



Adam <chhoradam@gmail.com>

Update and Questions about Yeast Automated Conservation Tool

8 messages

Adam Chhor <achhor@uw.edu>

Mon, Nov 25, 2024 at 6:25 AM

To: Emily Parnell <emily.parnell@biochem.utah.edu>

Cc: "Matt Miller (Biochemistry)" <matt.miller@biochem.utah.edu>, Brent Lagesse <lagesse@uw.edu>

Hi Dr. Emily,

I hope you are doing well. I wanted to give you an update on the tool and want to confirm that the result I am getting is accurate.

On the status of the tool, it can now with one click given the UNIPROT sequence entry like MP51 do the things listed below automatically:

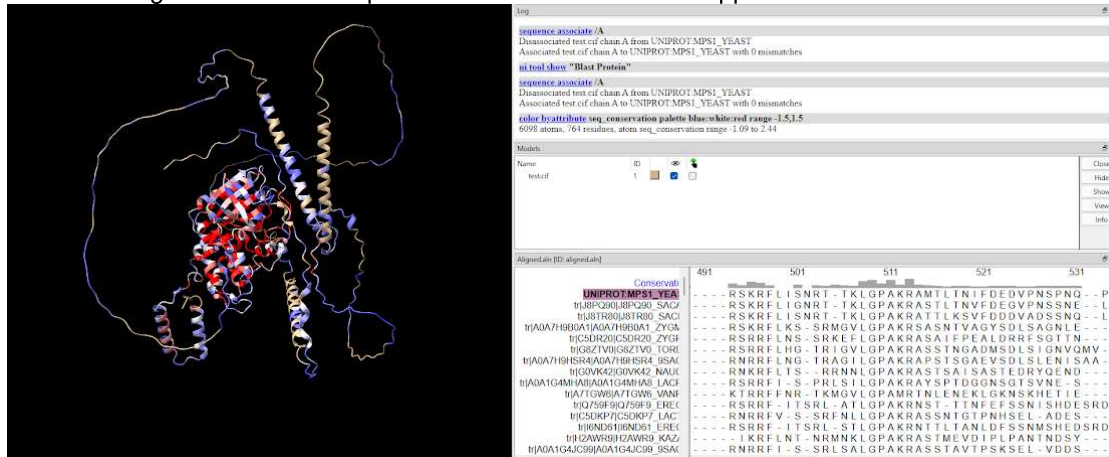
- Get the protein model from AlphaFold
- BLAST, and from the 1000 results, filter to the minimum similarity specified (i.e. $\geq 50\%$). Then, take every nth sequence until it is 20 sequences and put the sequence in one fasta file. The user can customize any of these values.
- Run Clustal Omega (Multi alignment) from the fasta file from the last step
- Open Jalview with the alignment from Clustal Omega
- Open ChimeraX with the model from AlphaFold along with the Clustal Omega alignment. Then, automatically run commands on the ChimeraX's command line, so the model associates with the sequence and runs the command that shows the spectrum of conservation from blue to red on the model.

Run time: 5-8 minutes (mostly due to BLAST)

I have a couple of clarifying questions and want to double-check check the results are accurate.

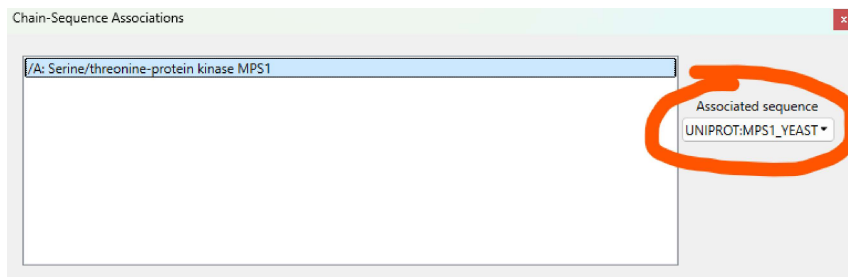
1. Previously, there was an error in the alignment result that caused the conversation to look very weird. I believe I fixed the error causing it and if possible, can you look at the Jalview file I attached? For all the files, I ran with MSP51 yeast.

