Aina Montalbán

Session 2: Orthology assignment

A.- Sequence based orthology assignment

Sequence similarity is often used as a substitute for orthology prediction. The assumption is that the closer two sequences are related based on sequence identity, the more likely it is that they are orthologs.

There are several approaches that allow us to predict orthology based on sequence similarity and that range from very simple to more complex. In this practical we will go through some of the most common ways: best bidirectional hits, inparanoid, and OrthoFinder.

1.- BRH (Best bidirectional hits or Best reciprocal hits):

1.1- How many orthologous pairs did you find?

We find 344 orthologous pairs.

1.2.- Which is the ortholog of Phy0042233_ACYPI?

The ortholog of Phy0042233_ACYPI is Phy007ATFF_MYZPE.

1.3.- Does Phy0042233_ACYPI have any other homologs in MYZPE? What kind of relationship do they have regarding Phy0042233_ACYPI?

Yes, with the BLAST method we find three homologs in MYZPE:

Phy007ATFF_MYZPE
Phy007AV73_MYZPE
Phy007ASLM_MYZPE

Regarding the relationship between them, we know that they are homologs, as we have performed a sequence similary method. However, we are not able to predict the specific (orthologs, paralogs, etc.) relationship with this methodology.

1.4.- Phy00BX2H1_ACYPI does not have a ortholog. Why?

Phy00BX2H1_ACYPI is found on the TRY1.unpaired.txt file, so it does not have an ortholog. The reason is that its best hit Phy007AWWE_MYZPE is the best hit for Phy0010BYG_ACYPI.

1.5.- Which advantages and drawbacks can you see of the BRH method?

The main **advantage** of the BRH method is that is very fast, as we only need to look for the sequence similarity. If the relationship is clear it will give correct outcomes.

Although, it has some **drawbacks**. It can miss to find an ortholog, in other words, it may no detect any ortholog. Moreover, the maximum size of orthologs we are going to find is 1, 1-to-1 relations. We will have more difficulties when we are dealing with multiple species.

2.- Inparanoid (http://inparanoid.sbc.su.se/).

Questions:

2.1.- How many groups of orthologs do we have? Are the groups the same as before?

We have 317 groups of orthologs. It seems that some groups are the same as before. However, there are some groups with more than two members.

2.2.- Are there any groups with more than two members? What are the evolutionary relationships between the proteins (select two examples)?

Yes, there are five groups with more than 2 members.

```
79      1062    Phy0042233_ACYPI 1.000      Phy007ATFF_MYZPE 1.000 Phy007AV73_MYZPE 0.138
130     861     Phy00BX36V_ACYPI 1.000      Phy007AYEX_MYZPE 1.000 Phy007AX6C_MYZPE 0.070
277     305     Phy00BX0GK_ACYPI 1.000      Phy007AOGX_MYZPE 1.000 Phy007AY06_MYZPE 0.051
298     232     Phy00BX47C_ACYPI 1.000      Phy007AYSI_MYZPE 1.000 Phy007AZ12_MYZPE 1.000 Phy007AMPM_MYZPE 0.566
307     208     Phy000XPLI_ACYPI 1.000     Phy00420LP_ACYPI 1.000  Phy007AM2D_MYZPE 1.000
```

We are going to select the groups 79 and 307.

79

Phy007ATFF_MYZPE and Phy007AV73_MYZPE are inparalogs. And both are co-orthologs to Phy0042233_ACYPI.

307

Phy000XPLI_ACYPI and Phy00420LP_ACYPI are inparalogs. And both are co-orthologs to Phy007AM2D_MYZPE.

2.3.- Go back to the BRH results and check what it says regarding this family. Which of the two results do you think is more reliable?

79

We find a BRH between Phy007ATFF_MYZPE and Phy0042233_ACYPI.

307

We find a BRH between Phy007AM2D_MYZPE and Phy000XPLI_ACYPI.

It seems more reliable InParanoid as we can observe inparalogs. The result from BRH is more simple than the InParanoid, that has more detailed information. In consequence, inParanoid result is more reliable.

2.4.- Which is the ortholog of Phy0042233_ACYPI? Is it the same one as before? Do we have any further information regarding the other paralogs we found for this family?

We find two orthologs: Phy007ATFF_MYZPE and Phy007AV73_MYZPE.

Yes, with the BRH we find the same ortholog: Phy007ATFF_MYZPE.

By looking into the output filel, we can find more information:

```
_____________________________________________________________________________
Group of orthologs #79. Best score 1062 bits
Score difference with first non-orthologous sequence - ACYPI.fa:359   MYZPE.fa:1062
Phy0042233_ACYPI     100.00%              Phy007ATFF_MYZPE     100.00%
                                          Phy007AV73_MYZPE     13.79%
Bootstrap support for Phy0042233_ACYPI as seed ortholog is 100%.
Bootstrap support for Phy007ATFF_MYZPE as seed ortholog is 100%.
```

2.5.- Does running inparanoid with an outgroup change the results? Why do you think that happens?

