**Design of Complex Homology Viewer**

**1. Functional requirements**

**1.1 Visualization**

Visualization tool shows users the following information;

1) Phylogenetic relationship of organisms in Complex Portal (CP);

2) Components of a complex curated by CP;

3) Presence and absence of one complex in each organism;

4) Curation status of one complex in each organism.

**1.2 Table view**

Table view shows users more detailed information about orthologs and paralogs of complex components. It contains the following fields:

1) Species name

2) UniProt protein id

3) Homology: ortholog/paralog

4) Complex prediction result

5) Complex curation status

**2. Data preparation**

**1.1 Phylogenetic tree**

Tree topology will be reconstructed from data of Tree of Life (TOL) web project (http://tolweb.org/tree/phylogeny.html). It will be changed **manually** if new organisms are added to CP. The phylogenetic tree is stored in a plain text file in Newick format. The actual branch lengths will not be included in the tree file.

**1.2 Orthologs and paralogs**

Ortholog and paralog data will be fetched from Compara using its API. It takes hours to run the Perl script in the first time, but the running time can be reduced within minutes after that as only dozens of new proteins will be added in each CP release.

**1.3 Complex distribution data**

The presence and absence of a complex in an organism is computed from the ortholog data of 1.2.

A complex is marked to be present if the protein component number (*N*) and its presence ratio (*R*) meet the following criteria:

*R ≥ 0.5, when N ≠ 2*

*R > 0.5, when N = 2*

Otherwise, a complex will be marked to be absent from an organism. Table 1 shows some examples.

Table 1. Examples of complex prediction from orthologs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Complex** | **#Componential protein in sp A** | **Orthologs in sp B** | **Present ratio** | **complex prediction in sp B** |
| 1 | 1 | 1 | 1 | √ |
| 2 | 1 | 0 | 0 | - |
| 3 | 2 | 1 | 0.50 | - |
| 4 | 2 | 2 | 1 | √ |
| 5 | 3 | 1 | 0.33 | - |
| 6 | 3 | 2 | 0.67 | √ |
| 7 | 4 | 1 | 0.25 | - |
| 8 | 4 | 2 | 0.5 | √ |

**1.4 Complex curation data**

If a complex is curated in an organism in CP, its homologous complex(s) are checked to see if they exist in other organisms. This is computed from ortholog and paralog data of 1.2. If all the orthologs of the componential proteins of a curated complex are all present in another curated complex, which contains no other proteins, then these two curated complexes are called **orthologous complexes**. If some of the proteins are paralogous, then they are called **paralogous complexes**.

But there will be exceptions. Homologous complexes may contains different numbers of proteins. So complex homology will be checked manually, using information such as complex names.

**3. Design**

**3.1 Frontend**

There are two pages for viewing the phylogenetic information of complexes. One is for **visualization tool**, and the other one is for a **table** showing detailed information of componential proteins of curated and predicted complexes.

**3.1.1 Visualization tool**

The tool will be implemented using **javascript**, **D3 library** (https://d3js.org/) and **newick.js** tree file parser (https://github.com/jasondavies/newick.js).

**3.1.1.1 Overview**

There are two mostly-used arrangements for showing phylogenetic trees, circular and right-angled. Circular trees are more faddish and suitable for phylogenetic trees with a great number of species, but they are not very obvious when users compare the complex components and prediction result. Right-angled trees are more compact than the circular trees and can show comparison better, but they need more space when the number of species grows larger so they are suitable for trees with relatively small number of species. So two types of arrangements will be used to display the complex homology information, since both of them have their own advantages and disadvantages. Users can switch between the two types of arrangements.

**3.1.1.2 Circular view**

Tree structure is at the centre. There are several tracks outside of the tree structure. Species names are shown as the inner-most track. The second track is for organism icons. The third track is for complex pie charts. The two outermost tracks are star marks for complex prediction and curation information.