2. Is the image below what the protein model on ChimeraX supposed to look like? I've attached the ChimeraX file also.



3. On the word document provided about ChimeraX, I was wondering about step 7 what the Associated sequence should be? Should it be the protein you are interested like MP51? I've attached a picture below after I click "Association on ChimeraX."


6. Open the Jalview alignment in ChimeraX (just drag it in)
7. Right click while hovering over the alignment window & choose Structure, then Associations, then choose one of the models (top one is probably fine) – this will associate the alignment with a sequence/structure
8. Then type into the command line at the bottom: `color byattr seq_conservation palette blue:white:red range -1.5,1.5`



4. From my understanding and testing using BLAST using the same sequence, the results are certainly not the same due to how BLAST retrieves the different protein sequences. Is this a problem for your research as if you BLAST twice, the proteins you will get will not be the same?

Best,
Adam

2 attachments

 **mp51_yeast.cxs**
422K

 **mp51.jvp**
160K

Emily Parnell <u0463089@umail.utah.edu>

Mon, Dec 2, 2024 at 3:13 PM

To: Adam Chhor <achhor@uw.edu>

Cc: "Matt Miller (Biochemistry)" <u6021934@umail.utah.edu>, Brent Lagesse <lagesse@uw.edu>

Hi Adam,

Thanks very much for your work on this. It looks like you've made good progress.

Here are some answers to your questions.

For question 1 on the alignment, unfortunately the alignment you sent me doesn't look correct. Can you tell me which species you used for the alignment? (I can't understand the list on the left side - which is where this information would be) Then I can run one myself & give you an image of what it should look like for MPS1.

For question 2, Yes, the chimera model looks correct. However, I only see one of them (test.cif - See Models at the side), and normally AlphaFold sends back 5. Ideally all 5 would be present; four of them can be unchecked in the Models box so that only one is viewed (either by your program or by the user). Having the other four present would allow the user to examine the others if desired, which could be important.

For question 3, Yes, you're correct; the protein you want to select in the Chimera menu is the original protein name of interest. Are you getting other options that are confusing?

For question 4, certainly it's not ideal that this is what happens (2 different BLAST searches give different results), but I don't know of a way to change this, and I think it will just be a limitation for us. I don't think it's a huge issue, but something for us to keep in mind. Using your tool will make it easy for us to run multiple times and compare the results.

If what I have written is unclear, let me know. We could set up a zoom call if that would be helpful. It looks like the biggest hurdle right now is the alignment problem in question 1. Maybe talking it through would be easier than email.

Also, if you'd like me to run the tool and tell you how it's working for me, I can do so.

Thanks again!
Emily

Adam Chhor <achhor@uw.edu>

Thu, Dec 5, 2024 at 10:47 AM

To: Emily Parnell <u0463089@umail.utah.edu>

Cc: "Matt Miller (Biochemistry)" <u6021934@umail.utah.edu>, Brent Lagesse <lagesse@uw.edu>

Hi Dr. Emily

Thanks for your really clear and detailed responses!

After writing this email, it got very long, so I am putting the more time-sensitive/important parts at the top.

Yes, I think it would be very helpful and a lot easier to meet on zoom to further discuss this and especially the stuff below. I am free next **Monday- Wednesday** (12/9 - 12/11) from **10am-5pm mountain time**. Let me know what times work best for you.

From what I remember, your lab uses Macs, so yesterday I was able to run the program on my friend's Mac. Later today, I will send you some detailed instructions later today on how to run it as it involves installing python and running some commands. If you encounter a problem, we can try to figure it out during the zoom call. Hopefully in the future, this will be easier.

