# Iterative profile search (PSI BLAST/SAM)

# Introduction

PFAM and Prosite are databases that contain motif representation of conserved protein domains (HMM profile and regular expression). The questions is how are such profiles built.

Normally we start from a multiple alignment. However to generate a multiple alignment you need to retrieve homologs that are likely to belong to the same gene family. As shown in the previous exercise we can do this with blastP. However if we want to build a profile, we need to ensure that we have a representative trainingsset (or input set). This is not trivial because with a standard p blast we tend to pick up the more related sequences but might miss sequences that are too distantly related because their overall degree of conservation with the query sequence is too low to pick them up (e.g. if in those sequences only the region of functional relevance is conserved).

This is why psi blast and sam were invented. They blast combines the advantages of a regular blast with a PWM search. It starts with the query sequence, performs a regular pblast. Based on this first iteration a multiple alignment is made and converted in a profile (PSSM or HMM representation of the multiple alignment). In a subsequent iteration the algorithm uses the PSSM or HMM to screen for novel hits. This allows to recover the more remote members of the protein family. By adding the more remote members the multiple alignment obtained faster each iteration becomes more representative (ideally). So you can now start iterating this process in order to refine the motif model and gradually add the less conserved homologs.

The procedure summarized:

1) detecting close homologs with blast

2) generating a multiple alignment,

3) deriving a motif model from the alignment

4) screening the genome for remote members of the family using the motif model

5) using the detected remote members to update the motif model.

This iterative procedure is performed by tools such as PsiBlast or SAM.

Once sufficient representative members of the family are recovered (seed alignment), a motif model representative for this protein family can be reconstructed. Curated motif models are subsequently stored in databases. Each of the databases differs from each other in the motif model representation they use. (PROSITE => regular expressions and profiles, PFAM => HMM, BLOCKS =>PSSM).

https://www.ncbi.nlm.nih.gov/books/NBK2590/

## PSI-BLAST

### Introduction

All the aforementioned steps can be performed automatically with the PSI BLAST server that itself allows to detect remote homologs (https://www.ncbi.nlm.nih.gov/books/NBK2590/):

The procedure PSI-BLAST uses can be summarized in five steps:

(1) PSI-BLAST takes as an input a single protein sequence and compares it to a protein database, using the gapped BLAST program.

(2) The program constructs a multiple alignment, and then a profile (weight matrix), from any

significant local alignments found. The original query sequence serves as a template for the multiple alignment and profile, whose lengths are identical to that of the query. Different numbers of sequences can be aligned in different template positions.

(3) The profile is compared to the protein database, again seeking local alignments. After a few minor modifications, the BLAST algorithm can be used for this directly.

