In Dialogue with Mathieu Harb

iGEM CONCORDIA

ASTROYEAST 2020

Mathieu Harb
Concordia University, Hallet Lab
August 5th 2020

Interviewed by iGEM Concordia Montreal
(Benjamin Clarke, Maher Hassanain, Hajar El Mouddene)

SUMMARY KEYWORDS
genes, study, upregulated, chromosome, algorithm, normalization, machine learning, heat maps, data, software, cluster, compare, microarrays, database, heat map, microgravity, upregulation, python

iGEM Concordia 00:33
The first thing we're doing this year is we're building software, where scientists can just go and look up certain genes and see if they're upregulated or downregulated. And fetches data from different studies, published to our database. Using the information that we have in our database, we are selecting reporters, our stress reporters, and we're going to be building our resistant strains. Which are strains that will be resistant to microgravity-induced stress specifically in yeast, or *S. cerevisiae*. I'll show you a working model of the software, we're pretty happy about what it will be.

Mathieu Harb 03:23
So you are working with more microarrays and RNA bulk?

iGEM Concordia
We've only started with microarrays, because it's a lot easier to deal with.

Mathieu Harb
Did you deal with single cell stuff, but bulk RNA is pretty simple, right? It's very similar to microarrays?

iGEM Concordia 03:55
Yeah, we just don't really have the facilities at hand yet for that. As you're aware we're collecting a lot of data from different organisms and different studies and different types of assays and so on. These are the eight criteria that users can search based on any of them. For example, working on certain species, organisms, will display all results that has to do with that organism. Trying to display basic information like the gene symbol, the adjusted p-value, the fold change, the generation for the study, the organism, species... And for each we also put all the relevant information when it comes to the data itself, GO functions. GO components and so on, which GEOD study and experiment information that was linked to it. Yeah. With a link to that study. It was pretty hectic programming all these combinations.Using Node and Express JS for back end along with MongoDB Database and EJS (Embedded Javascript Template) for front end. So HTML, CSS, Jquery and AJAX.

Mathieu Harb 05:58
Very cool. Okay, so what do you guys, what do you guys need from me? If I may ask.

iGEM Concordia 06:57
We had some questions about more data. About the EFA Chain, and how we are looking at our studies and evaluating them. Maybe it'll be clear if I can show some of the questions we have.Okay, so, for example, if we display current data, let's say, the current way of doing it that I'm sorting it based on adjusted p-value. From there it comes in, then everything is listed. And that way, since we have so much data, for now, I'm just limiting the fish number, okay, till I'm able to paginate it or infinite scroll depending on what I am going to implement. Let's say we choose two criteria this time. Let's say we choose 'organism' and 'fold change'. I'm still sorting based on adjusted p value. But it only shows upregulated fold changes instead of negative ones. Initially, we thought fold change was the main thing scientists would be looking for, because that would show how much difference in expression happened. But then we learned that it's more like a combination of the adjusted p value plus the log FC, what scientists would use as a good indicator. So we're just wondering how does this actually play, high adjusted p values are bad studies?

Mathieu Harb 09:25
I don't work with adjusted p value. So for me I, to me, that means nothing, right? The way, like the way I've been taught with Dr. Hallet and all that stuff, and how we look at data is more when you cluster that's the thing, I worked single cell I don't work with bulk or microarrays and even with microarrays, the thing that we normally do is you look at something that would be the norm. And then, you have to have your control and then off of your control, you look at what the upregulated things are, which can tell you, PCA and then with all the paths and mechanisms… like for breast cancer, for example, if you have estrogen, that's upregulated, there's going to be a brick load of genes that are going to be upregulated with it, because they're all next to each other on the chromosome.

The pathology, exactly, it comes down to more of like, oh, if this chromosome, this part of the chromosome's upregulated, what are all the genes that are associated to it, and then you'll be able to know like, well, that's what's getting affected and whatnot. And stress responses are normally, for us, I mean, I know humans more than anything else, it's mitochondrial genes, the more mitochondrial genes you have, the more stressed the cell is. And that's the way you read how cells are stressed. For upregulation, I mean, the majority of the study are pushing for whatever gene to be upregulated, right, they want to see how it affects. Like if I were to upregulate, RGI1, like the first gene there, and then I look at and then I do microarrays on on that, on that species of yeast that I just upregulated that gene, then you compare it to the one that wasn't, and you just look at which ones are all higher, and then you can say, well, this gene affects all of these. And you can also see the ones that are downregulated. That's how you can basically figure out what the mechanism and paths are for whatever genes you're upregulating one at a time.The fold change, yeah. And then and then it needs to be linked, I think an important thing would be linked to studies. Right? For whatever geo accession number you're getting, all the genes that were associated with that same accession number would have to be compared to one another, right? Do these come all from the same study?