The stuff below talks about the alignment not looking right and responding to your responses on the questions I had:

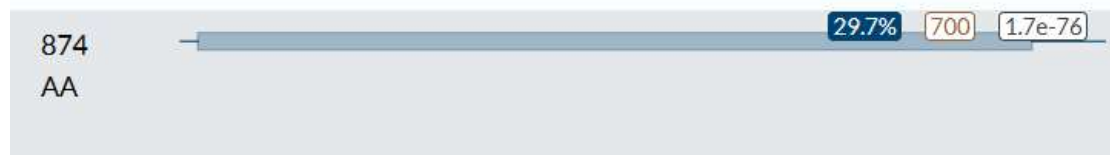
That is weird that the alignment looks off as the way I programmed is that I basically do the same steps of first BLASTing and individually looking up the 20 proteins sequence, combining it into a single fasta, and run multiple alignment on it. I ran a new test and attached the alignment file (.aln) for jalview. and also an excel sheet with the proteins chosen, which clicking the link on the "Entry" column opens the protein on uniprot. Because I am using the same service that is used on uniprot's web interface, funny enough we can view the results on uniprot's website despite not running it on there (though it does expire and become inaccessible after a certain amount of days, so the url might be blank the time you open it).

Note: the .aln /jalview file format I found to be very weird as the left side panel on jalview cannot use spaces and has a maximum character width. I replaced the spaces with '~' and the ends of each column might be cut off unfortunately and there seems to be no work around for this.

BLAST results: <https://www.uniprot.org/blast/uniprotkb/ncbiblast-R20241204-071257-0805-15129465-p1m/overview>

My 3 theories on why it might look wrong are (most likely a combination of all 3 though):

- The results might be too diverse as how its run right now is after filtering the 1000 results determined by % similarity, it automatically finds an optimal number to skip by. For example, if there are 500 results left after filtered by >50% similarity, it would automatically calculate to skip by 25 as since we want 20 proteins, to get even diversity it calculates: 500 results / 20 proteins we want = 25 to skip by.
- From a previous email, I believe you mentioned less than half the rectangle are less interesting. I am able to get the % the rectangle takes up and that is how I filter them i.e. >50%. One of the proteins from the excel file that was chosen was A0A094CS32 and the image below is what uniprot gives from blasting it from MPS1 yeast. Does the % being low matter or is all we care about is the rectangle part?



- A bug in my program

In regards to your responses:

- I am able to fetch a premade alphafold model for any protein from uniprot (it takes <5 second) but it only returns one. The alphafold web interface does return 5 models as it basically calculates and generates new models each time. It seems important to have multiple models to look at, so I'll investigate a way to retrieve different alphafold models.
- Nope, all good! I was just making sure I was selecting the right one.
- Yeah unfortunately, from what I can tell is that BLAST, even the web interface, results are slightly random. Due to skipping every 10 or so, the results may vary very widely, but yeah you can run as many times as you want.

Best,

Adam

[Quoted text hidden]

 **aligned.aln**
82K

Adam Chhor <achhor@uw.edu>

Thu, Dec 5, 2024 at 2:56 PM

To: Emily Parnell <u0463089@uemail.utah.edu>

Cc: "Matt Miller (Biochemistry)" <u6021934@uemail.utah.edu>, Brent Lagesse <lagesse@uw.edu>

Forgot to attach the excel file.

[Quoted text hidden]

 **protein infomation.xlsx**
15K

Emily Parnell <u0463089@uemail.utah.edu>

Fri, Dec 6, 2024 at 9:04 AM

To: Adam Chhor <achhor@uw.edu>

Cc: "Matt Miller (Biochemistry)" <u6021934@uemail.utah.edu>, Brent Lagesse <lagesse@uw.edu>

Hi Adam,

Thanks for your email. Yes, let's meet next week. How about if we target Tues afternoon? You can just let me know specifically what time would be best for you. I am good for the times you listed, with the exception of Wed morning.

Having the list of proteins that you used for the alignment was very helpful. I think the problem is not with the alignment itself but with the proteins that are being chosen. Some don't appear to me to actually be Mps1. I will need to do some testing myself and see if I can figure out something to tell you to change about how the proteins are chosen from the BLAST search. I will do this before we meet.

