

BGGN-213: FOUNDATIONS OF BIOINFORMATICS

https://bioboot.github.io/bggn213_W22/

Test Questions

Instructions: This open-book, open-notes test consists of **15 required questions** and one optional bonus point question (question 16). The number of points for each question is indicated in green font at the beginning of each question.

No communication (electronic or otherwise) with your fellow students regarding this test until after the due date.

Please remember to:

- Download the PDF version and open in Preview (Mac) or Acrobat Reader (Windows).
- Type all your answers directly in the space provided below each question.
- Remember that concise answers are preferable to wordy ones.
- Make sure your name, UCSD email and PID number are on the first page.
- Save and upload your completed test to gradescope.
- **Good luck!**

Name:

UCSD email:

PID:

GitHub username:

Q1. [10pt] The following sequences resulted from an analysis of a patients healthy and tumor tissue. You can also download these sequences from <https://tinyurl.com/mutantseqs>

```
>wt_healthy
MGPWSRSL SALLLLLQVSSWLCQEPEPCHPGFDAESYFTTVPRRHLEGRVLRVNFEDC
TGRQRTAYFSLDTRFKVGTGCVITVKRPLRFHNPQIHFLVYAWDSTYRKFS TKVTLNTVG
HHHRPPPHQASVSGIQAELLTFPNSSPGLRRQKRDWVIPPISC PENEKGPFPKNLVQIKS
NKDKEGKV FY SITGQGADTPPVGVFIIERETGWLKVTEPLDRERIATYTLF SHAVSSNGN
AVEDPMEILITVTDQNDNKPEFTQEVFKGSVMEGALPGTSVMEVTATDADDDVNTYNAAI
AYTILSQDPELPDKNMFTINRNTGVISVVT TGLDRESFPTYTLVVQAADLQEGGLSTTAT
AVITVTD TNDNPPIFNPPTYKGV PENEANVVITTLKVT DADAPNTPAWEAVYTILNDDG
GQFVVTTNPVNNDGILKTAKGLDFEAKQQYILHVAVTNVVPFEVSLTTSTATVTVDVLDV
NEAPIFVPPEKRVEVSEDFGVGQEITSYTAQEPDTFMEQKITYRIWRDTANWLEINPDTG
AISTRAELDREDFEHVKNSTYTALIIATDNGSPVATGTG TLLLILSDVNDNAPIPEPRTI
FFCERNPKPQVINIIDADLPNTSPFTAELTHGASANWTIQYNDPTQESIILKPKMALEV
GDYKINLKLMDNQNKDQVTTLEVSVCDCEGAAGVCRKAQPVEAGLQIPAILGILGGILAL
LILILLLLLFLRRRAVVKEPLLPEDDTRDNVYYYDEEGGGEEDQDFDLSQLHRGLDARP
EVTRNDVAPTLMSVPRYLPRPANPDEIGNFIDENLKAADTDPTAPPYDSSLVFDYEGSGS
EAASLSSLNSESDDKDQDYDYLNEWGNRFKKLADMYGGGEDD

>mutant_tumor
MGPWSRSL SALLLLLQVSSWLCQEPEPCHPGFDAESYFTTVPRRHLEGRVLRVNFEDC
TGRQRTAYFSLDTRFKVGTGCVITVKRPLRFHNPQIHFLVYAWDSTYRKFS TKVTLNTVG
HHHRPPPHQASVSGIQAELLTFPNSSPGLRRQKRDWVIPPISC PENEKGPFPKNLVQIKS
NKDKEGKV FY SITGQGADTPPVGVFIIERETGWLKVTEPLDRERIATYTLF SHAVSSNGN
AVEDPMEILITVTVQNDNKPEFTQEVFKGSVMEGALPGTSVMEVTATDADDDVNTYNAAI
AYTILSQDPELPDKNMFTINRNTGVISVVT TGLDRESFPTYTLVVQAADLQEGGLSTTAT
AVITVTD TNDNPPIFNPPTYKGV PENEANVVITTLKVT DADAPNTPAWEAVYTILNDDG
GQFVVTTNPVNNDGILKTAKGLDFEAKQQYILHVAVTNVVPFEVSLTTSTATVTVDVLDV
NEAPIFVPPEKRVEVSEDFGVGQEITSYTAQEPDTFMEQKITYRIWRDTANWLEINPDTG
AISTRAELDREDFEHVKNSTYTALIIATDNGSPVATGTG TLLLILSDVNDNAPIPEPRTI
FFCERNPKPQVINIIDADLPNTSPFTAELTHGASANWTIQYNDPTQESIILKPKMALEV
GDYKINLKLMDNQNKDQVTTLEVSVCDCEGAAGVCRKAQPVEAGLQIPAILGILGGILAL
LILILLLLLFLRRRAVVKEPLLPEDDTRDNVYYYDEEGGGEEDQDFDLSQLHRGLDARP
EVTRNDVAPTLMSVPRYLPRPANPDEIGNFIDENLKAADTDPTAPPYDSSLVFDYEGSGS
EAASLSSLNSESDDKDQDYDYLNEWGNRFKKLADMYGGGEDD
```

- What protein do these sequences correspond to?

