Bi/BE/CS183 2020
Profs. Matt Thomson and Lior Pachter
Problem Set 2
Due on Moodle by Thursday January 23rd at 12:00pm

Homework policy: You may collaborate with anyone on the homework and consult/use any resources. Please write down on the homework who you worked with and explain what, if any, help you received. Also list all resources that were used for the homework and explain how they helped.

Problem 1 (15 points)

Let m, x and s be fixed parameters for global alignment, and consider two sequence s_1 and s_2 that are to be aligned. Describe an algorithm that will output a random global alignment of s_1 and s_2 with probability proportional to its score.

Here m, x and s are the match, mismatch, and spacing penalties. Note there are many possible answers to this problem. A sufficient answer is one that would allow someone to completely implement the algorithm akin to following a recipe and be convinced that it should work.

To make things simpler, you can restrict all match, mismatch and spacing scores to be all negative (or all positive).

Problem 2 (10 points)

Last week we mentioned in class (and uploaded to moodle) the 2007 article *What is a gene, post-ENCODE? History and updated definition.* Read it (it's a fun read!) and answer the following:

a) According to the article, what would be a one sentence definition of the gene for 1860s-1900s, 1910s, 1940s, 1950s, 1960s, 1970s-1980s, 1990s, 2000s?

b) Describe how you'd think of a gene

c) Looking at table 1 from the article, can you identify some phenomena that would be "exceptions" to your answer in b?

Problem 3 (40 points)

In this problem you'll implement the Needleman-Wunsch algorithm. To help you with the implementation we uploaded two resources on moodle: Chapter 2 (Pairwise Alignment) of the book *Biological sequence analysis: Probabilistic models of proteins and nucleic acids* that was mentioned in class.
Chapter 2 (Computation) of Lior's book *Algebraic Statistics for Computational Biology*

a) Implement the Needleman-Wunsch algorithm, allowing for real valued match, mismatch and space parameters.

b) Use your implementation of the Needleman-Wunsch algorithm to find all optimal *alignments* for all possible parameters for the sequences `CAGATA` and `TCAGAC`.

Clarification: When you explore the parameter space, even if you were to explore all parameter values, not all kinds of alignment would arise. For example, if you scale all scores by a constant factor, the alignment does not change. It turns out that you cannot obtain any arbitrary alignment, regardless of the scores that you use. You can work out all possible optimal alignments by hand, but you can also use your implementation to do some brute force search over many parameters and see what alignments arise. For this question, all that you need to do is provide a list of these different alignments that you can obtain.

c) Use your algorithm to empirically determine the average length of the longest increase subsequence in a permutation of length 10.

E.g. if the permutation is $10, 2, 3, 1, 6, 7, 4, 5, 9, 8$ the longest increasing subsequence has length 5 ($2, 3, 4, 5, 8$ or $2, 3, 6, 7, 8$). You can do this by aligning the permutation with $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ where a match gives score $+1$, a mismatch gives score 0 and a gap gives score 0, then create random permutations and measure the average score. This is not just a math exercise - the longest increasing subsequence is part of the MUMmer alignnment software (`https://mummer4.github.io/`), first described in 1999. If you are curious, see the original paper *Alignment of whole genomes*, available at `https://mummer4.github.io/publications/MUMmer.pdf`

Problem 4 (15 points)

Visit the Ensembl website database page at `https://uswest.ensembl.org/info/data/ftp/index.html`. Navigate to the FTP download section and download the human cDNA (`FASTA` format). The file you want is named `Homo_sapiens.GRCh38.cdna.all.fa.gz` and is 65MB in size. To understand the naming convention, read the `README` file that is on the same FTP directory. Write some code and answer the questions below.

Hint: Biopython `https://biopython.org/` and Bioconductor `https://bioconductor.org/` are your friends. See the Bioconductor package Biostrings.

a) How many total bases of sequence are there in the transcriptome?

b) How many transcripts are there?

c) What is the longest transcript in the human genome?

d) Pick your favourite random entry from the human cDNA file. Explain what each thing means.

Problem 5 (15 points)

Here we look at some samples from the 2018 article *Sirt1 protects from K-Ras-driven lung carcinogenesis*, DOI 10.15252/embr.201643879. Take your time to understand what they did, and how to navigate the data repositories, because we'll revisit this and other datasets in future homeworks.

a) Install kallisto. For instructions see `http://pachterlab.github.io/kallisto`.

b) Download the reads for sample `GSM3168965` from RNA-Seq experiment `GSE115179`.
The GEO ("Gene Expression Omnibus") page for this experiment is `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115179`, but you should poke around and try to find it on your own.

c) Quantify the experiment with kallisto and determine the most highly abundant transcript. Explain what this transcript is and what it does. Does the abundance make sense given the experiment?

