From Student Site
< User:Thomas Hollis

Thomas Hollis' Journal

(Thomas Hollis' personal Journal)

Contents

0. Course preparations

This section outlines the work done on all the courses that have to do with course preparation.

0.1 Wiki Editing (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Wiki_editing)

Objective

The main objective is to gain an in depth understanding of how to navigate and edit the student wiki as well as comprehend what will be expected of me throughout the course.

> Time estimated: 2 h; taken: 6 h; date started: 2018-09-14; date completed: 2018-09-15

Activities

I accessed and created my main StudentWiki user page, journal page and insights page. This was not too challenging as I have been contributing to Wikipedia's public body of information for a few years now and I am used to the WikiMarkup syntax. As an avid fan of the open-source community, I also created and committed my current progress to my own personal "Bioinformatics" Github repository (https://github.com/PsiPhiTheta) for the rest of the world to use. What was the most time consuming was to explore the entire wiki to see everything that this course has to offer, including all the activities that occurred in previous years. I must confess I did get a little carried away in my exploration and the boundary between "fun" and "work" is already starting to blur. However, I do believe that this is a good sign and I am very excited for the course ahead.

Conclusion and Outlook

The next unit that I plan to take is Journal Writing (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Journal) .

0.2 Journal Writing (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Journal)

Objective

The main objective is to start off my Journal with a strong baseline on which I can build the rest of my discoveries in this course.

Time estimated: 1 h; taken: 1.5 h; date started: 2018-09-15; date completed: 2018-09-15

Activities

I started writing my journal, including this precise section. I also edited my main user page and added some information to help my classmates identify who I am and what my background is. This will undoubtably be useful for collaboration in the future, especially given my computer science background with very little biology knowledge.

Conclusion and Outlook

The next unit that I plan to take is Plagiarism (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-Plagiarism) .

0.3 Plagiarism (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-Plagiarism)

Objective

The main objective is to revise concepts related to Plagiarism. Collaboration can all to easily turn into plagiarism in academic institutions and the only way to stamp it out is through education.

Time estimated: 20 min; taken: 1 h; date started: 2018-09-15; date completed: 2018-09-15

Activities

Read in detail the wiki page and the websites referred to in the main body. Checked all existing content to make sure it is fully cited. Added a citation for Google Deep Dream algorithms on my main User: page. Added the following practice APA citations:

- a procedure in the methods section of a journal article, as you would cite it in a technical report:

Gruber, H., Holzer, M., & Ruepp, O. (2016). Sorting the Slow Way: An Analysis of Perversely Awful Randomized Sorting Algorithms. Lecture Notes in Computer Science Fun with Algorithms, 4475, 183-197. doi:10.1007/978-3-540-72914-3_17

- a piece of code you found in a StackOverflow article, as you would put it as a comment into computer code:

```
# BioGeek. (2015, April 24). Are there any worse sorting algorithms than Bogosort (a.k.a Monkey Sort)? Retrieved September 16, 2018, from https://stackoverflow.com/questions/2609857/are-there-any-worse-sorting-algorithms-than-bogosort-a-k-a-monkey-sort
```

- some contents from a classmate's journal that you incorporate into your own journal:

Moskal, N. (2013) Open Project Concept. Retrieved September 16, 2018, from http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Natalia_Moskal/Open_Project

Conclusion and Outlook

The next unit that I plan to take is Insights (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-Insights) .

0.4 Insights (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-Insights)

Objective

The main objective is to learn how to create and maintain the "insights!" page in BCH441.

Time estimated: 0.5 h; taken: 1 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Edited my insights page.

Conclusion and Outlook

The next unit that I plan to take is Netiquette (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Netiquette) .

0.5 Netiquette (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Netiquette)

Objective

The main objective is to learn how to interact with other users on this StudentWiki network and in the mailing list.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Read the Netiquette (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Netiquette) page carefully and came up with an insight for future me.

Conclusion and Outlook

The next unit that I plan to take is Backups (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Backups) .

0.6 Backups (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Backups)

Objective

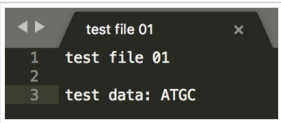
The main objective is to refresh existing knowledge on how to securely backup data to prevent accidental data loss.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Read the Backups (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Backups) page carefully and verified the effectiveness of my current backup solution. I currently store all of my personal files (excluding application data) in iCloud drive, which is continually backed up on Apple servers. So when is the last time I made a backup? Approximately three seconds ago when I saved my most recent file. Apple servers themselves are geographically distributed, redundant and backed up (often using RAID or similar technology). In addition, as I do not trust anyone else but myself with my own data, I also make differential local backups of my entire drive on the 1st of each month. This is done on two separate external hard drives (Time Machine really is a good utility). This allows me to have an on-site backup and off-site backup which protects me in the case that there is a fire and both my laptop and my hard drive get destroyed. Indeed the risk of this is non-negligeable as there are 24 000 house fires per year in Canada [1] for a total of 14 million houses [2]. The probability of fire is therefore around 0.17% per year, roughly the same as the SSD failure rate listed in the Backups (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Backups) page.

As a quick sanity check I created and deleted a test file (containing the data: "ATGC") that I successfully recovered, as shown aside.



Recovered data from deleted test file

Conclusion and Outlook

While friends may laugh at my backup strategy and fail to understand why I get so excited on the first day of each month, I will persist in backing up my data reliably. The next unit that I plan to take is Cargo Cult (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Cargo_Cult) .

0.7 Cargo Cult (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Cargo_Cult)

Objective

The main objective is to understand how to better identify examples of Cargo Cult science.

Time estimated: 0.5 h; taken: 1 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Read the 1974 Richard Feynman (http://en.wikipedia.org/wiki/Richard_Feynman) Caltech Commencement address (http://calteches.library.caltech.edu/51/2/CargoCult.htm) . Contributed to the Cargo Cult Science (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/Cargo_Cult_Science) and Not (quite) Cargo Cult Science (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/Not_(quite)_Cargo_Cult_Science) pages by adding the example of SCIgen (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/Cargo_Cult_Science#SCI_gen) and commenting on the Spongebob Cargo Cult (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/Cargo_Cult_Science#SpongeBob_and_his_bubble_dance) example.

Conclusion and Outlook

The next unit that I plan to take is Technical Questions (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Technical_questions) .

0.8 Technical questions (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Technical_questions)

Objective

The main objectives are to understand how to ask technical questions and to learn the concept of Minimal Working Examples (MWEs).

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Read the three recommended documents and bookmarked them for later use.

Conclusion and Outlook

The next unit that I plan to take is Information Sources (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Info_sources) .

0.9 Information Sources (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Info_sources)

Objective

The main objectives are to build a database of information sources that will be of invaluable help for subsequent modules and to learn to navigate these resources.

Time estimated: 0.5 h; taken 1 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Bookmarked all useful resources and spend some time on each to create an account and/or learn to navigate the website.

Conclusion and Outlook

The next unit that I plan to take is Biocomputing Setup (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Biocomputing_setup) .

0.10 Biocomputing Setup (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Biocomputing_setup)

Objective

The main objectives are to set up my computer with the required IDEs, languages and utilities needed for bioinformatics.

Time estimated: 1 h; taken: 10 min; date started: 2018-09-16; date completed: 2018-09-16

Activities

After reading the entire recommended setup, I made a total of 0 changes to my setup. Indeed, I already had installed all the recommended tools, already structured my filesystem well and of course all my file extensions are always displayed by default. Yay!

I very much like that Prof. Steipe is on my side in the macOS vs. Windows war... take that Bill (https://en.wikipedia.org/wiki/Bill_Gates) .

Conclusion and Outlook

The next unit that I plan to take is Installation (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Installation) .

1. R Programming

This section outlines the work done on all the courses related to learning the R programming language.

1.1 Installation (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Installation)

Objective

The main objectives are to remind myself on how to install and run packages in R.

Time estimated: 1 h; taken: 0.5 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Indeed as I have had prior experience using R and RStudio this module was more of a refresher. Nonetheless, I found it to be a useful review - especially the script that checks if a package is already installed before installing it. I also executed all the code suggested and had a good play around with the new packages discussed.

Conclusion and Outlook

The next unit that I plan to take is Setup (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Setup) .

1.2 Setup (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Setup)

Objective

The main objectives are to check that my R environment is setup as recommended for bioinformatics.

Time estimated: 0.5 h; taken: 1 h; date started: 2018-09-16; date completed: 2018-09-16

Activities

Indeed as I have used RStudio in the past only for machine learning and statistics, it was not optimally configured for Bioinformatics so I slightly tweaked my setup. I also learnt all about connecting RStudio to GitHub and did all the suggested exercises.

Conclusion and Outlook

The next unit that I plan to take is Console (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Console) .

1.3 Console (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Console)

Objective

The objective here was to learn the difference between executing entire scripts and executing commands individually from the console.

Time estimated: 0.5 h; taken: 0.1 h; date started: 2018-09-17; date completed: 2018-09-17

Activities

This was already known to me but reading the full wiki page is always good to serve as a refresher.

Conclusion and Outlook

The next unit that I plan to take is Help (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Help) .

1.4 Help (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Help)

Objective

The main objective here is to gain familiarity with all the different sources of help for R programming outside what has already been mentioned.

Time estimated: 0.5 h; taken: ?? h; date started: 2018-09-17; date completed: 2018-09-17

Activities

Read all the links recommended and tried out the code excerpts.

Conclusion and Outlook

The next unit that I plan to take is Syntax basics (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Syntax_basics) .

1.5 Syntax basics (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Syntax_basics)

Objective

The objective is to refresh on some R basics through examples.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-09-18; date completed: 2018-09-18

Activities

Running all the suggested code was a good opportunity for me to check that I am building my knowledge on good foundations. I was also able to make sure I don't have any bad R-programming habits to shake off. One particular thing I brushed up on was the difference between the "<-" and "=" in R. A really good explanation was read here (https://renkun.me/2014/01/28/difference-between-assignment-operators-in-r/) . I also found out in this week's lecture that you never need to explicitly define parameters that have the "=" sign in R documentation.

The code I wrote can be found here: [Expand]

Conclusion and Outlook

The next unit that I plan to take is Vectors (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Objects-Vectors) .

1.6 Vectors (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Objects-Vectors)

Objective

My objective for this unit is to build on my existing knowledge of R vectorisation.

Time estimated: 1 h; taken: 1 h; date started: 2018-09-18; date completed: 2018-09-18

Activities

Ran through all the code suggested and found out something that I completely overlooked in the past: the difference between `mode(x)` and `class(x)`. A fascinating extension to the comments from Prof. Steipe can be found here (https://stats.stackexchange.com/questions/3212/mode-class-and-type-of-r-objects) .

Conclusion and Outlook

The next unit that I plan to take is Data Frames (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Objects-Data_frames) .

1.7 Data Frames (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Objects-Data_frames)

Objective

My main goal in this section is to remind myself what an R data frame is and how it operates.

Time estimated: 1 h; taken: 1 h; date started: 2018-09-18; date completed: 2018-09-19

Activities

I ran through the code and spend a significant amount of time getting lost in rabbit wholes about how R *actually* does things... (see more on this in my insights! (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/insights!#Insight_.236_-_Why_you_should_never_trust_someone_who_says_.22Trust_Me.22) page).

Oh and I actually wrote some code too...

One line but it counts anyway right? [Expand]

Conclusion and Outlook

The next unit that I plan to take is Lists (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Objects-Lists) .

1.8 Lists (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Objects-Lists)

Objective

My objective here is to understand some concepts about lists that I may have overlooked.

> Time estimated: 0.5 h; taken: 1 h; date started: 2018-09-18; date completed: 2018-09-19

Activities

This unit was actually very interesting as it went over in detail an aspect of R that I had always overlooked: the detail of lists. I learn lots of new R tricks!

Here is the script I wrote: [Expand]

Conclusion and Outlook

The next unit that I plan to take is Subsetting (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Subsetting) .

1.9 Subsetting (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Subsetting)

Objective

The main goal for me here is to learn the details of subsetting in R which I may not already be aware of.

> Time estimated: 1 h; taken: 2 h; date started: 2018-09-19; date completed: 2018-09-19

Activities

I was surprised to learn a lot of new things about subsetting when completing all the required exercises. For example, it was interesting to find out the difference between `order(x)` and `sort.list(x)` simply by reading R documentation! I absolutely love how organised the R documentation is - makes coding in R extremely smooth! Lots of new things covered here. I must make sure to try to practice what I have learnt soon to help engrain it into long term memory.

Reminder from past-thomas to future-thomas:

Do not write code where you use the magic numbers: name your rows and columns instead!

Here is the code I wrote for this unit: [Expand]

Conclusion and Outlook

The next unit that I plan to take is Control Structures (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Control_structures) .

1.10 Control Structures (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Control_structures)

Objective

I think that my main task for this section will be to learn about the `seq_along()` function as the other material I believe will only serve as a reminder.

> Time estimated: 1 h; taken: 1 h; date started: 2018-09-19; date completed: 2018-09-19

Activities

Updated my knowledge on control structures - especially now I know that using `seq_along()` prevents unexpected behaviour for `null` sequences.

Code developed: [Expand]

Conclusion and Outlook

The next unit that I plan to take is Functions (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Functions) .

1.11 Functions (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Functions)

Objective

This unit will mostly be practice in consolidating my skills at writing functions in R.

> Time estimated: 1 h; taken: 1.5 h; date started: 2018-09-19; date completed: 2018-09-19

Activities

My main activity was to have my mind blown: inputs to R functions can be abbreviated up to their smallest unique form! I never knew that! Oh and I had an extended practice by modifying to the extreme some of the functions provided.

Code developed: [Expand]

Conclusion and Outlook

The next unit that I plan to take is Plotting (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Plotting) .

1.12 Plotting (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Plotting)

Objective

Here I would like to learn new methods of plotting and displaying data in R as it is a crucial skill.

> Time estimated: 1 h; taken: 1.5 h; date started: 2018-09-19; date completed: 2018-09-20

Activities

Sat in mesmerisation as all the nice colours of R plots flowed from my screen into my eyes. Then saved a local copy of the amazing tutorial I just completed because that was absolutely fantastic! Future-me is going to love this.

Conclusion and Outlook

The next unit that I plan to take is Coding Style (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Coding_style) .

1.13 Coding Style (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Coding_style)

Objective

Learn a consistent coding style to use throughout the Wiki!

> Time estimated: 0.5 h; taken: 1 h; date started: 2018-09-20; date completed: 2018-09-20

Activities

Well let's just say I may have re-editted all the code I have written till now to make it conform to the standards expected...

Conclusion and Outlook

The next unit that I plan to take is Introduction to R (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Introduction) .

1.14 Introduction to R (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Introduction)

Objective

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-09-24; date completed: 2018-09-25

Activities

Git cloned the ABC-units repository and ran the sample task to ensure R studio is set up correctly. All good.

Conclusion and Outlook

The next unit that I plan to take is Cell Cycle (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Cell_cycle) .

1.15 Testing R Code (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Unit_testing)

Objective

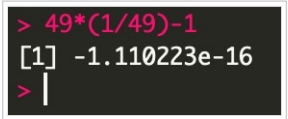
My main objective here is to follow a rigorous protocol for testing my future R code.

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

I followed the learning unit page thoroughly and ran through the full R script. Floating point errors always made me very very sad (see aside). But what makes me even more sad is that `(0.1 + 0.7) == 0.8` returns `FALSE` ! Absolute abomination! To make things worse the reason behind this is just as tragic: FAQ section 7.31 (https://cran.r-project.org/doc/FAQ/R-FAQ.html#How-can-I-clean-up-my-workspace_003f) .

Oh and I also wrote some code [Expand]



Floating Point Issues

Conclusion and Outlook

The next unit that I plan to take is Probability (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Probability) .

1.16 FASTA (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-FASTA)

Objective

My main objective here is to understand the structure of FASTA files.

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-23; date completed: 2018-10-24

Activities

I went through the entire programming tutorial.

Sadly there was no code for me to write here... [Expand]

Conclusion and Outlook

The next unit that I plan to take is Biostrings (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Biostrings) .

1.17 Biostrings (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Biostrings)

Objective

My main objective here is to gain familiarity with the Biostrings R package.

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-24; date completed: 2018-10-24

Activities

The usual `init()` and run through the tutorial goodness.

Sadly there was no code for me to write here either... [Expand]

Conclusion and Outlook

The next unit that I plan to take is Measuring Sequence Similarity (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Similarity) .

1.18 R RegEx (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-RegEx)

Objective

Here I want to practice using RegEx in R.

> Time estimated: 0.5 h; taken: 1 h; date started: 2018-10-26; date completed: 2018-10-26

Activities

Had fun reading through the wiki. I was already briefly familiar with what RegEx is although I had avoided it all my life until now...

Here is the code I wrote [Expand]

Conclusion and Outlook

RegEx is a powerful tool in R. The next unit that I plan to take is Sequence Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Concepts) .

1.19 Extra learning units

Objective

The main objective is to read through extra programming modules that are not necessary to complete the course.

> Time estimated: 1 h; taken: 1 h; date started: 2018-09-25; date completed: 2018-09-25

Activities

I had a good read through of the Debugging (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Debugging) and Literate Programming (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Literate_programming) courses! Sadly, the unit that appealed most to me, Optimization (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Optimization) , is still a stub.

Conclusion and Outlook

Although this was not required, it was definitely worthwhile and good fun.

2. Fundamentals

This section outlines the work done on all the courses that have to do with bioinformatics fundamentals.

2.1 Cell Cycle (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Cell_cycle)

Objective

Understand the fundamentals behind the cell cycle, in particular for the case of yeast and the role of the Mbp1 protein.

Time estimated: 1 h; taken: 2 h; date started: 2018-09-25; date completed: 2018-09-25

Activities

Very intensive reading followed by more intensive reading. All the sources provided (as well as a few visits to Wikipedia) where explored. Much of the content was not understood so I am going to come back to this. Wow, that's a lot of biology I need to catch up on! Thanks to Alana Man (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Alana_Man/BCH441_Journal) for making me realise how much biology knowledge I need to work to catch up on! Her journal is a good sanity check when I find a certain unit challenging.

Conclusion and Outlook

The next unit that I plan to take is Data Models (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Data_models) .

2.2 Data Models (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Data_models)

Objective

Learn how to sketch data models.

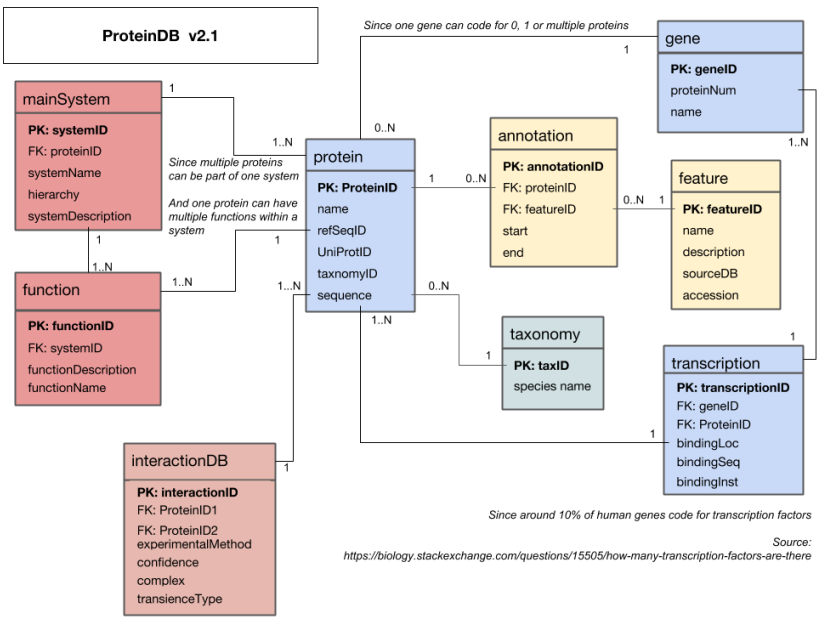
Time estimated: 1 h; taken: 2 h; date started: 2018-09-30; date completed: 2018-10-04

Activities

I wrote my first bioinformatics sketch! Also I was the first one in the class to do so, so I am happy with the pace at which I am currently working. I don't think my sketch is very good but I look forward to all the feedback from the mailing list! Here is the first version:

I am so ashamed of it that I have hidden it inside this collapsible box. [Expand]

Indeed, from all the feedback I received I found out that I had completely missed a critical part of the unit which was a link to a pdf document about data models. Since I had totally not seen this I was a bit confused as I guessed my sketch out of the blue. One humiliating lecture painfully revealed how much I have missed. This was however a surprisingly a great learning opportunity to ensure I am more careful in the future. I redid this unit and wrote an amended version of my initial abomination:



I feel like the toughest part here was understanding the biology behind it rather than actually building the model... I had to google A LOT of things. I may revisit this to improve the model at a later date.

Conclusion and Outlook

The next unit that I plan to take is My Species (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-MYSPE) .

2.3 Software Development (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Software_development)

Objective

Learn the formal theory behind software development, something that I have taken for granted for all my years as a software engineer.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

Read through in detail the contents of the learning unit page.

Conclusion and Outlook

The next unit that I plan to take is Test Driven Development (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Test_driven_development) .

2.4 Test Driven Development (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Test_driven_development)

Objective

Learn what test driven development is and how to leverage it in bioinformatics.

Time estimated: 1 h; taken: 0.5 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

Read through in detail the contents of the learning unit page. Indeed I had actually done some test-driven development before but I was simply not aware that this development approach is called like that. I find it to be often useful but tend to overlook this approach for smaller projects. I can totally understand why it would be critical for bioinformatics and look forward to returning to this development style in subsequent work.

Conclusion and Outlook

The next unit that I plan to take is Testing R Code (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Unit_testing) .

2.5 Graphs & Networks (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-MAT-Graphs_and_networks) (submitted: 6 marks)

Objective

The objective in this unit is to understand how relationships between genes are presented in various graphs & networks. Much time was dedicated here as I plan to maximise my performance in the quiz that will take on this unit.

Time estimated: 6 h; taken: 6 h; date started: 2018-10-13; date completed: 2018-10-23

Activities

I will take the in class quiz for this learning unit (yay, I got 100%)! I also attended a quiz before on the NCBI and EBI, just to get a feeling of how they are.

I read through the notes on graphs and networks and went through the R script tutorial, which I particularly enjoyed. I also had a gander at the pubmed paper (https://www.ncbi.nlm.nih.gov/pubmed/21527005) on analysing biological networks using graph theory. A very interesting read even if some of the details went over my head.

- The 8 main categories of graphs covered (in lectures and in my external readings) were: DAGs (directed acyclic graphs), cyclic graphs (like DAGs but with infinite response), trees (top to bottom direction only graph), random graphs, hypergraphs (with sets), dual graphs (a.k.a Voronoi diagrams), De Bruijn graphs (with amino acid sequences), Euler cycles (each edge visited once), Hamilton Cycles (each node visited exactly once).
- The 5 main types of graph set metrics covered were: node number (size), edge number (in-degree, out-degree or simply degree), histogram of degrees, topology (degree centrality leads to high degree nodes in centre of network), state (random, scale-free i.e. no meaningful average node or hierarchical).
- The 5 main topologic metrics covered are: shortest path, centrality, diameter, spanning trees (all the vertices covered with minimum possible number of edges - cannot be disconnected), causality (cannot have time delay units violating principles of causality)
- The 4 main methods for automatic graph generation were: Erdos-Renyi (random connection), Gilbert, Barabasi-Albert (degree drop as node number increases), Price.
- The 4 main techniques for achieving scale-free graphs were: Erdos-Renyi (as before), copy model (add new nodes by copying a fraction of links of existing nodes), hierarchal model, hyperbolic geometric graphs (nodes connected if distance between them is smaller than a hyperbolic threshold)
- The 2 main algorithms for structure analysis covered were: Dijkstra's algorithm (find shortest path in $O(V^2)$), Floyd-Warshall algorithm (all the shortest paths in $O(V^3)$)

The main equations to remember here are:

- Scale free state probability (where k is usually between 2 and 3):

$P(k) \propto k^{-\gamma}$ (blimey, wiki renders LaTeX poorly...)

- Betweenness centrality:

$$C_b(w) = \sum_{i \neq j \neq w} \frac{\sigma_{ij}(w)}{\sigma_{ij}}$$

Conclusion and Outlook

Graph theory is so awesome. I was familiar with Dijkstra before (a personal hero of mine) but I had no idea about many metrics discovered today. A true gem of information. I can see how these graphs could be leveraged frequently in bioinformatics!

The next unit that I plan to take is Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Databases) .

2.6 Genetic Code (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Genetic_code)

Objective

The main objective here was to learn how DNA leads to protein synthesis through the use of mRNA and codons.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-06; date completed: 2018-10-06

Activities

Thankfully I read a lot of the supplementary materials including much of the wikipedia content. I think without this, the coding exercise would have been very confusing. I find all this stuff so cool. My world of computer science crashing into my genome, it's a beautiful sight.

I was surprised to find a lot of M codons (start codons) and did some digging as to why this is the case. In fact I found on the Biology StackExchange (https://biology.stackexchange.com/questions/46427/multiple-start-and-stop-codons-in-mrna-and-pre-mrna) that in about 5% of genes the first AUG (start codon) is skipped and translation starts at one of the other AUG sequences (although this is still poorly understood).