Yes, when we define an outgroup, we find 299 groups. Because the outgroup interfers in the blast scores. If you include a third specie, it is clear that the result and the number of orthologs will be different.

2.6.- Is it worth it then to run inparanoid with an outgroup?

Yes, it can produce more detailed results as we have more information.

3. OrthoFinder

3.1.- How many orthogroups do you have in this case? Are they comparable to the previous results?

We have find 350 orthogroups. Compared to the other results, we can observe that there are more groups with more than two members.

3.2.- Which proteins are orthologous to Phy0042233_ACYPI? Is this result congruent with the previous results?

```
OG0000016: Phy0042233_ACYPI Phy007ATFF_MYZPE Phy007AV73_MYZPE Phy00BWVQD_ACYPI
```

Yes, this result is congruent with the previous results. However, there is one new paralog Phy00BWVQD_ACYPI.

3.3.- How many orthologs have a one-to-one relationship?

There are 322 orthologs with 1-to-1 relationship.

	ACYPI	MYZPE
ACYPI	0	322
MYZPE	322	0

3.4.- How many duplication events are shared between ACYPI and MYZPE? Are there species specific duplications?

There are 10 duplication events shared between ACYPI and MYZPE. This information can be found in the *duplications.csv.* Yes, there are species specific duplication.

3.5.- How many orthogroups do you have in this case?

We have 349 orthogroups.

3.6.- Which proteins are orthologous to Phy0042233_ACYPI? Is this result congruent with the previous results?

The orthologous proteins are: Phy007ATFF_MYZPE, Phy007AV73_MYZPE and Phy00F1D6_SIPHA.

Yes, the result is congruent. However, there is a new ortholog, which is the outgroup: Phy00F1D6_SIPHA.

3.7.- How many orthologs have a one-to-one relationship? Is the number similar to the previous analysis? Why?

The next table shows the orthologs that have 1-to-1 relationship. It is more or less similar, although we have influence of the outgroup.

	ACYPI	MYZPE	SIPHA
ACYPI	0	318	247
MYZPE	318	0	243
SIPHA	247	243	0

3.8.- How many duplication events are shared between ACYPI, MYZPE, and SIPHA?

There are two duplication events shared between ACYPI, MYZPE and SIPHA.

B.- Phylogeny based orthology assignment

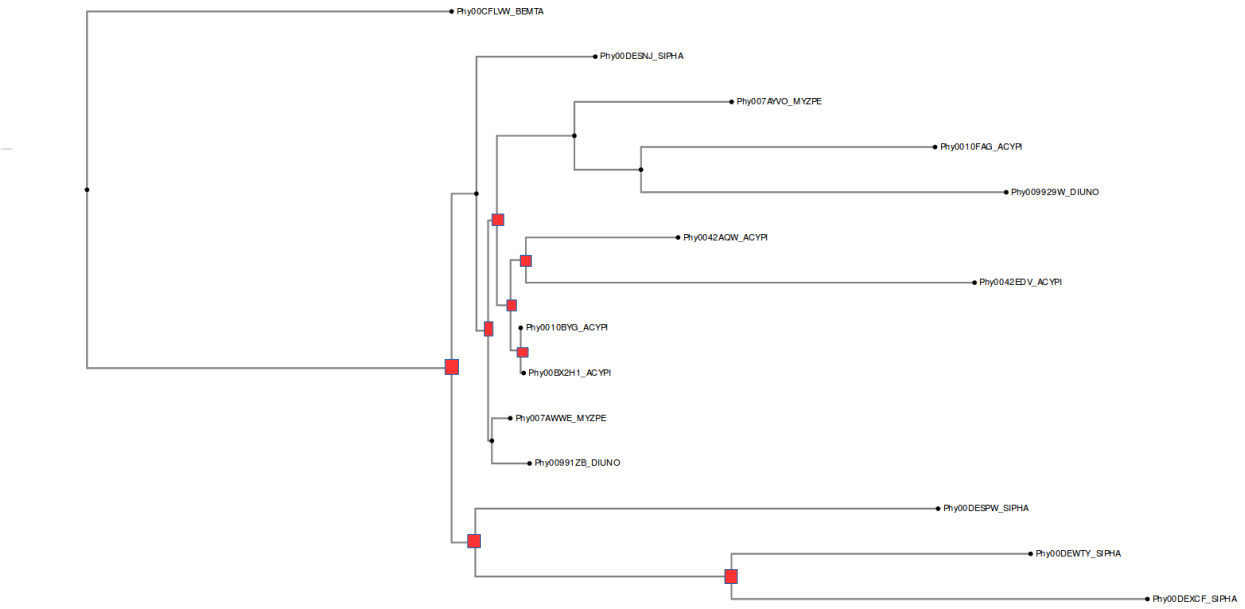
4.1 Fill in the following table of orthology and paralogy relationships when referring to Phy007AWWE_MYZPE.