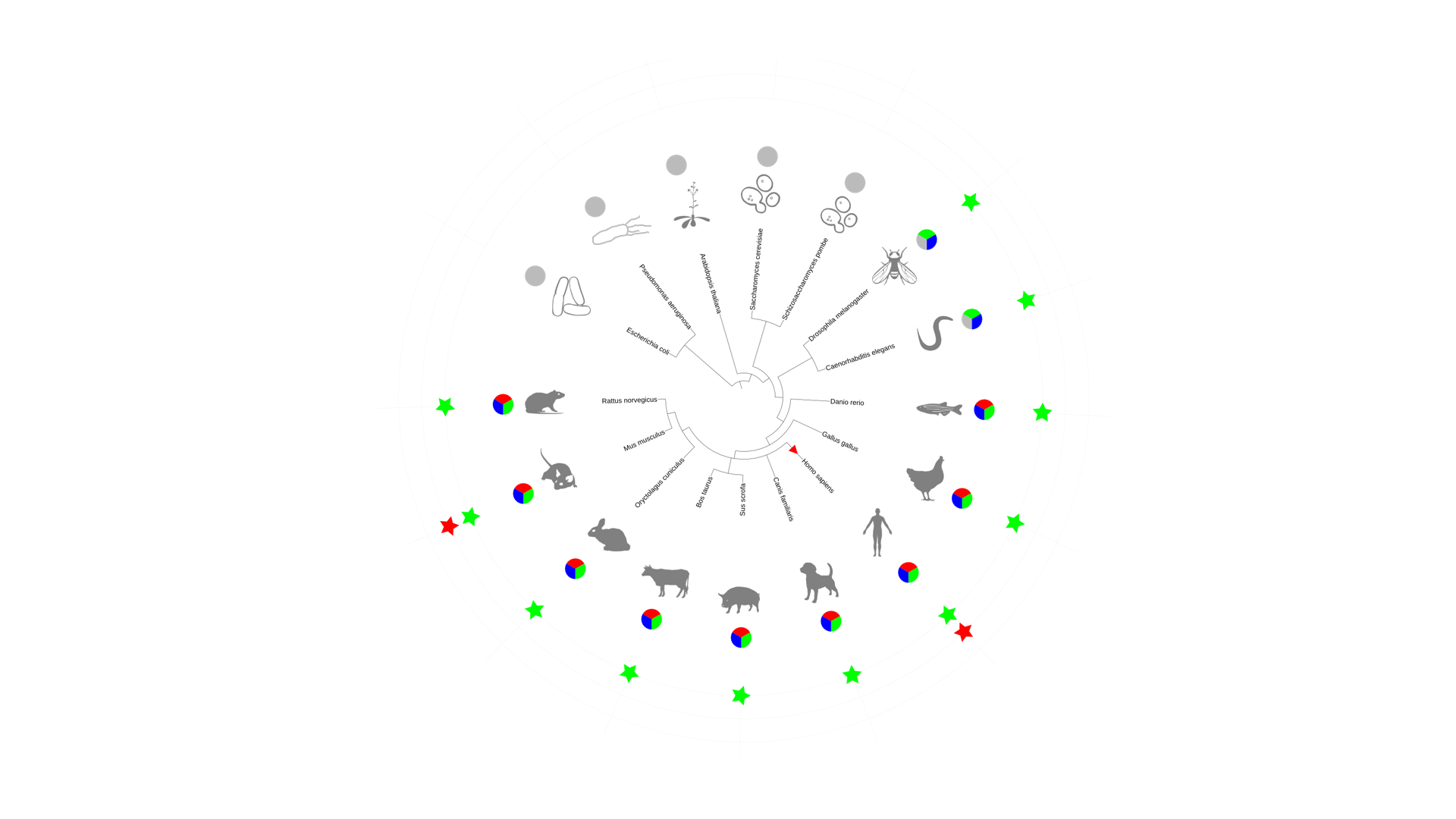


Figure 1. Circular view of complex homology

1) Phylogenetic tree

The Phylogenetic tree is implemented by d3.layout.cluster.

2) Organism icons

Organism icons are in SVG format transformed from CP icons and stored in a separate file.

3) Complex pie charts

Each color in the pie represents a protein component. Grey color indicate that the respective protein is missing in that organism. Pie charts are implement by d3.layout.pie and d3.svg.arc.

4) Star marks for complex presence and absence

A filled star indicate that the complex is predicted to be present in that organism. There will be a blank if the complex is predicted not to be present in that organism.

5) Star marks for curated complexes

A filled star indicate that the complex (or its homologous complex) is curated in that organism. There will be a blank if the complex (or its homologous complex) is not curated in that organism.

6) Starting point

A red filled triangle is added as the prediction starting point to the branch of the organism whose complex is used as a query to predict.

**3.1.1.3 Right-angled view**

Information provided by right-angled view is similar to that of circular view except that organism icons are eliminated as there is not enough space for them.

Tree structure is at the left side with species names near the leaf nodes. The three columns at the right side are complex pie charts, star marks for complex prediction and curation information. The red triangle prediction starting point mark is similar to that of the circular view.