Conclusion and Outlook

The next unit that I plan to take is Sequence (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Sequence)

2.7 Extra learning units

Objective

The main objective is to read through extra fundamental modules that are not necessary to complete the course to make sure I am fully equipped.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-01; date completed: 2018-10-05

Activities

So far most the extra units I have checked out (Systems Models (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Systems_models) , Integration testing (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Integration_testing) , Bayes Theorem (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Bayes_theorem) , Enrichment (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Enrichment) and Hypothesis Testing (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Hypothesis_testing)) are all still stubs. However, Structured Process Notation (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-SPN) was a very interesting read!

Conclusion and Outlook

Although this was not required, it was definitely worthwhile!

3. Bioinformatics

This section outlines the work done on all the courses that have to do with bioinformatics fundamentals.

3.1 My Species (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-MYSPE)

Objective

Acquire my own species that I will work on for the rest of the course.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

Read through the learning unit and ran the R code. My species is "Babjeviella inositovora" and it's BiCode is "BABIN". This is awesome, I love it! I even created a Wikipedia page (https://en.wikipedia.org/wiki/Babjeviella_inositovora) for it! Why does El Salvador have a page for it and not the english wiki? According to its NCBI page (https://www.ncbi.nlm.nih.gov/genome/?term=Babjeviella+inositovora) , my species has potential applications in biotechnology.

Update: due to an issue in the BABIN genome sequence being incomplete, Prof. Steipe has asked me to abandon it and pick a new one (by incrementing my student number by 1). This was quite sad as I was starting to get attached to BABIN. Oh well, my new species is "Isaria fumosorosea" and it's BiCode is "ISAFU". It already has a Wikipedia page (https://en.wikipedia.org/wiki/Isaria_fumosorosea) and I look forward to adopting it. It looks like a lot of research has already been done on it, as it shows promise of being a good pesticide!

Conclusion and Outlook

The next unit that I plan to take is Abstractions (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Abstractions) .

3.2 Abstractions (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Abstractions)

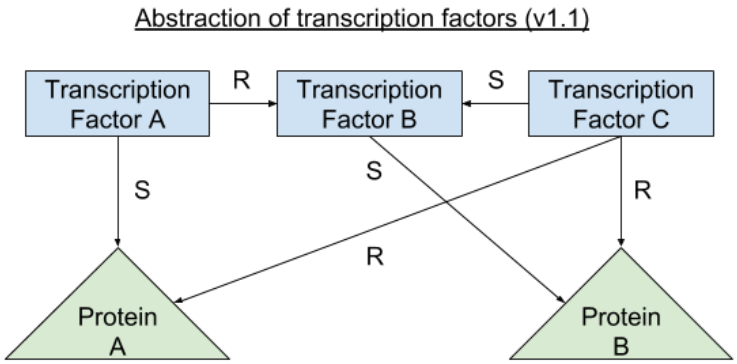
Objective

Learn how abstractions are made in the field of bioinformatics.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

Interestingly enough there are many similarities between the abstractions I am used to making in computer science (such as those made by Directed Acyclic Graphs (https://en.wikipedia.org/wiki/Directed_acyclic_graph)) and those made in bioinformatics. This was my first bioinformatics abstraction attempt:



Although it is not very complete it came closer than I expected to the actual abstraction that is commonly used: Gene Regulatory Networks (https://en.wikipedia.org/wiki/Gene_regulatory_network) .

Conclusion and Outlook

The next unit that I plan to take is Software Development (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-CSC-Software_development) .

3.3 Sequence (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Sequence)

Objective

Learn how sequences work in bioinformatics, what types of sequences exist and how to use them in my own work.

Time estimated: 1 h; taken:1 h; date started: 2018-10-06; date completed: 2018-10-06

Activities

I must confess I have not yet memorised all of the IUPAC one letter amino acid codes. I also was not able to answer all of the past exam questions (http://steipe.biochemistry.utoronto.ca/abc/index.php/Amino_Acid_Exam_Questions) but I believe this might be because of my lack of biological knowledge. The R code was certainly interesting to go through too!

Conclusion and Outlook

The next unit that I plan to take is Genetic Code Optimality (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Genetic_code_optimality)

3.4 Genetic Code Optimality (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Genetic_code_optimality)

Objective

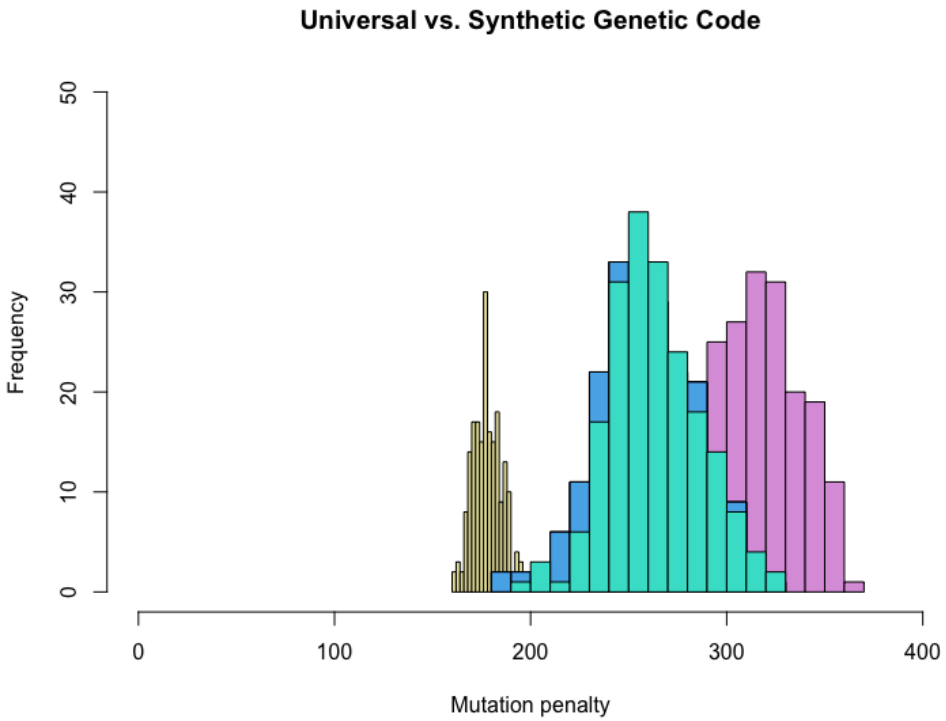
The main objective here is to look at alternative genetic codes and evaluate their robustness and optimality with regard to code changes (mutations).

Time estimated: 1 h; taken: 1 h; date started: 2018-10-06; date completed: 2018-10-06

Activities

Followed the code in depth. Very interesting stuff. A very powerful demonstration of genetic robustness!

The code I wrote can be found here: [Expand]



Interpretation: Both modified alternative genetic codes yielded distributions with significantly more mutations (based on mutation frequency & penalty as defined in the code). Interestingly, the variance of these mutation distributions was also wider for the modified genetic codes (in blue and purple) than for the original (in yellow). This suggests that the original genetic code is particularly good at resisting the impact of mutations. Indeed, even after some changes, these mutations will not necessarily be as expressed in the original genetic code than it would be in the alternative genetic codes.

Conclusion and Outlook

The genetic code is remarkably mutation proof. This is because the impact of the mutations (once translated back) on average causes less penalty (tangible impact) than in randomly generated variations of the genetic code. This would be in line with my expectations as I would expect natural selection to leave room for some mutation but not an excessive and unmanageable amount (with my limited knowledge of biology though, I am not making this statement with much confidence or authority...).

The next unit that I plan to take is Storing data (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Storing_data)

3.5 Storing Data (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Storing_data) (submitted: 6 marks)

Objective

The main goal here is to learn about multiple concepts regarding data storage and ACID databases (for example JSON, mySQL and their R wrappers).

Time estimated: 6 h; taken: 6 h; date started: 2018-10-07; date completed: 2018-10-13

Activities

My first learning unit to be submitted for credit! I intend on spending a lot of time on this to make sure I get it right. I created the evaluation submission page at: User:Thomas_Hollis/Eval_Storing_Data (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_Storing_Data) .

The data formats covered are: **FASTA** (simple, commonly used), **GBFF** (extensively annotatable), **BLAST** (output file parsers heavily worked on for many years), **PDB** (apparently impossible to read all information from all PDB files with a single parser).

The data grammars covered are: **ASN.1** (abstract syntax notation v1 - used by the NCBI), **XML** (extended markup language - widely used online), **JSON** (Java Script Object Notation - popular for key:value pairs, more human readable than XML).

The data model implementations covered are: **text files in filesystems** (using for example a folder hierarchy containing JSON files), **spreadsheets** (do not scale well, convert to .csv files for R imports), **R lists and dataframes** (what will be used mostly in this course), **relational databases** (such as MySQL, Maria DB, or Postgres).

I also went through all the R code thoroughly with some extra background reading to make sure everything is in tip top shape. I had to change MYSPE as BABIN had missing information. Information on my new MYSPE can be found here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Journal#3.1_My_Species) .

Code developed:

```
# Thomas Hollis (BCH441, University of Toronto) -v2.1
#     - Purpose: Add an entry to a database and output information in a desired format
#     - Bugs & issues: no bugs, no issues, no warnings

# --- SECTION 1: Add entry to database -----------------
#    Write and submit code that adds another philosopher to the datamodel:
#       Immanuel Kant, (1724 - 1804), Enlightenment Philosophy.
```

```
#       Works: Critique of Pure Reason (1781), Critique of Judgement (1790)

tempPersonID <- autoincrement(philDB$person)
temp <- data.frame(id = tempPersonID,
                   name = "Immanuel Kant",
                   born = "1724",
                   died = "1804",
                   school = "Enlightenment Philosophy",
                   stringsAsFactors = FALSE)
philDB$person <- rbind(philDB$person, temp)

rm(temp) #I like my code like I like my room: clean and tidy :)

tempBookID1 <- autoincrement(philDB$books)
temp <- data.frame(id = tempBookID1,
                   title = "Critique of Pure Reason",
                   published = "1781",
                   stringsAsFactors = FALSE)
philDB$books <- rbind(philDB$books, temp)

rm(temp) #I like my code like I like my room: clean and tidy :)

tempBookID2 <- autoincrement(philDB$books)
temp <- data.frame(id = tempBookID2,
                    title = "Critique of Judgement",
                    published = "1790",
                    stringsAsFactors = FALSE)
philDB$books <- rbind(philDB$books, temp)

rm(temp) #I like my code like I like my room: clean and tidy :)

temp <- data.frame(id = autoincrement(philDB$works),
                   personID = tempPersonID,
                   bookID = tempBookID1,
                   stringsAsFactors = FALSE)
philDB$works <- rbind(philDB$works, temp)

rm(temp) #I like my code like I like my room: clean and tidy :)

temp <- data.frame(id = autoincrement(philDB$works),
                    personID = tempPersonID,
                    bookID = tempBookID2,
                    stringsAsFactors = FALSE)
philDB$works <- rbind(philDB$works, temp)

rm(temp) #I like my code like I like my room: clean and tidy :)
rm(tempPersonID) #I like my code like I like my room: clean and tidy :)
rm(tempBookID1) #I like my code like I like my room: clean and tidy :)
rm(tempBookID2) #I like my code like I like my room: clean and tidy :)

# --- SECTION 2: Output information in desired format -----------------
#    Write and submit code that lists the books in alphabetical order,
#    followed by the author and the year of publishing. Format your output like:
#    "Analects" - Kongzi (220 BCE)
#    Show the result.

sel <- order(philDB$books$title)
pID <- philDB$books$id[sel]
sel <- numeric()
for (ID in pID) {
  sel <- philDB$works$personID[which(philDB$works$bookID == ID)]
  cat(sprintf("\"%s\" - %s (%s)", philDB$books$title[ID], philDB$person$name[sel], philDB$books$published[ID]))
  cat("\n")
}

rm(sel) #I like my code like I like my room: clean and tidy :)
rm(pID) #I like my code like I like my room: clean and tidy :)
rm(ID) #I like my code like I like my room: clean and tidy :)

#   The output of this is:
#   "Analects" - Kongzi (220 BCE)
#   "Being and Time" - Martin Heidegger (1927)
#   "Critique of Judgement" - Immanuel Kant (1790)
#   "Critique of Pure Reason" - Immanuel Kant (1781)
#   "Daodejing" - Laozi (530 BCE)
#   "On the Way to Language" - Martin Heidegger (1959)
#   "Zhuangzi" - Zhuangzi (300 BCE)

# END
```

My E-value is: 6e-39.

The NCBI protein database page link is: here (https://www.ncbi.nlm.nih.gov/protein/XP_018701899.1?report=genbank&log$=protalign&blast_rank=1&RID=W4ASBWSC015)

My `MBP1_BABIN.json` file is:

```
[
  { "name" : "MBP1_ISAFU",
    "RefSeqID" : "XP_018701899",
    "UniProtID" : "A0A167PNW1",
    "taxonomyID" : 1081104,
    "sequence" : [
       "MVKAAPPPPP TGPGIYSATY SGIPVYEYQF GHELKEHVMR RRHDDWINAT",
       "HILKAAGFDK PARTRILERD VQKDVHEKIQ GGYGKYQGTW IPLESGEALA",
       "HRHSVYDRLR PMFEYVPGDQ SPPPAPRHAS KPKAPKVKPP LPKWGASKPK",
       "KSPAVATETH TLVPGTVPMR DDFENTDSMI VDDDTPDNIT IASASYMADE",
       "DRYDMAHVST GHRRRKREEH LHDLAEQQHA MYGDELLDYF LLSRNGGQAI",
       "VKPDPPPNFQ SDWPIDAENH TALHWACAMG DISVVRQLKR FNASISAKNA",
       "RGETPFMRSV NFTNCYERQT FSDVMKELWD TVTAQDVSGC TVIHHAAIMK",
       "NCRLYSPTCS RFYLDSILTR LDSHLSPSAL QQMLDIQDSD GNTALHLAAQ",
       "RNARKCVRSL LGRNASTDIP NNEGIRAEDL IAELNAAKKD PGPQRSSSPF",
       "APDSQRHASF RDALKDKPPK KTRKIFGSAA ATTVQSRIAP LLEEKFADLA",
       "KSYDEEWNVK DRAESEARRI LTNTQAELQA AHEQIAKLET QLEPDEVATQ",
       "VTNEANLAKH HVLSLITHQN RLHISQTADG ELSRINGDGG QDESYEERLN",
       "LARQLSHMIA EQRIAETEYV DALSMVGVGD KIDKYRRLLK RCLDPKDSES",
       "LDNNIDSVIG LMEEDRDLQN RHGPAEEPDP MDIPVGI"
       ]
  }
]
```

My `MYSPEtaxonomy.json` file is:

```
[
  { "ID" : 1081104,
    "species" : "Isaria fumosorosea"}
]
```

The result of the `biCode(myDB$taxonomy$species) %in% biCode(MYSPE)` command is:

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

The result of the `myDB$protein$taxonomyID %in% myDB$taxonomy$ID[(myDB$taxonomy$species == MYSPE)]` command is:

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[14] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[27] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[40] FALSE FALSE FALSE FALSE FALSE  TRUE
```

Conclusion and Outlook

As expected the only parts similar are the end. This entire unit was very fun as it helped show how we can make a rudimentary database in R.

The next unit that I plan to take is Graphs & Networks (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-MAT-Graphs_and_networks) .

3.6 Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Databases)

Objective

The main objective here is to expand my knowledge of Bioinformatics databases to help introduce later databases such as the NCBI and EBI.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-23; date completed: 2018-10-23

Activities

I read the introductory notes and all of the required NAR articles and issues. It's really cool, I never expected biologists to take so many leaves out of the CS book!

Conclusion and Outlook

The next unit that I plan to take is Molecular Structures (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Concepts) .

3.7 Molecular Structures (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Concepts)

Objective

The main objective here is to learn about how molecules are structured and their computational analysis.

Time estimated: 0.5 h; taken: 1 h; date started: 2018-10-23; date completed: 2018-10-23

Activities

This unit was tough. There is a lot I didn't understand and had to do external reading about (biologically and chemically, especially for different types of bonding)! But in the end I think I've tackled this unit well.

The 2 main analysis methods examined were: X-ray crystallography and NMR spectroscopy (neither of these are imaging, they are methods of building consensus models).

Other analysis methods are: homology models, electron diffraction, neutron diffraction and single molecule diffraction with X-ray lasers (atomic resolution is not trivial to implement).

It was also fun to read about certain algorithms overfitting the data creating less plausible models (perhaps ML advances may help address this in the future, such as the use of regularisers). I am excited to use ChimeraX for visualisations! I am especially excited to learn how to view stereo images.

Conclusion and Outlook

The next unit that I plan to take is Structure Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PDB)

3.8 Structure Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PDB)

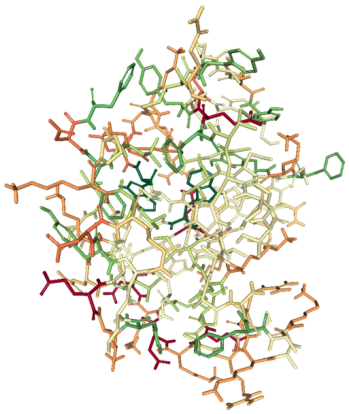
Objective

The main purpose here is to learn how to understand the purpose of and learn how to navigate PDB databases (in particular the RCSB PDB).

Time estimated: 0.5 h; taken: 1 h; date started: 2018-10-23; date completed: 2018-10-23

Activities

I read through all the nodes and navigated the RCSB PDB thoroughly. I identified the following three Saccharomyces cerevisiae Mbp1 transcription factor structures (listed by PDB ID): 1MB1 (http://www.rcsb.org/structure/1MB1) , 1L3G (http://www.rcsb.org/structure/1L3G) & 1BM8 (http://www.rcsb.org/structure/1BM8) . The 1BM8 with all the settings as desired (coloured by hydrophobicity with liquorice style) looks like this:



Conclusion and Outlook

The next unit that I plan to take is Information Theory (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Information_theory) .

3.9 Biomolecular Function Concepts (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Concepts)

Objective

Here I want to learn the main concepts behind biomolecular functions (in particular representation, annotation and prediction).

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-23; date completed: 2018-10-23

Activities

A short and sweet unit. Bioinformatics function does not exist. I'm happy with that. Covered all the lecture notes with some extra wiki searches as usual.

Conclusion and Outlook

The next unit that I plan to take is FASTA (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-FASTA) .

3.10 Measuring Sequence Similarity (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Similarity)

Objective

Here I want to figure out how to measure similarity of amino acid sequences in order to build the groundwork for sequence alignment.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-24; date completed: 2018-10-26

Activities

Went through the notes in great detail. Scoring matrices are used in sequence alignment to represent a variety of different models of similarity (identity, biophysical similarity, genetic code similarly, PAM or BLOSUM). Here my answers to the Q&A:

Q: Compare an identical match of histidine with an identical match of serine. What does this mean?

A: This means we have two sequences that we want to analyse and we have detected histidine in one area of the first sequence and serine in another area of the second sequence. We want to compare these matches for similarity to help us with sequence alignment. In this table we can see that histidine and serine have a similarity of -1.

Q: How similar are lysine and leucine, as compared to leucine and isoleucine? Is this what you expect?

A: Lysine and leucine have a similarity of -2 while leucine and isoleucine have a similarity of 2. This is indeed what I expected.

Q: PAM matrices are sensitive to an interesting artefact. Since W and R can be interchanged with a single point mutation, the probability of observing W→R and R→W exchanges in closely related sequences is much higher than one would expect from the two amino acid's biophysical properties. Why?

A: This is known as the extrapolation problem (as explained in Prof. Steipe's notes). This is because extrapolation to large PAM distances causes problems. If two amino acids have similar codons then mutating from one to the other is likely at the very close evolutionary distance of proteins in the Dayhoff dataset. Also, since evolution favours secondary mutation, introducing amino acids that are biophysically more compatible causes R->W to be unlikely in distantly related pars. However, in the Dayhoff model large evolutionary distances are extrapolated by repeatedly multiplying the matrix with itself so this discrepancy gets amplified and bring the result to almost identity.

Q: PAM matrices were compiled from hypothetical point exchanges and then extrapolated. Therefore these matrices assign a relatively high degree of similarity to (W, R), that is not warranted considering what actually happens in nature. Do you see this problem in the BLOSUM matrix?

A: No I do not.

Q: If BLOSUM does not have this issue, why not?

A: BLOSUM does not have this issue since is compiles matrices directly from blocks of ungapped alignments of sequences at given evolutionary distances.

Conclusion and Outlook

Different methods of similarity are absolutely critical for sequence alignment and I should always use BLOSUM unless otherwise specified! The next unit that I plan to take is System models (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SYS-Concepts) .

3.11 System Models (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SYS-Concepts)

Objective

Here I want to practice building a system model in bioinformatics.

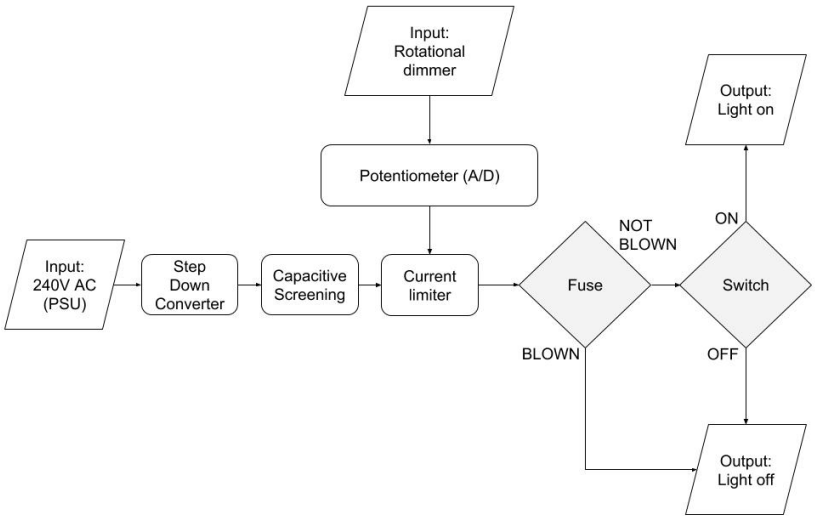
Time estimated: 0.5 h; taken: 1 h; date started: 2018-10-26; date completed: 2018-10-26

Activities

Having already taken a formal course on signals and systems hopefully this course should not bee too challenging.

Important definition: a system (with conceptual levels of bottom up, middle out and top down) is a collection of collaborating genes that have more significant relationships among each other than to genes that are not system members.

Similarly to Prof. Steipe, I don't like UML conventions either. So here is my architectural diagram of a simple desk light:



Conclusion and Outlook

Systems are a powerful abstraction of many fields of knowledge including bioinformatics. The next unit that I plan to take is R RegEx (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-RegEx) .

3.12 Sequence Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Concepts)

Objective

Here I want to learn the basics of sequence analysis in bioinformatics.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-26; date completed: 2018-10-26

Activities

Learning all about the seven "Cs" of bioinformatics: Composition (e.g. molecular weight), Concatenation (e.g. hydrophobicity), Comparison (e.g. binding sites), Correspondence (e.g. multiple alignment), Conservation (e.g. evolutionary pressure), Context (e.g. domain annotation), Collaboration (e.g. co-expression via "guilt by association").

Conclusion and Outlook

The next unit that I plan to take is Protein Protein Interaction (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PPI-Concepts) .

3.13 Protein Protein Interaction (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PPI-Concepts)

Objective

Here I want to learn the basics of sequence analysis in bioinformatics.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-26; date completed: 2018-10-27

Activities

Critical information:

- Think of biomolecular analysis as what happens when A and B interact in manner C
- Complex membership can be interpreted with spoke model, matrix model or complex model (Venn-like diagram)

Conclusion and Outlook

This unit was useful to understand the high level abstractions of sequence analysis in bioinformatics. The next unit that I plan to take is Physical vs Genetic Interactions (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-PPI-Physical_vs_genetic) .

3.14 Physical vs Genetic Interactions (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-PPI-Physical_vs_genetic)

Objective

Here I want to learn the basics of physical and genetical interactions, how they differ and what impact this has in bioinformatics.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-27; date completed: 2018-10-27

Activities

I must say I like the name of "synthetic lethal" protein-protein interactions. Next time I break both wrists at the same time I know what I will tell my friends...

Conclusion and Outlook

A short and sweet high-level unit. The next unit that I plan to take is PPI Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PPI-Databases) .

3.15 PPI Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PPI-Databases)

Objective

The main goal here is to learn where to find PPI databases and how to use them.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-27; date completed: 2018-10-27

Activities

There are 283 interactions from 4 databases for MBP1. Most of the interactions come from the regulatory protein SWI6 which makes sense. Most of the interactions are genetic in nature (83%).

Conclusion and Outlook

This unit was very useful. Examining PPI databases will be invaluable to understand MYSPE better. The next unit that I plan to take is Sequence Composition (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Composition) .

3.16 Sequence Composition (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Composition)

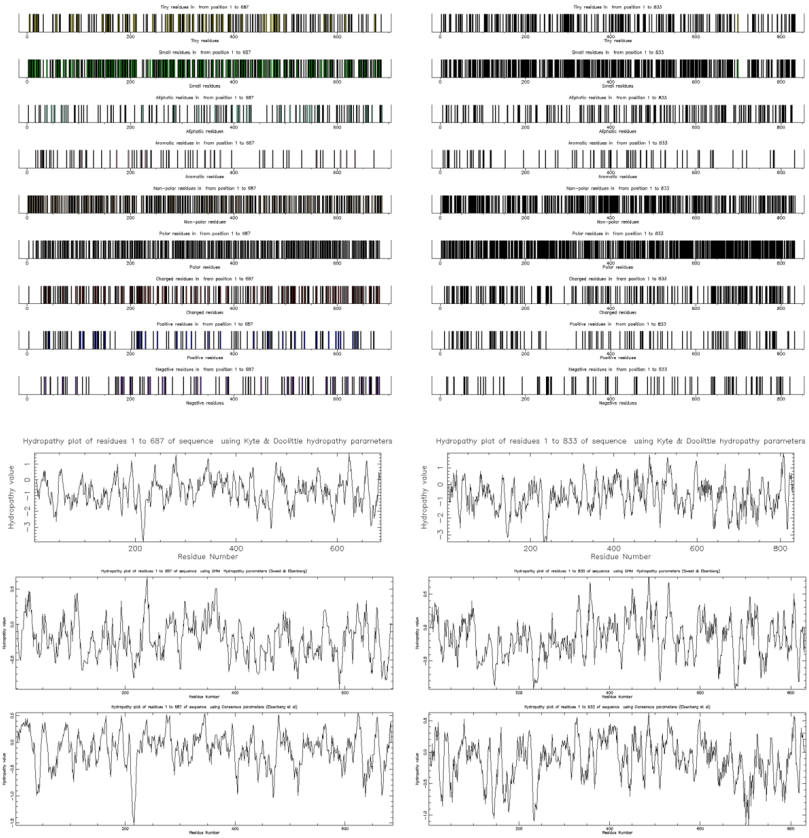
Objective

I want to investigate here how to analyse sequences using the first C of bioinformatics: Composition.

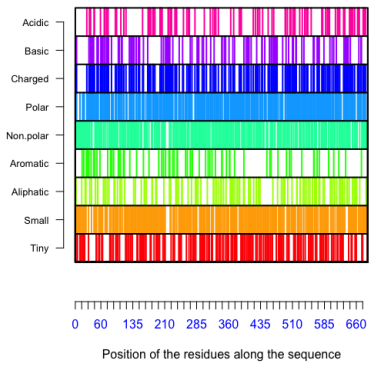
Time estimated: 1 h; taken: 1 h; date started: 2018-10-27; date completed: 2018-10-27

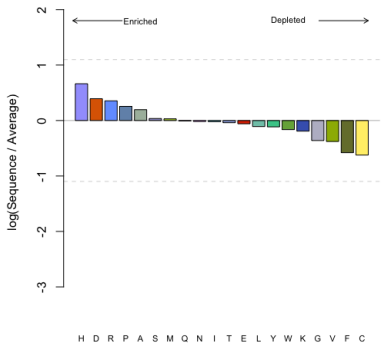
Activities

Using the EMBOSS explorer I ran MYSPE and Baker's Yeast MBP1 through. Indeed it is quite hard to compare for now since there is no expectation, overlay or formal techniques. Here is the data I exported:



From the R tutorial I found out that MYSPE has an isoelectric point of 5.973908 and a molecular weight of 77054.5. The aggregate properties of MYSPE extracted using R can be seen as follows:





Although this data is not fully interpretable as I do not have a perfect baseline I can infer a few characteristics. Indeed the three most enriched amino acids are histidine, aspartic acid and arginine. Similarly, the three most depleted amino acids are cysteine, phenylalanine and valine.

Conclusion and Outlook

Wow. Sequence analysis by composition can go a long way in predicting the structure/role of a protein... The next unit that I plan to take is Sequence Comparison (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Comparison) .

3.17 Sequence Comparison (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Comparison)

Objective

I want to investigate here how to analyse sequences using the second C of bioinformatics: Comparison.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-27; date completed: 2018-10-27

Activities

I like the idea of the consensus sequence. There are strong parallels from this topic to topics in the machine learning realm that I am currently studying. Indeed, already being familiar with Markov models and Neural Networks was of good use for this unit.

When passing MYSPE and MBP1 on EMBOSS' pepcoil program I found out both MYSPE and MBP1 have coiled-coil sequences. MYSPE has 1 sequence with a probability of 1.0 while MBP1 has 2 sequences with probabilities of around 0.5 each (approximately comparable regions).

Similarly when passing MYSPE through EMBOSS' tmap program, it correctly identified that there are no TM helices. In addition, tmap did find all ten TM-helices in Gef1.

Conclusion and Outlook

Sequence Comparison is more complicated than the composition method but also provides significant insight. The next unit that I plan to take is Sequence Collaboration (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Collaboration) .

3.18 Sequence Collaboration (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SEQA-Collaboration)

Objective

I want to investigate here how to analyse sequences using the third C of bioinformatics: Collaboration.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-27; date completed: 2018-10-27

Activities

I did not find MYSPE in KEGG unfortunately. The full lineage is here: *cellular organisms; Eukaryota; Opisthokonta; Fungi; Dikarya; Ascomycota; saccharomyceta; Pezizomycotina; leotiomyceta; sordariomyceta; Sordariomycetes; Hypocreomycetidae; Hypocreales; Cordycipitaceae; Cordyceps*.

COXPRESdb did not return any coexpressed genes. I believe this is because the database is not particularly complete for yeast genes...

Conclusion and Outlook

I don't think KEGG is particularly well designed but this was a fun unit nonetheless. The next unit that I plan to take is Molecular Function Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Databases) .

3.19 Molecular Function Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Databases)

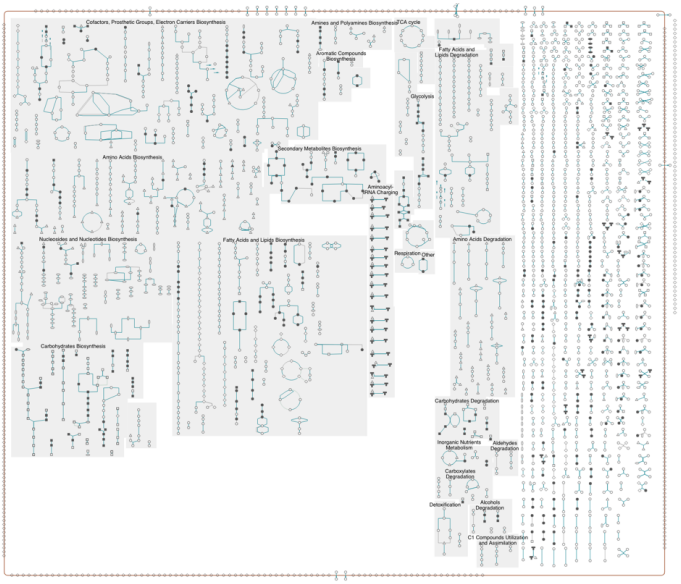
Objective

I just want to know about some of the best databases for molecular function as I am sure this will be useful later!

Time estimated: 1 h; taken: 1 h; date started: 2018-10-27; date completed: 2018-10-27

Activities

Wow... The payway diagram of Saccharomyces cerevisiae astounded me!



Conclusion and Outlook

The amount of databases I come across in bioinformatics never ceases to amaze me... The next unit that I plan to take is Miscellaneous Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Miscellaneous_DB) .

3.20 Miscellaneous Databases (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Miscellaneous_DB)

Objective

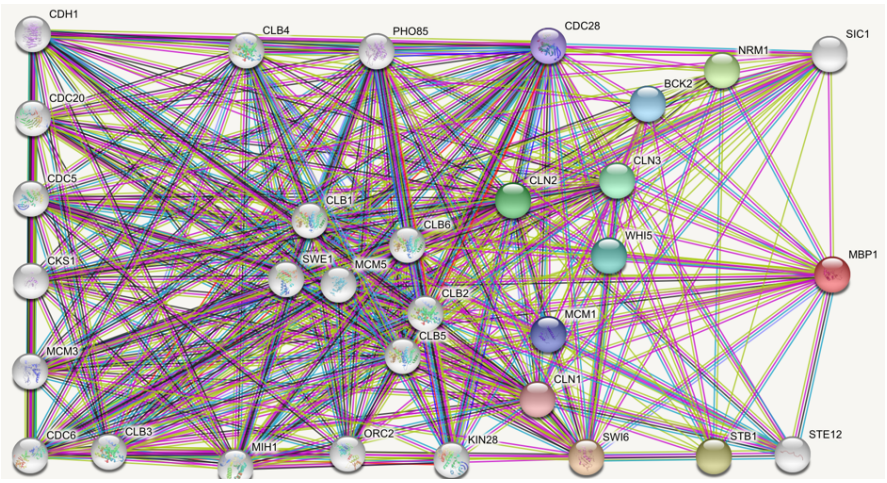
More databases! I want all the databases logged in this journal for future reference when I need information.

Time estimated: 1 h; taken: 1 h; date started: 2018-10-27; date completed: 2018-10-27

Activities

I liked that SGD gives the option to save as a tab-delimited file because I can read those easily in R!

The STRING database is so amazing! It can do k-means clustering (an algorithm I've studied in ML classes) and a whole host of other really cool features! Look what it comes up with for MBP1 (I tried to make 2 levels readable but it is not an easy task):



Conclusion and Outlook

I would never have expected so much database computing in biology... The next unit that I plan to take is Mutation Impact (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Mutation_impact) .

3.21 PPI Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PPI-Analysis) (submitted: 6 marks)

Objective

Another learning unit for submission! Objective is to get full marks of course!

Time estimated: 2 h; taken: 1 h; date started: 2018-10-31; date completed: 2018-10-31

Activities

The subpage created for submission can be found here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_PPI_Analysis) . The code developed is as follows:

```
# Thomas Hollis (BCH441, University of Toronto) -v2.1
#     - Purpose: PPI Analysis learning unit (printing ensembleID info)
#     - Bugs & issues: no bugs, no issues, no warnings

for (ID in ENSPsel) {
  CPdefs[[ID]] <- getBM(filters = "ensembl_peptide_id",
                        attributes = c("hgnc_symbol",
                                       "wikigene_description",
                                       "interpro_description",
                                       "phenotype_description"),
                        values = ID,
                        mart = myMart)

  symb <- CPdefs[[ID]]['hgnc_symbol'][1,1]
  gDesc <- CPdefs[[ID]]['wikigene_description'][1,1]
  pDesc <- CPdefs[[ID]]['phenotype_description'][1,1]

  cat(" The ID is:", ID, "\n",
      "The first row's HGNC symbol is:", symb, "\n",
      "The first row's wikigene description is:", gDesc, "\n",
      "The first row's phenotype is:", pDesc, "\n \n")
}

# END
```

Conclusion and Outlook

This was an easy learning unit which offered a mere glimpse of all the possibilities for PPI analysis in R. The next unit that I plan to take is Homology (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Homology) .

3.22 Homology (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Homology)

Objective

Prof. Steipe says homology is the most important thing in bioinformatics so my objective is to make sure I have a solid understanding of it!

Time estimated: 1 h; taken: 1 h; date started: 2018-10-31; date completed: 2018-10-31

Activities

Homology - diverged from a common ancester either after speciation (orthologous - close analogues in different species) or after duplication (paralogous - neofunctionalisation or subfunctionalisation). This is a commutative and transitative property.

Analogous - similar (i.e. sharing a property) but not homologous

Conclusion and Outlook

MYSPE (ISAFU) is orthologous to MBP1, as expected. The next unit that I plan to take is Alignment (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Alignment) .

3.23 Alignment (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Alignment)

Objective

Most sequence similarity analysis requires alignment so my main goal here is to find out how to align sequences! (Note: the exceptions being algorithms that measure similarity through an alignment-free approaches)

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-02; date completed: 2018-11-02

Activities

I must say this unit troubled me. It obliterated my pre-existing belief that sequence alignment was mathematical and fairly straightforward. In fact, it presents alignment in its messy and imperfect true light. I wonder what solutions and algorithms currently exist for such major alignment issues!

Conclusion and Outlook

A worrying and sad unit where the dreams of beautiful and perfect sequence alignment was shattered. The next unit that I plan to take is Optimal Sequence Alignment (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Optimal_sequence_alignment) .

3.24 Optimal Sequence Alignment (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Optimal_sequence_alignment)

Objective

After the reality check of the previous unit, I am curious to see the optimal solutions for the various issues with sequence alignment.

Time estimated: 0.5 h; taken: 1 h; date started: 2018-11-02; date completed: 2018-11-02

Activities

Okay, so even if we can never be certain that we have the *correct* alignment we can still make fair assumptions. If the alignment of two sequences is such that their random similarity is highly unlikely, we can assume homology as the most probably reason.

The optimal alignment is the closest we can ever get to the true alignment. Therefore if even the optimal alignment does not support homology then we can state confidently that the correct alignment does not support it either.

So how to we do we do optimal sequence alignment?

Well not easily. For starters it having an NP-hard algorithm whose permutation space is more than the number of particles in the universe is not a good start. Enter: divide and conquer, recursion, momoization and a whole bunch of techniques to make this problem tractable.

Path matrices are an awesome way of showing alignments in a computationally efficient manner. As proved by the Needleman-Wunsch algorithm, the optimal sequence is that which leads to the highest possible sum of all pair scores it contains. Have touched on dynamic programming before, I find this solution elegant and beautiful.

Cargo cult warning here as aligning non-homologous sequences is meaningless.

EMBOS was fun to use!

Here are the results for different programs and different gap opening and gap extensions between ISAFU and SACCE [Expand]

I also know how to do this in R which will be very useful in the future no doubt!

Conclusion and Outlook

This was a very lengthy unit but absolutely fascinating stuff! I feel comfortable in sequence alignment science (and associated pseudoscience). The next unit that I plan to take is Gene Ontology (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-GO) .

3.25 Gene Ontology (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-GO)

Objective

The term ontology is still not crystal clear in my mind so my objective is to really understand what GO is and how GOA is undertaken and stored in databases.

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-03; date completed: 2018-11-03

Activities

I think the main takeaway here is that GO is split into three categories: cellular components, molecular function and biologic components. All annotations in GO have GOA evidence codes to specify the explicit source of knowledge. This seems like a crazy web of knowledge.

Having now visited amiGO and quickGO, the tree graphs and general organisation is truly admirable!

Conclusion and Outlook

Wow GO seems like something that will become very powerful in subsequent units. The next unit that I plan to take is Semantic similarity (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Semantic_similarity) .

3.26 Semantic Similarity (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Semantic_similarity)

Objective

How are genes similar? <- my objective is to be able to answer this question in considerable detail

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-03; date completed: 2018-11-03

Activities

I think the following definition is critical so I will note it down in my journal:

Two genes are considered similar if the nodes they are annotated to are close in the gene ontology (GO).

I think the concept of information gain (or most informative common ancestor) and entropy cropping up here is interesting as it reminds me a lot of my decision tree classifiers in Machine Learning.

It was amazing to run through the R script and see how well bioinformaticians have supported and documented R packages! Want to get the genetic similarity between genes (using optimal assignment)? Easy, just run `getGenSim("gene1", "gene2", similarity = "OA")` and voila!

Conclusion and Outlook

This was very informative and now I know how to compare similarity between genes in R! The next unit that I plan to take is BLAST (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-BLAST) .

3.27 BLAST (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-BLAST)

Objective

I've already seen BLAST databases in a previous learning unit so I have a rough idea how BLAST searches for homologues but I want to know the details!

Time estimated: 0.5 h; taken: 1 h; date started: 2018-11-04; date completed: 2018-11-04

Activities

Homology algorithms having a complexity of O(n^2) means they are too slow for database-wide searches.

Enter: Reciprocal Best-Match (RBM). RBM is defines orthologues by finding the best match in another specie's genome being identical it's own.

BLAST (a non-global heuristic alternative to exact alignment) builds on this by adding extra enhancement and different implementations of RBM to find "High Scoring Pairs".

Cargo-cult warning: high E-values do not mean that the hit is not a homologue, but small E-values do indicate that a pure-chance alignment is unlikely.

Hmm there seems to be a bug in the R-script... *types commands in the console furiously for 30min* ... aha! I found the source of my error! All fixed!

Conclusion and Outlook

This unit took longer than expected as I was bug-hunting. However, at least now I understand what BLAST is (and more importantly what it is not). It is simply put a heuristic algorithm. The next unit that I plan to take is Genome Sequencing (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-Sequencing) .

3.28 Genome Sequencing (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-Sequencing)

Objective

As I wanted to take a break from alignment and BLAST, I moved to this Genome Sequencing unit. My main objective here is to find out how my 23andme test (and other similar sequencing methods) were carried out!

Time estimated: 0.5 h; taken: 1 h; date started: 2018-11-04; date completed: 2018-11-04

Activities

The main methods of sequencing are (note that Mba/Gba is used here to represent the Megabase/Gigabase unit rather than Megabit/Gigabit):

- First generation: traditional sequencing (2001 - 2 Mba/day, up to $6000 (400h 0m)/Mb)
- Second generation: parallel sequencing (2005 - 800+ Mba/day, depending on the technique $25 (1h 41m)-0.05/Mb)
- Third generation: single molecule sequencing (2014 - 10+ Gba/day, $0.01 (0h 1m)-0.05/Mb)

However, a much more complete list can be found on the DNA sequencing Wikipedia page (https://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods) .

Prof. Steipe said, "the golden age of the desktop bioinformaticians may be coming to an end" for cross-genome comparisons but a silver lining is that our desktops get more and more powerful, our DNA complexity remains stable.

Conclusion and Outlook

This unit was absolutely fascinating. My favourite so far! I am even considering buying my own full genome sequence in the future - rather than simply my partial genome that I purchased from 23andme! The earlier methods required a lot of preparation which is no longer needed today. The next unit that I plan to take is Genome Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-Annotation) .

3.29 Genome Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-Annotation)

Objective

So I know how sequence annotation is done but what about full genome annotations? This is the question I want to be able to answer in this unit!

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-04; date completed: 2018-11-04

Activities

Well I could have guessed but the first step is to split the entire genome in subsequences and annotate the subsequences.

Databases: ENCODE (human, mouse, worm & fly data), GOLD (info about current status of genome projects and their datasources).

We must also find the genes for genome annotation. This can be achieved by four methods: analysis by signal (look for translation start sites, attachment sites and boundary signals), analysis by contents (look for characteristic trinucleotide patterns), analysis by homology (look for homologues in other species - most accurate if available) or interpretation of the transcriptome (look at RNASeq data).

Conclusion and Outlook

This was a short but informative unit. Thankfully full genome annotation does not seem too far fetched from previous annotation techniques seen in this course. The next unit that I plan to take is PSI BLAST (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-PSI-BLAST) .

3.30 PSI BLAST (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-PSI-BLAST)

Objective

My objective here is to learn what the PSI BLAST algorithm improves on from the original BLAST, how it was designed and how to run it.

Time estimated: 1 h; taken: 3 h; date started: 2018-11-04; date completed: 2018-11-04

Activities

It seems the Position Specific Iterated (PSI) BLAST is a more sensitive version of BLAST that reduces neutral drift and works as follows:

```
1. BLAST a query against protein DB
```

```
2. Construct multiple sequence alignments from BLAST hits to create a Position Specific Scoring Matrix (PSSM profile).
```

```
3. Query the DB with the PSSM
```

```
4. Estimate statistical significance (E-values) and propose significant hits for inclusion in next iteration.
```

```
5. Go back to step 3 (usually 5 iterations).
```

Helpful advice: don't be overly vague or overly specific, use PSI-BLAST for weak but biologically meaningful relationships between proteins, irrelevant proteins can cause profile corruption (so filter if low complexity regions appear to cause specificity issues by introducing an E-value limit, visually inspect output after each iteration and remove suspicious hits by unchecking certain boxes) and false positive rate around 5%.

On the other hand DELTA-BLAST builds a profile from CDD domains in the query and searches for this in the database. Other BLAST flavours include bl2seq and the PROCAIN server's improved BLAST. Finally we could also use EMBOSS for optimal alignments.

During the search I had some results with particularly low E-values for coverages above 80% so I omitted them. As recommended I noted them here:

```
iter = 2, removed: XP_018706125.1     KilA/APSES-type HTH DNA-binding domain protein [Cordyceps fumosorosea ARSEF 2679]
38.3  38.3    97%        9e-04        20%     XP_018706125.1
iter = 2, removed: Transcription regulator HTH, APSES-type DNA-binding domain protein [Cordyceps fumosorosea ARSEF 2679]
48.7  48.7    86%     2e-07   26%        XP_018703649.1
```

```
iter = 3, removed: XP_018706125.1     KilA/APSES-type HTH DNA-binding domain protein [Cordyceps fumosorosea ARSEF 2679]
40.2  40.2    97%             2e-04   19%             XP_018706125.1
iter = 3, removed: Transcription regulator HTH, APSES-type DNA-binding domain protein [Cordyceps fumosorosea ARSEF 2679]
52.6  52.6    90%     1e-08   24%             XP_018703649.1
iter = 3, removed: Select seq XP_024513299.1  hypothetical protein CNH00150 [Cryptococcus neoformans var. neoformans JEC21]
39.9  39.9    83%     3e-04   20%             XP_024513299.1
```

```
iter = 4, removed: hypothetical protein CC1G_13964 [Coprinopsis cinerea okayama7#130]
39.1  39.1    88%     6e-04   19%     XP_002911924.1
iter = 4, removed: XP_024513299.1     hypothetical protein CNH00150 [Cryptococcus neoformans var. neoformans JEC21]
39.1  39.1    83%     5e-04   19%     XP_024513299.1
```

```
iter = 5, removed: XP_024513299.1     hypothetical protein CNH00150 [Cryptococcus neoformans var. neoformans JEC21]
40.3  40.3    88%     2e-04   17%     XP_024513299.1
iter = 5, removed: XP_002911924.1     hypothetical protein CC1G_13964 [Coprinopsis cinerea okayama7#130]
39.5  39.5    88%     4e-04   19%     XP_002911924.1
```

The above log was useful until convergence after which I proceeded without logging. The results of my PSI-BLAST are as follows:

Cordyceps fumosorosea ARSEF 2679 [ascomycetes]	▼ Next ▲ Previous ⏫ First		
start control protein cdc10 [Cordyceps fumosorosea ARSEF 2679]	146	9e-42	XP_018704740
Transcription regulator HTH, APSES-type DNA-binding domain protein [Cordyceps fumosorosea ARSEF 2679]	142	2e-40	XP_018701899
APSES transcription factor [Cordyceps fumosorosea ARSEF 2679]	110	6e-29	XP_018701244
KilA/APSES-type HTH DNA-binding domain protein [Cordyceps fumosorosea ARSEF 2679]	101	2e-26	XP_018706125
Transcription regulator HTH, APSES-type DNA-binding domain protein [Cordyceps fumosorosea ARSEF 2679]	95.8	4e-24	XP_018703649
Protein kinase-like domain protein [Cordyceps fumosorosea ARSEF 2679]	75.0	1e-16	XP_018703356

Conclusion and Outlook

This unit was very lengthy (waiting for PSI-BLAST and copying data into JSON files was a true test of patience) but worthwhile as I love to have this new skill of being able to PSI-BLAST smartly! The next unit that I plan to take is Multiple Sequence Alignment (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-MSA) .

3.31 Multiple Sequence Alignment (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-MSA) (submitted: 6 marks)

Objective

After all the units before, I am ready to tackle the MSA learning unit and I want to submit it. Again, the objective here is to hoover up as many marks as possible (and of course apply what I have learnt while learning about MSA).

Time estimated: 2 h; taken: 12 h; date started: 2018-11-04; date completed: 2018-11-05

Activities

First things first, my submission page is available here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_MSA) .

Well this unit started off with a two disheartening facts:

1. Optimal pairwise alignments are less good than multiple sequence alignments!

2. Optimal multiple alignment is intractable for more than about 10 sequences!

Objective functions here remind me of cost functions in machine learning... Are we going to minimise/maximise these with gradient descent/ascent? Sadly, no gradient surfing for today. MSA algorithms covered:

- Progressive alignment (e.g. Clustal W): fundamental but superseded with far better algorithms
- Consistency-based alignment (e.g. MUSCA & MEME): fundamental and incorporated into modern algorithms
- Probabilistic alignment (e.g. HMMER): fundamental
- Profile-based alignment (e.g. Gapped BLAST & PSI-BLAST): results can be displayed as an MSA
- Consistency-based progressive alignment (e.g. TCoffee): Prof. Steipe's favourite in terms of usefulness and usability
- General purpose (e.g. MUSCLE & MAFFT): very easy to use

The file formats covered are:

- Aligned FASTA
- CLUSTAL (not the same as the CLUSTAL algorithm)
- MSF (used by EMBOSS' EMAA)

Software examined was:

- JALVIEW (MSA editor)

Finally, the code developed for my submission is as follows:

```
# Thomas Hollis (BCH441, University of Toronto) -v9.3
#     - Purpose: generate overview plot of a full-length alignment in one image
#     - Bugs & issues: no bugs, no issues, no warnings
#     - Acknowledgements: thanks to Prof. Steipe's learning unit on R MSA which was of great help

###### Section 0. Excluded code used to call the function while testing ######
if (FALSE) {
  source(file = "BIN-ALI-MSA.R")
  fullAliPlot(msaT)
  fullAliPlot(msaT, lCol = "firebrick", fCol = "black", colGrad = FALSE)
  fullAliPlot(msaT, lCol = "black", fCol = "orange", colGrad = TRUE)
  fullAliPlot(msaW, lCol = "lightgrey", fCol = "skyblue", colGrad = FALSE)
  fullAliPlot(msaW, lCol = "firebrick", fCol = "black", colGrad = FALSE)
  fullAliPlot(msaW, lCol = "black", fCol = "orange", colGrad = TRUE)
}

###### Section 1. Full alignment plot function ######
fullAliPlot <- function(MsaAAMultipleAlignment, lCol = "lightgrey", fCol = "skyblue", colGrad = FALSE) {

