Re: Bachelor Project

Richard Röttger <roettger@imada.sdu.dk>
Mon 2/1/2021 2:34 PM

To: Niels Peter Roest <niroe18@student.sdu.dk>; Martin Enemark Pedersen <mbpe01@student.sdu.dk>; Kasper Hoflund <kahof18@student.sdu.dk>

2 attachments (11 MB)
13058_2016_Article_724.pdf; molecules-24-00631.pdf;

Hi All,

sorry, was busy with a phone call. As promised, I have prepared some data and additional Information for you.

You can download the prepared data set from here:
http://imada.sdu.dk/~roettger/teaching/resources/bachelor_project/all_data.zip

The zip-File contains 4 files:
* BRCA_clinicalMatrix - Here, you find all relevant clinical data for the samples. We probably won't need the file, this is just more FYI what Information we have.
* BRCA_survival.txt.gz - Here, you find a table with the survival times of the various patients, etc. We might need this file later, when we have time to also look at the unsupervised problem and see whether our groups differentiate in survival times.
* HiSeqV2.gz - normalized read counts, i.e., how much each transcript (that is the mRNA part) is encountered in the sample.
* PAM50_labels.xlsx - These are the assignment of the correct lables to the samples of the cancer subtypes. We will look at the following subtypes (look at the column PAM50): LumA, LumB, Normal, Basal, Her2.

As you can already see, Bioinformatics is quite a messy job, we have here various file types and conventions mixed wildly :-). Have a look at the files, then you should be able to figure rather quickly what is where and you might consider transferring your data to a more accessible and clean format; but that is up to you.

Alrighty. Additionally, I have attached two papers, which broadly describe how the data is processed. Please read through the papers (large parts are not relevant for you, but you should have seen them).

The papers are as follows:
* 13058_2016_Article_724: This is an example what type of machine-learning / data analysis is performed on these data. This should just give you an overview that we are working on actual problems and not just made up stuff. Also note, they combine this with Methylation data (something else we can measure in a cell, not important for us). Nevertheless, they base their work basically on the same dataset as we do, so read carefully through the preprocessing steps and data handling descriptions. When you google the paper, you can also have a look at the Supplemental Material and get further Inspiration.

* molecules-24-00631, similar story, uses more or less our data and does patient stratification. Again,

read through it, get a hang of it but not necessary to understand the details.

* Finally, have a look at this website:
https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html
Here, it is explained, how we get from the actual machine output to the completed data analysis. Note, again, many parts are not relevant, since you are not dealing with the individual reads (basically, what pops out of the machine) but already with the normalized read counts. Basically steps 1 & 2 are irrelevant (but very interesting to read to further understand the matter), it get important from point 3 onwards.

Your tasks:
*** Read through the material**
*** Have a look at the data and start playing with it. Think of what visualizations might be most useful to understand the data at hand**
*** Come up with a rough plan of what and how you want to perform the analysis (what models, how to evaluate, how to interpret, build an ensemble of classifiers (how), etc.)**
*** Write your Project Description**

Then we meet again (probably together with the second group) and discuss your next steps.

Cheers,
Richard

On 06.01.21 16:09, Niels Peter Roest wrote:

> Hi Richard,
>
> Sounds good. See you then.
>
> Best regards,
> Niels and group
>
> Hent Outlook til iOS
>
> ---
>
> **Fra:** Richard Röttger <roettger@imada.sdu.dk>
> **Sendt:** Wednesday, January 6, 2021 10:11:20 AM
> **Til:** Niels Peter Roest <niroe18@student.sdu.dk>; Martin Enemark Pedersen <mbpe01@student.sdu.dk>; Kasper Hoflund <kahof18@student.sdu.dk>
> **Emne:** Re: Bachelor Project
>
> Hi All,
>
> yeah, it looks like I have slot at 13:00 on Feb 1st.

Cheers,
Richard

On 05.01.21 17:19, Niels Peter Roest wrote:

> Hello Richard
>
> We have talked together and found that we would rather move this meeting to the 1st of feburary, would this be suitable for you?
> Best regards,
> Niels and group
>
> Hent Outlook til iOS
>
> ---
>
> **Fra:** Richard Röttger <roettger@imada.sdu.dk>
> **Sendt:** Tuesday, January 5, 2021 5:00:08 PM
> **Til:** Martin Enemark Pedersen <mbpe01@student.sdu.dk>; Kasper Hoflund <kahof18@student.sdu.dk>; Niels Peter Roest <niroe18@student.sdu.dk>
> **Emne:** Bachelor Project
>
> Hi All,
>
> I have not made a mistake, you should have been assigned to the Bachelor project with me. First of all, welcome to the Bioinformatics group and I hope you are looking forward to working on the project with me.
>
> I would suggest that we have meeting after the exams but before the lectures start. Would Friday, 29th of January at 10:30 suit you?
>
> Cheers,
> Richard
>
> --
> Richard Röttger
> Associate Professor, Dr. rer. nat.
>
> Dept. of Mathematics and Computer Science (IMADA)
> University of Southern Denmark (SDU)
> Campusvej 55
> DK-5230 Odense M
> Denmark

--
Richard Röttger
Associate Professor, Dr. rer. nat.

Dept. of Mathematics and Computer Science (IMADA)

University of Southern Denmark (SDU)
Campusvej 55
DK-5230 Odense M
Denmark

--
Richard Röttger
Associate Professor, Dr. rer. nat.

Dept. of Mathematics and Computer Science (IMADA)
University of Southern Denmark (SDU)
Campusvej 55
DK-5230 Odense M
Denmark