Reminder: To process the sample with kallisto, first you need to use kallisto to build an index. For a given transcriptome and *kmer* size the transcriptome index needs to be built only once. The default *kmer* size is 31, and you don't need to change that. In order to build the index, first you need to download from Ensembl the appropriate transcriptome. In problem 4 you learned how to download the human transcriptome from

Ensembl. You should determine (from the paper, sample GEO page or sample SRA page) which species transcriptome to download, and then build the index. You should be able to build the index on a laptop with 8GB RAM, it usually takes 15-30 min if you have enough memory. The resulting index file will be about 2GB.

Problem 6 (5 points)

Take the Google colab notebook for HMM shown in class and extend it for 4 observed states.

Appendix

"Ah wait!"

You say, after trying to do 5(b).

"I see no download button! Not even an FTP link!"

Correct. That would be too easy.

The GEO help page (`https://www.ncbi.nlm.nih.gov/geo/info/download.html`) simply says:
All the data in GEO can be downloaded in a variety of formats using a variety of mechanisms.

In our case (and very often) the data that we want is actually hosted on the Sequence Read Archive (SRA), `https://www.ncbi.nlm.nih.gov/sra`, so the sample `GSM3168965` gets another identifier associated to it. In the SRA it's known as `SRR7244429`. And on the sample's SRA page you can actually look at individual reads from the sequencing run (click on the `Reads` tab):
`https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR7244429`

But we can't download it directly. We need to use the `NCBI SRA Toolkit` available at:
`https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software`
And development information is on their github repository: `https://github.com/ncbi/sra-tools`
After the toolkit is installed, we can download the files using the tool `fastq-dump`, documented at:
`https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software`

Specifically, for this sample we'd use the command:
```
fastq-dump -I --split-files SRR7244429
```
Because it's a paired end read Illumina run, so it produces two `fastq` files

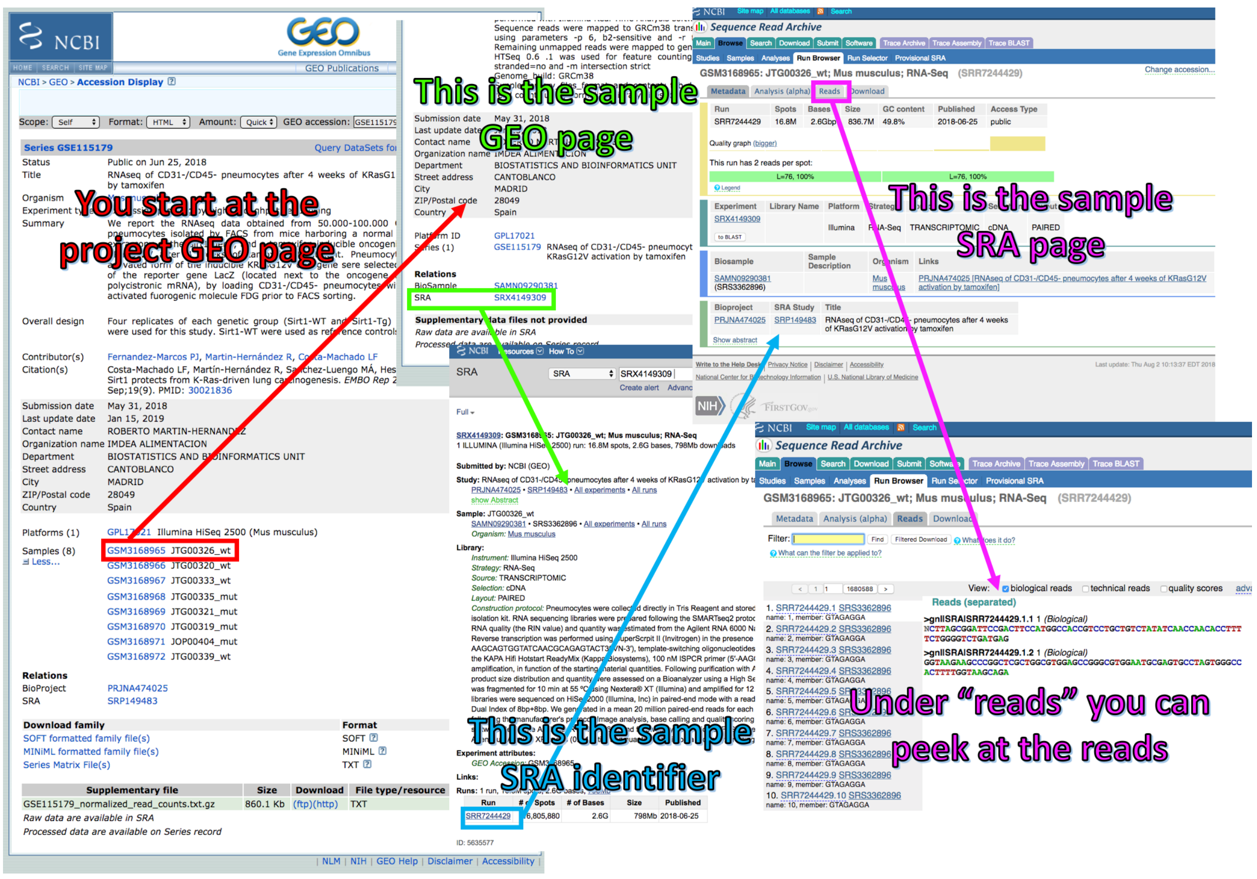
This downloads two files of 3.7GB each, `SRR7244429_1.fastq` and `SRR7244429_2.fastq`. You have to hope your connection doesn't get interrupted. If you are unhappy with the official documentation (as many are) here is an additional helpful resource: `https://edwards.sdsu.edu/research/fastq-dump/`