  # 1.1 Copy the input object to local data object to avoid accidental modifications
  data <- MsaAAMultipleAlignment

  # 1.2 Create the background plot
  lenSeq <- nchar(data) # get the sequence length (here: 1269)
  xAxis <- c(1,lenSeq) # create the x axis
  numSeq <- length(data@unmasked) # get the number of sequences being aligned (here: 11)
  yAxis <- c(1,numSeq) # create the y axis
  plot(x = xAxis, y = yAxis, type = "n", ylab = "Sequences", xlab = "Position", main = "Full MSA Plot") # plot

  # 1.3 Pre-compute and create the colour palette
  if (colGrad) { # if colour gradient option is chosen, create the score colour palette
    aliScore <- msaConservationScore(data, substitutionMatrix = BLOSUM62)
    lev <- cut(aliScore, labels = FALSE, breaks = 10)
    myPal <- colorRampPalette(c("lightgrey", "firebrick")) #colour from grey to red
  } else { # else use the fCol to fill the rectangle
    myCol <- fCol
  }

  # 1.4 Add the rectangles and lines to the plot
  for (i in 1:numSeq) { # run through all lines
    currSeq <- unlist(strsplit(as.character(data@unmasked[i]),"")) # split current line to vector of char

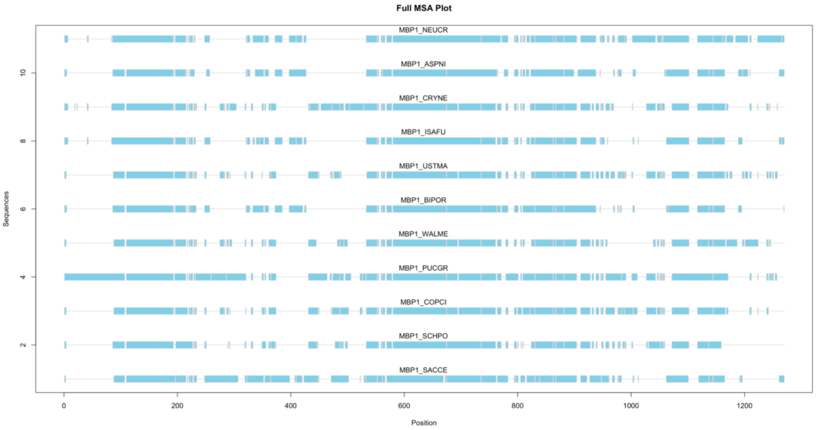
    segRLE <- rle(currSeq == "-") # I forgot about the existance of RLE & spent 3h trying to implement it manually.
    segRLEidx <- cumsum(c(1,segRLE$lengths)) # Eventually gave up & checked Alana Man's Journal which saved me here!
    # Citation: Alan Man, available at: steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Alana_Man

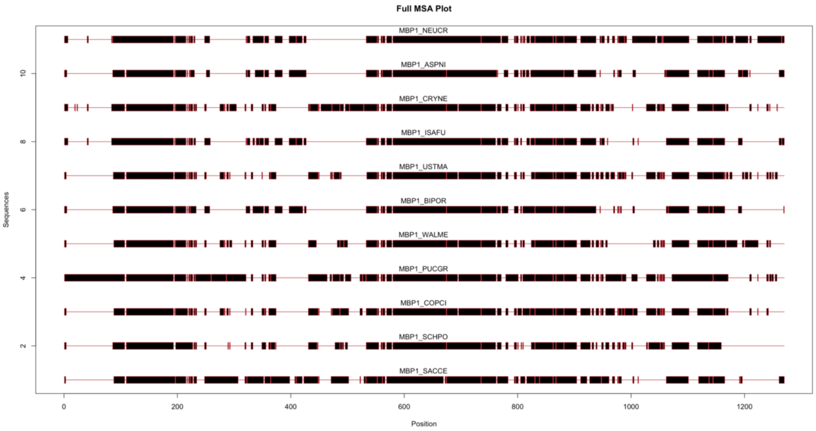
    for (j in 1:length(segRLE$lengths)) { # run through all positions in current line
      if (segRLE$values[j]) { # returns true only for locations with segments
        segments(segRLEidx[j], i, (segRLEidx[j]+segRLE$lengths[j]), i, col = lCol) # plot segments
        } else { # only runs for locations with rectangles
        if (colGrad) { # if colour gradient option is chosen, set the colour
          currScore <- ceiling(mean(lev[segRLEidx[j]:(segRLEidx[j]+segRLE$lengths[j])])) # ceiling for more vibrancy
          myCol <- myPal(10)[currScore] # pick the colour from the colour palette
        }
        rect(segRLEidx[j], i-0.1, (segRLEidx[j]+segRLE$lengths[j]), i+0.1, col = myCol, border = lCol) # plot rect
      }
    }
  }

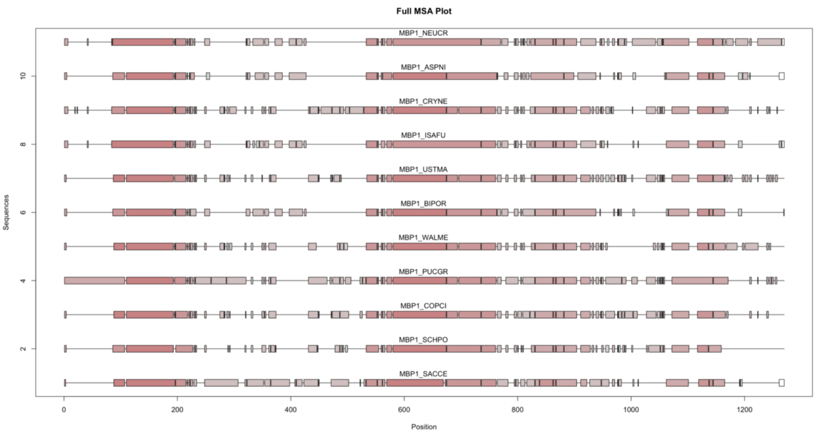
  # 1.5 Add labels to the plot
  labels <- names(data@unmasked)
  for (k in 1:numSeq) { # run through all lines
    text(floor(lenSeq/2), (k+0.25), labels = labels[k]) # place labels in middle
  }
}

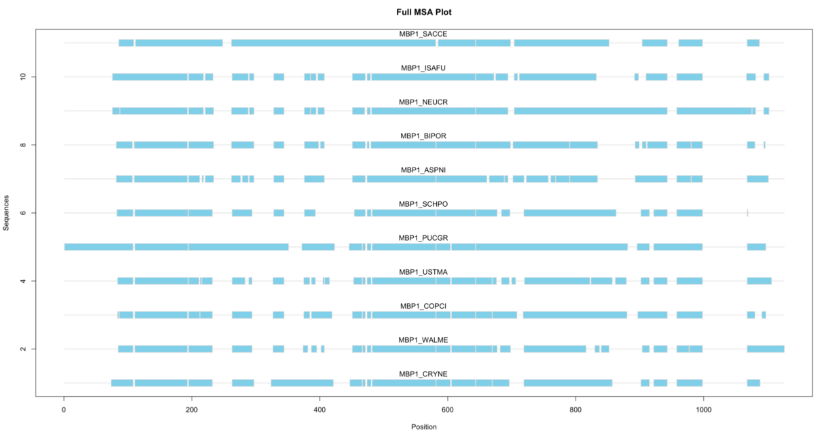
# END
```

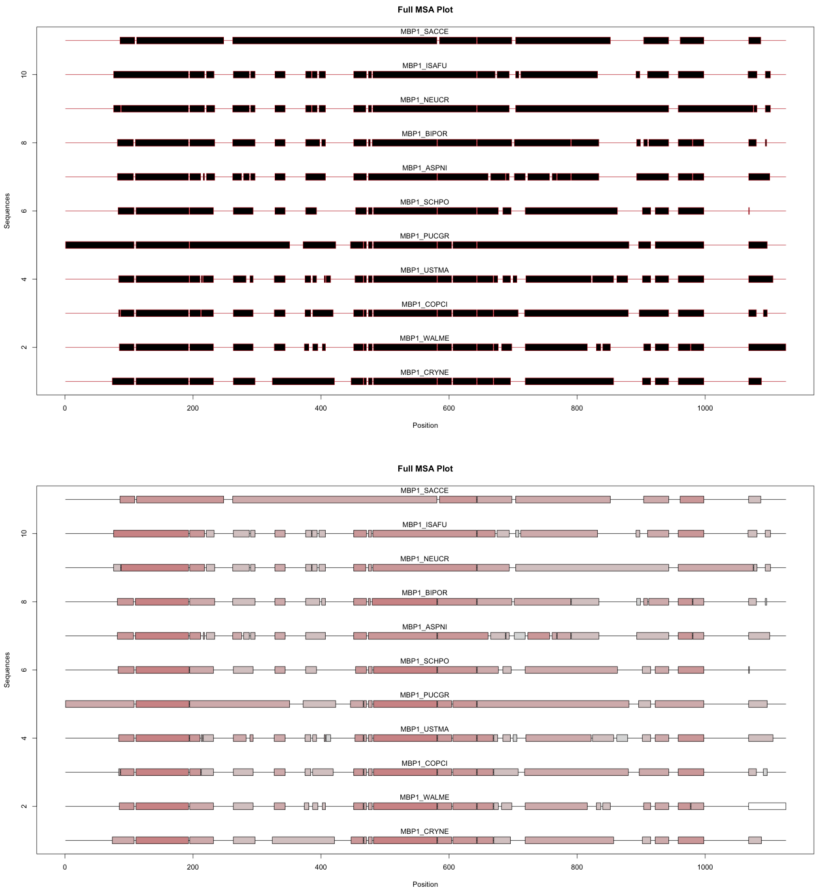
That code took me over 10 hours of non-stop coding to complete! Damn... Here are the beautiful outputs:


Full MSA Plot
MBP1_NEUCR
MBP1_ASPNI
MBP1_CRYNE
MBP1_ISAFU
MBP1_USTMA
MBP1_BIPOR
MBP1_WALME
MBP1_PUCGR
MBP1_COPCI
MBP1_SCHPO
MBP1_SACCE
Sequences
Position
0
200
400
600
800
1000
1200


Full MSA Plot
MBP1_NEUCR
MBP1_ASPNI
MBP1_CRYNE
MBP1_ISAFU
MBP1_USTMA
MBP1_BIPOR
MBP1_WALME
MBP1_PUCGR
MBP1_COPCI
MBP1_SCHPO
MBP1_SACCE
Sequences
Position
0
200
400
600
800
1000
1200


Full MSA Plot
MBP1_NEUCR
MBP1_ASPNI
MBP1_CRYNE
MBP1_ISAFU
MBP1_USTMA
MBP1_BIPOR
MBP1_WALME
MBP1_PUCGR
MBP1_COPCI
MBP1_SCHPO
MBP1_SACCE
Sequences
Position
0
200
400
600
800
1000
1200


Full MSA Plot
MBP1_SACCE
MBP1_ISAFU
MBP1_NEUCR
MBP1_BIPOR
MBP1_ASPNI
MBP1_SCHPO
MBP1_PUCGR
MBP1_USTMA
MBP1_COPCI
MBP1_WALME
MBP1_CRYNE
Sequences
Position
0
200
400
600
800
1000



Conclusion and Outlook

Wow this learning unit was content-heavy and the R code submission was also very challenging! Thankfully I did lots of note taking and I am now happy I successfully implemented MSA in R. The next unit that I plan to take is Concepts of Phylogenetic Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Concepts) .

3.32 Concepts of Phylogenetic Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Concepts)

Objective

I want to be able to understand what phylogeny is and how it is used in bioinformatics for analysis.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-05; date completed: 2018-11-05

Activities

Some important definitions I want to note in my journal:

- Monophyletic group: a clade (node and all its descendants).
- Paraphyletic group: a clade minus some of its members.
- Polyphyletic group: members of several clades, perhaps grouped by a convergence feature or other superficial similarity.
- Cladogram: phylogenetic relationship graph showing only branching pattern, no meaningful branch length.
- Phylogram: most popular phylogenetic relationship graph showing branching where branch lengths indicate amount of genetic change.
- Ultrametric tree: cladograms but where branching point height is proportional to amount of time passed.
- Split network: phylogenetic relationship graph showing collections of splits.
- Horizontal gene transfer: transfer of genetic material other than by parent-offspring transmission (e.g. genetic engineering or gene-modifying viruses like those found in Chagas disease)

Conclusion and Outlook

Great, I got the basics, now only 34'459'425 possible trees to explore! The next unit that I plan to take is Preparing Data for Phylogenetic Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Data_preparation) .

3.33 Preparing Data for Phylogenetic Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Data_preparation)

Objective

I have covered the basic concepts now I want to learn how to implement them in R!

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-06; date completed: 2018-11-06

Activities

A few key facts I wish to note down in my journal before building phylogenetic trees:

0. (Optional) Assume a molecular clock

1. A good tree relies on a good MSA. Do it right.

2. Rows of sequences have to have the same number of characters and must hold aligned characters in the right columns.

3. Data must not include fragments of sequence which have evolved under a different evolutionary model.

4. Add an outgroup to root the tree (i.e. a last common ancestor).

Conclusion and Outlook

Machine Learning methods are currently taking over human inspection for Phylogenetic Data Preparation techniques! It is however useful to know how to implement a quick and dirty approach in R. The next unit that I plan to take is Tree Building (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Tree_building) .

3.34 Tree Building (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Tree_building)

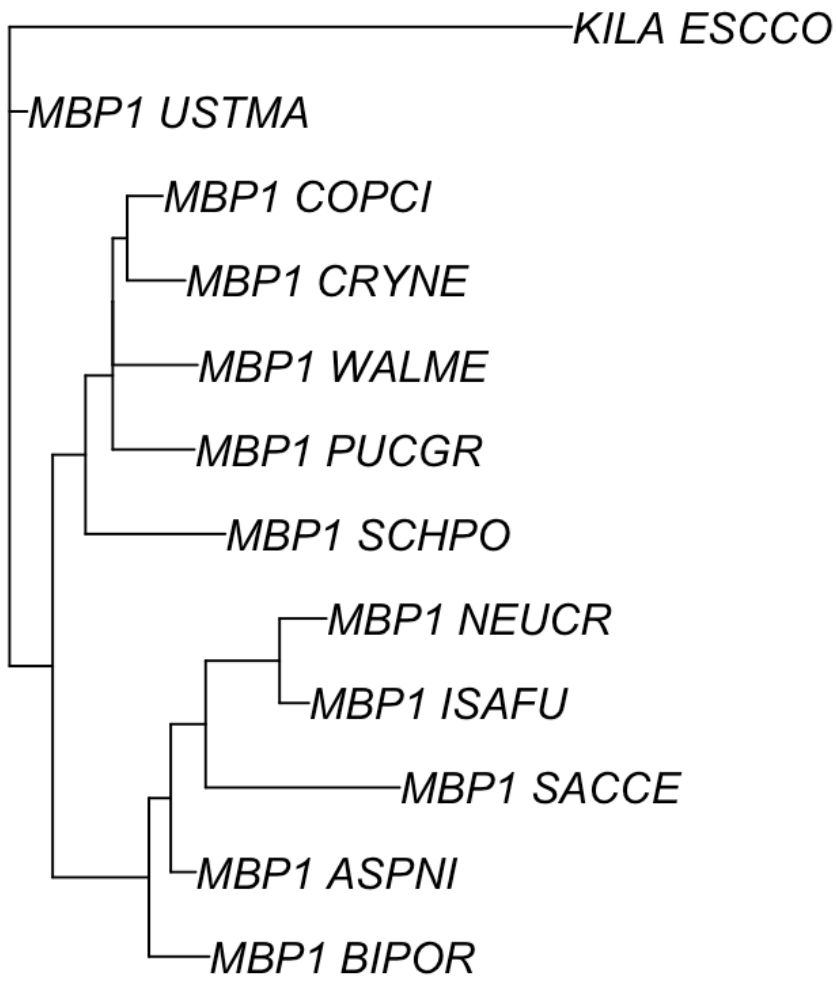
Objective

Now that I know how to prepare the data, my objective is to learn how to build my first phylogenetic tree!

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-06; date completed: 2018-11-06

Activities

Covered all the lectures notes and installed the software as required. Successfully built my first phylogenetic tree using maximum likelihood from the 'proml' package! Very cool! Here it is in all it's glory:



Conclusion and Outlook

A short and straight-to-the-point unit. It looks like MYSPE (ISAFU) is very close (in this tree) to SACCE (as expected) and NEUCR but far from ESCCO (which makes sense as this is the root). Further analysis here is needed and will be done in subsequent units. As the tree was successfully built, I am moving onto the next unit that I plan to take which is Tree Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Tree_analysis) .

3.35 Tree Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Tree_analysis)

Objective

Having finally built my phylogenetic tree my objective is to learn to analyse it and others!

Time estimated: 0.5 h; taken: 1.5 h; date started: 2018-11-06; date completed: 2018-11-06

Activities

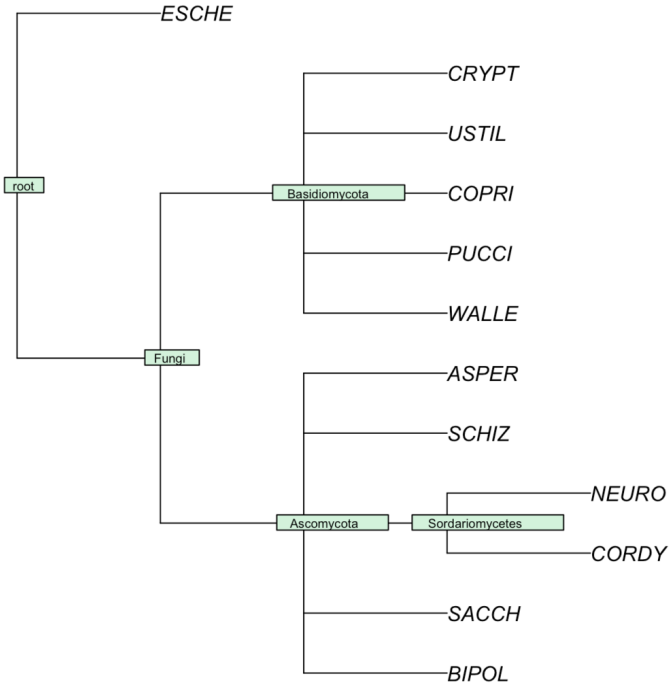
After going through the notes, I fixed an error with the learning unit and sent out the update to the mailing list (pending Prof. Steipe's approval). My fix is inspired from information on Biostars and is as follows:

```
if (!require(treeio, quietly=TRUE)) {
  if (! exists("biocLite")) {
    source("https://bioconductor.org/biocLite.R")
  }
  biocLite("treeio")
  library(treeio)
}

treeText <- readLines("phyliptree.phy")
treeText <- paste0(treeText, collapse="")
fungiTree <- read.tree(text = treeText)
distMat <- cophenetic(fungiTree)

fungiTree$tip.label <- apsTree2$ (0h 8m)tip.label #run this when the first error crops up at the end of the script
```

The learning unit says to follow the script, print a copy of the tree and bring it to class but the script no longer says to do this... So I decided to save some trees and print it here on the Wiki instead:



Conclusion and Outlook

I hope Prof. Steipe accepts my hacky-fix but this made the unit more fun. The next unit that I plan to take is Phylogeny (Intergrator Unit) (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Phylogeny) .

3.36 Genome Browsers (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-Browsers)

Objective

While waiting to finish my phylogeny integrator unit, let's move back to Genome Browsers for a while! I want to find which are the main online resources for this and how to use them.

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-06; date completed: 2018-11-06

Activities

A short and sweet unit. So I took advantage of this to curate my own database of databases! Great stuff I am sure I will find useful in the future. I added the fantastic UCSC Genome Browser Project to the database. An easy choice to make (albeit maybe not as carefully considered as NAR compilations).

Conclusion and Outlook

I know how to use the UCSC Genome Project - sweet. The next unit that I plan to take is NCBI (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-NCBI) .

3.37 NCBI (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-NCBI)

Objective

More databases to look at with the main objective again of understanding what they contain and how to use them.

> Time estimated: 0.5 h; taken: 1 h; date started: 2018-11-06; date completed: 2018-11-06

Activities

Went through the database instructions. Also did a major overhaul of the Entrez page (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/Entrez) .

- MYSPE refSeq ID: `XP_018701899`
- MYSPE UniProt ID: `A0A167PNW1`

Conclusion and Outlook

The NCBI is vast, I still cannot believe the things I find on there. The next unit that I plan to take is EBI (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EBI) .

3.38 EBI (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EBI)

Objective

The NCBI's sibling the EBI also must be learned and loved. My objective here is again to find out what it contains and how to use it.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-06; date completed: 2018-11-06

Activities

Thoroughly went through the EBI website. Not much more to add here.

Conclusion and Outlook

The EBI also seems vast but I don't seem to come across it as often (yet). The next unit that I plan to take is Data Integration (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Data_integration) .

3.39 Data Integration (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Data_integration)

Objective

Now that we have all this data available, my objective is to find out how it all integrates together!

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-06; date completed: 2018-11-07

Activities

Databases interface with... identifiers (a.k.a. IDs), cross references, federated databases (distinct DB, distributed query, merged result), programmable APIs (like in Entrez)!

Here is a snipped of code that I wrote for this unit:

```
myMappedIDs[3,2] <- myIDs$name[match("NP_010038", myIDs$refID)]
myMappedIDs[3,1] <- "NP_010038"
myMappedIDs[4,2] <- myIDs$name[match("NP_013629", myIDs$refID)]
myMappedIDs[4,1] <- "NP_013629"
```

Conclusion and Outlook

Having worked with APIs before this unit was not too surprising but was fun anyway. The next unit that I plan to take is Scripting Data Downloads (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Scripting_data_downloads) .

3.40 Scripting Data Downloads (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Scripting_data_downloads)

Objective

Automation is often key. My goal is to figure out how to script and automate my data downloads!

Time estimated: 0.5 h; taken: 1 h; date started: 2018-11-08; date completed: 2018-11-08

Activities

I had seen status codes before but I guess 200 was new to me since it always defaulted as OK. Interesting!

The two minor functions that I wrote in this unit resemble those by Prof. Steipe so much that I have omitted them here.

Conclusion and Outlook

Scripting data downloads is a powerful tool not to be ignored. It is however worth noting that different websites are often best scraped in different ways so these data mining techniques need to be approached with a flexible mindset. The next unit that I plan to take is Expression Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-Analysis) .

3.41 Expression Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-Analysis)

Objective

The main objective here is to see how genes and expression levels are collected, stored and analysed from databases.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-08; date completed: 2018-11-08

Activities

Slides gone through thoroughly. A bit of extra Wiki searching was helpful here!

Conclusion and Outlook

A short and sweet unit! The next unit that I plan to take is NCBI GEO (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-GEO) .

3.42 NCBI GEO (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-GEO)

Objective

What is the NCBI GEO and how does it work? My objective is to answer this question.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-08; date completed: 2018-11-08

Activities

Well it is a database of gene expressions resulting from measurements. Pretty cool.

Conclusion and Outlook

Another very short unit, interesting nonetheless. The next unit that I plan to take is Discovering Differentially Expressed Genes (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-DE) .

3.43 Discovering Differentially Expressed Genes (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-DE)

Objective

The main objective here is to learn how to use GEO tools to evaluate differentially expressed genes.

Time estimated: 0.5 h; taken: 1 h; date started: 2018-11-08; date completed: 2018-11-08

Activities

Thoroughly went through the short lecture notes. Interesting stuff.

Thoroughly explored GEO2R. Differential expression is when a particular gene is expressed different amounts in two groups suggesting a certain condition could suppress the expression of a particular gene.

Conclusion and Outlook

This was somewhat confusing at first but a bit of wikipedia saved me. I can see why differentially expressed genes can be revelatory in bioinformatics. The next unit that I plan to take is Multiple Testing (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-Multiple_testing) .

3.44 Multiple Testing (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-EXPR-Multiple_testing)

Objective

The main objective here is to learn the fundamentals behind multiple testing.

Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-08; date completed: 2018-11-08

Activities

A very short but concise and interesting unit.

I must say I really loved the light-hearted introduction which reveals the issue of multiple testing with a bar bet!

Conclusion and Outlook

The next unit that I plan to take is GEO2R (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-GEO2R) .

3.45 GEO2R (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-GEO2R) (submitted: 6 marks)

Objective

The main objective here is to get full marks in my GEO2R, R code submission! :D

Time estimated: 6 h; taken: 8 h; date started: 2018-11-08; date completed: 2018-11-16

Activities

My submission page can be found here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_GEO2R) .

Task 1 - What are the data in each cell, column and row?

Here we are looking at the results from a series of experiments looking at the expression levels inside Saccharomyces cerevisiae.