No. So this one came from this 25th generation. This was the same study, but 5th generation, and this was a different study.

Mathieu Harb 12:43
That's a thing, depending on what the study was looking at, the log fold change can mean something completely separate right? The biggest factor is that whatever the log fold change is, it has to have a control, you need to be able to know what the control is. If it's an increase in log fold change, but you were increasing that chromosome, you should see all the other genes within that chromosome also be upregulated. Whereas another study that just did an upregulation of chromosome 10, but this is on chromosome 11. And you're seeing that upregulated. That saying that whatever was affected on the 10 is affecting whatever is on 11. So then you can make a link and understand more about what's going on within that. Right, you see what I mean by that?

iGEM Concordia 13:34
Yeah. Is it the same as this almost like can you know, just from the platform ORF itself? Or more about the chromosome location?

Mathieu Harb 13:43
I know, for humans, it's chromosomes a lot. I'm not sure but the platform ORF. I think it does play a role, if I remember, right, I did a bit of yeast stuff, because I have somebody in my lab that does that. The biggest thing for me,, if I were to look up a gene, right, it has to be the same study. So if you could categorize it by study, that's massive, right? Because while you're categorizing it by study, you can say, okay, all these genes that came from the same study, they're based on the same control. If they're all based on the same control, then you can know the log FC is all relative to the same thing, right? Because without a baseline, numbers mean nothing. Yeah.

iGEM Concordia 14:40
Let me ask you this, Mathieu, because exactly the argument that somebody else he's a microgravity researcher in bacteria, and he brought up the same point. He had said that if we include the a heat map for connectivity, it kind of gives a like a better idea. In comparing that gene versus all the other genes that were in that study.

Mathieu Harb 15:05
That's super good as well. Heat maps are insanely telling.

iGEM Concordia 15:13
Heat Map for every GEO study in the list, and then be able to display...

Mathieu Harb 15:19
You'd have to cluster them by study, right? That would be the important thing. Yeah, you have your heat map...You know what, I'm going to share really fast. I will share a whiteboard screen share. All right. Okay, so you know how heatmap works, right? Like this. And then, let's say here are the genes, this whole thing is genes, right? You know how heat maps work, right?

iGEM Concordia 16:12
I know they look, but yet so much how they work.

Mathieu Harb
This is a heat map, for example. Right? Yep. So the coloring, whatever, doesn't matter. The coloring is a set. Each column is a cell. Whereas each row is a gene, right? But each column would have to be the studies. The thing is, you're gonna have repeated genes, right? Because for study one, let's say this is one study and it has all these genes. You have to keep them all clustered together. And then this one would be another study where, this would be study two, right? So on, but the genes within them are pretty much all the same because otherwise, you're comparing random genes to one another.

If you have all these studies that are just showing what their genes were. I don't know the genes of yeast, but you know that they have to be all the same genes so that you can see what each study compared to one another would show. You might see study one, just assume that these are exactly the same, right? These genes are highly increased, right? Blue means downregulated. Red means upregulated. And then you say, oh, wow, it's study two, you can just compare and contrast. That's the fun part about heat maps is that it allows you to literally see if there's differences between studies. So you'd have all the genes here, and then you'd have studies at the top. The studies that you're going to be using here, right, like study one, study two, study three, the studies that you're going to be using.

The problem with that is that they need to be looking at the same thing. If they're not, you need to know that they're not, because if they're not, if, let's say study, one was the upregulation in chromosome 11. And then study two is looking at chromosome upregulation. Three, well, you're going to expect stuff like this, right? You're gonna expect study one to have upregulation of those genes, whereas in study two, you had upregulation of the genes in the middle, which are different, but that's expected, which means that makes sense. Right? If one study is looking at chromosome 11, and another studies looking at chromosome 3, you're not going to expect the same genes to be upregulated. Because they're fully different chromosomes. But if you end up seeing study one upregulated, chromosome 3, and then study two upregulated, chromosome 11. And yet chromosome 3 has an upregulation in it, then now you could you can start making links because you're saying if you touch this chromosome, it's messing with the other one. That's how it lets us see differences. This is what heat maps allow you to see. It lets you cluster stuff. And it lets you see clear as day between some of the things.