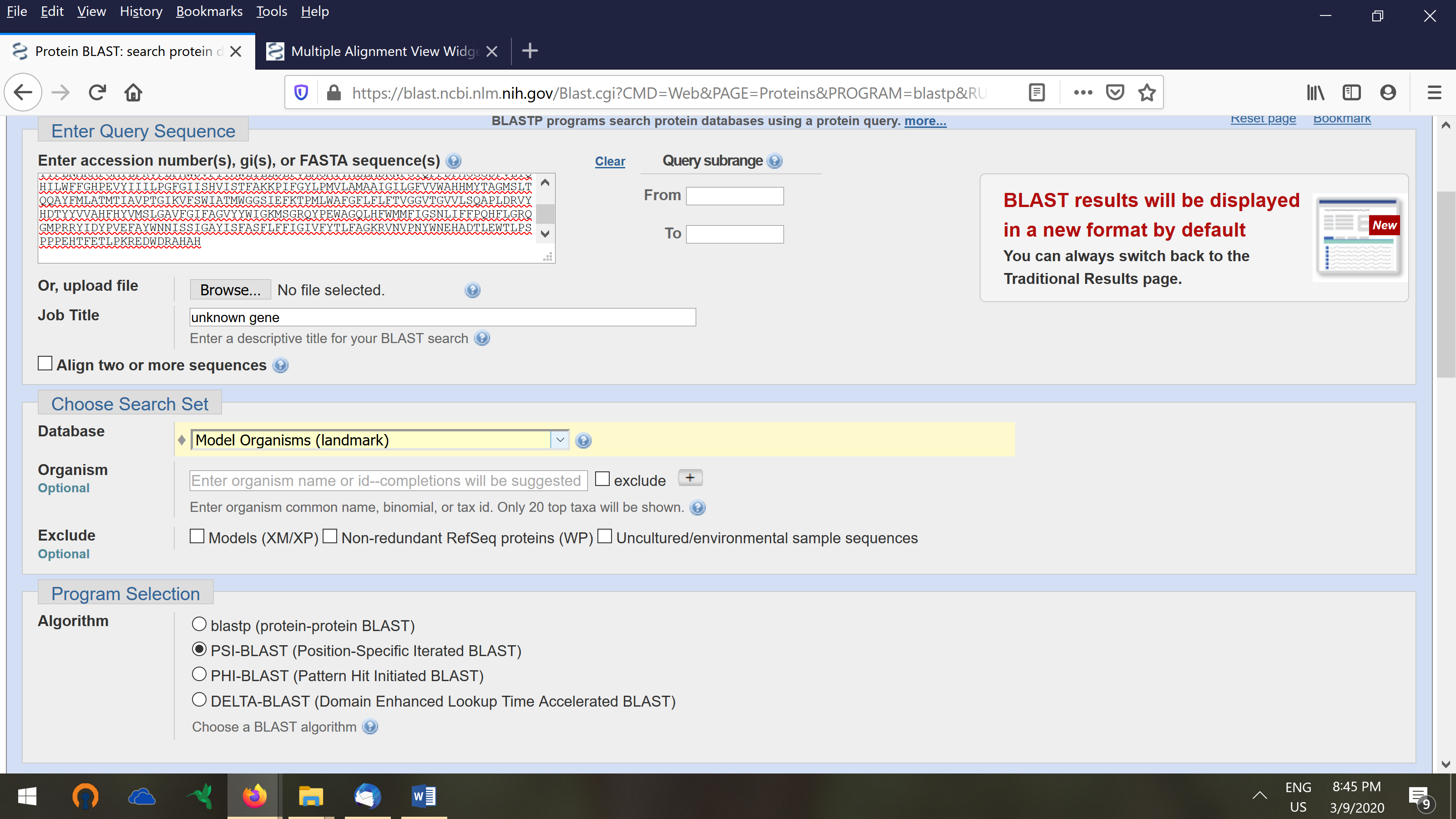
(4) Finally, PSI-BLAST iterates, by returning to step (2), an arbitrary number of times or until convergence.

### PSIBLAST the cox family

# try to find a part of the original sequence located around the site involved in binding the Cu ligands. This site is well conserved in all members of the family (see the multiple alignment). Using this part of the sequence to construct a profile will be very informative.

>unknown gene

MADAAVHGHGDHHDTRGFFTRWFMSTNHKDIGILYLFTAGIVGLISVCFTVYMRMELQHPGVQYMCLEGARLIADASAECTPNGHLWNVMITYHGVLMMFFVVIPALFGGFGNYFMPLHIGAPDMAFPRLNNLSYWMYVCGVALGVASLLAPGGNDQMGSGVGWVLYPPLSTTEAGYSMDLAIFAVHVSGASSILGAINIITTFLNMRAPGMTLFKVPLFAWSVFITAWLILLSLPVLAGAITMLLMDRNFGTQFFDPAGGGDPVLYQHILWFFGHPEVYIIILPGFGIISHVISTFAKKPIFGYLPMVLAMAAIGILGFVVWAHHMYTAGMSLTQQAYFMLATMTIAVPTGIKVFSWIATMWGGSIEFKTPMLWAFGFLFLFTVGGVTGVVLSQAPLDRVYHDTYYVVAHFHYVMSLGAVFGIFAGVYYWIGKMSGRQYPEWAGQLHFWMMFIGSNLIFFPQHFLGRQGMPRRYIDYPVEFAYWNNISSIGAYISFASFLFFIGIVFYTLFAGKRVNVPNYWNEHADTLEWTLPSPPPEHTFETLPKREDWDRAHAH