If you could also send me the info for installing python and running commands by Mon afternoon or so, that would be great (and yes, I am using a Mac). Then I can test it before we talk.

For other things, I think we can just discuss next week.

Thanks again,
Emily

On Dec 5, 2024, at 11:47 AM, Adam Chhor <achhor@uw.edu> wrote:

Hi Dr. Emily

Thanks for your really clear and detailed responses!

After writing this email, it got very long, so I am putting the more time-sensitive/important parts at the top.

Yes, I think it would be very helpful and a lot easier to meet on zoom to further discuss this and especially the stuff below. I am free next **Monday- Wednesday** (12/9 - 12/11) from **10am-5pm mountain time**. Let me know what times work best for you.

From what I remember, your lab uses Macs, so yesterday I was able to run the program on my friend's Mac. Later today, I will send you some detailed instructions later today on how to run it as it involves installing python and running some commands. If you encounter a problem, we can try to figure it out during the zoom call. Hopefully in the future, this will be easier.

The stuff below talks about the alignment not looking right and responding to your responses on the questions I had:

That is weird that the alignment looks off as the way I programmed is that I basically do the same steps of first BLASTing and individually looking up the 20 proteins sequence, combing it into a single fasta, and run multiple alignment on it. I ran a new test and attached the alignment file (.aln) for jalview. and also an excel sheet with the proteins chosen, which clicking the link on the "Entry" column opens the protein on

uniprot. Because I am using the same service that is used on uniprot's web interface, funny enough we can view the results on uniprot's website despite not running it on there (though it does expire and become inaccessible after a certain amount of days, so the url might be blank the time you open it).

Note: the .aln /jalview file format I found to be very weird as the left side panel on jalview cannot use spaces and has a maximum character width. I replaced the spaces with '~' and the ends of each column might be cut off unfortunately and there seems to be no work around for this.

BLAST results: <https://www.uniprot.org/blast/uniprotkb/ncbiblast-R20241204-071257-0805-15129465-p1m/overview>

My 3 theories on why it might look wrong are (most likely a combination of all 3 though):

- a. The results might be too diverse as how its run right now is after filtering the 1000 results determined by % similarity, it automatically finds an optimal number to skip by. For example, if there are 500 results left after filtered by >50% similarity, it would automatically calculate to skip by 25 as since we want 20 proteins, to get even diversity it calculates: 500 results / 20 proteins we want = 25 to skip by.
- b. From a previous email, I believe you mentioned less than half the rectangle are less interesting. I am able to get the % the rectangle takes up and that is how I filter them I.e. >50%. One of the proteins from the excel file that was chosen was A0A094CS32 and the image below is what uniprot gives from blasting it from MPS1 yeast. Does the % being low matter or is all we care about is the rectangle part?
<image.png>
- c. A bug in my program

In regards to your responses:

2. I am able to fetch a premade alphafold model for any protein from uniprot (it takes <5 second) but it only returns one. The alphafold web interface does return 5 models as it basically calculates and generates new models each time. It seems important to have multiple models to look at, so I'll investigate a way to retrieve different alphafold models.
3. Nope, all good! I was just making sure I was selecting the right one.
4. Yeah unfortunately, from what I can tell is that BLAST, even the web interface, results are slightly random. Due to skipping every 10 or so, the results may vary very widely, but yeah you can run as many times as you want.

Best,
Adam

On Mon, Dec 2, 2024 at 3:13 PM Emily Parnell <u0463089@umail.utah.edu> wrote:
Hi Adam,

Thanks very much for your work on this. It looks like you've made good progress.

Here are some answers to your questions.

For question 1 on the alignment, unfortunately the alignment you sent me doesn't look correct. Can you tell me which species you used for the alignment? (I can't understand the list on the left side - which is where this information would be) Then I can run one myself & give you an image of what it should look like for MPS1.