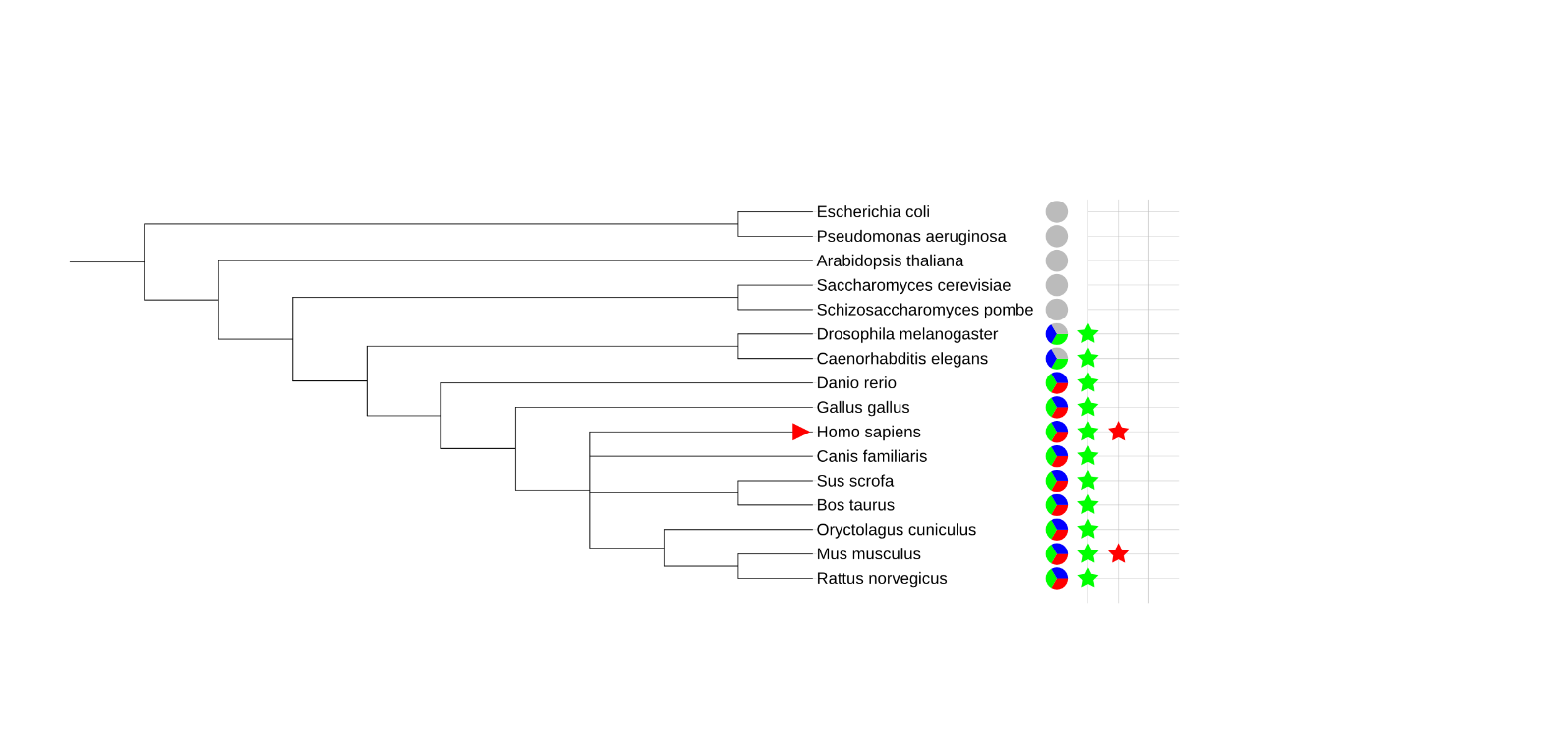


Figure 2. Right-angled view of complex homology

**3.1.2 Table view**

Table 2. Details of example prediction of complex SOSS2 (CPX-614)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Species name** | **Complex**  **Prediction** | **Curated**  **Complex** | **Protein 1** | **Protein 2** | **Protein 3** |
| *Homo sapiens* | CPX-614 | Query | Q68E01 | Q9NRY2(1:1) | Q96AH0 |
| *Mus musculus* | √ | CPX-615 | Q7TPD0(1:1\*) | Q3TXT3(1:1) | Q8BGW5(1:1) |
| *Rattus norvegicus* | √ |  | D3ZUT9(1:1) | D3Z914(1:1) | Q5FVP2(1:1) |
| *Oryctolagus cuniculus* | √ |  | G1SNY3(1:1) | G1TC81(1:1) | G1TDX1(1:1) |
| *Canis familiaris* | √ |  | E2RFL4(1:1) | F1Q3P8(1:1) | J9P956(1:1) |
| *Sus scrofa* | √ |  | F1SFW4(1:1) | F1SNA6(1:1) | F1SN32(1:1) |
| *Bos Taurus* | √ |  | E1BN70(1:1) | Q2NKT2(1:1) | A5D7P8(1:1) |
| *Gallus gallus* | √ |  | R4GLP5(1:n)  A0A1L1RPP9(1:n)  A0A1D5PPE8(1:n) | F1NAD7(1:1) | R4GHL6(1:1) |
| *Danio rerio* | √ |  | Q1LXC9(1:1) | Q7ZV26(1:1) | F1QH28(1:n)  F1RA89(1:n) |
| *Drosophila melanogaster* | √ |  | Q7PLS8(1:1) |  | Q9VM17(1:1) |
| *Caenorhabditis elegans* | √ |  | W6RQY9(1:1) |  | Q17749(1:1) |
| *Saccharomyces cerevisiae* | - |  |  |  |  |
| *Schizosaccharomyces pombe* | - |  |  |  |  |
| *Arabidopsis thaliana* | √ |  | O23315(1:1) |  | Q9FFP5(1:1) |
| *Escherichia coli* | - |  |  |  |  |
| *Pseudomononas aeruginosa* | - |  |  |  |  |

\*1:1, one-to-one ortholog; 1:n, one-to-many ortholog; m:n, many-to-many ortholog; p, paralog

The number of columns will grow if there are more proteins in the query complex. Some of the protein columns will hide and users can choose to show them if the number of proteins is larger than five. Linkout to UniProt will be added to proteins and linkout to CP will be added to complex IDs.

**3.2 Backend**

**3.2.1 Files**

1) File organization

├── cph\_tab.html

├── cphviewer.css

├── cphviewer.js

├── icon.js

├── index.html

└── tree.nwk

File names may be changed and new files may be needed.

2) Tree file

The content of the tree file tree.nwk is like the following:

(((((((Homo sapiens:0.3,(Oryctolagus cuniculus:0.2,(Mus musculus:0.1,Rattus norvegicus:0.1)1:0.1)11:0.1,Canis familiaris:0.3,(Sus scrofa:0.2,Bos taurus:0.2)2:0.1)12:0.1,Gallus gallus:0.4)3:0.1,Danio rerio:0.5)4:0.1,(Drosophila melanogaster:0.5,Caenorhabditis elegans:0.5)5:0.1)6:0.1,(Saccharomyces cerevisiae:0.6,Schizosaccharomyces pombe:0.6)7:0.1)13:0.1,Arabidopsis thaliana:0.8)8:0.1,(Escherichia coli:0.8,Pseudomonas aeruginosa:0.8)9:0.1)10:0.1;

The numbers following the colon are artificial branch lengths, which are used to optimize visualization. The numbers following the right parentheses are node ids, which are not used here.

The Newick tree will be parsed and transformed to JSON format by Newick.js.

3) Organism icon file

Organism icons are stored as a object in file icon.js. The following is part of the file:

var icons = [

{name: "Homo sapiens", pathD: ["M578 1268 c-19 -16 -33 -84 -18 -93 5 -3 10 -19 10 -34 0 -24 -7 -31 -49 -49 -46 -19 -52 -26 -85 -94 -37 -78 -101 -185 -136 -230 -12 -15 -23 -40 -26 -55 -3 -14 -7 -29 -9 -33 -2 -3 0 -12 6 -20 7 -13 9 -13 9 0 0 13 1 13 10 0 8 -12 10 -10 10 10 0 18 3 21 9 12 6 -10 12 0 19 30 7 24 39 81 73 128 57 79 99 150 99 166 0 4 -12 -3 -26 -17 l-26 -24 23 47 c14 28 31 49 44 52 12 3 34 15 49 27 25 20 27 26 24 87 -3 57 0 67 14 70 24 5 44 -46 29 -70 -28 -45 -3 -88 68 -117 21 -9 35 -27 51 -64 12 -29 33 -69 47 -90 13 -22 23 -41 20 -43 -5 -5 -46 55 -71 104 -21 41 -53 43 -62 5 -3 -16 -10 -42 -14 -59 -4 -18 -5 -51 -2 -75 10 -78 -2 -476 -15 -509 -4 -8 -13 57 -20 145 -17 210 -16 202 -30 216 -10 10 -13 9 -17 -2 -10 -32 -33 -277 -29 -310 3 -20 1 -39 -5 -43 -6 -3 -10 31 -11 92 -1 53 -6 123 -13 154 -8 44 -8 86 2 169 12 110 8 200 -10 229 -17 27 -21 2 -10 -59 8 -45 8 -78 -1 -127 -8 -45 -9 -101 -4 -168 5 -56 9 -144 10 -196 2 -52 6 -128 10 -168 7 -62 5 -75 -10 -92 -23 -26 -13 -37 23 -24 24 8 27 15 29 59 1 28 7 109 12 180 6 72 11 170 11 218 0 114 16 106 19 -10 1 -48 4 -131 8 -183 3 -52 6 -129 7 -170 1 -85 7 -100 42 -100 29 0 28 -2 8 40 -15 29 -15 41 -4 104 7 39 12 105 11 146 -1 41 5 106 13 144 11 54 12 91 4 175 -11 125 -11 131 -1 196 l8 50 33 -58 c18 -32 53 -83 78 -113 25 -31 50 -75 57 -101 12 -48 43 -75 58 -52 10 17 1 59 -17 80 -17 18 -97 163 -131 234 -13 28 -27 57 -32 65 -4 8 -29 25 -55 38 -55 26 -63 40 -43 79 16 33 7 81 -20 100 -21 16 -35 16 -55 1z"]},

