Manual to the Co-regulation Data Harvester for *T. thermophila*

Lev M Tsypin

June 29, 2017

Abstract

Identifying co-regulated genes can provide a useful approach for defining pathway-specific machinery in an organism. To be efficient, this approach relies on thorough genome annotation, which is not available for most organisms with sequenced genomes. Studies in *Tetrahymena thermophila*, the most experimentally accessible ciliate, have generated a rich transcriptomic database covering many well-defined physiological states. Characterized genes that are involved in the same pathway show significant co-regulation, and screens based on gene co-regulation have identified novel factors in specific pathways, for example in membrane trafficking. However, a limitation has been the relatively sparse annotation of the *Tetrahymena* genome, making it impractical to approach genome-wide analyses. We have therefore developed an efficient approach to analyze both co-regulation and gene annotation, called the co-regulation Data Harvester (CDH). The CDH automates identification of co-regulated genes by accessing the *Tetrahymena* transcriptome database, determines their orthologs in other organisms via reciprocal BLAST searches, and collates the annotations of those orthologs' functions. Inferences drawn from the CDH reproduce and expand upon experimental findings in *Tetrahymena*. The CDH, which is freely available, represents a powerful new tool for analyzing cell biological pathways in *Tetrahymena*. Moreover, to the extent that genes and pathways are conserved between organisms, the inferences obtained via the CDH should be relevant, and can be explored, in many other systems.

Contents

1 The Basics

1.1 Welcome

When you launch the program, it opens a terminal window with welcome text:

```
Welcome to the Coregulation Data Harvester for T. thermophila!
```

Next, it notifies you where the data analysis will be saved. There are three options, based on which operating system you are using.

Linux

When launched in Linux, the co-regulation Data Harvester saves all files (see Section 4 on page 22) in the same location as the program is saved. Namely, the program will display:

```
Your reports will be written to './csvFiles'.
```

In UNIX systems, './' simply means 'here'.

Windows

The Windows distribution saves the analysis output and intermediary data files in separate places. The analysis data is saved to the folder 'csvFiles', which is a subfolder to 'CoregulationDataHarvester', which itself will live in your 'My Documents' folder.

```
Your reports will be located in 'Documents/CoregulationDataHarvester/csvFiles'.
You can use Excel or an equivalent program to view the data.
```

Mac OS

The Mac OS distribution is much like the Windows one. It saves the files in entirely analogous places, but since the operating systems are different, these places have somewhat different designations:

```
Your reports will be written to
Documents/CoregulationDataHarvester/csvFiles in your Home directory.
You can use Excel or an equivalent program to view the data.
```

In essence, this means that you can open Finder, select Documents in the left directory list, and find the data analysis from there.

1.2 Interface

This section is a step-by-step walkthrough for using the CDH. Your first choice is the Gene ID (TTHERM_ID) to initialize your query:

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas:
```

You respond by typing in the whole ID (e.g. TTHERM_00216010) for your gene of interest. You can also use deprecated IDs, such as "3723.m00724". However, most often, genes with deprecated IDs have no cDNA or protein sequence listed, so they will result in an empty CDH report, since it will not be able to find any homologs for such genes. If you enter a gene ID that does not include 'TTHERM', the CDH will ask you to confirm that you want to continue with the entered search in the following way:

```
Your query does not have "TTHERM" in it. Please make sure that
there are no typos. Please note that you can no longer simply
enter a string of digits. Proceed (y/n)?
```

If you enter 'y', you will proceed to the next questions. If you enter 'n', you will be asked to enter a gene ID again.

Your next choice is the lower-bound threshold for co-regulation z-scores, This effectively determines how many of the top co-regulated genes will be subjected to homology analysis. The CDH reports will always include information on all genes that are co-regulated with the query, but will only provide annotation predictions down to this lower bound:

```
To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation:
```

Here you can respond with digits or decimal values. Next come the less trivial options. Suppose you want to re-run just the BLAST search and analysis, but not spend any time on searching *TGD* or *TetraFDG* again. Or, maybe, you want to do just the opposite. Or suppose that you just want to fill in any missing files without overwriting old ones. Such options, when properly planned, can make your workflow (see Section 2 on page 6) lot neater and faster. To allow you these options, the CDH asks the following:

```
How should I process your query?
  (1) overwrite all associated files,
  (2) overwrite just the BLASTs and analysis, as well
  as fill in any missing files,
  (3) overwrite only the analysis and fill in any
  missing files, or
  (4) sanitize database errors, or
  (5) run only the FGD/TGD search
  Your choice:
```

You respond by typing the number corresponding to your choice. The next choice allows you to opt to use Dropbox:

```
Send to Dropbox?
  (1) Yes, and also write new results locally.
  (2) Yes, but do not write new results locally.
  Remark: if you chose option (2) or (3) above,
  some files may still be synchronized between the
  Dropbox and local directories.
  (3) No, run everything locally.
  Your choice:
```

At the moment, the CDH does not make use of a true Dropbox application-program interface. However, if you have the desktop application for Dropbox installed with the necessary directories