```
> exprs(tmp)   # exprs() gives us the actual expression values.
           GSM81064     GSM81065    GSM81066    GSM81067   GSM81068     GSM81069
YAL001C -0.103485383 -0.045432985  0.03036515 -0.07880245 0.02152000  0.064690128
YAL002W -0.080561906 -0.065335117  0.05850999  0.10747982 0.03228810  0.068920076
YAL003W -0.072594196 -0.013590488 -0.01922754  0.03102694 0.02831172 -0.000978841
YAL004W -0.110246792 -0.020650784 -0.03424255  0.10437975 0.11602340  0.044769045
YAL005C  0.009980532  0.006532799  0.07304193  0.20764822 0.24090600  0.165025279
YAL007C -0.179559127  0.003669272  0.19000509  0.14934748 0.06447802 -0.018602740
```

- Each cell contains the expression value (i.e. how much a particular gene of a particular sample is expressed).
- Each column contains the sample name (representing one experiment) and all the expression levels associated with that sample in particular genes.
- Each row contains feature names (in this case genes) and all the expression levels associated with that gene across multiple experiments' sample.

Task 2 - What are these experiments?

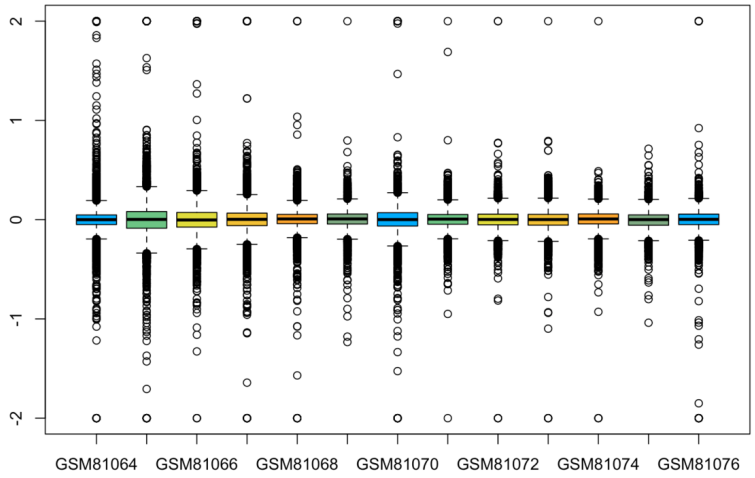
These experiments are (in order):

- GSM81064 - Yeast cell cycle-time point 0 min 2001-05-03_0000.rfm
- GSM81065 - Yeast cell cycle-time point 10 min 2001-05-03_0010.rfm
- GSM81066 - Yeast cell cycle-time point 20 min 2001-05-03_0020.rfm
- GSM81067 - Yeast cell cycle-time point 30 min 2001-05-03_0030.rfm
- GSM81068 - Yeast cell cycle-time point 40 min 2001-04-11_0040.rfm
- GSM81069 - Yeast cell cycle-time point 50 min 2001-04-11_0050.rfm

In each experiment the yeast cell cycle microarray was used to identify global trancription profile. The full description is as follows: "Cells were synchronized with alpha factor and sampled every 10 min across 2 cell cycles. A total of 13 samples were analyzed. No replicates were included. cDNA of the cell cycle samples were labelled with Cy5. For Cy3 labelling, asynchronous yeast population was used."

(Full citation: Pramila T, Miles S, GuhaThakurta D, Jemiolo D et al. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. Genes Dev 2002 Dec 1;16(23):3034-45. PMID: 12464633)

Task 3 - Study this boxplot



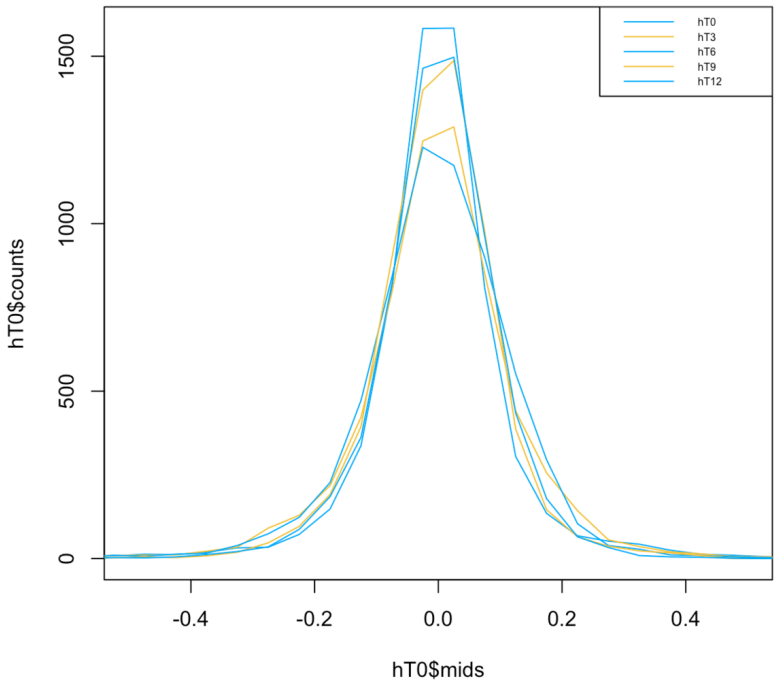
- *Q3.1: What's going on? Are these expression values?*

This box plot shows the distribution of gene expression values for the different samples detailed earlier. Box plots work by labelling the median in the middle, the inter-quartile range on either side of the median and the maximum and minimum quartiles. Each quartile contains 25% of the data points. This shows us how spread the data is and where it is concentrated.

- *Q3.2: What do the numbers that exprs() returns from the dataset mean?*

The `exprs()` command returns the expression measurements as well as error measurements (with controls) of the experiment data (as per the package description). We can confirm this by inspecting the function output. Indeed, we see expression levels for the genes as well as some levels for "blank", "empty" and various controls (like "E. coli control") which.

Task 4 - Study this plot



- *Q4.1: What does it tell you? Is there systematic, global change in the values over time? Within a cycle? Over the course of the experiment?*

This line plot shows that the gene expression distribution is roughly Gaussian-distributed, centred about 0. Initially at T=0, around 3000 genes have expression levels very near 0. There is a systematic decrease in this number of genes near zero from T=0 to T=6 (inclusive). Then from T=9 to T=12 (inclusive) the gene expression values tend to go back to 0.

I could infer from this that certain genes start being more expressed or suppressed between times of 30min and 60min in this experiment.

Task 5 - Row-wise Analysis: Expression Profiles

- *Q5.1: Are all rows genes?*

No. As described earlier, not all rows are genes. Many rows contain "controls", some are "blank" and some are "empty".

- *Q5.2: What identifiers are being used?*

We are using the SGD Nuclear ORF systematic nomenclature convention and format here. Y stands for Yeast, A is the chromosome number (1), and the third letter is either "L" or "R" for the chromosome arm, as described in the SGD page (https://sites.google.com/view/yeastgenome-help/community-help/nomenclature-conventions) .

- *Q5.3: Are all rows/genes unique?*

Indeed all rows/genes are unique since `unique(rownames(ex)) == rownames(ex)` only returns `TRUE` values.

- *Q5.4: Are all yeast genes accounted for?*

As described in the code, we cannot know this yet as we have no information other than the gene identifiers, for now. (answers to come in future tasks no doubt...)

Task 6 - Read a table of features

Code developed:

```
# Thomas Hollis (BCH441, University of Toronto) -v2.3
# Purpose: Read a table of features

# ==   4.1  Task - Read a table of features  ==================================

# This data file is rather typical of datasets that you will encounter "in the
# wild". To proceed, you need to write code to read it into an R-object. Develop
# the code in your script file according to the following specification:
#
#    -  read "./data/SGD_features.tab" into a data frame
#         called "SGD_features"

SGD_original <- read.table(file = "./data/SGD_features.tab", quote = "", header = FALSE,
                           fill = TRUE, sep = "\t", stringsAsFactors = FALSE) #for debugging

SGD_features <- read.table(file = "./data/SGD_features.tab", quote = "", header = FALSE,
                           fill = TRUE, sep = "\t", stringsAsFactors = FALSE)

#    -  remove unneeded columns - keep the following information:
#         -   Primary SGDID (1)
#         -   Feature type (2)
#         -   Feature qualifier (3)
#         -   Feature name - (the systematic name !) (4)
#         -   Standard gene name (5)
#         -   Description (16)

SGD_features <- SGD_features[,-(6:15)]

#    -  give the data frame meaningful column names:

colnames(SGD_features) <- c("SGDID", "type", "qual", "sysName", "name", "description")

#    -  remove all rows that don't have a systematic name. (You'll have to check
#         what's in cells that don't have a systematic name)

(nrow(SGD_features)) #16454
```

```
SGD_features <- SGD_features[!((SGD_features$sysName)==""), ]

(nrow(SGD_features)) #8061, checked visually - OK

#    -  check that the systematic names are unique (Hint: use the duplicated()
#          function.)

anyDuplicated(SGD_features$sysName) == "0" # returns TRUE - OK

#    -  assign the systematic names as row names

rownames(SGD_features) <- SGD_features$sysName

#    -  confirm: are all rows of the expression data set represented in
#                  the feature table? Hint: use setdiff() to print all that
#                  are not.
#                  Example:  A <- c("duck", "crow", "gull", "tern")
#                            B <- c("gull", "rook", "tern", "kite", "myna")
#                            setdiff(A, B)
#                            setdiff(B, A)

feature_names <- SGD_features$sysName
length(featureNames(GSE3635)) # 6228
length(feature_names) # 8061 - at least 1833 missing some that are missing in GSE3635
length(setdiff(feature_names, featureNames(GSE3635))) # 1907 missing

#       If some of the features in the expression set are not listed in the
#       systematic names, you have to be aware of that, when you try to get
#       more information on them. I presume they are missing because revisions
#       of the yeast genome after these experiments were done showed that these
#       genes did not actually exist.
#    -  confirm: how many / which genes in the feature table do not
#                have expression data?

# Answer: Since we saw above we had 1907 missing, thus 1907 genes in SGD_features table
# will not have corresponding expression data in GSE3635.
# All the names of the missing genes can be found by running:
# setdiff(feature_names, featureNames(GSE3635))

#  How should we handle rows/columns that are missing or not unique?

# Answer: Could simply drop them if they are not important. Else if they are
# we could pull the data from another database

# END
```

Task 7 - Selected Expression Profiles

```
# Thomas Hollis (BCH441, University of Toronto) -v2.3
# Purpose: Plot expression profiles for certain genes and study them

# ==   4.2  Selected Expression profiles  ====================================


# Here is an expression profile for Mbp1.

gName <- "MBP1"
(iFeature <- which(SGD_features$name == gName))
(iExprs   <- which(featureNames(GSE3635) == SGD_features$sysName[iFeature]))
plot(seq(0, 120, by = 10),
     exprs(GSE3635)[iExprs, ],
     main = paste("Expression profile for", gName),
     xlab = "time (min)",
     ylab = "expression",
     type = "b",
     col= "maroon")
abline(h =  0, col = "#00000055")
abline(v = 60, col = "#00000055")

# Print the description
SGD_features$description[iFeature]

# Here is a list of gene names that may be involved in the cell cycle switch,
# and some genes that are controls (cf. BIN-SYS-Concepts):

# Turning it on
# Cdc14, Mbp1, Swi6, Swi4, Whi5, Cdc28, Cln1, Cln2, Cln3

# Turning it off
# Rad53, Cdc28, Clb1, Clb2, Clb6, Nrm1

# Housekeeping genes
# Act1, and Alg9

#TASK>  Plot expression profiles for these genes and study them. What do you
#TASK>  expect the profiles to look like, given the role of these genes? What
#TASK>  do you find? (Hint: you will make your life much easier if you define
#TASK>  a function that plots and prints descriptions with a gene name as input.
#TASK>  Also: are the gene names in the feature table upper case, or lower case?
#TASK>  Also: note that the absolute values of change are quite different.
#TASK>  Also: note that some values may be outliers i.e. failed experiments.)

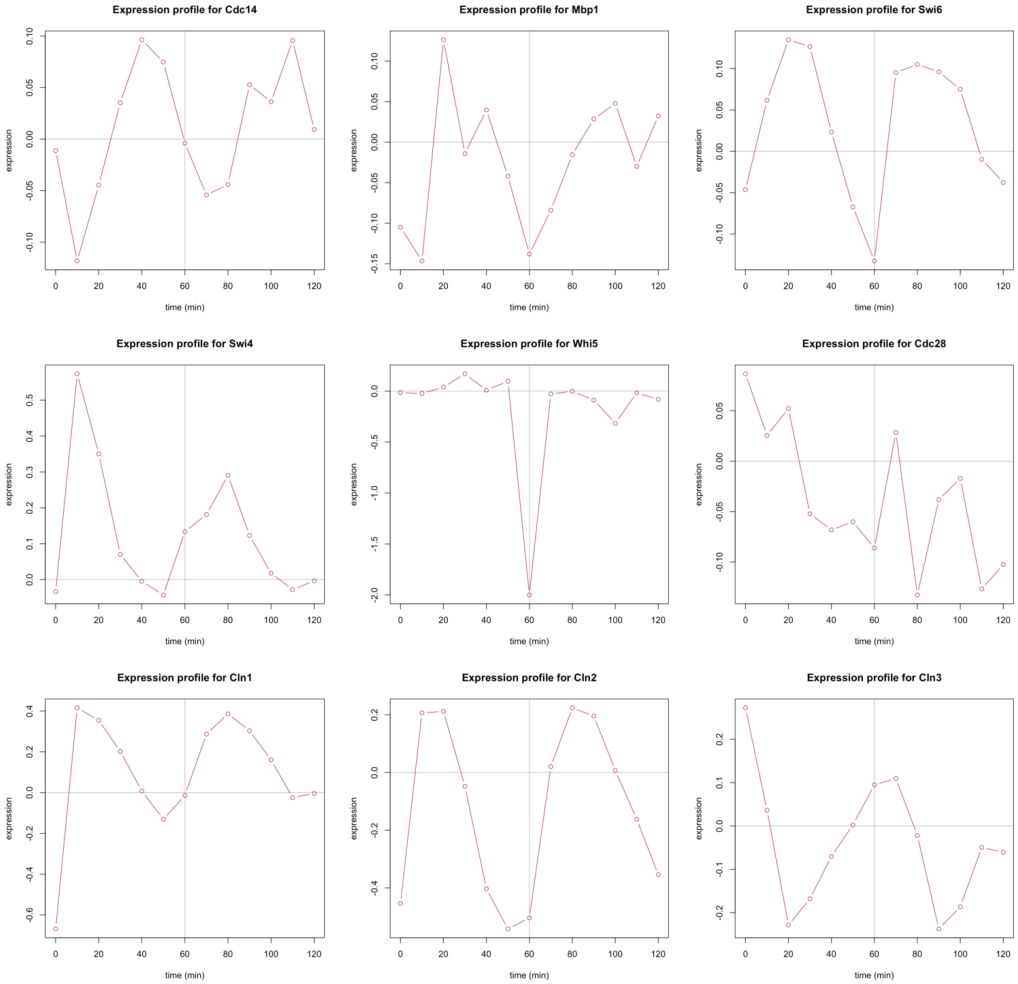
# Indeed the names are in uppercase so I have accounted for this in my function

printExpProf <- function(gName) {
  iFeature <- which(SGD_features$name == toupper(gName))
  iExprs   <- which(featureNames(GSE3635) == SGD_features$sysName[iFeature])
  plot(seq(0, 120, by = 10),
       exprs(GSE3635)[iExprs, ],
       main = paste("Expression profile for", gName),
       xlab = "time (min)",
       ylab = "expression",
       type = "b",
       col= "maroon")
  abline(h =  0, col = "#00000055")
  abline(v = 60, col = "#00000055")

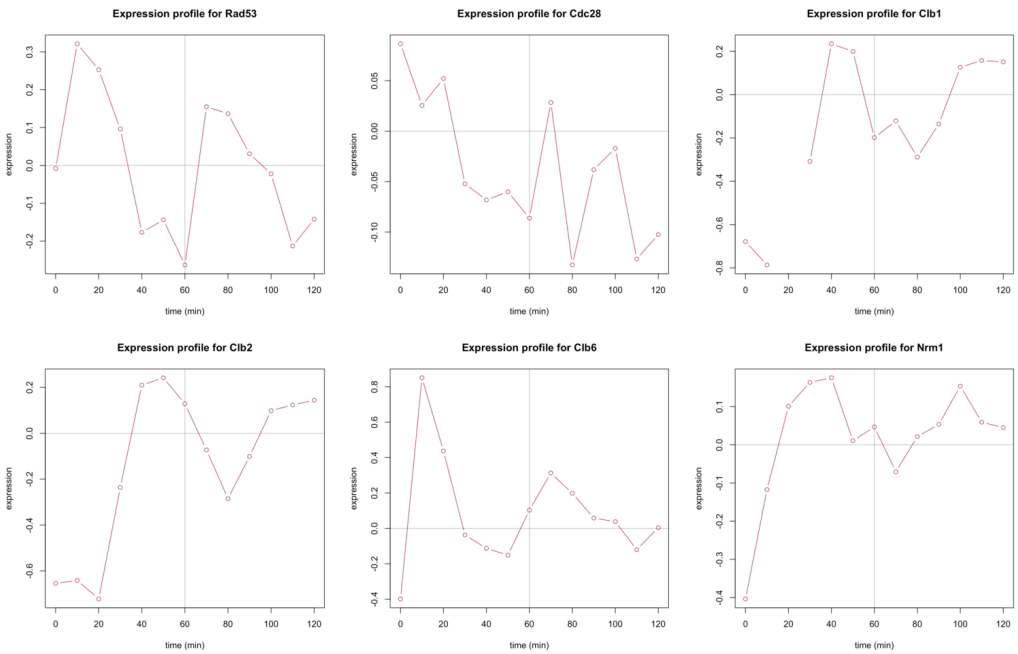
  # Print the description
  SGD_features$description[iFeature]
}

# END
```

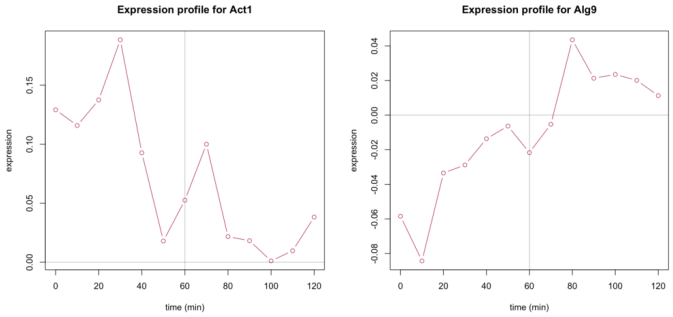
Turn on gene plots can be found below. Indeed, from a pretty lengthy read of wikipedia I would expect turn on genes to have a cyclical pattern. I expect them to rise throughout the cycle and fall back again before the start of the next cycle (roughly 60min cycles here). This seems to be consistent with what most of the expression plots show (albeit Cdc14 and Cln3 seem to have a slightly delayed onset). The only gene that does not seem consistent with my expectations would be Cdc28 as it does not seem to be cyclical. Indeed this might be because it is both a turn on and a turn off gene (as will be discussed in the next section). I also hypothesise that Whi5 might have an odd point at T=60min which, if removed, might become inconsistent with my expectations.



Turn off gene plots can be found below. Indeed, I would expect turn off genes to initially have a low expression (as they are suppressing) before returning to a higher expression. All these plots seem moderately consistent with this hypothesis except Cdc28 which seems to behave unexpectedly again. Once again, I hypothesise there that it is because it is both a turn on and a turn off gene that it has such unusual expression behaviour.



Housekeeping genes plots can be found below. Indeed, I have read that housekeeping genes are needed to maintain basic cellular function. The housekeeping gene wiki page (https://en.wikipedia.org/wiki/Housekeeping_gene) also explains that we can expect genes to be expressed at relatively constant rates (i.e. non-cyclical) in most non-pathological situation, even if this varies depending on experimental conditions. Therefore, I would expect them to be roughly expressed constantly in a non-cyclical manner throughout the cycle. This seems to be consistent with what their expression actually does as neither is cyclical (albeit alg9 has a slightly strange low start which increases expression throughout the cycle). This deviation is nonetheless consistent with literature as expression rates vary on experimental conditions, as seen in the wiki page.



Task 8 - Gene descriptions

Here is the code I developed:

```
# Thomas Hollis (BCH441, University of Toronto) -v2.3
# Purpose: Print descriptions of differentially expressed genes

# ==   5.1  Final task: Gene descriptions  ===================================

#    Print the descriptions of the top ten differentially expressed genes.

top10 <- myTable[1:10,1]
for (g in top10) {
  cat("====================================== \n")
  des <- SGD_features$description[rownames(SGD_features) == g]
  if (length(des) != 0) {
    cat("Gene", g, "has description: \n", des, "\n")
  } else {
    cat("Gene", g, "has description: \n", des)
    cat("Error 404: No description found \n")
  }
}