Mathieu Harb 21:37
When you're doing something like that, across studies you have to group them by studies. Me I'm grouping them by cells, you would be doing it by study basically, instead, and you'd have your list of genes on the right. So that you can see across studies, what's being fiddled with and what's not being fiddled with. And if you're dealing with the microgravity ones, it's especially important because you can compare a bunch of microgravity studies with the exact same genes that we're all looking at and see if there's a trend or not. It allows you to basically see trends and whatnot. And then if you see some trends, then you look up those genes, where they are, what their effects are. If I look up a gene, let's say CDH3, if you look up this gene,some random gene, it has a bunch of studies, right? It's going to have study, on study, on study, on study,it's going to have a bunch of different studies.

The important part is, you have all these different studies that would be looking at it in different ways. You have to be able to compare the studies that did it in the same way, right? Because if I'm looking at this CDH3 in breast tissue and then I look at it in liver and I'm like, wow, it's so downregulated, well obviously, it's two fully different types…[timeframe resets] Basically, the goal is you need to have a basis that compares everything relative to something. You can't be comparing two studies that look at two fully different things and look at the fold change. It's going to be meaningless. You could look maybe, at chromosome, that's one that you had as a factor. So sorting it by, you can group them by that, or you can have heat maps, where it groups them by study, and heat maps are, trust me, they're way simpler than they look. If you want to do an R or Python, like both of them are super simple to do. Basically, you have to try and figure out a way to have a baseline that compares everything with without comparing everything to one another with no bias. You can't have a bias. Because if you have a bias, it's useless to compare things. It needs to be based on the same thing.

iGEM Concordia
To make the heatmap I'm trying to think of how to do it in our but you basically just column binding the two studies together, but telling it not to cluster the studies together, right?

Mathieu Harb
Yeah, let's say your rows are genes and your columns are studies, you just turn off column clustering, so that when you make your array or your matrix, which has a study, the first column is study one, the second column is study two, study three, all the way to study, nothing gets moved around. Basically, it's in that order, it stays in that order. That's the important part.

iGEM Concordia
Would you do hierarchical clustering on the rows?

Mathieu Harb
Hierarchical clustering on genes is normally a good thing. Because it ends up saying like, all the red ones will end up together, right? It's a little easier on the eyes.

iGEM Concordia
And people do that by correlation, I think, or is it by something else?

Mathieu Harb
Hierarchical clustering?

iGEM Concordia
Yeah, like the distance function.

Mathieu Harb
There's so many different algorithms, each one has their own. For Python, you have Seaborn and for making the heat map, I’m trying to see what the argument is again, it’s called Method. Look at all those methods, you have single which for all points I in cluster U and J and cluster, like its nearest point algorithm you have complete, which is farthest point algorithm, you have average. You have weighted, you have centroid, you have median, there's all these different kinds of clustering algorithms.

iGEM Concordia
I think for micro arrays, people were doing a Euclidean distance. I'm not sure though.

Mathieu Harb 04:04
Okay. I've seen the Ward used a lot. And that's what Mike loves. So that's just the one I've been using lately. It’s a variance minimization algorithm.

iGEM Concordia 04:31
Now I'm having problems understanding how we can actually integrate this into the software, right? Image, it's simple, but if you want to make somehow correlation for every gene...

Mathieu Harb 04:45
That algorithm is implemented in making the image it's in the function. It's just a parameter and in the function, all you do is you're doing whatever- heat map- parentheses- this data- comma- method- equals bam, and then it'll use that method for the clustering. Don't worry about that. It's literally just a parameter. You don't have to program in that clustering algorithm yourself.There's so many different heat mapping libraries, making a heat map is a lot like once you have your first one, you're just like, oh, that was it. All you have to do is you make your 2D array. And that's it. I mean, it's not that hard to make your 2D arrays, if you have the data, right? You already have your data.

So you just have to combine it to make one massive array. And then you just, you just flop that. It's a plot, right? Think about a plot, except you have a lot of dots in it. That's about it. And merging the data. I mean, how much data manipulation have you done?

iGEM Concordia 06:34
Not a lot. I mean, we're using limma. And we're just pulling log FC values from GEO.

Mathieu Harb 06:47
You have been able to load the microarray data?

iGEM Concordia
Most of the work I've been doing is looking at specific studies and figuring out what the experiment is. Because for every study, there's a ground control, and then there's the space flown or simulated space, simulated microgravity. I just look into that, since there's several different ways of doing microarrays, right. There's two colors and single channels and everything like that.

