

PART 2: Sequence Alignments are a BLAST!

Annotate the *M. grisea* genome (cause of the Rice Blast Disease)

A. Choose the Protein Sequence to Analyze:

- Go to the *M. grisea* genome project website:
<http://www.broad.mit.edu/annotation/fungi/magnaporthe/>
- Click on **Genes**. This form allows researchers to find genes using a variety of methods. One is based on how a gene has been annotated – or assigned a function. Keep in mind that most assigned functions are simply best guesses.
- Click on **Search for Genes**.
- In the space for “Name:” type in the word “hypothetical” and click Submit. These genes, along with those called “predicted” are ones that the computer has predicted to be a gene, but is not sure of its function. These all need further investigation by better computer tools or humans.

Look at the list of genes found. How many are there? How does this compare to about 11,000 genes scientists predict are in *M. grisea*?

- Pick a gene and record its, MG number, starting point, stop point, length and strand. The first time through we will all do the same gene and then you can go back and do your own.

| | Locus | Start | Stop | Length | Strand |
|----------|-------|-------|------|--------|--------|
| Class | | | | | |
| Your Own | | | | | |

- Click on the hypothetical gene MG00040.4 and scroll down to get the sequence information.
- Copy the protein sequence at the bottom of the page - *not the DNA sequence*. The protein sequence should start with the letter M.

>(MG00040.4) hypothetical protein (translated coding sequence)

```

MLGGKSIRVNGAECGVESLFLGAVTSLGGFLFGYDTGQISGMLIFEDFQRRFATGPVGE
NGIREWVPIIQSTMVSLMSIGTLIGALSGAYTADWWGRRRSLAFGVILFIIIGNIIQITA
MESWIHMMMGRLIAGFGIGNLSVGVPFMFQSECAPREIRGAVVASVYQLLITFGILISNII
NYGVRNIQGS DASWRIVIGLGIFFSVPLGIGILLVPE SPRWLAGRQDWDGARMSIARLR
GMKHPNNVLVETDISEMYKIIKESSVGVGSWAECFTGKGGSEGIPKVVYR TILGMFM
HFTQQWTVGVNYFFYYGATIFQSAGVDDPIVTQLILGAVNVAMTFLGLYIVEKFGRRGPL
FIGGAWQAVWLAVFAAIGTALPPTENRVSGIVMIVSACMFIAFASTWGPICWV VIGET
FPLRTRAKQASLSTAGNWLGNFMISFLTPIATDGISFSYGFVFAAVNLCGALGVYFFLY
ESRMLSLENVDRMYGDPSIKPWNSRKWTPPGYINRRTKDEKYVPEGEHVQGGV RGSVGS
DNTAVPGEADVNH DHAKQRPGSDGVPPTTEQHEQAVLRNAKLTGLAASESIYIQT*

```

B. Compare the gene to other sequenced genes:

Go to the *National Center for Biotechnology Information (NCBI) home page at <http://www.ncbi.nlm.nih.gov/>*. This is the home of the national archive of gene sequences called GenBank. This database contains virtually all sequenced genes. To determine what our “hypothetical” gene is, we will compare it to everything previously sequenced.

8. At the top of the page, click on **BLAST** and then on the next page **Protein-protein (blastp)**. If we were using the DNA (nucleotide) sequence, we could use the blastn.
9. Paste in the protein sequence you copied in step 7. Click **BLAST**. The algorithm is now comparing your query sequence to all others in the database to find the most likely matches. This is a complex search since most matches will not be identical. Thus it may take a few minutes.
10. Click **FORMAT** to see if the results are ready. After clicking once, the results page will refresh itself until the search is complete – this will take a few minutes.
11. Look at the list of matches and the e-values in the far right hand column. The e-value is the probability that your sequence randomly matches the sequence listed. If the e-value is 0.00 the two sequences match perfectly. Anything smaller than e^{-10} is considered good.

If you get a match that is not a “hypothetical protein”, that has an e-value smaller than e^{-10} , click on the matching gene name and find and record what it does. Then proceed to part D.

If not, get another protein and start over or look for sub-regions of the protein with specific activities via Interpro in part C.

For MG00040.4 you should get the list on the next page.