For question 2, Yes, the chimera model looks correct. However, I only see one of them (test.cif - See Models at the side), and normally AlphaFold sends back 5. Ideally all 5 would be present; four of them can be unchecked in the Models box so that only one is viewed (either by your program or by the user). Having the other four present would allow the user to examine the others if desired, which could be important.

For question 3, Yes, you're correct; the protein you want to select in the Chimera menu is the original protein name of interest. Are you getting other options that are confusing?

For question 4, certainly it's not ideal that this is what happens (2 different BLAST searches give different results), but I don't know of a way to change this, and I think it will just be a limitation for us. I don't think it's a huge issue, but something for us to keep in mind. Using your tool will make it easy for us to run multiple times and compare the results.

If what I have written is unclear, let me know. We could set up a zoom call if that would be helpful. It looks like the biggest hurdle right now is the alignment problem in question 1. Maybe talking it through would be easier than email.

Also, if you'd like me to run the tool and tell you how it's working for me, I can do so.

Thanks again!
Emily

<aligned.aln>

Adam Chhor <achhor@uw.edu>

Fri, Dec 6, 2024 at 2:25 PM

To: Emily Parnell <u0463089@uemail.utah.edu>

Cc: "Matt Miller (Biochemistry)" <u6021934@uemail.utah.edu>, Brent Lagesse <lagesse@uw.edu>

Hi Emily,

Yes, I think Tuesday at 12pm Mountain time works perfectly. And yes, we can discuss everything during the call.

Sorry for the delay. I ran into problems involving running on Macs when I did some more complicated tests and I'm working on resolving them right now. I expect I can send you instructions over the weekend.

Thank you,
Adam

[Quoted text hidden]

Brent Lagesse <lagesse@uw.edu>

Fri, Dec 6, 2024 at 5:47 PM

To: Adam Chhor <achhor@uw.edu>

Cc: Emily Parnell <u0463089@uemail.utah.edu>, "Matt Miller (Biochemistry)" <u6021934@uemail.utah.edu>

I have a thesis defense at that time, but if you can send me the Mac instructions, I can test them on my Mac. Albeit mine is old, so I have an Intel CPU and not an M1 or M2 CPU.

lirc, y'all have to get it to install software on your computers, is that still the case? There are ways to do local installs on some software to get around this if so, but I'm not sure it'll be the case for everything in the pipeline.

[Quoted text hidden]

Adam Chhor <achhor@uw.edu>

Mon, Dec 9, 2024 at 8:55 AM

To: Brent Lagesse <lagesse@uw.edu>, Emily Parnell <u0463089@uemail.utah.edu>

Cc: "Matt Miller (Biochemistry)" <u6021934@uemail.utah.edu>

Hi Dr. Emily,

I was able to fix the problems with Mac M1, and it should work smoothly if you use an intel one. The only thing not working is ChimeraX not automatically loading the Jalview file and running the commands to show the conservation. I have a temporary solution in mind, but I'm looking for a better solution to fix it so decided to wait for the next version for it.

Program source code and instructions: <https://drive.google.com/drive/u/1/folders/1OkIblzXnjhu39rz1liviC3BbyQ18AUM2>

Let me know if there is something wrong with the instructions or have any questions. I'll be checking my email frequently today, but I'll also we can talk about it over Zoom tomorrow.

Here is the Zoom link to tomorrow's meeting at 12 pm Mountain Time: <https://washington.zoom.us/j/5561180934>

Hi Professor Brent,

Feel free to install the zip in the Google Drive link above or can also git clone my adam-ui branch and then follow the instructions on running the program

Also if you are referring to installing ChimeraX and Jalview, I believe the biologist should have them installed, so it shouldn't be a problem, but I'll discuss that with Dr. Emily during the call. Technically the tool could be run without installing them as the data is stored in the output folder, and having the programs installed just means the tool can automatically load the data to the respective programs (Though currently, the program might crash at the end). If it is about installing the Python program and converting it into an executable without having to install Python, I'm currently investigating to get it to work, and hopefully, with GitHub Actions, it can automatically release versions.

Best,
Adam

[Quoted text hidden]