Mathieu Harb
Is all your data on our server? Our server, as in my labs server, because last years, iGEM used our server. Where are you?

iGEM Concordia
Not yet. Everything's on AWS (Amazon Web Services, Amazon cloud) On the cloud.

Mathieu Harb 07:41
How much data do you guys have right now?

iGEM Concordia 07:50
4000. Roll. It's pretty small. I can tell you exactly how much...It's .013 Gigabytes so far.

Mathieu Harb 08:00
Yeah. Because I'm sure there's not hundreds of studies on microgravity, it's very little, tiny amounts of data. Okay.

iGEM Concordia
We only have like five studies working right now. We have two in yeast. One Arabidopsis and three or three or four bacteria ones.

Mathieu Harb
Okay. And that's the microarray data, though.

iGEM Concordia
Yeah, we'll try to get to RNA seq later. But we need server access.

Mathieu Harb
RNA seq is a lot of data. much bigger data.

iGEM Concordia
I didn't know before going into this that it’s a couple hundred gigs.

Mathieu Harb
Yeah, I mean, I did four breast cancer cell lines. And let me see how many Gigs it was like the raw reads from the sequencing was 35 Gigs. And that's just one sequencing. All four samples went on the same sequencer together, like, 35 that's just one study, right?

iGEM Concordia
And people don't upload count files?

Mathieu Harb
People just upload raw fastq, or I don't know what it would be for microarrays.

iGEM Concordia
Oh no, they do for microarrays for some reason they upload FC values, so the fluorescence values, that's pretty small because it's just one value per gene.

Mathieu Harb
That's the thing it's all bulk right? You have all that population into one big sum. It's at least it's easier. What is it uploaded it in? Is it just a text file or a CSV?

iGEM Concordia
I make the CSV, but then we store it in MongoDB.

Mathieu Harb
Okay, so what do you get the like when you look at the GEO accession number? I've never worked with microarray data.

iGEM Concordia
It's a very specific file format. There's like three or four going around, but I'm using an R package called GEOquery, then it's making it fairly simple for now. It stores it into an actual R object called an Expression Set and then we can work with that. I think the basic one is the CEL file.

Mathieu Harb
Then the files you get, have they been normalized or do you get really raw, raw data.

iGEM Concordia
It should be normalized. Often you get metadata information, and it tells you it's been normalized, and it tells you how it's been done. But then you get it. And then you look at if it's median centered or not, and it's not at all. It's kind of like up to your discretion, or like, I don't really know exactly when they uploaded it. It was before normalization or after. Most of the time it's normalized.

Mathieu Harb
If it's normalized, you'll see the numbers all weird, and not just solid numbers anyways. If it's point, whatever, you'll know, if it's all like integers, then yeah, it's not normalized. Normalization should obscure the numbers, right?

iGEM Concordia
Yeah, it should be like, median center. It should be around zero for everything.

Mathieu Harb
That’s good. There's so many normalizations. That's the biggest issue right now. I think in single cells that there is no baseline. So every study you read normalizes a different way.

Okay, you have all you have some data, which is good. Honestly, what I recommend right now, play with your data, like learn to manipulate it a little bit? Because you are doing actual biological experiments, right? Yeah. And you will get data from those experiments, so learn how to play around with data that you already get. So that once you get that new data, you can process it as fast as possible and get your basic understanding of your data as fast as possible. Have a pipeline from getting that CEL file and txt file, to doing your QC on it, to doing your normalization and doing really basic differential expression analysis on it. So that once you get your data, you can be like, okay, let's do a quick little breakdown, you can see it and be like, okay, this is what we have.

And then that's where the longer part comes in, where you have to look at every individual thing to see like, Okay, this was pushed up, and this is linked to this…You need to have like, you need to have that pipeline ready to go. As soon as you get your data. That's what would be to me the most important thing so that once you get your data, you can process it fast.

iGEM Concordia
When I came into this, I assumed that there was a pipeline that you could make that would do that for you. But it seems like for every single study, it changes every time.

Mathieu Harb
So you have to figure out what you want your pipeline to be. And the way to do that is playing with other data and seeing what you prefer. Is there a normalization method that you'd rather have? For me, I did an experiment where I tried four different types of normalization are my single cell data. I did some breast cancer stuff classification on them and saw which ones were more accurate...I looked at a bunch of different breast cancer like known breast cancer cell lines, I took all the raw data and normalized them all, with four different types of ways. And then with all those normalizations, I classified what subtype of cancer it would be. And then I looked at what the actual one was, and the ones that were the most accurate. Those were the ones I took. You have to figure out a way to find what processes and know what methods you think is best for you. Since there's not a lot of microgravity studies, try and see if there's some similarities in the way they processed their data. That would be obviously helpful. Because if they're all using the same basic baseline, then obviously that's probably the standard within that field, so use whatever they use.