	Orthologs	**Paralogs**
BRH[1]	Phy0010BYG_ACYPI	Phy00BX2H1_ACYPI Phy0042AQW_ACYPI Phy0010FAG_ACYPI Phy0042EDV_ACYPI Phy00BX2N6_ACYPI
InParanoid	Phy0010BYG_ACYPI	Outparalogs: Phy007AYVO_MYZPE

1 In BRH it is assumed that all proteins that have a blast hit against another protein but it is not reciprocated are paralogs.

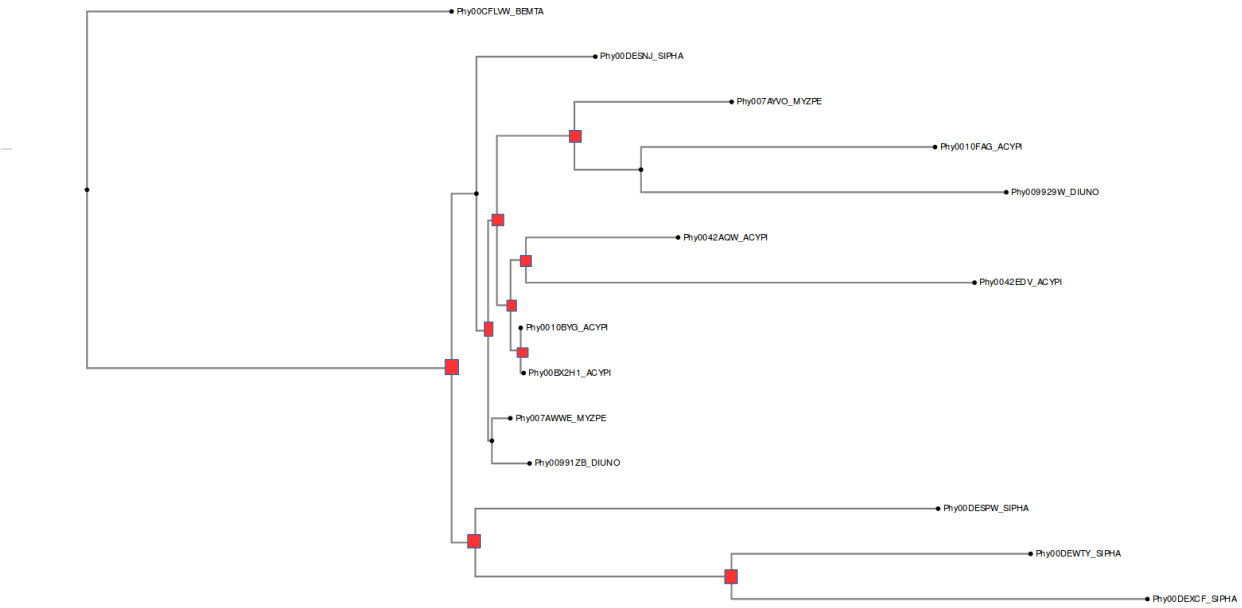
OrthoFinder	Set_a (two species): Phy0010BYG_ACYPI Phy0010FAG_ACYPI Phy0042AQW_ACYPI Phy0042EDV_ACYPI Phy00BX2H1_ACYPI Set_b (three species): Phy0010BYG_ACYPI Phy0010FAG_ACYPI Phy0042AQW_ACYPI Phy0042EDV_ACYPI Phy00BX2H1_ACYPI Phy00DESNJ_SIPHA	Set_a (two species): Phy007AYVO_MYZPE Set_b (three species): Phy007AYVO_MYZPE
Tree based (Species Overlap)	Phy00991ZB_DIUNO	Outparalogs: Phy0010FAG_ACYPI Phy009929W_DIUNO Phy0042AQW_ACYPI Phy0042EDV_ACYPI Phy0010BYG_ACYPI Phy00BX2H1_ACYPI Phy007AYVO_MYZPE
Tree based (Reconciliation)	Phy00991ZB_DIUNO	Outparalogs: Phy0010FAG_ACYPI Phy009929W_DIUNO Phy0042AQW_ACYPI Phy0042EDV_ACYPI Phy0010BYG_ACYPI Phy00BX2H1_ACYPI Phy007AYVO_MYZPE

Species Overlap



Duplication

Reconciliation



Duplication

Discuss the differences between the different predictions and why they happened. Which method do you think is the most reliable?

There are two types of methods: the sequence similarity and the phylogeny-based methodology.

BRH InParanoid OrthoFinder	Tree Reconciliation Species Overlap

We can observe that we the first two methods (BRH and InParanoid), we have obtained only one ortholog and some paralogs, but it is not clear what kind of paralogs are. In the third method, we start to have more information, having more orthologs than paralogs.

Finally, with the tree-based methods, there is more detailed information about the duplications and speciations. Therefore, we obtained some out-paralogs and orthologs of the interested gene.

To conclude, it is clear that the phylogeny-based methods are more reliable. It is essential to say that, in order to perform the tree-based methods, we should have a correct species tree. However, if we want to get results fast, we can use sequence similarity to study gene or protein relationships.