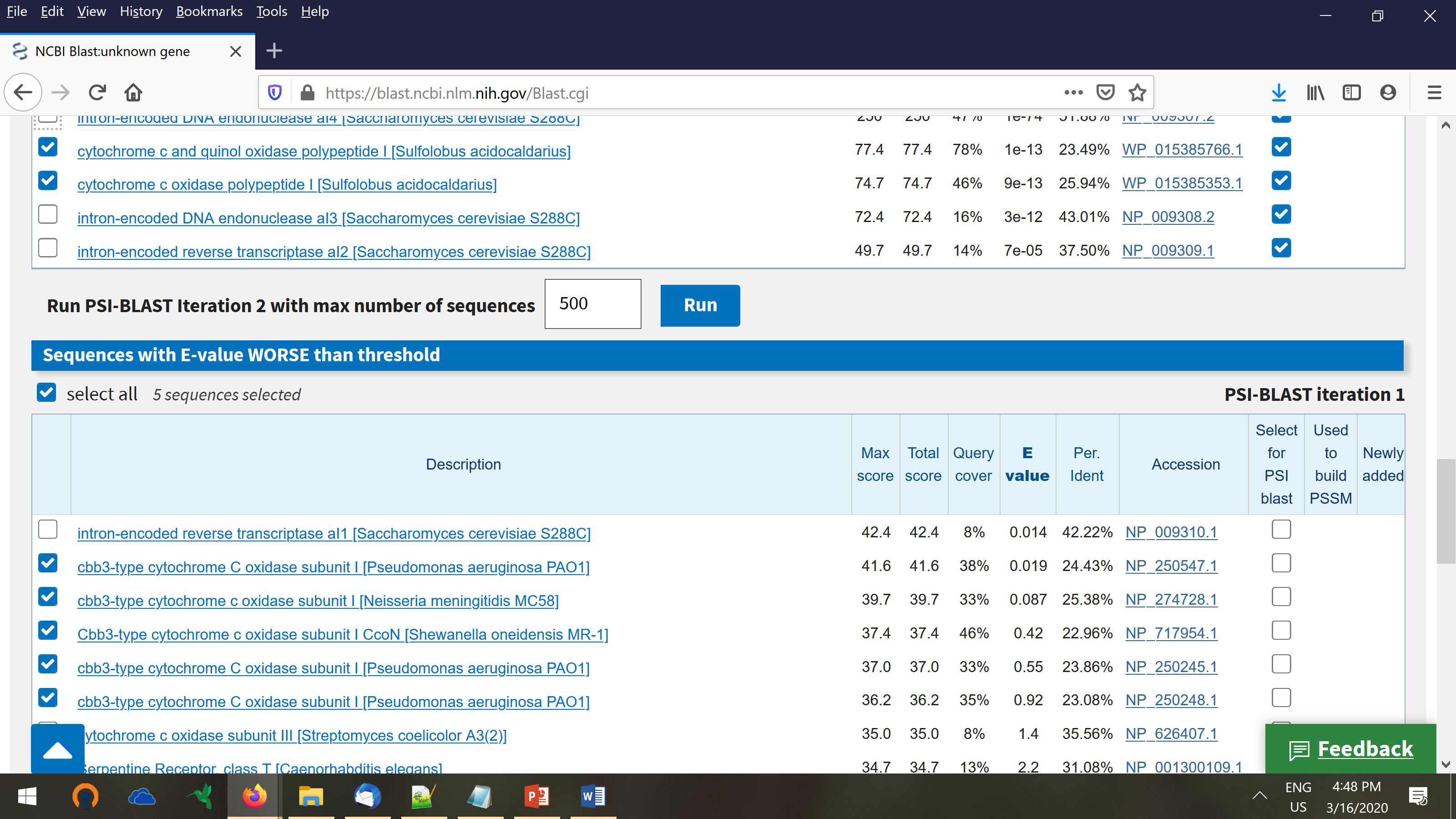
Paste this sequence in PSIblast, restrict the blast results to the Model organisms (landmark) to only have representative hits in model organisms.

Perform the first iteration:

The psi blast report in the top panel the hits that have an E value above the threshold (remember what the E value means). By using psi blast we can find members of the cytcaa3 oxidases (oxidase that can work at normal oxygen pressure).