# END
```

And here is the output:

```
======================================
Gene YMR215W has description:
 Putative 1,3-beta-glucanosyltransferase; has similarity go other GAS family members; low
abundance, possibly inactive member of the GAS family of GPI-containing proteins; localizes
to the cell wall; mRNA induced during sporulation
======================================
Gene YEL032W has description:
 Protein involved in DNA replication; component of the Mcm2-7 hexameric helicase complex
that binds chromatin as a part of the pre-replicative complex
======================================
Gene YBL003C has description:
 Histone H2A; core histone protein required for chromatin assembly and chromosome
function; one of two nearly identical (see also HTA1) subtypes; DNA damage-dependent
phosphorylation by Mec1p facilitates DNA repair; acetylated by Nat4p
======================================
Gene YIL123W has description:
 Protein of the SUN family (Sim1p, Uth1p, Nca3p, Sun4p); may participate in DNA replication;
promoter contains SCB regulation box at -300 bp indicating that expression may be cell
cycle-regulated; SIM1 has a paralog, SUN4, that arose from the whole genome duplication
======================================
Gene YGR098C has description:
 Separase, a caspase-like cysteine protease; promotes sister chromatid separation by
mediating dissociation of the cohesin Scc1p from chromatin; inhibits protein phosphatase
2A-Cdc55p to promote mitotic exit; inhibited by Pds1p; relative distribution to the nucleus
increases upon DNA replication stress
======================================
Gene YML117W-A has description:
 Error: No description found
======================================
Gene YNL031C has description:
 Histone H3; core histone protein required for chromatin assembly, part of
heterochromatin-mediated telomeric and HM silencing; one of two identical histone H3
proteins (see HHT1); regulated by acetylation, methylation, and phosphorylation; H3K14
acetylation plays an important role in the unfolding of strongly positioned nucleosomes
during repair of UV damage
======================================
Gene YJL137C has description:
 Glycogenin glucosyltransferase; self-glucosylating initiator of glycogen synthesis, also
glucosylates n-dodecyl-beta-D-maltoside; similar to mammalian glycogenin; GLG2 has a
paralog, GLG1, that arose from the whole genome duplication
======================================
Gene YOR066W has description:
 Activator of G1-specific transcription factors MBF and SBF; involved in regulation of the
timing of G1-specific gene transcription and cell cycle initiation; localization is cell-cycle
dependent and regulated by Cdc28p phosphorylation; MSA1 has a paralog, MSA2, that
arose from the whole genome duplication
======================================
Gene YDR224C has description:
 Histone H2B; core histone protein required for chromatin assembly and chromosome
function; nearly identical to HTB2; Rad6p-Bre1p-Lge1p mediated ubiquitination regulates
reassembly after DNA replication, transcriptional activation, meiotic DSB formation and H3
methylation
```

Conclusion and Outlook

This was a really tough submission but I learnt a lot as I had to read so much information from Wikipedia! The next unit that I plan to take is Domain Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Domain_annotation) .

3.46 Domain Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Domain_annotation)

Objective

My objective here is to understand the complexities behind domain annotation.

Time estimated: 1 h; taken: 1 h; date started: 2018-11-08; date completed: 2018-11-09

Activities

A domains are the natural unit of analysis of protein structure (used in separate folding, distinct function and modular inheritance). They allow:

- identification of folding regions
- identification of gene fusion or insertion (evolutionary history)
- understand protein mechanism (domain architecture)
- classification of proteins

Some information to note down can be found here:

- 1. From the section on "Confidently predicted domains ..."

The start and end coordinates of the KilA-N domain (...according to SMART, not Pfam, in case the two differ): start = 35, end = 118

All start and end coordinates of low complexity segments: start = 6, end = 14 & start = 131, end = 143

All start and end coordinates of ANK (Ankyrin) domains: start = 268, end = 297 & start = 390 & end = 419

Start and end coordinates of coiled coil domain(s) I expect only one: start = 515, end = 546

Start and end coordinates of AT hook domain(s) I expect some but not all not all Mbp1 orthologues have one: none here

- 2. From the section on "Features NOT shown ..."

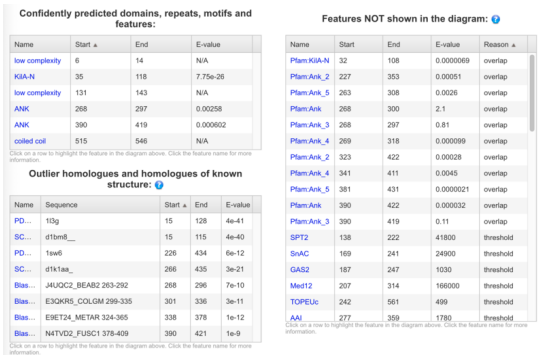
All start and end coordinates of low complexity segments for which the Reason is "overlap": none here

Any start and end coordinates of overlapping coiled coil segments: none here

- 3. From the section on "Outlier homologues ..."

Start and end coordinates of a PDB:1SW6|B annotation (if you have one): start = 226, end = 434

(click here to check the source) (http://smart.embl-heidelberg.de/smart/job_status.pl?jobid=741414813043610154173906 3TbEgRdmLYH)



Confidently predicted domains, repeats, motifs and features:

Name	Start	End	E-value
low complexity	6	14	N/A
KilA-N	35	118	7.75e-26
low complexity	131	143	N/A
ANK	268	297	0.00258
ANK	390	419	0.000602
coiled coil	515	546	N/A

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

Outlier homologues and homologues of known structure:

Name	Sequence	Start	End	E-value
PD...	1l3g	15	128	4e-41
SC...	d1bm8__	15	115	4e-40
PD...	1sw6	226	434	6e-12
SC...	d1k1aa_	266	435	3e-21
Blas...	J4UQC2_BEAB2 263-292	268	296	7e-10
Blas...	E3QKR5_COLGM 299-335	301	336	3e-11
Blas...	E9ET24_METAR 324-365	338	378	1e-12
Blas...	N4TVD2_FUSC1 378-409	390	421	1e-9

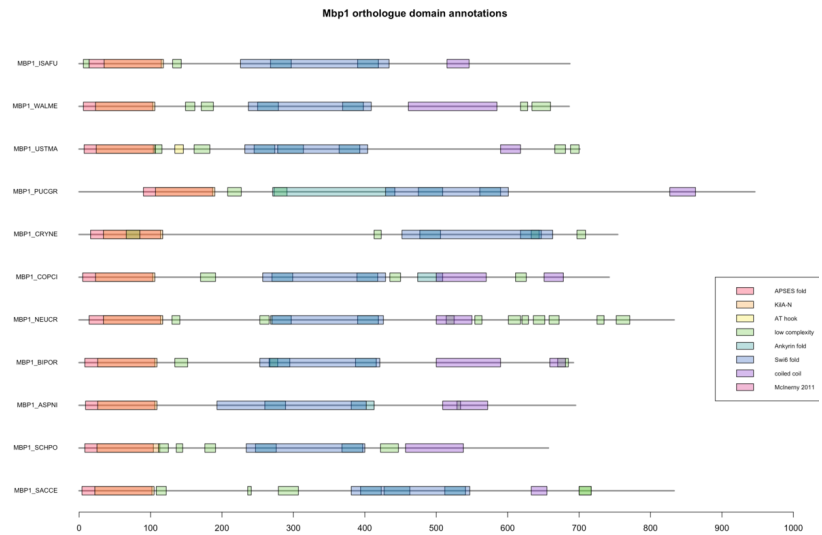
Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

Features NOT shown in the diagram:

Name	Start	End	E-value	Reason
Pfam:KilA-N	32	108	0.0000069	overlap
Pfam:Ank_2	227	353	0.00051	overlap
Pfam:Ank_5	263	308	0.0026	overlap
Pfam:Ank	268	300	2.1	overlap
Pfam:Ank_3	268	297	0.81	overlap
Pfam:Ank_4	269	318	0.000099	overlap
Pfam:Ank_2	323	422	0.00028	overlap
Pfam:Ank_4	341	411	0.0045	overlap
Pfam:Ank_5	381	431	0.0000021	overlap
Pfam:Ank	390	422	0.000032	overlap
Pfam:Ank_3	390	419	0.11	overlap
SPT2	138	222	41800	threshold
SnAC	169	241	24900	threshold
GAS2	187	247	1030	threshold
Med12	207	314	166000	threshold
TOPEUc	242	561	499	threshold
AAI	277	359	1780	threshold

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

Wow, after a lot of editing of the source code I finally got my plot to fit in my screen... I guess RStudio and 4k resolution are not good friends. Here it is in all its glory:



As for my interpretation of this... Well it is clear from the plot that all MBP1 orthologue domains share similarities in the KilA-N region which makes sense as this is the common ancestor. They also have some similarities in the APSES fold which again also makes sense. The other similarities do not quite line up but are still there. I assume more interpretation can be made here which is why I plan to come back to this after having completed further units.

Conclusion and Outlook

Wow, domains are really powerful in helping sequence analysis in bioinformatics. Although inputting the data manually here was tedious. Specificity is an asset to avoid algorithms taking into account irrelevant data (this it turns out is true both for bioinformatics and a whole host of machine learning algorithms that I am studying at the moment). The next unit that I plan to take is Function Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Annotation) .

3.47 Function Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Annotation)

Objective

My objective here is to recap all the units in this area.

Time estimated: 0.5 h; taken: 0 h; date started: 2018-11-10; date completed: 2018-11-10

Activities

This unit seems empty. Since this unit is a recap unit anyway, I didn't expect much material. I was a bit mislead by the "do not work on this unit until it is live" but Wiki history saved me as it confirmed this unit has looked like this since September 2017 so I can safely skip past it!

Conclusion and Outlook

The next unit that I plan to take is Genome Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Genome_annotation) .

3.48 PDB files (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-SX-PDB)

Objective

My objective here is to learn how to read PDB files and how to interpret their contents.

Time estimated: 0.5 h; taken: 1 h; date started: 2018-11-18; date completed: 2018-11-18

Activities

The first file has a resolution of 1.71 Angstroms. The the first residue in the SEQRES section "GLN" also the first residue with an ATOM record ("GLN" too). I think 66 water molecules of solvent were used while there are 248 in the structure. Remark 525 just lists two particular water molecules that are further than 5 Angstroms away.

Conclusion and Outlook

Honestly I struggled a lot with this unit because there was no introductory theory. It was straight into the PDB files with no biological context. I did learn loads of awesome new plotting techniques though! I might revisit this unit later once I have more perspective. The next unit that I plan to take is Chimera (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Chimera) .

3.49 Chimera (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Chimera)

Objective

I want to learn how to use Chimera! Seems like an amazing piece of kit!

Time estimated: 1 h; taken: 2 h; date started: 2018-11-18; date completed: 2018-11-18

Activities

Chimera? ChimeraX? After thinking for a few minutes about the pro's and con's of each I just downloaded and installed both. What is the worst than could happen?

Took all my tutorials with Chimera since this fits much more nicely with the notes. Will rely on this mostly in future units.

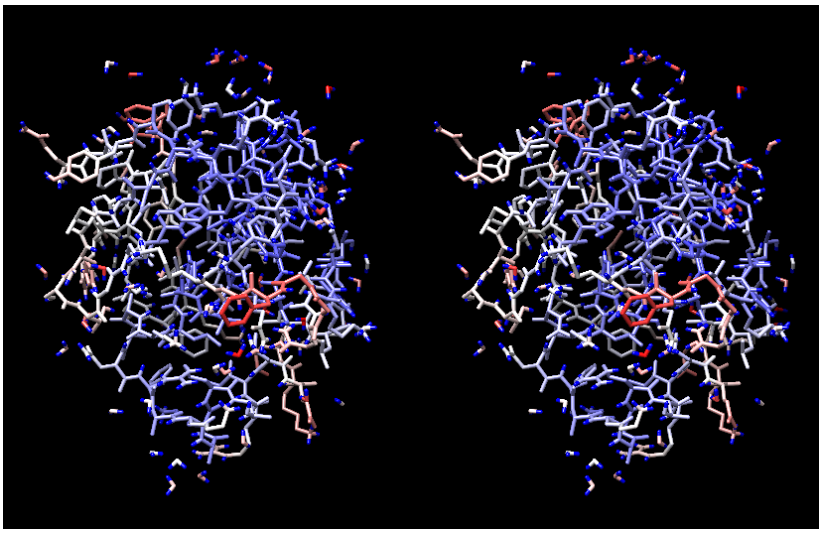
Measuring my pupil separation distance: 6.4 cm.

Molecule seperation distance between equivalent points: 1.15*6.4 = 7.36cm.

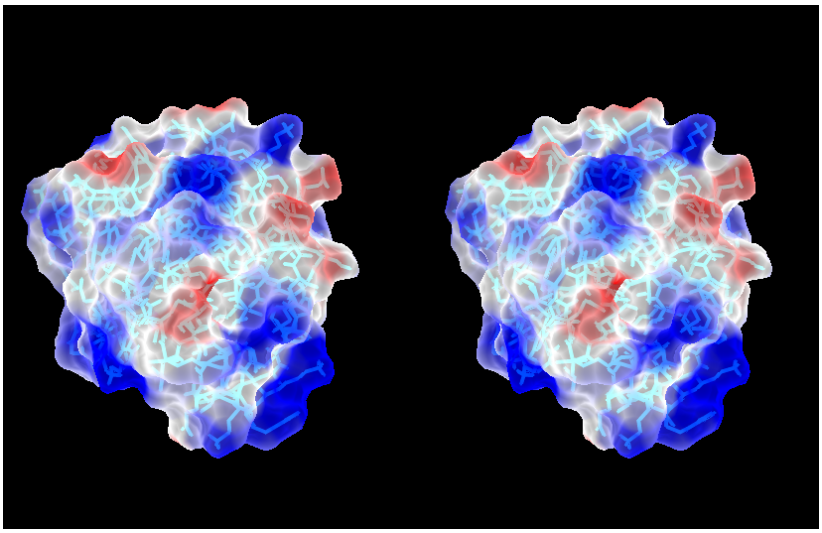
Great now my screen is full of sweat and snot. Fantastic!

In fact, I am actually already getting really close to 'seeing' in 3D. I suspect it has to do with me using Google Cardboard in the past...

Anyway, here is the lovely stereo vision of the structure of the yeast transcription factor Mbp1 DNA binding domain:



And now a quick look at Mbp1's Coulomb (electrostatic) potential



The picture changed so... Chimera uses the implicit sequence!

Conclusion and Outlook

Wow, this program is so mesmerising, I spent hours procrastinating with it! Like Google Maps for molecules... The next unit that I plan to take is Small Molecules (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Small_molecules) .

3.50 Small Molecules (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Small_molecules)

Objective

What is so special about small molecules? How do I use them in Bioinformatics? <- My goal is to answer all of these

> Time estimated: 1 h; taken: 1 h; date started: 2018-11-18; date completed: 2018-11-18

Activities

Apparently, not so much except that I can build their structure myself and get close to the original! Cool! Chimera is so versatile!

Conclusion and Outlook

Another really fun unit. I love using Chimera! The next unit that I plan to take is Molecular Forcefields (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Molecular_forcefields) .

3.51 Molecular Forcefields (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Molecular_forcefields)

Objective

I already know a little about intra and inter molecular forces. I expect this will really stretch my knowledge so my objective is to learn new stuff!

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-18; date completed: 2018-11-18

Activities

Never thought Van der Waals forces would be so significant... Going through the notes was pretty interesting!

Conclusion and Outlook

A short but complex unit requiring lots of wiki background reading (I grew my bookmarks tree quite a lot)! The next unit that I plan to take is Structure Domains (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Domains) .

3.52 Structure Domains (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Domains)

Objective

My main objective here is to find out how domains are manifested in molecular structure!

> Time estimated: 1 h; taken: 1 h; date started: 2018-11-18; date completed: 2018-11-19

Activities

Domains and subdomains... I wonder why when we get to the increasingly small (electron flying around a nucleus) we see resemblances with the increasingly large (galaxy about to be engulfed by a black hole)...

Domain swapping!?!? Biology is so much messier than I thought when you get down to it. It's sad but I wrote a cool insight about it so that's a silver lining.

Conclusion and Outlook

Another very conceptually tough unit that I will definitely come back to (especially for definition divergence analysis with Chimera). The next unit that I plan to take is Structure Superposition (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Superposition) .

3.53 Structure Superposition (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Superposition)

Objective

Wow we can not only superpose sequences but also structure - my objective is to find out how this magic works (I suspect I will have to be careful not to fall into Cargo Cult traps here).

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-11-19; date completed: 2018-11-19

Activities

Went through the lecture notes and had a great read of the wiki article. Root Mean Square Deviation metric can be a good tool to avoid Cargo-Cult here! However, other interesting metrics include VAST and the FATCAT server for structure comparison.

Conclusion and Outlook

Lots of learning here. I don't think I will remember this all so I will probably come back to this unit at the end! The next unit that I plan to take is Homology Modelling (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Homology_modelling) .

3.54 Homology Modelling (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Homology_modelling)

Objective

This looks like a BIG unit so my objective is to put together everything I learnt, maybe go back on some earlier units and understand how to do homology modelling!

> Time estimated: 2 h; taken: 2 h; date started: 2018-11-19; date completed: 2018-11-19

Activities

Energy refinement of homology is Cargo Cult! Cargo Cult seems to spread deep and wide...

I liked the nuance between alignment and superposition so I will write it down here:

- Alignment based on an evolutionary model recovers information on an evolutionary event.
- Superposition shows how the event has been structurally accommodated.

'Slipped' alignments can really mess things up...

I really want to avoid Cargo Cult here so I will also note down the prototype cases for using homology modelling results:

- Analytical: residues involved in catalysis, electrostatic properties, shape, flexibility and dynamics
- Comparative: context specific patterns, predict effect of sequence variation, distinguish physiological and non physiological interactions

I also want to note down some important terminology:

- Target: The protein that you are planning to model.

- Template: The protein whose structure you are using as a guide to build the model.
- Model: The structure that results from the modelling process. It has the Target sequence and is similar to the Template structure.

I chose "4UX5_A" as my template since it has a higher sequence similarity to the target!

The final comparison was pretty impressive...

There was also a type on the last line of the script. It should be `write.pdb(pdb = MYSPEmodel, file=PDB_OUTFILE).`

Conclusion and Outlook

I really struggled here so I think I will go over this unit again as part of my recap (only 5 or so units need recapping). The next unit that I plan to take is Homology Modelling (Integrator Unit) (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Homology_modelling) .

3.55 Extra learning units

Objective

The main objective is to read through extra fundamental modules that are not necessary to complete the course to make sure I am fully equipped.

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

So far the only units I have checked out are Human Genomics (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-Human_genomics) , Dotplot (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Dotplot) , Internal Repeats (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-ALI-Internal_repeats) , Phylogenetic Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Selective_pressure) , Conservation Scores (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PHYLO-Conservation_scores) , Structure Motifs (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Motifs) , Molecular Dynamics (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-MD) , Prediction (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-SX-Ab_initio_prediction) and Next Generation Sequencing (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-NGS_bioinformatics) , which are still stubs. So I decided instead to go ahead and get lost in wikipedia!

Conclusion and Outlook

Although this was not required, it was definitely worthwhile!

4. Statistics

This section will be focussing on learning units that concern mathematical statistical theory and it implications in bioinformatics.

4.1 Probability (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Probability)

Objective

As a student in Machine Learning, the aim of this course is to refresh my knowledge on basic probability theory.

> Time estimated:0.5 h; taken: 0.5 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

This was fun. The touch of humour of Prof. Steipe is definitely appreciated! I liked the examples chosen too. Worked through the entire unit without too much trouble.

Conclusion and Outlook

The next unit that I plan to take is Probability Distribution (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Probability_distribution) .

4.2 Probability Distribution (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Probability_distribution)

Objective

As a student in Machine Learning and having already covered KL divergence, the aim of this course is to refresh my knowledge on probability distributions.

> Time estimated: 1 h; taken: 1 h; date started: 2018-10-05; date completed: 2018-10-05

Activities

I went through the R code thoroughly. There was actually a lot of very useful material that I had certainly forgotten or not previously considered (taken for granted). And as always I love all those colourful R graphs.

Conclusion and Outlook

The next unit that I plan to take is Significance (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Significance) .

4.3 Significance (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Significance)

Objective

Since I did my bachelor thesis on granger causality relations of blockchain financial time series hopefully this chapter should be more of a consolidation than tackling new material. The main aim is to ensure I fully understand how to use and manipulate p-values.

> Time estimated: 1 h; taken: 1 h; date started: 2018-10-05; date completed: 2018-10-06

Activities

I went through the R code thoroughly and read the recommended papers. The p value I got using `p = (Nobs + 1) / (N + 1)` was `0.03297888` . This suggests statistically significant evidence against the null hypothesis (which in this case is that positively charged residues are not closer to negatively charged residues). The code also asks to mark down the sum of all the "chs" in binary hex format (I assume for Prof. Steipe to verify). Mine is: `00 00 00 00 8c 9a e1 40` .

The first paper (https://www.ncbi.nlm.nih.gov/pubmed/26961635) reviewed basically describes the tears of the American Statistical Association (ASA) which are concerned that papers only get published when p-values indicated statistical significance while papers showing no statistical significance do not get published. This yields a falsely robust methodology that the ASA are hoping to reverse.

The second paper (https://www.ncbi.nlm.nih.gov/pubmed/21878926) reviewed criticises neuroscience publications finding that around 50% of them use the incorrect procedure for evaluating statistical significance between two events, say A and B. The main error being that often A is shown to be statistically significant and control event B is said to be statistically significant but the difference between A and B is not examined to see if it itself is statistically significant. This paper also criticises the Cliff effect whereby people are overly swayed once the artificial boundary of $p < 0.05$ (0h 1m) is crossed. It however concludes than in many cases this lack of statistic rigour would not however change the conclusions of the papers.

Conclusion and Outlook

The next unit that I plan to take is Genetic Code (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-Genetic_code) .

4.4 Information Theory (http://steipe.biochemistry.utoronto.ca/abc/index.php/FND-STA-Information_theory)

Objective

Since I am already quite familiar with information theory hopefully this should be a refresher.

> Time estimated: 0.5 h; taken: 0.5 h; date started: 2018-10-23; date completed: 2018-10-23

Activities

I went through the notes thoroughly and made sure I knew everything. This was not too challenging for me given my background.

Conclusion and Outlook

The next unit that I plan to take is Biomolecular Function Concepts (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-FUNC-Concepts) .

5. Integrator Units

5.1 Mutation Impact (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Mutation_impact) (submitted: 8 marks)

Objective

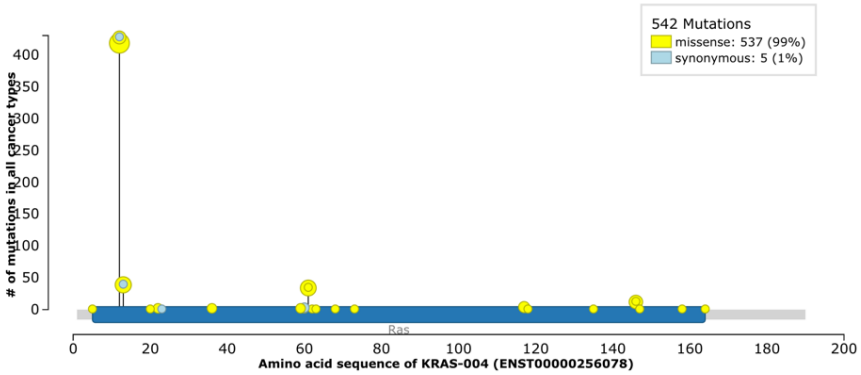
My first integrator unit on mutations. Objective? 100%.

> Time estimated: 6 h; taken: 12 h; date started: 2018-10-27; date completed: 2018-10-30

Activities

I really want to nail this integrator unit so I have set aside a large number of hours for it. I also made sure to wait until I explored a lot of the bioinformatics map to be as informed as possible before tackling this. The R code option tickles my fancy more than anything else so here goes: evaluation page (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_Mutation_Impact) .

A lot of information about the cancer driving KRAS gene can be found on IntOGen (https://www.intogen.org/search?gene=KRAS) . The following graph from their website gives me a good idea of what a definite cancer driving gene mutation distribution looks like:



The work I submitted for evaluation is here:

```
# Thomas Hollis (BCH441, University of Toronto) -v2.1
#     - Purpose: Mutation Impact Evaluation (load two functions used for mutation effect analysis)
#     - Bugs & issues: no bugs, no issues, no warnings
#     - Acknowledgements: thanks to Prof. Steipe's learning unit on R Genetic Code Optimality which was of great help

###### Section 1. Load packages if not already installed ######
if (!require(Biostrings, quietly=TRUE)) {
  if (!exists("biocLite")) {
    suppressMessages(source("https://bioconductor.org/biocLite.R")) #slight modification to load silently (I love the clear R documentation)
  }
  suppressMessages(biocLite("Biostrings")) #slight modification to always load packages silently (nb: biocLite does not play nice)
  suppressMessages(library(Biostrings, warn.conflicts = FALSE, quietly=TRUE)) #slight modification to always load packages silently
}

if (!require(readr, quietly=TRUE)) {
  install.packages("readr")
  suppressMessages(library(readr, quietly=TRUE)) #slight modification to always load packages silently
}

if (!require(testthat, quietly=TRUE)) {
  install.packages("testthat")
  suppressMessages(library(testthat, quietly=TRUE)) #slight modification to always load packages silently
}

###### Section 2. Load evalMut function ######
evalMut <- function(FA, N) {
  # Purpose: evaluate the distribution of silent, missense and nonsense
  # codon changes in cDNA read from FA for N random mutation trials.
  # Parameters:
  #     FA   chr      Filename of a FASTA formatted sequence file of cDNA
  #                     beginning with a start codon.
  #     N    integer  The number of point mutation trials to perform
  # Value:   list     List with the following elements:
  #                     FA     chr  the input file
  #                     N      num  same as the input parameter
  #                     nSilent    num  the number of silent mutations
  #                     nMissense  num  the number of missense mutations
  #                     nNonsense  num  the number of nonsense mutations

  #2.1. Load and process data
  nSilent <- 0 #initialise silent mutation object
  nMissense <- 0 #initialise missense mutation object
  nNonsense <- 0 #initialise nonsense mutation object

  seq <- paste0(readLines(FA)[-1], collapse = "") #load and clean the sequence
  codons <- as.character(codons(DNAString(seq))) #convert to codons, can drop stop codon using  codons <- codons[-length(codons)]

  AA <- character(length(seq)) #initialise amino acid object
  for (j in seq_along(codons)) {
    AA[j] <- GENETIC_CODE[codons[j]] #iterate through all codons, set AA
  }

  #2.2. Mutate for N random mutations & count the silent, missense and nonsense codon changes
  nuc <- c("A", "C", "G", "T") #initialise nucl names

  #Note this loop was greatly inspired by work done by Prof. Steipe
  #Available at: http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-Genetic_code_optimality
  for (i in 1:N) {
    mutCodons <- codons #initialise the mutated codons object
    loc <- sample(1:length(mutCodons), 1) #pick random mutation location

    #2.2.1. Mutate
    triplet <- unlist(strsplit(mutCodons[loc], "")) #split into three nucl.
    iNuc <- sample(1:3, 1)                          #choose one of the three
    mutNuc <- sample(nuc[nuc != triplet[iNuc]], 1)  #chose a mutated nucleotide
    triplet[iNuc] <- mutNuc                         #replace the original
    mutCodons[loc] <- paste0(triplet, collapse = "")  #collapse it to a codon

    #2.2.2. Translate back
    mutAA <- character(length(seq)) #initialise mutated amino acid object
    for (j in seq_along(mutCodons)) {
      mutAA[j] <- GENETIC_CODE[mutCodons[j]] #iterate through all mutated codons, set mutAA
    }

    #2.2.3. Count the mutation types
    if (mutAA[loc] == AA[loc]) {    #check if silent
      nSilent <- nSilent + 1
    } else if (mutAA[loc] == "*" | AA[loc] == "*") { #check if nonsense
      nNonsense <- nNonsense + 1
    } else {                        #else missense (assuming single point)
      nMissense <- nMissense + 1
    }
  }

  #2.3. Create output list
  list_out <- list(FA, N, nSilent, nMissense, nNonsense)

  return (list_out)

}

###### Section 3. Load readIntOGen function ######
readIntOGen <- function(IN) {
  # Purpose: read and parse an IntOGen mutation data file. Return only the
  #            number of silent, missense, and nonsense point mutations.
  #            All indels are ignored.
  # Parameters:
  #     IN   chr      Filename of an IntOGen mutation data file.
  # Value:   list     List with the following elements:
  #                     nSilent    num the number of silent mutations
```

```
  #                     nMissense  num the number of missense mutations
  #                     nNonsense  num the number of nonsense mutations

  #3.1. Load and process data
  DataIntOGen <- read_tsv(IN)

  nSilent <- 0 #initialise silent mutation object
  nMissense <- 0 #initialise missense mutation object
  nNonsense <- 0 #initialise nonsense mutation object

  #3.2. Count the mutations
  for (i in 1:length(DataIntOGen$MOST_SEVERE)) {
    if (DataIntOGen$MOST_SEVERE[i] == "synonymous_variant") {
      nSilent <- nSilent + 1
    } else if (DataIntOGen$MOST_SEVERE[i] == "missense_variant") {
      nMissense <- nMissense + 1
    } else if (DataIntOGen$MOST_SEVERE[i] == "stop_gained") {
      nNonsense <- nNonsense + 1
    }
  }

  list_out <- list(nSilent, nMissense, nNonsense)

  return (list_out)

}

###### Section 4. Test code & analysis script ######

#4.1. Protected code for testing
if (FALSE){
  #Test 4.1.1: Use this test to check the file sources cleanly & the right functions get loaded
  expect_silent(source("tom_INT1.R"))
  source("tom_INT1.R")

  #Test 4.1.2: Use this test to check that data can be loaded cleanly from paths and written to FA
  FA <- "./data/KRAS_HSa_coding.fa"
  data_KRAS <- read_file("./data/KRAS_HSa_coding.fa")
  (data_KRAS)
  data_PTPN11 <- read_file("./data/PTPN11_HSa_coding.fa")
  (data_PTPN11)
  data_OR1A1 <- read_file("./data/OR1A1_HSa_coding.fa")
  (data_OR1A1)
  rm(FA, data_KRAS, data_PTPN11, data_OR1A1) #always clean up after yourself!