But yeah, the most important thing in the end is you have to find a baseline for comparing stuff. That website, you need to have some baseline because a lot of studies will have a bunch of genes. But some did something where it affects the gene one way, and the others did it another way. So you have to find what that baseline is, to me is that I think that's the big component right now. I don't know if that's clear. If it's not, tell me I'll try and make it clear. I'm not the best with words. I'm really good with things in my head, but I'm never good at articulating.

iGEM Concordia
Have you worked with bulk RNA seq. before?

Mathieu Harb
No, I've literally just done single cell stuff. We just got bulk data, and I will be getting more, I will be getting both data with the new stuff that I'm doing right now. You know what, that's a lie. I have experience with bulk RNA seq. Yes. Because now that I think about it, do you know what the CCLE is?It's the Cancer Cell Line Encyclopedia. It's just a bunch of cancer stuff. And the majority of the data is all bulk RNA seq and I've downloaded their data. Yeah, I've taken their data and used it so yes, I have experience with that.

iGEM Concordia
I don't want to bog you down with questions because this is the first time I've done this stuff.

Mathieu Harb
It's alright dude, it takes time to understand so bog me down with questions whenever you need.

iGEM Concordia
If there's anything that comes up, because we don't have server access yet, hopefully we'll get that soon. But if that comes up, hopefully if you're around if I need some help.

Mathieu Harb
Yeah, absolutely. Processing it now for me is instant. It doesn't take any time.

iGEM Concordia
It doesn't seem that complicated, but I've never used R remotely before.

Mathieu Harb
Doing SSH is for a portal to something and then you can use it on your computer. All those little things. No, don't worry. I've been doing it for two years now. I've gotten the hang of it. At first, though, it's a little overwhelming. I get it. I fully understand.

I mean, yeah. Don't worry about it. I'll give you my Facebook so you can just message me there whenever you need help. Right. That's great. That's fine. I don't mind you can all add me. So yeah, bulk RNA seq data, depending on which one you get. If you get the count matrix, right, which is nice, because you don't have to go through all the processing of that file from the sequencer. Perfect.

iGEM Concordia
The count matrix files are still pretty big thing, no?

Mathieu Harb
Count matrix is no, not that big because they're text files or CSV. I mean, there'll be a couple hundred Megs but right. For the CCLE. Recounts 52 Megs and that's 1000 cell lines. I think right now you should try for the grouping on the website, try and group it by study as well, like instead of log fold, because it's harder... if you have the same gene, but one is downregulated and the others upregulated. There's clearly a difference in the study that that happened. Or you could add another column where it says like, it's a different study. You know what I mean?

iGEM Concordia 21:24
That's what I was thinking initially, because there's already a search for like, the egd number, the accession number.

Mathieu Harb 21:31
So you could just add that as a column. So when you're looking, you're like, Okay, well, this one's downregulated. And this one's upregulated. But there come from two different studies. Let's see what the difference was in the study. Oh, one affected this one affected that. Okay, that makes sense. That's not a bad idea, either.

iGEM Concordia 21:53
And I was planning to maybe link the map to each one with that. So if you lets say you clicked a certain study for, let's say, it just be HSP30. And then let's say, you get the display, and then whatever you click on each one of them, it will show you like the whole map for that study.

Mathieu Harb 22:16
Yeah. Exactly. Because your microarray, I'm assuming the data you get, it's more than just one sample, that's how you'd have it, you'd have your sample 1,2,3, all the way to n, for study 1-n. Like all together, then you'd have study 2 with sample 1 - n, sample three with 1-n, and then the genes. The genes are your rows and have a blast with putting clustering on them. That's fine. That's basically how you should do it. For the heat map.

[shares an image of a heat map]

If you look at this, this is much prettier in my opinion. If you look at that heat map, right? Clearly you see all the sample IDs at the bottom. And then you see how one's red and one blue. Let's just say red is sample study one and blue is study two. Yeah. So that's, that's really nice. Whereas the genes can flip and whatnot. But you see how there's clearly a difference, right?

iGEM Concordia 23:51
Yeah. Which, is this made an R or is it..?