- What are the tumor specific mutations in this particular case (e.g. A130V)?

- List one RCSB PDB identifier with 100% identity to the *wt_healthy* sequence?

- Using the [NCI-GDC](#) list the observed top 4 missense mutations in this protein?

Q2. [6pts] Suggest databases to answer the following questions along with the answer you obtain using the *wt_healthy* sequence in **Q1**. Only one database suggestion per question please. The first one has been done for you for demonstration purposes.

- I want to find the genomic location of some human gene?

NCBI GENE database;
Chromosome 16 (16q22.1)

- I want to determine what protein domains my novel protein contains?

- I want to find if an NMR structure has been solved for my protein of interest or it's homologues?

RCSB PDB sequence search w/ filter 'solution NMR' under experimental methods. For this protein, the first result is 1SUH.

- I want to find out about the human mendelian disorders associated with this gene?

Q3. [3pts] What is the name of the major bioinformatics file formats used for storing the following biomolecular data:

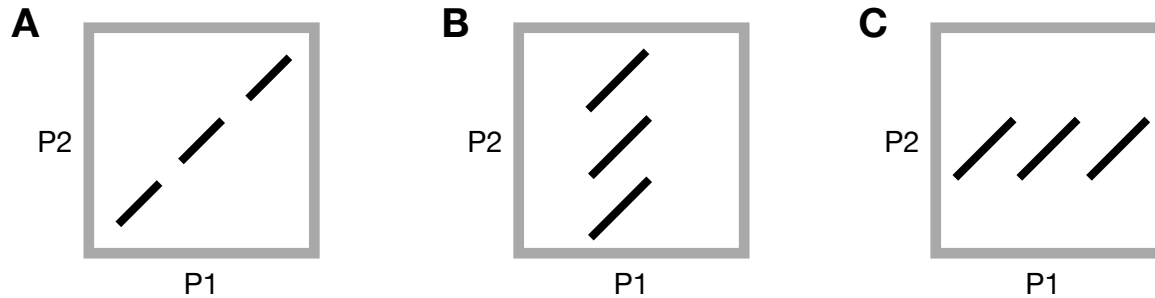
- Protein (and other) sequence data that is used by many bioinformatics tools?

- Protein (and other) atomic structure data?

- Sequence reads from NGS?

Q4. [1pts] Represent the following three sequences FGMAKLQGD, FGMGRLQGD and GHAMKLQGD (named **seqA**, **seqB** and **seqC**) in the format you answered for the first part of question **Q3**.

Q5. [1pts] Which schematic figure (**A-C**) below best represents a dotplot for two proteins, **p1** and **p2**, whose only significant similarity is a common domain family, and where the domain is occurring once in **p2** and three times in **p1**.



Your Selection (**A, B** or **C**):

Q6. [3pts] Consider the following multiple alignment of Transcription Factor Binding site DNA sequences

	1	2	3	4	5
Sequence 1	-	T	A	G	C
Sequence 2	C	T	G	G	A
Sequence 3	C	T	A	-	A
Sequence 4	A	T	A	G	T

Give the (average) profile of the above alignment by filling out the table below (overleaf). The first column (i.e. position in the alignment) has been done for you, now complete the rest. You will use this table for answering questions 7-9 below.

	1	2	3	4	5
A	0.25				
C	0.5				
T	0				
G	0				
-	0.25				

Q7. [2pts] What is the highest scoring sequence match to your profile above (question Q6) and what is it's score?

Sequence:

Score:

Q8. [2pts] Using your completed profile table above (from question 6) score the following two sequences (S1 and S2):

S1. CTAGC:

S2. AGAGA

Q9. [2pts] Following the heuristic threshold for a positive match proposed in Harbison et al. [Nature (2004) 431:99-104.] namely using the threshold for a positive match = 60% x Max Score. Are either of the two sequences in question 12 potential transcript factor binding sites? If so why?

Q10. [3pts] Name two major repositories from which R packages can be obtained and the base R function that can be used to install the package bio3d?