  #Test 4.1.3: Use this test to make sure the 1st function runs as expected on KRAS, PTPN11 & OR1A1
  N <- 100000 #point mutations, modifiable (will take 20s + to run)
  evalMut("./data/KRAS_HSa_coding.fa",N) #expected output is nSilent = 20501, nMissense = 73927, nNonsense = 5572
  evalMut("./data/PTPN11_HSa_coding.fa",N) #expected output is nSilent = 21038, nMissense = 74066, nNonsense = 4896
  evalMut("./data/OR1A1_HSa_coding.fa",N) #expected output is nSilent = 23382, nMissense = 73007, nNonsense = 3611
  #Note: the expected outputs above will change on every iteration since we have not set a seed
  #Sanity check: outputs are indeed close to those of Prof. Steipe (with his seq, function outputs nSilent = 24021, nMissense = 67998, nNonsense = 79

  #Test 4.1.4: Use this test make sure the 2nd function runs as expected
  IN <- "./data/intogen-KRAS-distribution-data.tsv"
  readIntOGen("./data/intogen-KRAS-distribution-data.tsv") #expected output = 16 synonymous, 160 missense, 0 nonsense (as per the file)
  readIntOGen("./data/intogen-PTPN11-distribution-data.tsv") #expected output = 23 synonymous, 82 missense, 5 nonsense (as per the file)
  readIntOGen("./data/intogen-OR1A1-distribution-data.tsv") #expected output = 17 synonymous, 28 missense, 3 nonsense (as per the file)
  #Note: the expected outputs above should not change as they are reading from a file
  #Sanity check: the expected outputs above are the same as those in the file (but not the same as those online, must be much smaller)

  #Test 4.1.5: Nuclear option - don't run this unless you really have to
  rm(list=ls())
}

#4.2. Protected script for simulating 10000 point mutations of PTPN11 and analysing the results
if (FALSE){
  #4.2.1. Acquire the data and store in objects
  simulated <- evalMut("./data/PTPN11_HSa_coding.fa",10000)
  literature <- readIntOGen("./data/intogen-PTPN11-distribution-data.tsv")

  #4.2.2. Calculate the totals
  simTotal <- sum(simulated[[3]], simulated[[4]], simulated[[5]])
  litTotal <- sum(literature[[1]], literature[[2]], literature[[3]])
  simPercentageSilent <- (simulated[[3]]/simTotal)*100
  litPercentageSilent <- (literature[[1]]/litTotal)*100
  simPercentageMissense <- (simulated[[4]]/simTotal)*100
  litPercentageMissense <- (literature[[2]]/litTotal)*100
  simPercentageNonsense <- (simulated[[5]]/simTotal)*100
  litPercentageNonsense <- (literature[[3]]/litTotal)*100

  #4.2.3. Display the totals side by side for comparison (note as before expected values will fluctuate)
  cat(" Our simulations predict", simPercentageSilent, "% silent mutations, while we oberserved", litPercentageSilent, "%.\n",
      "Our simulations predict", simPercentageMissense, "% missense mutations, while we oberserved", litPercentageMissense, "%.\n",
      "Our simulations predict", simPercentageNonsense, "% nonsense mutations, while we oberserved", litPercentageNonsense, "%.")

  # Output:
  # Our simulations predict 21.78 % silent mutations, while we oberserved 20.90909 %.
  # Our simulations predict 73.2 % missense mutations, while we oberserved 74.54545 %.
  # Our simulations predict 5.02 % nonsense mutations, while we oberserved 4.545455 %.
}

# END
```

Conclusion and Outlook

This was a great unit to help make sure I fully understand the previous unit on gene optimality. It seems we can conclude that PTPN11 does indeed play a role in cancer. The next unit that I plan to take is PPI Analysis (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-PPI-Analysis) .

5.2 Phylogeny (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Phylogeny) (submitted: 16 marks)

Objective

I want to review all the concepts I have learnt in phylogeny and phylogenetic tree analysis to be ready for questions by Prof. Steipe during my oral test.

Time estimated: 6 h; taken: ? h; date started: 2018-11-07; date completed: 2018-11-15

Activities

Went through the notes carefully (also did a second pass through of previous learning units to help remember better the content). Submission page can be found here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_Phylogeny) . Code developed is shown as follows:

```
# Thomas Hollis (BCH441, University of Toronto) -v2.1
#     - Purpose: Phylogeny Oral Evaluation (build & analyse aphylogenetic tree)
#     - Bugs & issues: no bugs, no issues, no warnings
#     - Acknowledgements: thanks to Prof. Steipe's learning units on Phylogeny which were of great help

###### 1. Load packages and data ######
init()

if (! require(msa, quietly=TRUE)) {
  if (! exists("biocLite")) {
    source("https://bioconductor.org/biocLite.R")
  }
  biocLite("msa")
  library(msa)
}

if (! require(Biostrings, quietly=TRUE)) {
  if (! exists("biocLite")) {
    source("https://bioconductor.org/biocLite.R")
  }
  biocLite("Biostrings")
  library(Biostrings)
}

if (!require(Rphylip, quietly=TRUE)) {
  install.packages("Rphylip")
  library(Rphylip)
}

if (!require(phangorn, quietly=TRUE)) {
  install.packages("phangorn")
  library(phangorn)
}

source("makeProteinDB.R")

######  2. Produce the phylogenetic tree ######
  # 2.1 Data Preparation
    # 2.1.1 Align all sequences in the database + KILA_ESSCO (outgroup, for rooting the tree)
    mySeq <- myDB$protein$sequence # add all DB sequences
    names(mySeq) <- myDB$protein$name # add names of all DB sequences

    mySeq <- c(mySeq,
               "IDGEIIHLRAKDGYINATSMCRTAGKLLSDYTRLKTTQEFFDELSRDMGIPISELIQSFKGGRPENQGTWVHPDIAINLAQ")
    names(mySeq)[length(mySeq)] <- "KILA_ESCCO" # add KILLA_ESSCO sequence and name

    (mySeqMSA <- msaClustalOmega(AAStringSet(mySeq))) # ClustalOmega used, too many seq for MUSCLE

    # 2.1.2 Get the sequence of the SACCE APSES domain ######
    sel <- myDB$protein$name == "MBP1_SACCE" # selector for SACCE protein
    proID <- myDB$protein$ID[sel] # store SACCE protein in proID for later

    sel <- myDB$feature$ID[myDB$feature$name == "APSES fold"] # selector for APSES
    fanID <- myDB$annotation$ID[myDB$annotation$proteinID == proID &
                                myDB$annotation$featureID == sel] # SACCE & APSES
    (start <- myDB$annotation$start[fanID]) # set start point (4)
    (end   <- myDB$annotation$end[fanID]) # set the end point (102)

    (SACCEapses <- substring(myDB$protein$sequence[proID], start, end)) #SACCE APSES seq

    # 2.1.3 Extract the APSES domains from the MSA
    (APSESmsa <- fetchMSAmotif(mySeqMSA, SACCEapses)) # stored in .utilities.R, returns matching seq
    # note this includes PSI-BLAST results which can be found in MYSPE_APSES_PSI-BLAST.json
    # note APSESmsa is of type "AAStringSet" not "MsaAAMultipleAlignment"

    # 2.1.4 Process the APSESmsa for Tree Building
    (numAli <- length(names(APSESmsa))) # get the number of alignments (52)
    (lenAli <- length(APSESmsa$MBP1_SACCE)) # get the length of the alignments (269)

    msaMatrix <- matrix(nrow = numAli, ncol = lenAli) # initialize matrix to hold all chars
    (rownames(msaMatrix) <- APSESmsa@ranges@NAMES) # assign the correct rownames
    for (i in 1:numAli) {
      msaMatrix[i, ] <- unlist(strsplit(as.character(APSESmsa[i]), "")) # fill matrix
    }
    msaMatrix[1:7, 1:14] # check result is a well defined and formatted matrix

    colMask <- logical(lenAli) # initialize a mask (logical vector to trim cols)
    limit <- round(numAli * (2/3)) # define the 2/3 threshold for rejecting a column
    for (i in 1:ncol(msaMatrix)) {  # iterate over all columns
      count <- sum(msaMatrix[ , i] == "-") # count hyphens in col
      colMask[i] <- count <= limit # write TRUE if less-or-equal to limit, FALSE if not
    }
    colMask # check mask makes sense
    sum(colMask) # check how many collumns are kept (should be 101)
    cat(sprintf("Masking %4.2f %% of columns.\n", 100*(1 - (sum(colMask)/length(colMask)) )))
    maskedMatrix <- msaMatrix[ , colMask] # remove masked cols
    ncol(maskedMatrix) # check how many cols left (should be 101)

    tomAPSESphyloSet <- character() # initialise char object, namechange to avoid confusion
    for (i in 1:nrow(maskedMatrix)) { # collapse back to string
      tomAPSESphyloSet[i] <- paste(maskedMatrix[i, ], collapse="")
    }
    names(tomAPSESphyloSet) <- rownames(maskedMatrix) # add names

    writeALN(tomAPSESphyloSet) # inspect final result
    writeMFA(tomAPSESphyloSet, myCon="tom_APSESphyloSet.mfa") # save aligned & masked to multi-FASTA

  # 2.2 Tree Building
  PROMLPATH <- "/Users/tom/Applications/phylip-3.695/exe/proml.app/Contents/MacOS" # set PROMLPATH
  list.dirs(PROMLPATH) # confirm one directory is listed only
  list.files(PROMLPATH) # confirm two files listed only "proml" and "proml.command"
  apsIn <- read.protein("tom_APSESphyloSet.mfa") # import aligned & masked from multi-FASTA
  myTree <- Rproml(apsIn, path=PROMLPATH) # run PROML to build tree (warning: 4-8h) start:6pm,end:?pm
  plot(myTree) # have a quick look before further analysis
  save(myTree, file = "tom_myTreeRproml.RData") # save locally to RData file

  # 2.3 Tree Analysis

    # 2.3.1 A few quick exploratory views of the tree
    load(file = "tom_myTreeRproml.RData")
    plot(myTree) # default type: "phylogram"
    plot(myTree, type = "unrooted") # unrooted type
    plot(myTree, type = "fan", no.margin = TRUE) # concentric circle type

    str(myTree) # inspect the tree object
    myTree$tip.label # inspect the tree object
    myTree$edge # inspect the tree object
    myTree$edge.length # inspect the tree object

    plot(myTree) # plot the tree with labels
    tiplabels(cex = 0.5, frame = "rect") # useful to find the outgroup (should be 17 here)
    edgelabels(cex = 0.5) # add edge labels - looks messy, not very useful here
    nodelabels(cex = 0.5, frame = "circle") # add node labels - looks messy, not very useful here

    Nnode(myTree) # number of nodes (should be 50)
    Nedge(myTree) # number of edges (should be 101)
    Ntip(myTree) # number of tips (should be 52)
```

```
    write.tree(myTree) # show the tree in console in Newick format

    # 2.3.2 Rearrange the tree & root it with the outgroup
    myTree <- root(myTree, outgroup = 17, resolve.root = TRUE) # change this for outgroup number
    plot(myTree) # check it out
    is.rooted(myTree) # make sure the tree is now rooted

    myTree$edge.length[1] <- 0.1 # rescale the tree to a reasonable size
    plot(myTree, cex = 0.7) # check it out

    myTree$root.edge <- mean(myTree$edge.length) * 1.5 # make the root visable
    plot(myTree, cex = 0.7, root.edge = TRUE) # check it out

    # 2.3.3 Rotate the clades (useful to compare with cladogram of species)
    nodelabels(cex = 0.5, frame = "circle") # add node labels - looks messy, not very useful here
    myTree <- rotate(myTree, node = 100)
    myTree <- rotate(myTree, node = 54)
    myTree <- rotate(myTree, node = 60)
    myTree <- rotate(myTree, node = 61)
    myTree <- rotate(myTree, node = 62)
    myTree <- rotate(myTree, node = 77)
    myTree <- rotate(myTree, node = 74)
    myTree <- rotate(myTree, node = 69)
    myTree <- rotate(myTree, node = 70)
    myTree <- rotate(myTree, node = 71)
    myTree <- rotate(myTree, node = 98)
    myTree <- rotate(myTree, node = 96)
    myTree <- rotate(myTree, node = 95)
    myTree <- rotate(myTree, node = 81)
    myTree <- rotate(myTree, node = 64)
    myTree <- rotate(myTree, node = 58)
    myTree <- rotate(myTree, node = 80)
    myTree <- rotate(myTree, node = 84)
    myTree <- rotate(myTree, node = 93)
    myTree <- rotate(myTree, node = 88)
    myTree <- rotate(myTree, node = 99) # beautify
    plot(myTree, cex = 0.7, root.edge = TRUE)

    # 2.3.4 Computing Tree Distance with reference tree
    # This is bonus tree analysis and was omitted here as it is not included in instructions.

# END
```

I also put this into an R markdown for better presentation in the oral!

```
---
title: 'Phylogeny: An introductory R Notebook by T. Hollis'
output:
  html_document:
    df_print: paged
  pdf_document: default
---

# 1. Introduction
This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code. I dec

Thomas Hollis (BCH441, University of Toronto) -v2.1

   - Purpose: Phylogeny Oral Evaluation (build & analyse aphylogenetic tree)

   - Bugs & issues: no bugs, no issues, no warnings

   - Acknowledgements: thanks to Prof. Steipe's learning units on Phylogeny which were of great help

Disclaimer: I used a patch designed by myself (see mailing list) in earlier units so I hope this will not hinder the oral test.

# 2. Code

## 2.1 Import required libraries

As always, lets import our required libraries and data before starting:

```{r eval=FALSE}
init()

if (! require(msa, quietly=TRUE)) {
  if (! exists("biocLite")) {
    source("https://bioconductor.org/biocLite.R")
  }
  biocLite("msa")
  library(msa)
}

if (! require(Biostrings, quietly=TRUE)) {
  if (! exists("biocLite")) {
    source("https://bioconductor.org/biocLite.R")
  }
  biocLite("Biostrings")
  library(Biostrings)
}

if (!require(Rphylip, quietly=TRUE)) {
  install.packages("Rphylip")
  library(Rphylip)
}

if (!require(phangorn, quietly=TRUE)) {
  install.packages("phangorn")
  library(phangorn)
}

source("makeProteinDB.R")
```

2.2 Data Preparation

2.2.1 Alignment

Let's first align some sequences (all sequences in the database + KILA_ESSCO for rooting the tree):

```{r eval=FALSE}
mySeq <- myDB$protein$sequence # add all DB sequences
names(mySeq) <- myDB$protein$name # add names of all DB sequences

mySeq <- c(mySeq,"IDGEIIHLRAKDGYINATSMCRTAGKLLSDYTRLKTTQEFFDELSRDMGIPISELIQSFKGGRPENQGTWVHPDIAINLAQ")
names(mySeq)[length(mySeq)] <- "KILA_ESCCO" # add KILLA_ESSCO sequence and name

(mySeqMSA <- msaClustalOmega(AAStringSet(mySeq))) # ClustalOmega used, too many seq for MUSCLE
```

2.2.2 Get the sequence of the SACCE APSES domain

```{r eval=FALSE}
sel <- myDB$protein$name == "MBP1_SACCE" # selector for SACCE protein
```
```

```
proID <- myDB$protein$ID[sel] # store SACCE protein in proID for later

sel <- myDB$feature$ID[myDB$feature$name == "APSES fold"] # selector for APSES
fanID <- myDB$annotation$ID[myDB$annotation$proteinID == proID & myDB$annotation$featureID == sel] # SACCE & APSES
(start <- myDB$annotation$start[fanID]) # set start point (4)
(end   <- myDB$annotation$end[fanID]) # set the end point (102)

(SACCEapses <- substring(myDB$protein$sequence[proID], start, end)) #SACCE APSES seq
```

2.2.3 Extract the APSES domains from the MSA

It is worth noting that this includes PSI-BLAST results which can be found in MYSPE_APSES_PSI-BLAST.json. It addition, APSESmsa is of type "AAStringS

```{r eval=FALSE}
(APSESmsa <- fetchMSAmotif(mySeqMSA, SACCEapses)) # stored in .utilities.R, returns matching seq
```

2.2.4 Process the APSESmsa data for Tree Building

We need to mask some of the collumns. To do this we must first convert to a matrix of characters, then mask, then convert back and export to multi-FA

```{r eval=FALSE}
(numAli <- length(names(APSESmsa))) # get the number of alignments (52)
(lenAli <- length(APSESmsa$MBP1_SACCE)) # get the length of the alignments (269)

msaMatrix <- matrix(nrow = numAli, ncol = lenAli) # initialize matrix to hold all chars
(rownames(msaMatrix) <- APSESmsa@ranges@NAMES) # assign the correct rownames
for (i in 1:numAli) {
  msaMatrix[i, ] <- unlist(strsplit(as.character(APSESmsa[i]), "")) # fill matrix
}
msaMatrix[1:7, 1:14] # check result is a well defined and formatted matrix

colMask <- logical(lenAli) # initialize a mask (logical vector to trim cols)
limit <- round(numAli * (2/3)) # define the 2/3 threshold for rejecting a column
for (i in 1:ncol(msaMatrix)) {  # iterate over all columns
  count <- sum(msaMatrix[ , i] == "-") # count hyphens in col
  colMask[i] <- count <= limit # write TRUE if less-or-equal to limit, FALSE if not
}
colMask # check mask makes sense
sum(colMask) # check how many collumns are kept (should be 101)
cat(sprintf("Masking %4.2f %% of columns.\n", 100*(1 - (sum(colMask)/length(colMask)) )))
maskedMatrix <- msaMatrix[ , colMask] # remove masked cols
ncol(maskedMatrix) # check how many cols left (should be 101)

tomAPSESphyloSet <- character() # initialise char object, namechange to avoid confusion
for (i in 1:nrow(maskedMatrix)) { # collapse back to string
  tomAPSESphyloSet[i] <- paste(maskedMatrix[i, ], collapse="")
}
names(tomAPSESphyloSet) <- rownames(maskedMatrix) # add names

writeALN(tomAPSESphyloSet) # inspect final result
writeMFA(tomAPSESphyloSet, myCon="tom_APSESphyloSet.mfa") # save aligned & masked to multi-FASTA
```

2.3 Build the tree using PROML

Import the clean, aligned & masked data. Use this to build the tree. (WARNING: took around 8h to run)

```{r eval=FALSE}
PROMLPATH <- "/Users/tom/Applications/phylip-3.695/exe/proml.app/Contents/MacOS" # set PROMLPATH
list.dirs(PROMLPATH) # confirm one directory is listed only
list.files(PROMLPATH) # confirm two files listed only "proml" and "proml.command"
apsIn <- read.protein("tom_APSESphyloSet.mfa") # import aligned & masked from multi-FASTA
myTree <- Rproml(apsIn, path=PROMLPATH) # run PROML to build tree (warning: 4-8h) start:6pm,end:?pm
plot(myTree) # have a quick look before further analysis
save(myTree, file = "tom_myTreeRproml.RData") # save locally to RData file
```

2.4 Analyse the tree

2.4.1 A few quick exploratory views of the tree

```{r eval=FALSE}
load(file = "tom_myTreeRproml.RData")
plot(myTree) # default type: "phylogram"
plot(myTree, type = "unrooted") # unrooted type
plot(myTree, type = "fan", no.margin = TRUE) # concentric circle type

str(myTree) # inspect the tree object
myTree$tip.label # inspect the tree object
myTree$edge # inspect the tree object
myTree$edge.length # inspect the tree object

plot(myTree) # plot the tree with labels
tiplabels(cex = 0.5, frame = "rect") # useful to find the outgroup (should be 17 here)
edgelabels(cex = 0.5) # add edge labels - looks messy, not very useful here
nodelabels(cex = 0.5, frame = "circle") # add node labels - looks messy, not very useful here

Nnode(myTree) # number of nodes (should be 50)
Nedge(myTree) # number of edges (should be 101)
Ntip(myTree) # number of tips (should be 52)

write.tree(myTree) # show the tree in console in Newick format
```

2.4.2 Rearrange the tree & root it with the outgroup

```{r eval=FALSE}
myTree <- root(myTree, outgroup = 17, resolve.root = TRUE) # change this for outgroup number
plot(myTree) # check it out
is.rooted(myTree) # make sure the tree is now rooted

myTree$edge.length[1] <- 0.1 # rescale the tree to a reasonable size
plot(myTree, cex = 0.7) # check it out

myTree$root.edge <- mean(myTree$edge.length) * 1.5 # make the root visable
plot(myTree, cex = 0.7, root.edge = TRUE) # check it out
```

2.4.3 Rotate the clades

This is useful to better compare with cladogram of species.

```{r eval=FALSE}
nodelabels(cex = 0.5, frame = "circle") # add node labels - looks messy, not very useful here (except for rotations)
myTree <- rotate(myTree, node = 100)
myTree <- rotate(myTree, node = 54)
myTree <- rotate(myTree, node = 60)
myTree <- rotate(myTree, node = 61)
myTree <- rotate(myTree, node = 62)
myTree <- rotate(myTree, node = 77)
myTree <- rotate(myTree, node = 74)
myTree <- rotate(myTree, node = 69)
myTree <- rotate(myTree, node = 70)
myTree <- rotate(myTree, node = 71)
myTree <- rotate(myTree, node = 98)
myTree <- rotate(myTree, node = 96)
```

```
myTree <- rotate(myTree, node = 95)
myTree <- rotate(myTree, node = 81)
myTree <- rotate(myTree, node = 64)
myTree <- rotate(myTree, node = 58)
myTree <- rotate(myTree, node = 80)
myTree <- rotate(myTree, node = 84)
myTree <- rotate(myTree, node = 93)
myTree <- rotate(myTree, node = 88)
myTree <- rotate(myTree, node = 99) # beautify
plot(myTree, cex = 0.7, root.edge = TRUE)
```

```
### 2.4.4 Computing Tree Distance with reference tree

This is bonus tree analysis and was omitted here as it is not included in instructions.
```

Conclusion and Outlook

This was a fun and productive roundup putting together all the units in this part of the course map. The next unit that I plan to take is Genome Browsers (http://steipe.biochemistry.utoronto.ca/abc/index.php/BIN-Genome-Browsers) .

5.3 Genome Annotation (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Genome_annotation) (submitted: 8 marks)

Objective

As per, objective 100% here.

Time estimated: 6 h; taken: 7 h; date started: 2018-11-18; date completed: 2018-11-21

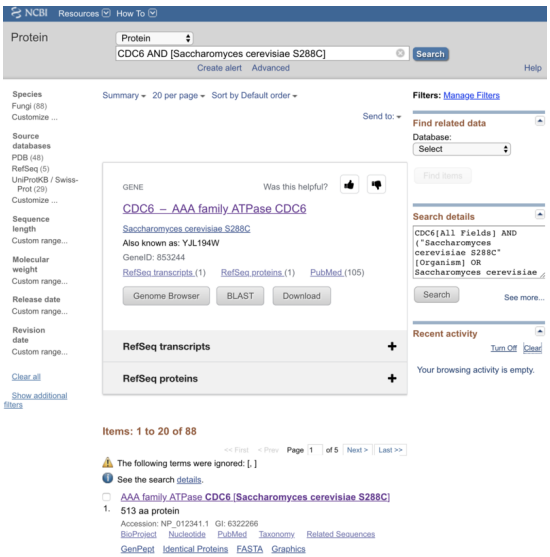
Activities

You can find my submission page here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php?title=User:Thomas_Hollis/Eval_Genome_Annotation&action=edit&redlink=1) .

Task 1 - Finding the ISAFU orthologue for Cdc6

- Find the MYSPE orthologue for yeast Cdc6.

1.1. First and foremost I need to find the RefSeqID for SACCE Cdc6:

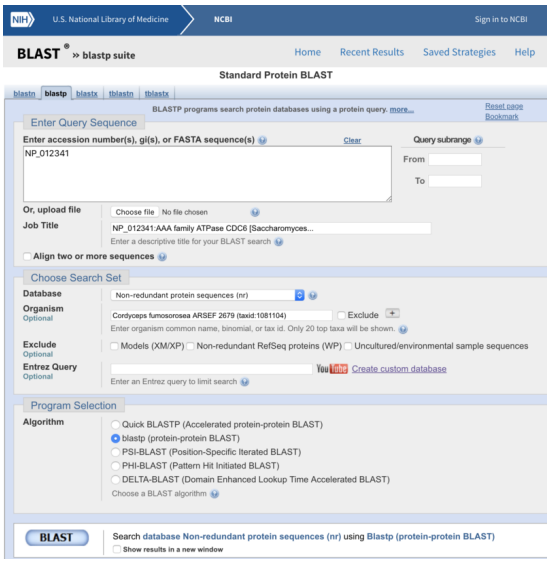


The RefSeq is `NP_012341.1` (NCBI page here (https://www.ncbi.nlm.nih.gov/protein/NP_012341.1))

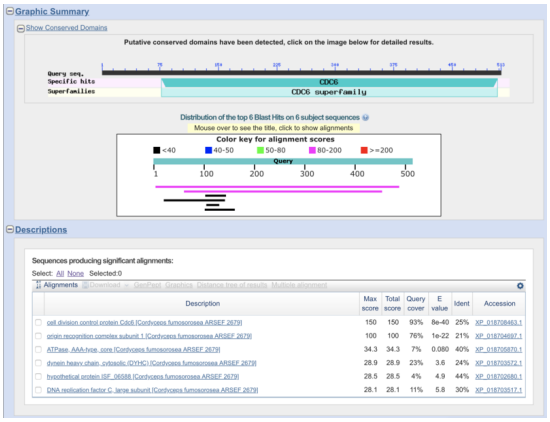
FASTA is:

```
>NP_012341.1 AAA family ATPase CDC6 [Saccharomyces cerevisiae S288C]
MSAIPITPTKRIRRNLFDDAPATPPRPLKRKKLQFTDVTPESSPEKLQFGSQSIFLRTKALLQKSSELVN
LNSSDGALPARTAEYEQVMNFLAKAISEHRSDSLYITGPPGTGKTAQLDMIIRQKFQSLPLSLSTPRSKD
VLRHTNPNLQNLSWFELPDGRLESVAVTSINCISLGEPSSIFQKIFDSFQDLNGPTLQIKNMQHLQKFLE
PYHKKTTFVVVLDEMDRLLHANTSETQSVRTILELFLLAKLPTVSFVLIGMANSLDMKDRFLSRLNLDRG
LLPQTIVFQPYTAEQMYEIVIQKMSSLPTIIFQPMAIKFAAKKCAGNTGDLRKLFDVLRGSIEIYELEKR
FLLSPTRGSLNSAQVPLTPTTSPVKKSYPEPQGKIGLNYIAKVFSKFVNNNSTRTRIAKLNIQQKLILCT
IIQSLKLNSDATIDESFDHYIKAITKTDTLAPLQRNEFLEICTILETCGLVSIKKTKCKGKTKRFVDKID VDLDMREFYDEMTKISILKPFLH
```

1.2. Now I need to do a protein BLAST:



The results of this BLAST are:



The top hit for the ISAFU (a.k.a. CORFU) orthologue is a "cell division control protein". (Accession: XP_018708463.1 (https://www.ncbi.nlm.nih.gov/protein/XP_018708463.1?report=genbank&log$=protalign&blast_rank=1&RID=Z8RJWMNP013) , E-value: 8e-40, Coverage: 93%)

FASTA is:

```
>XP_018708463.1 cell division control protein Cdc6 [Cordyceps fumosorosea ARSEF 2679]
MASSALGKRSRRAVIYDSDESPSKRPRRSCAATVDYNDENEDPEDIVSVATTNPIIIKTEVDDSPSKRNR
SLSQSSARNPVTPSTPRHHDALAAYPTTPRHAVMSAGKLFKRLTPHSPLSPSTIQTVYHSARQLFARGAE
PGQLVGRDEERQQLTAFLERCTSDSPNGCLYVSGPPGTGKSAMITGLIQKYANRDDVKSAYINCMSVKSS
KDLYMTLLDGMGLEAEGVTEADAMDALQDAFYPRDDNATTYLITLDEIDHILTMGLESLYRVFEWALHKS
SKLLLVGIANALDLTDRFLPRLKSKNLKPELLPFLPYTATQVKNIITTRLKSLMPQGKENFVPFIHPAAI
ELCSRKVSSQTGDLRKAFEICRRALDLIENETRSKHEEEAREAMLQMTPSKRPLGENINGAMGGRSIVQI
MGNSLKALTAETAPRASIAHLNKVTAAAFSNGTTQRLKTLNLQQKAALCALVAYENHLRSKVKVGMPTTP
SKTQLLAPTVKKLFGAYCRLCTRDSVLHPLSSSEFREVMGSLETLGLVNAVDGKNGSFITPQTPSKRGRK AAVLASGDDKRVASCVTEKDMESVVAGLGSGILASILHGEALD
```

Task 2 - Fetching nucleotides

- Fetch 500 nucleotides of upstream genome sequence. (Demonstrate that this is the correct sequence by showing the first 10 translated Cdc6 codons with your sequence.)

Using the NCBI website, the upstream genome sequence can be extracted with some casual URL modification. The original URL (https://www.ncbi.nlm.nih.gov/nuccore/NW_017387301.1?report=fasta&from=1232756&to=1234623) is modified to decrease the "from" parameter by 500, effectively shifting upstream.

Original URL (https://www.ncbi.nlm.nih.gov/nuccore/NW_017387301.1?report=fasta&from=1232756&to=1234623) output FASTA:

```
>NW_017387301.1:1232756-1234623 Cordyceps fumosorosea ARSEF 2679 chromosome Unknown scaffold_1, whole genome shotgun sequence
ATGGCGTCCTCAGCTCTGGGAAAGAGATCCCGCCGCGCGGTAATCTATGGTACGTACAAACAGGCACGGT
GCCTTCTAGACAATCGTACTTACCTGTTACCCCAGACTCAGACGAGTCTCCTTCCAAACGGCCTCGCCGG
TCGTGCGCCGCCACTGTCGACTACAACGACGAGAACGAGGATCCCGAGGACATTGTCAGCGTTGCTACCA
CCAATCCCATTATTATCAAGACCGAAGTCGACGACTCGCCTAGCAAGAGAAATCGATCTCTTTCTCAGTC
TTCAGCTCGGAACCCCGTCACACCGTCGACGCCACGCCATCACGATGCCTTGGCCGCGTATCCCACGACG
CCTCGACACGCCGTCATGTCGGCCGGCAAGCTCTTCAAGAGACTGACGCCGCACTCTCCCCTGTCTCCCA
GCACCATCCAAACCGTTTACCACTCGGCTCGCCAATTATTCGCCCGCGGGGCGGAGCCCGGTCAGCTCGT
CGGCCGTGACGAAGAGCGTCAGCAGCTCACGGCTTTCCTGGAGCGGTGCACATCAGACTCGCCAAATGGG
TGCCTCTACGTCAGCGGCCCGCCTGGTACCGGCAAGAGCGCCATGATTACCGGGCTCATTCAAAAGTACG
CAAATAGGGACGATGTCAAGTCGGCATACATCAACTGCATGAGCGTCAAGTCCTCCAAGGATCTCTACAT
GACGCTACTCGACGGCATGGGGCTTGAGGCCGAGGGCGTAACCGAGGCTGATGCCATGGACGCTCTGCAA
GACGCTTTCTACCCCAGAGACGACAACGCCACGACCTATCTGATCACTCTCGATGAGATTGATCATATTC
TCACTATGGGCCTGGAGAGCTTGTATCGTGTATTTGAATGGGCCCTCCACAAGTCGTCAAAGCTTCTCCT
TGTCGGCATTGCCAACGCGCTAGATCTCACGGACCGATTTCTCCCTCGTCTCAAGTCCAAGAACCTTAAA
```

```
CCTGAGCTTCTTCCGTTTCTTCCCTATACCGCTACGCAAGTGAAGAATATCATCACAACTCGTCTGAAGT
CTTTGATGCCCCAGGGTAAAGAAAACTTCGTTCCATTTATTCACCCTGCTGCCATTGAACTGTGCTCCCG
CAAAGTGTCGAGTCAGACGGGGGATCTTCGCAAGGCGTTTGAGATTTGCAGAAGGGCTTTGGATCTGATT
GAGAACGAAACACGGTCGAAGCACGAGGAAGAGGCAAGGGAAGCTATGCTTCAGATGACACCATCGAAGC
GACCTCTTGGCGAGAATATCAACGGGGCCATGGGAGGACGAAGCATTGTACAAATTATGGGCAATTCACT
CAAGGCATTGACGGCCGAGACGGCCCCTCGCGCTTCCATAGCCCATCTGAACAAGGTGACGGCAGCAGCG
TTCAGCAACGGCACCACACAGAGGCTCAAGACTCTCAACCTTCAGCAAAAGGCAGCCCTGTGCGCTCTGG
TTGCGTACGAGAATCACTTGCGATCAAAGGTTAAGGTCGGCATGCCCACTACACCAAGCAAGACCCAGTT
ACTGGCACCAACTGTGAAGAAGCTCTTTGGCGCGTACTGCCGCCTGTGCACGCGCGACTCTGTGCTCCAC
CCTCTCTCGAGCTCCGAGTTCCGCGAGGTCATGGGTAGCTTGGAGACGCTGGGCCTTGTCAACGCGGTCG
ATGGCAAGAATGGAAGCTTCATTACGCCGCAGACGCCTAGCAAGCGTGGCCGCAAAGCCGCCGTGTTGGC
GAGCGGAGACGACAAGCGAGTTGCGAGCTGTGTGACCGAGAAGGACATGGAGTCGGTTGTTGCCGGCCTC GGCTCAGGCATTCTAGCCAGCATTTTGCATGGAGAGGCGTTGGATTAA
```

Modified URL (https://www.ncbi.nlm.nih.gov/nuccore/NW_017387301.1?report=fasta&from=1232756&to=1234623) output FASTA (upstream and sequence):

```
>NW_017387301.1:1232256-1234623 Cordyceps fumosorosea ARSEF 2679 chromosome Unknown scaffold_1, whole genome shotgun sequence
CTATCAATTGAGCCAGCTTTTCAAATATTCATTCTTTCCTTTTTTCAATAGTTTCCAGACCCTGCCGTTA
ACAGCTCTCCCTCCTCCCAGGTATTGAGCCTTGGGCTGTGGAGCGGCACTAGCCACGCCCGCGCCACACC
CAGGTTTTGCCATATCCACACACTGACCAGGGCTTTAATAGATTTTGTCGCGTACTGGCGCGTTGAAACG
CGCTCGACACGGTTCGTCCCGCCCTTGCGCGACCCTGGTCGTGGCGCGAGCTGCTTTACTTAATCCTTTC
GTGCCCTTGTGAGACTTTCCACAGCAACCAGCTTTTTCTCATCTACCAATTCCCCAGCGACGCATTTCCA
TCTCGACATTGCCTCTTCACATCTGTTTATATTAGTCGCCCGCATTCTGCATCTCGCACTACCAGCATTC
TACACTTAGCCGGCAAGAACATATCCGTATCAGCATTTTTCGATTGGCAGGAAGCTTCCTGCCCCAGGTG
TTCCGCATCAATGGCGTCCTCAGCTCTGGGAAAGAGATCCCGCCGCGCGGTAATCTATGGTACGTACAAA
CAGGCACGGTGCCTTCTAGACAATCGTACTTACCTGTTACCCCAGACTCAGACGAGTCTCCTTCCAAACG
GCCTCGCCGGTCGTGCGCCGCCACTGTCGACTACAACGACGAGAACGAGGATCCCGAGGACATTGTCAGC
GTTGCTACCACCAATCCCATTATTATCAAGACCGAAGTCGACGACTCGCCTAGCAAGAGAAATCGATCTC
TTTCTCAGTCTTCAGCTCGGAACCCCGTCACACCGTCGACGCCACGCCATCACGATGCCTTGGCCGCGTA
TCCCACGACGCCTCGACACGCCGTCATGTCGGCCGGCAAGCTCTTCAAGAGACTGACGCCGCACTCTCCC
CTGTCTCCCAGCACCATCCAAACCGTTTACCACTCGGCTCGCCAATTATTCGCCCGCGGGGCGGAGCCCG
GTCAGCTCGTCGGCCGTGACGAAGAGCGTCAGCAGCTCACGGCTTTCCTGGAGCGGTGCACATCAGACTC
GCCAAATGGGTGCCTCTACGTCAGCGGCCCGCCTGGTACCGGCAAGAGCGCCATGATTACCGGGCTCATT
CAAAAGTACGCAAATAGGGACGATGTCAAGTCGGCATACATCAACTGCATGAGCGTCAAGTCCTCCAAGG
ATCTCTACATGACGCTACTCGACGGCATGGGGCTTGAGGCCGAGGGCGTAACCGAGGCTGATGCCATGGA
CGCTCTGCAAGACGCTTTCTACCCCAGAGACGACAACGCCACGACCTATCTGATCACTCTCGATGAGATT
GATCATATTCTCACTATGGGCCTGGAGAGCTTGTATCGTGTATTTGAATGGGCCCTCCACAAGTCGTCAA
AGCTTCTCCTTGTCGGCATTGCCAACGCGCTAGATCTCACGGACCGATTTCTCCCTCGTCTCAAGTCCAA
GAACCTTAAACCTGAGCTTCTTCCGTTTCTTCCCTATACCGCTACGCAAGTGAAGAATATCATCACAACT
CGTCTGAAGTCTTTGATGCCCCAGGGTAAAGAAAACTTCGTTCCATTTATTCACCCTGCTGCCATTGAAC
TGTGCTCCCGCAAAGTGTCGAGTCAGACGGGGGATCTTCGCAAGGCGTTTGAGATTTGCAGAAGGGCTTT
GGATCTGATTGAGAACGAAACACGGTCGAAGCACGAGGAAGAGGCAAGGGAAGCTATGCTTCAGATGACA
CCATCGAAGCGACCTCTTGGCGAGAATATCAACGGGGCCATGGGAGGACGAAGCATTGTACAAATTATGG
GCAATTCACTCAAGGCATTGACGGCCGAGACGGCCCCTCGCGCTTCCATAGCCCATCTGAACAAGGTGAC
GGCAGCAGCGTTCAGCAACGGCACCACACAGAGGCTCAAGACTCTCAACCTTCAGCAAAAGGCAGCCCTG
TGCGCTCTGGTTGCGTACGAGAATCACTTGCGATCAAAGGTTAAGGTCGGCATGCCCACTACACCAAGCA
AGACCCAGTTACTGGCACCAACTGTGAAGAAGCTCTTTGGCGCGTACTGCCGCCTGTGCACGCGCGACTC
TGTGCTCCACCCTCTCTCGAGCTCCGAGTTCCGCGAGGTCATGGGTAGCTTGGAGACGCTGGGCCTTGTC
AACGCGGTCGATGGCAAGAATGGAAGCTTCATTACGCCGCAGACGCCTAGCAAGCGTGGCCGCAAAGCCG
CCGTGTTGGCGAGCGGAGACGACAAGCGAGTTGCGAGCTGTGTGACCGAGAAGGACATGGAGTCGGTTGT TGCCGGCCTCGGCTCAGGCATTCTAGCCAGCATTTTGCATGGAGAGGCGTTGGATTAA
```

Remember, from the second FASTA file in Task 1 we can see that the first 10 codons are "MASSALGKRS".

Therefore, we can check that our process is correct so far by executing the following code (shows the first 10 translated Cdc6 codons with my sequence):

```
###### TASK 2 ######

# Thomas Hollis (BCH441, University of Toronto) -v9.3
#     - Purpose: Genome Annotation Learning Unit (Task 2)
#     - Bugs & issues: no bugs, no issues, no warnings
#     - Acknowledgements: thanks to Prof. Steipe's learning units which were of great help

if (! require(Biostrings, quietly=TRUE)) {
  if (! exists("biocLite")) {
    source("https://bioconductor.org/biocLite.R")
  }
  biocLite("Biostrings")
  library(Biostrings)
}

Task1Seq <- readAAStringSet("./data/tom_orth.fasta")[[1]]
cat("Ensure that the next output is equal to: \n", as.character(Task1Seq[1:10]))
Task2Seq <- translate( (readDNAStringSet("./data/tom_orth_upGen.fasta")[[1]])[501:531] )
cat(as.character(Task2Seq))

# END
```

Output:

```
> Task1Seq <- readAAStringSet("./data/tom_orth.fasta")[[1]]
> cat("Ensure that the next output is equal to: \n", as.character(Task1Seq[1:10]))
Ensure that the next output is equal to:
 MASSALGKRS> Task2Seq <- translate( (readDNAStringSet("./data/tom_orth_upGen.fasta")[[1]])[501:531] )
Warning message:
In .Call2("DNAStringSet_translate", x, skip_code, dna_codes[codon_alphabet],  :
  last base was ignored
> cat(as.character(Task2Seq))
MASSALGKRS
```

Indeed, as the above strings match we can confirm our process thus far.

Task 3 - Analysing Genome using Regex

- The yeast Mbp1 canonical binding site is defined by the regular expression [AT]CGCG[AT]. Are there CGCG motifs present in your nucleotide sequence? Identify them using a regular expression search. Are there [AT]CGCG or CGCG[AT] motifs? What about [AT]CGCG[AT]? Where are they located? Do they cluster? Are they arranged in a similar way as the yeast binding sites that you visited at UCSC?

In order to answer all of these questions, I wrote an R script to compare both sequences.

The sequence to compare to is the 500-upstream SACCE sequence found with the same method of URL modification as before. Indeed the original URL (https://www.ncbi.nlm.nih.gov/nuccore/NC_001142.9?from=69338&to=70879&report=fasta) is modified by subtracting 500 from the "from" field to produce the modified URL (https://www.ncbi.nlm.nih.gov/nuccore/NC_001142.9?from=68838&to=70879&report=fasta) , from which the FASTA is extracted.

Here is the interleaved input and output:

```
> ###### TASK 3 ######
>
> # Thomas Hollis (BCH441, University of Toronto) -v9.3
> #     - Purpose: Genome Annotation Learning Unit (Task 3)
> #     - Bugs & issues: no bugs, no issues, no warnings
> #     - Acknowledgements: thanks to Prof. Steipe's helper code which was of great help
>
> mySeq <- as.character(readDNAStringSet("./data/tom_orth_upGen.fasta")[[1]])
> mySeq2 <- as.character(readDNAStringSet("./data/tom_Cdc6_upGen.fasta")[[1]])
>
> patt <- "..CGCG.."
> m <- gregexpr(patt, mySeq)
> cat(m[[1]], "\n")
128 187 197 207 236 253 532 833 963 1424 1857 2069 2089 2130 2171
> (out <- regmatches(mySeq, m)[[1]])
 [1] "CCCGCGCC" "GTCGCGTA" "GGCGCGTT" "AACGCGCT" "TGCGCGAC" "GGCGCGAG" "GCCGCGCG"
 [8] "GCCGCGTA" "CCCGCGGG" "AACGCGCT" "CTCGCGCT" "GGCGCGTA" "CACGCGCG" "TCCGCGAG"
[15] "AACGCGGT"
> m <- gregexpr(patt, mySeq2)
> cat(m[[1]], "\n")
95 284 296
> (out <- regmatches(mySeq2, m)[[1]])
[1] "GACGCGGG" "GACGCGAG" "CACGCGTC"
>
> patt <- "[AT]CGCG.."
> m <- gregexpr(patt, mySeq)
> cat(m[[1]], "\n")
188 208 1425 1858 2090 2172
> (out <- regmatches(mySeq, m)[[1]])
[1] "TCGCGTA" "ACGCGCT" "ACGCGCT" "TCGCGCT" "ACGCGCG" "ACGCGGT"
> m <- gregexpr(patt, mySeq2)
> cat(m[[1]], "\n")
96 285 297
> (out <- regmatches(mySeq2, m)[[1]])
[1] "ACGCGGG" "ACGCGAG" "ACGCGTC"
>
> patt <- "..CGCG[AT]"
> m <- gregexpr(patt, mySeq)
> cat(m[[1]], "\n")
187 197 236 253 833 2069 2091 2130
> (out <- regmatches(mySeq, m)[[1]])
[1] "GTCGCGT" "GGCGCGT" "TGCGCGA" "GGCGCGA" "GCCGCGT" "GGCGCGT" "CGCGCGA" "TCCGCGA"
> m <- gregexpr(patt, mySeq2)
> cat(m[[1]], "\n")
284 296
> (out <- regmatches(mySeq2, m)[[1]])
[1] "GACGCGA" "CACGCGT"
>
> patt <- "[AT]CGCG[AT]"
> m <- gregexpr(patt, mySeq)
> cat(m[[1]], "\n")
188
> (out <- regmatches(mySeq, m)[[1]])
[1] "TCGCGT"
> m <- gregexpr(patt, mySeq2)
> cat(m[[1]], "\n")
285 297
> (out <- regmatches(mySeq2, m)[[1]])
[1] "ACGCGA" "ACGCGT"
>
> # END
```

So we can see that there are indeed some CGCG, [AT]CGCG, CGCG[AT] and [AT]CGCG[AT] motifs throughout the sequence. This is true for both sequences.

Task 4 - Interpreting findings

- Interpret your finding. Does this support or refute the idea that MBP1_MYSPE has the same DNA sequence binding specificity as MBP1_SACCE?

I am not a biologist but I will do my best to interpret my findings. I took the sequences and aligned them in Sublime as best I could. From this information and since the regex has shown a high similarity in the clustering, it seems that MBP1_ISAFU (a.k.a. MBP1_CORFU) has the same DNA sequence binding specificity as MBP1_SACCE.

Conclusion and Outlook

I struggled with this one but the NCBI website saved me! Such good documentation! The next unit that I plan to take is PDB files (http://steipe.biochemistry.utoronto.ca/abc/index.php/RPR-SX-PDB) .

5.4 Homology Modelling (http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-INT-Homology_modelling) (submitted: 8 marks)

Objective

My objective here is to create a flawless publication image of MBP1_MYSPE APSES domain bound to DNA, based on the 4UX5 structure.

Time estimated: 6 h; taken: 8 h; date started: 2018-11-21; date completed: 2018-11-23

Activities

You can find my submission page here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php?title=User:Thomas_Hollis/Eval_Homology_Modelling&action=edit&redlink=1) .

This is what my final publication image looked like (it took me almost half a century to create but I did my best):

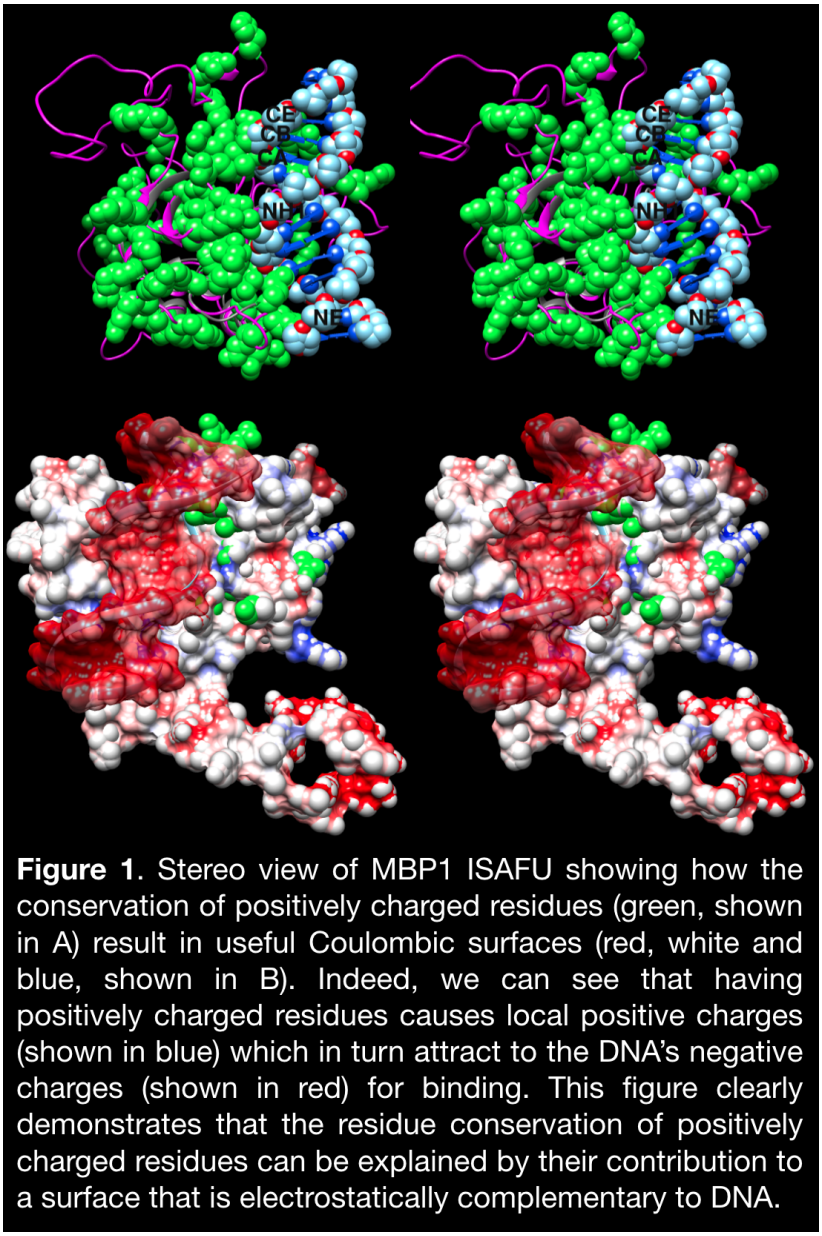


Figure 1. Stereo view of MBP1 ISAFU showing how the conservation of positively charged residues (green, shown in A) result in useful Coulombic surfaces (red, white and blue, shown in B). Indeed, we can see that having positively charged residues causes local positive charges (shown in blue) which in turn attract to the DNA's negative charges (shown in red) for binding. This figure clearly demonstrates that the residue conservation of positively charged residues can be explained by their contribution to a surface that is electrostatically complementary to DNA.

Conclusion and Outlook

Wow this is the final unit... Cannot quite believe it... This might be a good time to do a big review of my journal and run through any chapters that I feel are weak!

6. Custom Learning Unit

6.1 Cancer Detection using kNN (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_Cancer_kNN) (submitted: 14 marks)

Objective

My objective here is to write an absolutely fabulous introductory learning unit to Machine Learning applications (in this case kNN) in bioinformatics (in this case cancer detection).

Time estimated: 12 h; taken: 20 h; date started: 2018-11-09; date completed: 2018-11-15

Activities

You can access the page here (http://steipe.biochemistry.utoronto.ca/abc/students/index.php/User:Thomas_Hollis/Eval_Cancer_kNN) .

Please see the version history to see all the development that has gone on in building this unit from v0.1 (stub) to the latest version!

I am so proud of this unit. I designed it in the same style as prof. Steipe's units so that it could easily be added to future courses without too many changes. I absolutely loved creating it and I really hope some student stumbles across it some day (I must confess I spend a bit too many hours on this unit and I got really carried away...)

Conclusion and Outlook

This really was great fun. I hope my learning unit is taken by someone in the future! I am even considering having a wiki structure for any course that I may end up teaching if my career heads toward education.

Notes and References

1. ↑ SGI Canada. (2018). Fire prevention and safety. Retrieved September 16, 2018, from https://www.sgicanada.ca/news?title=fire-prevention-and-safety
2. ↑ Statistics Canada. (2016). Canada 2016 Census. Retrieved September 16, 2018, from https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016005/98-200-x2016005-eng.cfm