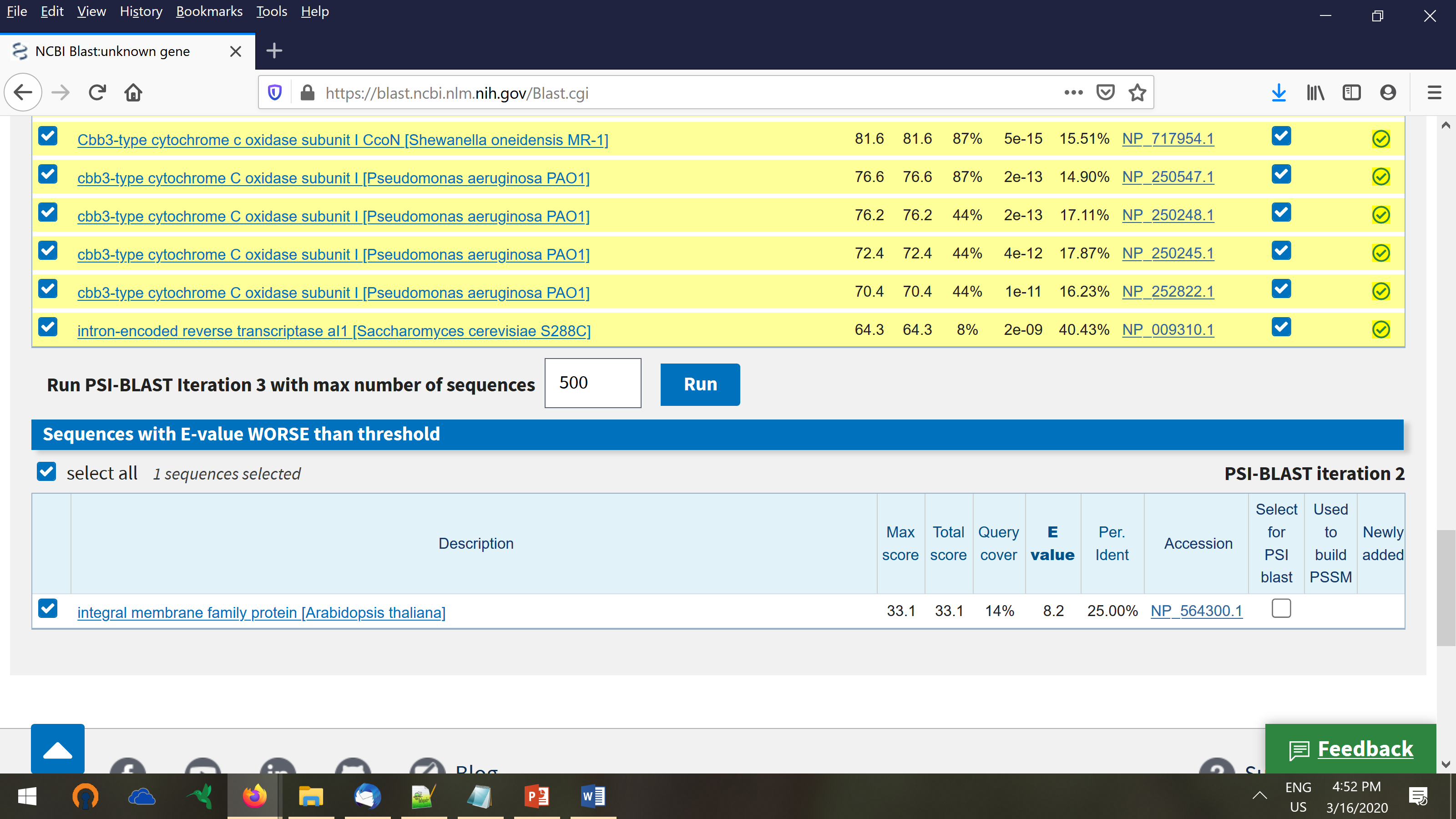
The lower panel gives the hits that are not passing the E value threshold of blast. You can now decide whether you want to include these additional members in the second iteration of blast. If you do, they will be accounted for the profile that will be generated of the multiple alignment, otherwise not.

The low scoring hits are here the remote members of this family of oxidases (cytcbb3). These are the ones that occur in microaerobic bacteria. You will notice that after the first iteration they are below the threshold but you can still decide to add them to the psiblast. I have added the cbb3 cytochrome c oxidase but have removed the intron encoded ones as I am not sure they truly belong to this protein family (by using the checkboxes).



What will happen if we add sequences that do not belong to the same family to the psi blast?

After the cytcbb3 oxidases were used to first build the profile (after the first iteration) and then screen the sequences in the database, the cytcbb3 oxidases scored significantly (which is to be expected, as the profile became more representative). Note that after the second iteration the psi blast converged. It only detected one additional sequence below the threshold on the E value (with a really bad E value) . So this sequence we won’t add as it might start deteriorating the profile.



You can also view the multiple alignment (with multiple alignment or multiple alignment viewer). This can help you to inspect whether you indeed recovered conserved regions.