{name: "Mus musculus", pathD: ["M85 1287 c-12 -44 1 -131 29 -192 27 -60 86 -140 177 -243 44 -50 56 -69 49 -82 -17 -32 -90 -99 -129 -119 -34 -18 -42 -29 -60 -84 -12 -35 -28 -73 -36 -83 -19 -24 -19 -45 -1 -38 9 3 17 -3 21 -15 8 -25 25 -28 25 -4 0 14 2 14 20 -2 19 -18 20 -17 20 3 0 29 30 35 98 22 88 -18 91 -19 112 -59 10 -21 26 -41 35 -44 12 -5 15 -23 15 -85 0 -58 4 -83 15 -92 25 -20 89 -24 123 -6 17 9 53 16 79 16 26 1 73 10 103 21 30 11 72 24 93 28 21 5 56 23 79 41 l41 33 -5 -31 c-3 -21 0 -32 8 -32 6 0 14 -7 18 -15 4 -13 9 -13 25 -3 15 9 23 9 34 -1 21 -17 30 2 16 33 -7 14 -21 26 -33 28 -16 2 -20 9 -18 30 5 43 -16 176 -34 211 -9 18 -27 43 -39 57 -13 14 -38 62 -55 107 -37 93 -80 144 ……", "M510 725 c-8 -9 -14 -53 -16 -108 -3 -88 -2 -93 23 -119 15 -16 30 -28 35 -28 18 0 98 88 98 108 0 29 -91 162 -111 162 -9 0 -22 -7 -29 -15z m78 -71 c43 -66 53 -104 27 -104 -8 0 -24 -13 -37 -30 l-22 -30 -24 22 c-19 19 -22 29 -16 58 3 19 6 56 5 83 -3 87 10 87 67 1z", "M619 357 c-19 -12 -35 -31 -37 -41 -6 -28 36 -66 73 -66 63 0 88 70 39 109 -33 26 -34 26 -75 -2z m81 -47 c0 -28 -26 -43 -60 -35 -36 8 -48 28 -29 51 32 39 89 29 89 -16z"]},

**3.2.2 Database**

Two new tables are needed by the tool, **homo\_protein** which stores ortholog/paralog data, and **homo\_complex** which stores complex curation status data.

Fields of table homo\_protein:

prot\_query\_id varchar(15) #UniProt ID of protein used as query to fetch homolog(s) from Compara

sp\_name\_query varchar(50) #Name of species to which prot\_query\_id belongs

cpx\_id varchar(10) #ID of complex to which prot\_query\_id belongs

prot\_target\_id varchar(15) #UniProt ID of protein fetched from Compara

sp\_name\_target varchar(50) #Name of species to which prot\_target\_id belongs

homo\_type varchar(10) #Type of homology, 1:1|1:n|m:n|paralog

Fields of table homo\_complex:

cpx\_id1 varchar(10) #ID of complex1

sp\_name1 varchar(50) #Name of species to which cpx\_id1 belongs

cpx\_id2 varchar(10) #ID of complex which is homologous to cpx\_id1

sp\_name1 varchar(50) #Name of species to which cpx\_id2 belongs

homo\_type varchar(10) #Type of homology, ortholog|paralog