Mathieu Harb 24:00
I could do that in Python. I could do it in R either. Or it works. I hate R so I majorly go to Python...

iGEM Concordia
We're kind of forced these days because there's the only packages available to process microarrays in R. I couldn't find a good one in Python. At least.

Mathieu Harb
What do you mean? What is it in? What are you trying to do? What do you mean with it's only available in R? What are you processing?

iGEM Concordia 25:07
Well, to just query the GEO database? I'm pretty sure that's available in Python, but the package Lemma, which is like a mixed model...

Mathieu Harb
Because you can wait when you put a GEO Accession number, and then it downloads the data straight into a data frame in R. I manually fetch the data from the GEO accession number and store it on my server. That's the difference. I have the data on my server, and then I can just load it because it's a file, right? You downloaded it directly? That's pretty cool, actually, I did not know that you can do that.

iGEM Concordia
Yeah, it's pretty convenient. Lemma, that's the differential expression equation. I mean, it's a mixed model.

Mathieu Harb 26:05
There's something called GEO Parse? In Python, GEO parse is a Python package that can be used to query and retrieve data from GEO. Inspiration is from the great R library GEO query. Everything exists in Python, because there are libraries, there are people who just do everything in Python.

iGEM Concordia
But there's no Lemma equivalent in Python that I found.

Mathieu Harb
What's so great about Lemma?

iGEM Concordia
It's just really great.

Mathieu Harb
I feel that you're hooked to a library. I understand that a lot.

iGEM Concordia
I mean, it's just you can do differential expression analysis in three lines of code if you want it to. Mathieu HarbSay no more. I get it. I know how it is. Once you love it. Once you love a file. You love it. And that's always it's how it was for me as well. So yeah, that's the thing for me. It's Python. I love it. I know how to do everything with Python. When I worked at Amazon, I did everything in Python. I think that's what set my baseline. It was that when I worked at a big company, and then I just stuck with that language.

iGEM Concordia 27:48
I haven't done much computational work in Python.

Mathieu Harb
Yeah, machine learning is literally based in Python. Everything in everything in machine learning is Python. AI is literally Python based right now. It was R very long ago. And now it's fully Python. My brother does AI like that's his. That's what he did his PhD in. That's all it is. It's pure, pure, pure, pure Python.

iGEM Concordia
You would freak out if you heard what the original proposal of our project was. So to find GSEs is the first part. Initially, the plan was to make everything automated. To find these studies, then perform differential expression analysis. Actually, we're supposed

to read the text yet, you're supposed to use, like a mix of NLP (Natural Language Processing) plus more machine learning to be able to find genes from the paper and basically provide all information about the gene, it was a bit vague. We're not going for the differential expression we're going for, let's use machine learning and NLP and be able to have a program,I don't know, it feels like it's going to be a bot or something. They'll be like reading all these papers, finding all relevant genes, how they're affected and their information and it just keeps doing this forever. Yeah, and automate this process pretty much.

Mathieu Harb
So you use something like Jenkins (Open source automation server) you can use that and that would do that. That's a cool idea. But it's a personal project honestly.

[...]

Mathieu Harb 30:40
Exactly. Now, machine learning is tough. It's not as easy as it looks. I still don't grasp it fully.

iGEM Concordia
Most of what I did I was able to just build the base algorithm. And basically, I was able to teach it because there were like, two available universities, I believe one in Australia, one in the United States, where they had this already, training data data. And one of them was 93%, and one was 94%. All I did was merge them both together. And that was like, 97%. But to make up my own data. I have no idea how to do that.

Mathieu Harb 31:27
That's the fun part of machine learning is making a training set. You have to manually do it. And it has to be millions of things long. Yeah. You have to manually, it's harsh.

iGEM Concordia
I just had like, one last question. Yeah, we, as you saw, like, we're categorizing these genes up or downregulated. So far, all ll we have done is, on the basis of our p value, well, we should maybe have a category of unaffected and Maher had the proposal, I don't know if you want to explain it, just to make sure you're doing it, right.

Mathieu Harb 32:18
So like just unaffected genes, how to categorize them?

iGEM Concordia 32:21
Yeah. How do we define unaffected genes out of these studies?

Mathieu Harb 32:26
You'd have to look at the log Fold Change and see if it's near zero, right. You have to make your own criteria, is it 0.05 to minus 0.05. If it's within that range, then you just put it as not affected because it's such a little change, that it could just be the cell was in a certain state or its environment was making it so that gene was just a little less used?

[End]