```
CoregulationDataHarvester/pickledData
```

and

```
CoregulationDataHarvester/BLASTresults,
```

then you have the ability to sync your intermediary data files with Dropbox. Various combinations of options for encountering previously analyzed genes and sending to Dropbox allow you to sync and overwrite your data both locally and in the cloud in different ways (again, see Section 2 on the following page). Next, you select whether to run BLASTp, BLASTx, or both:

```
What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice:
```

The two BLAST algorithms sometimes give slightly different results. Finally, you can specify in which taxonomic groups to run the NCBI BLASTs:

```
You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates, everywhere
or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice:
```

If you select option (4), you will see a continuation of the dialog. First,

```
Please enter your custom entrez query:
```

Next,

```
Please enter a short description of the taxonomic group that
you entrez query defines, such as NOTciliates or opisthokonts.
These descriptions should not include any spaces or punctuation,
as they will be used in file names. Take care to use something
succinct and informative:
```

Since this is the most customizable element in the query so far, it requires the most instructions for the sake of consistency, troubleshooting, and reliability. This will be discussed further in the example in section 2.2.8 on page 14 and section 4 on page 22.

2 Workflow, Tips, and Tricks

Below, are several examples of different CDH functionalities. First, here are a few brief notes about using the CDH effectively:

1. The *de novo* analysis of each gene co-regulated with a given query takes an average of 5-7 minutes, though this depends on the time of day (BLAST searches run faster outside of peak hours) and the size of the taxonomic group being BLASTed against. When specifying the lower-bound z-score for the strength of co-regulation, you are effectively choosing how many of the top co-regulated genes to BLAST, which will determine how long a given query will take to finish.

2. The CDH requires a stable internet connection.

3. When running large searches (over 50 co-regulated genes), please try to run them between 5pm and 9am US Eastern time or over the weekend: the NCBI can get upset and blacklist your IP address if you swamp their computer cluster. Please don't run parallel searches from one computer for the same reason.

4. You may interrupt a running CDH search with Ctrl-C (both on Mac and Windows). This corresponds to a keyboard interrupt error. When you do this the CDH will ask you if you want to start another search.

2.1 Oops, I messed up! How do I restart my query?

While you are still configuring your query, you have ability to easily restart. Once you reach the questions that have predefined options, you can just select something that isn't listed and the program will send you back to the beginning of the query. However, it is important that you use the same *type* of data that the CDH expects. Currently, all of the predefined choices are selected by inputting a number. So, when the CDH asks,

```
What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice:
```

if you were to type in the number "4", which isn't a listed option, the CDH would send you to the beginning of the query to start over. Once the CDH begins its web-searches and analyses, however, shutting it off is the only option if you want to restart.

2.2 Example queries

This section has some recipes, examples of how the CDH can be used.

2.2.1 I just want to run a quick test search to see how the *CDH* runs

This example has a runtime of about 10 mintutes, depending on how fast the BLAST goes.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 8

What should I do if I encounter a previously analyzed gene?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 1

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between
    the Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 3

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
```

```
    (2) BLASTx, or
    (3) both?
    Your choice: 1

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
        and instructions for formulating the entrez query)
    Your choice: 1
```

At the end of this search, you will have a report about TTHERM_00216010, which is a 14-3-3 protein. The selected stringency for the strength of co-regulation is too high to BLAST any other genes. This query instructs the CDH to search for homologs strictly outside of the Ciliates.

2.2.2 I want to run a search from scratch for all co-regulated genes, with the intermediary data stored locally, and use both BLASTs, searching for homologs strictly outside of the Ciliates

This example has a *long* runtime. The reason why it takes so long is because it runs both BLASTp and BLASTx independently for all genes in the co-regulated set.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 3.49

How should I process your query?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 1

Send to Dropbox?
    (1) Yes, and also write new results locally.
```

```
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 3

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 3

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 1
```

Once this search is done, you will have everything that the CDH can give you about TTHERM_00216010.

2.2.3 I want to decompose the previous report with both BLASTs into a report with only BLASTp results

The CDH can reuse previous BLAST results. In essence, all you do is run the previous query, but, in this case, you specify that you want to reuse previous data and that you only want BLASTp. Analogously, you could decompose a report with both BLASTp and BLASTx into one that has only BLASTx results. Or, you could take two separate reports, one with BLASTp and the other with BLASTx, and meld them into one.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 3.49

How should I process your query?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
```

```
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 3

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 3

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 1

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 1
```

2.2.4 I want to remove database errors from past search and re-run the analysis, writing new BLAST results locally

This is a more complicated example and requires some walking-through. As mentioned below (see Section 3 on page 19), sometimes when the cDNA or protein sequence is exceedingly long, NCBI BLAST times out and returns a "CPU limit error". This is marked in the CDH's reports as 'db error' in the place where the predictions normally go. There are sometimes other database errors that occur. If the BLAST for those specific cases is redone, then the problem often goes away. In order to clear database errors, you follow the same procedure as before: specify a unique query that identifies a report that already exists and has the database errors, and select the 'sanitize database errors' option. The CDH automatically deletes the faulty BLAST results, replaces them, and writes a new report. Depending on how many errors there were, this takes a variable amount of time. If you would like to test this, but haven't had any database errors come up, you can take a report you already have and edit one of the annotation predictions to say 'db error'. After you save the file, the CDH will recognize the error and can fix it.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 3.49