- Notice that the first match is with the same protein you were searching for and the E value is 0. this means it is a perfect match.
- Around the tenth sequence you see matches that are not hypothetical and still have very small E values (less than e^{-90} !). All of these transport sugars of some type (hexose, glucose etc. Anything ending in -ose is a sugar). What do you think this hypothetical protein does?

| Sequences producing significant alignments: | | Score | E |
|--|---------------------------------------|----------------------|-------|
| | | (bits) | Value |
| gi 38101416 gb EAA48382.1 | hypothetical protein MG00040.4 [... | 1156 | 0.0 |
| gi 32410003 ref XP_325482.1 | hypothetical protein [Neurospo... | 765 | 0.0 |
| gi 42551538 gb EAA74381.1 | hypothetical protein FG05042.1 [... | 757 | 0.0 |
| gi 46097856 gb EAK83089.1 | hypothetical protein UM02037.1 [... | 484 | e-135 |
| gi 42554326 gb EAA77169.1 | hypothetical protein FG07582.1 [... | 396 | e-108 |
| gi 40739179 gb EAA58369.1 | hypothetical protein AN5860.2 [A... | 395 | e-108 |
| gi 40739625 gb EAA58815.1 | hypothetical protein AN4277.2 [A... | 394 | e-108 |
| gi 27461193 gb AAL89824.1 | monosaccharide transporter [Aspe... | 380 | e-104 |
| gi 42548959 gb EAA71802.1 | hypothetical protein FG02978.1 [... | 369 | e-101 |
| gi 50313479 gb AAT74609.1 | hexose transporter [Sclerotinia ... | 362 | 2e-98 |
| gi 40744145 gb EAA63325.1 | hypothetical protein AN3357.2 [A... | 357 | 3e-97 |
| gi 19075239 ref NP_587739.1 | hexose transporter. [Schizosac... | 354 | 4e-96 |
| gi 19113217 ref NP_596425.1 | hexose transporter [Schizosacc... | 350 | 6e-95 |
| gi 19075247 ref NP_587747.1 | putative glucose transporter p... | 348 | 2e-94 |
| gi 2407189 gb AAB70519.1 | hexose transporter [Schizosacchar... | 347 | 6e-94 |
| gi 19075240 ref NP_587740.1 | hexose transporter. [Schizosac... | 345 | 2e-93 |
| gi 32415185 ref XP_328072.1 | hypothetical protein [Neurospo... | 345 | 2e-93 |
| gi 619164 emb CAA87389.1 | permease [Kluyveromyces lactis] >... | 340 | 7e-92 |
| gi 4098350 gb AAD00266.1 | sugar transporter 1 [Pichia stipi... | 337 | 4e-91 |
| gi 19075246 ref NP_587746.1 | putative glucose transporter p... | 335 | 1e-90 |
| gi 19111856 ref NP_595064.1 | putative glucose transporter p... | 334 | 5e-90 |

C. Try to identify the protein function:

You may go to another tool called **Interpro**. This site looks for functional motifs or sub-regions in your protein to help identify its function. Go to <http://www.ebi.ac.uk/InterProScan/>.

12. Paste in your protein and click on **submit job**. You can also enter your email and the results will be emailed to you. The results will also appear in the browser window as soon as the job is complete. This may take a few minutes. When results are displayed, click on the figure or PD number to discover the function of any sub-domain found.

D. Submit what you have found:

To add this annotation – gene function- to the list of other gene annotations of the *M. grisea* genome, go to **MGOS** at <http://www.mgosdb.org:9100/>. This is the web site where we are collecting all the experimental information about the rice blast disease.



13. Click on the green **Genes Page** button and then **Outreach** and **New**.

| | | |
|--|------------------------------|-----------------------------|
| Select Organism <input checked="" type="radio"/> MG <input type="radio"/> OS | | |
| Others | Add New Gene | Modify Gene |
| Outreach | Add Ne Genew | Modify Gene |
| Mutant Recovery* | Add New Gene | Modify Gene |

14. The user ID is “outreach” and password is “mgosoutreach”. Enter all the information you have on the gene you have annotated. Use the alternative name field for the MG00040.4 name and the comment field for the gene description. This data will be reviewed and used as the permanent annotation of this gene. If you are a student use your last name as the submitter, the school as the lab, and fill in all other information that you have uncovered.

| Details | |
|---|---|
| Gene Name * | AMG16287 (eg., AMG00001) |
| Alternative Gene Name | MG00040.4 |
| Submitted By | Harris |
| Lab | The Science House |
| Email | physics@hotmail.com |
| Chromosome * | II (eg., II[for MG] or Chr10[for Rice]) |
| Start Position * | 10705 (eg., 111) |
| End Position * | 12831 (eg., 211) |
| Strand * | 1 (eg., -1 or 0 or 1) |
| Swissprot ID (or Description) | (eg., predicted protein) |
| Exon Structure (Exon#:start-end) <i>Note: start and end are relative to itself</i> | (eg., 1:111..140) |
| Sequence | MLGGKSIRVNGAECGVESLFLGAVTS LGGFLFGYDTGQISGMLIFEDFQR RFATGPVGENGIREWVPIIQSTMVSLM SIGTLIGALSGAYTADWWGRRRS LAFGVILFIIGNIIQITAMESWIHMMMGR |
| Comment | I think this protein is a sugar transporter. |

Reset Submit