Q11. [2pt] What do the Eigenvalues from principal component analysis (PCA) represent and what PC will have the highest Eigenvalue?

Q12. [1pt] List one major advantage of hierarchical clustering over k-means clustering?

Q13. [2pts] Why do researchers often normalize read counts in gene expression studies and what features/effects are they normalizing for?

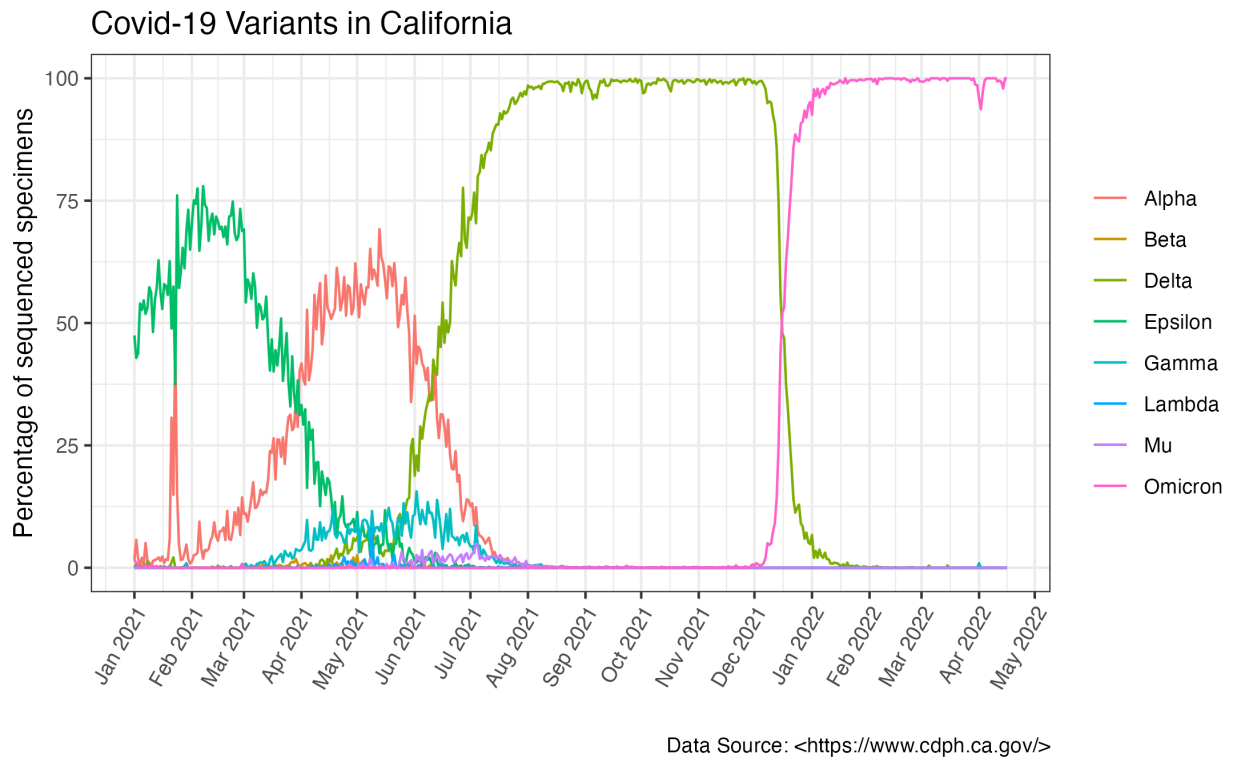
Q14. [2pts] Name two functional databases commonly used for gene set enrichment analysis?

Q15. [10pt] Obtain the most recently dated **COVID-19 Variant Data** from the California Health and Human Services (CHHS) open data site:

<https://data.chhs.ca.gov/dataset/covid-19-variant-data>

Upload to [gradescope](#) a PDF format report generated from an Rmarkdown document that demonstrates reading the above CSV file and generating the below visualization of this data.

NB. You can chose how to make this plot and whether you want to make improvements or stylistic changes. However, you are strongly encouraged to use the ggplot2, lubridate and dplyr packages for this task. Please make sure your **name** and **PID** number is on the first page and that your report contains all of your **code**, **text description**/narrative text of why you doing a particular task/code chunk and the resulting **figure**.



Q16. [10pt] Optional: This is not a required question but will yield you **10 extra bonus points**. Using git upload (a.k.a. *push*) your RStudio project containing your complete work for **Q15** to GitHub and provide a link to your project directory here:

—END OF TEST—