Thankfully there is a (much much) faster option: `fasterq-dump` (but it's only available for Mac and Linux)
`https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump`

If you decide to use `fasterq-dump`, you can download the data with the command `fasterq-dump SRR7244429`

Finally, to help you navigate the confusing and teacherous world of getting other people's data in bioinformatics (welcome!) below is a helpful diagram of how to find the sample SRA identifier, which is what you need to run `fastq-dump` or `fasterq-dump`. Get used to this gymnastics, because you'll have to do it a few times.

From GEO to SRA: a perilous journey


You start at the project GEO page
This is the sample GEO page
This is the sample SRA identifier
This is the sample SRA page
Under "reads" you can peek at the reads

NCBI > GEO > **Accession Display**

Scope: Self Format: HTML Amount: Quick GEO accession: GSE115179

Series GSE115179

Status	Public on Jun 25, 2018
Title	RNAseq of CD31-/CD45- pneumocytes after 4 weeks of KRasG1… by tamoxifen
Summary	We report the RNAseq data obtained from 50.000-100.000 … pneumocytes isolated by FACS from mice harboring a normal … of the reporter gene LacZ (located next to the oncogene … polycistronic mRNA), by loading CD31-/CD45- pneumocytes wi… activated fuorogenic molecule FDG prior to FACS sorting.
Overall design	Four replicates of each genetic group (Sirt1-WT and Sirt1-Tg) were used for this study. Sirt1-WT were used as reference controls
Contributor(s)	Fernandez-Marcos PJ, Martin-Hernández R, Costa-Machado LF
Citation(s)	Costa-Machado LF, Martín-Hernández R, Sanchez-Luengo MÁ, Hes… Sirt1 protects from K-Ras-driven lung carcinogenesis. *EMBO Rep* 2… Sep;19(9). PMID: 30021836
Submission date	May 31, 2018
Last update date	Jan 15, 2019
Contact name	ROBERTO MARTIN-HERNANDEZ
Organization name	IMDEA ALIMENTACION
Department	BIOSTATISTICS AND BIOINFORMATICS UNIT
Street address	CANTOBLANCO
City	MADRID
ZIP/Postal code	28049
Country	Spain
Platforms (1)	GPL17021 Illumina HiSeq 2500 (Mus musculus)
Samples (8)	GSM3168965 JTG00326_wt
	GSM3168966 JTG00320_wt
	GSM3168967 JTG00333_wt
	GSM3168968 JTG00335_mut
	GSM3168969 JTG00321_mut
	GSM3168970 JTG00319_mut
	GSM3168971 JOP00404_mut
	GSM3168972 JTG00339_wt

Relations

BioProject	PRJNA474025
SRA	SRP149483

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE115179_normalized_read_counts.txt.gz	860.1 Kb	(ftp)(http)	TXT

Raw data are available in SRA
Processed data are available on Series record

| NLM | NIH | GEO Help | Disclaimer | Accessibility |

Sequence reads were mapped to GRCm38 trans… using parameters -p 6, b2-sensitive and -r … Remaining unmapped reads were mapped to gen… HTSeq 0.6 .1 was used for feature counting stranded=no and -m intersection strict
Genome_build: GRCm38

Submission date	May 31, 2018
Department	BIOSTATISTICS AND BIOINFORMATICS UNIT
Street address	CANTOBLANCO
City	MADRID
ZIP/Postal code	28049
Country	Spain
Platform ID	GPL17021
Series (1)	GSE115179 RNAseq of CD31-/CD45- pneumocyt… KRasG12V activation by tamoxifen

