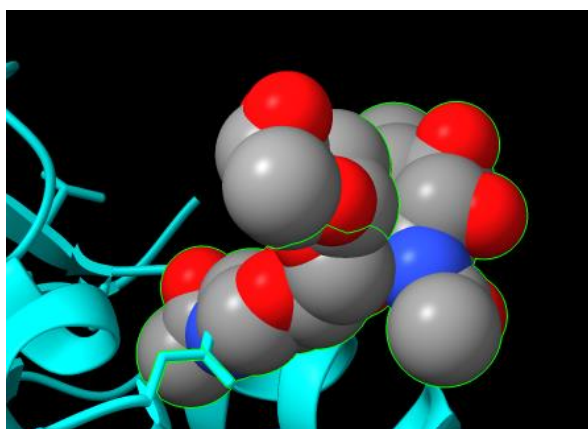
Παρουσιάστε την εικόνα που δημιουργήσατε- Απάντηση για το (5β)



γ) Σε συνέχεια του ερωτήματος 5β:

Επαναλάβετε ό,τι και στο 5β) μόνο που τώρα τα δύο σάκχαρα θα τα «ζωγραφίσετε» με την επιλογή “sphere” δηλαδή με σφαίρες Van der Waals

Παρουσιάστε την εικόνα που δημιουργήσατε- Απάντηση για το (5γ)



Σημείωση : Σε περίπτωση που θέλετε να αλλάξετε τα χρώματα για την αναπαράσταση των σακχάρων

Command: select :FUC μετά Actions → colour → all options

Επιλέξτε μόνο το Atoms/Bonds και χρωματίστε τα άτομα με το χρώμα της αρεσκείας σας. Θα έχουν όλα το ίδιο χρώμα. Αν θέλετε να δείξετε το άζωτο με μπλέ και το οξυγόνο με κόκκινο όπως συνηθίζεται στη συνέχεια χρησιμοποιήστε την επιλογή By Heteroatom ή εναλλακτικά By Element

Βιβλιογραφία

- ΕΙΣΑΓΩΓΗ ΣΤΟΥΣ ΑΛΓΟΡΙΘΜΟΥΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ, NEIL C. JONES, PAVEL A. PEVZNER.
- Βιοπληροφορική & Λειτουργική Γονιδιωματική, Jonathan Pevsner.
- Σημείωσης μαθήματος "Συστήματα Πολυμέσων" Διδάσκον Άγγελος Πικράκης, Ph.D.