Relations

BioSample	SAMN09290381
SRA	SRX4149309

Supplementary data files not provided
Raw data are available in SRA

SRA | SRA | SRX4149309 | Create alert | Advanc…

SRX4149309: GSM3168965: JTG00326_wt; Mus musculus; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2500) run: 16.8M spots, 2.6G bases, 798Mb downloads

Submitted by: NCBI (GEO)
Study: RNAseq of CD31-/CD45- pneumocytes after 4 weeks of KRasG12V activation by ta…
PRJNA474025 • SRP149483 • All experiments • All runs
show Abstract
Sample: JTG00326_wt
SAMN09290381 • SRS3362896 • All experiments • All runs
Organism: Mus musculus
Library:
Instrument: Illumina HiSeq 2500
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: PAIRED
Construction protocol: Pneumocytes were collected directly in Tris Reagent and stored … isolation kit. RNA sequencing libraries were prepared following the SMARTseq2 protoc… RNA quality (the RIN value) and quantity was estimated from the Agilent RNA 6000 Na… Reverse transcription was performed using SuperScrpit II (Invitrogen) in the presence … AAGCAGTGGTATCAACGCAGAGTACT30VN-3'), template-switching oligonucleotides … the KAPA Hifi Hotstart ReadyMix (Kappa Biosystems), 100 nM ISPCR primer (5'-AAG… amplification, in function of the starting material quantities. Following purification with … product size distribution and quantity were assessed on a Bioanalyzer using a High Se… was fragmented for 10 min at 55 °C using Nextera® XT (Illumina) and amplified for 12 … libraries were sequenced on HiSeq2000 (Illumina, Inc) in paired-end mode with a read … Dual Index of 8bp+8bp. We generated in a mean 20 million paired-end reads for each …
Experiment attributes:
GEO Accession: GSM3168965
Links:
Runs: 1 run, 16.8M spots, 2.6G bases, 798Mb

Run	# of Spots	# of Bases	Size	Published
SRR7244429	16,805,880	2.6G	798Mb	2018-06-25

ID: 5635577

NCBI Site map All databases Search
Sequence Read Archive
Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace BLAST
Studies Samples Analyses Run Browser Run Selector Provisional SRA

GSM3168965: JTG00326_wt; Mus musculus; RNA-Seq (SRR7244429) Change accession...

Metadata | Analysis (alpha) | Reads | Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR7244429	16.8M	2.6Gbp	836.7M	49.8%	2018-06-25	public

Quality graph (bigger)

This run has 2 reads per spot:
L=76, 100% | L=76, 100%

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX4149309		Illumina	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED

Biosample	Sample Description	Organism	Links
SAMN09290381 (SRS3362896)		Mus musculus	PRJNA474025 [RNAseq of CD31-/CD45- pneumocytes after 4 weeks of KRasG12V activation by tamoxifen]

Bioproject	SRA Study	Title
PRJNA474025	SRP149483	RNAseq of CD31-/CD45- pneumocytes after 4 weeks of KRasG12V activation by tamoxifen

Show abstract

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility
National Center for Biotechnology Information | U.S. National Library of Medicine
Last update: Thu Aug 2 10:13:37 EDT 2018

GSM3168965: JTG00326_wt; Mus musculus; RNA-Seq (SRR7244429)

Metadata | Analysis (alpha) | **Reads** | Download

Filter: Find Filtered Download What does it do?
What can the filter be applied to?

< 1 1 1680588 > View: [✓ biological reads] technical reads quality scores

Reads (separated)

1. SRR7244429.1 SRS3362896
name: 1, member: GTAGAGGA
2. SRR7244429.2 SRS3362896
name: 2, member: GTAGAGGA
3. SRR7244429.3 SRS3362896
name: 3, member: GTAGAGGA
4. SRR7244429.4 SRS3362896
name: 4, member: GTAGAGGA
5. SRR7244429.5 SRS3362896
name: 5, member: GTAGAGGA
6. SRR7244429.6 SRS3362896
name: 6, member: GTAGAGGA
7. SRR7244429.7 SRS3362896
name: 7, member: GTAGAGGA
8. SRR7244429.8 SRS3362896
name: 8, member: GTAGAGGA
9. SRR7244429.9 SRS3362896
name: 9, member: GTAGAGGA
10. SRR7244429.10 SRS3362896
name: 10, member: GTAGAGGA

```
>gnl|SRA|SRR7244429.1.1 1 (Biological)
NCTTAGCGGATTCCGACTTCCATGGCCACCGTCCTGCTGTCTATATCAACCAACACCTTT
TCTGGGGTCTGATGAG
>gnl|SRA|SRR7244429.1.2 1 (Biological)
GGTAAGAAGCCCGGCTCGCTGGCGTGGAGCCGGGCGTGGAATGCGAGTGCCTAGTGGGCC
ACTTTTGGTAAGCAGA
```