How should I process your query?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 4

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 3

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 1

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 1
```

2.2.5 I want to sync all intermediary data files for a past query between the local directory and Dropbox

Anytime that you choose to do this, the CDH compares the local and Dropbox directories. If a file exists locally but not in Dropbox, it is copied to Dropbox; if a file exists in Dropbox

but not locally, it is copied to the local directory; if the file is available in both places, the CDH uses the Dropbox version and overwrites the local one; if the file does not exist in either location, its corresponding search is run *de novo*. For a more in-depth discussion of file processing, see Section 4 on page 22. The key to syncing the data is to enter a query that defines a search that you have already run locally, but indicate that you want to make use of Dropbox.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 3.49

What should I do if I encounter a previously analyzed gene?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 3

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 1

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 1

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 1
```

Note that in this case, the assumption is that the data exists locally, but not in Dropbox. If the report that you are wanting to work with already exists locally, choosing either (1) or (2) will copy the data to Dropbox. Choosing (3) will keep the data in the local folder. If, on the other hand, the file already exists in Dropbox, but not locally, then several things might happen: choosing (1) will copy data locally; choosing (2) will keep everything in Dropbox; choosing (3) will run a new search locally.

2.2.6 I have a new search I want to do, but I do not have time to run all of the BLASTs right now. I still want to get through the FGD/TGD step, and I also want the search data available through Dropbox

This particular example will collect all data that are listed on the *TGD* and *TetraFGD* for the 180 genes that are co-regulated with actin (ACT1; TTHERM_00190950).

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00190950

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 3.49

How should I process your query?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 5

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 1
```

Note that the CDH does not ask you anything about BLASTs in this case.

2.2.7 I previously ran only the FGD/TGD search, and now I want to do both BLASTs, with the results saved both locally and in Dropbox

This example assumes that you are finishing the search for ACT1 from the above example.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00190950

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 3.49

What should I do if I encounter a previously analyzed gene?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 3

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 1

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 3

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 1
```

2.2.8 I want to run a search that finds homologs in a more specific group that just 'strictly outside or strictly within the Ciliates'

In such searches it is important to remember, as discussed in section 4 on page 22, that parameters in your query are used to generate unique file names. Hence, if you make use of

a custom homology search parameter, you need to keep in mind how you would translate it into a short word or phrase that makes sense in a file name.

Ascomycota

This search will output a report on TTHERM_00216010 and the two genes most closely co-regulated with it, with the BLASTp algorithm search for homologs only within the Ascomycota (sac fungi).

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 6

How should I process your query?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 3

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 1

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 1

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
```

```
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 4

Please enter your custom entrez query: Ascomycota

Please enter a short description of the taxonomic group that
you entrez query defines, such as NOTciliates or opisthokonts.
These descriptions should not include any spaces or punctuation,
as they will be used in file names. Take care to use something
succinct and informative: sacFungi
```

Opisthokonta

This search is analogous to the one above, but performs BLAST searches against all opisthokonts, and not only sac fungi.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 6

How should I process your query?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 3

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 1

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 1
```

```
You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 4

Please enter your custom entrez query: Opisthokonta

Please enter a short description of the taxonomic group that
you entrez query defines, such as NOTciliates or opisthokonts.
These descriptions should not include any spaces or punctuation,
as they will be used in file names. Take care to use something
succinct and informative: opisthokonts
```

Humans and (domestic) rabbits

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00216010

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 6

How should I process your query?
    (1) overwrite all associated files,
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 3

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between the
    Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 1
```

```
What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 1

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
    and instructions for formulating the entrez query)
    Your choice: 4

Please enter your custom entrez query: Homo sapiens[Orgn] OR
Oryctolagus cuniculus[Orgn]

Please enter a short description of the taxonomic group that
you entrez query defines, such as NOTciliates or opisthokonts.
These descriptions should not include any spaces or punctuation,
as they will be used in file names. Take care to use something
succinct and informative: humansORdomesticRabbits
```

2.2.9 I want to run a cross analysis to gather CDH predictions for genes that are all co-regulated with several genes of interest

This example allows you to specify several genes of interest to define a CDH query. This can be used to find genes that are all co-regulated with multiple known members of a given pathway. Below is a query for genes that are co-regulated with *NUP50* (TTHERM_00260700), *RBP81* (TTHERM_00549610), and *Importin beta* (TTHERM_01031040). It will retrieve the 43 genes that are co-regulated with all three, and run a full analysis using BLAST searches that strictly exclude the ciliates, writing all data locally. An advantage to this approach is that it necessarily limits the number of co-regulated genes and reduces the run-time of the analysis.

```
Please enter one or multiple TTHERM_IDs (for cross analysis),
separated by commas: TTHERM_00260700, TTHERM_00549610, TTHERM_01031040

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 3.49

What should I do if I encounter a previously analyzed gene?
    (1) overwrite all associated files,
```

```
    (2) overwrite just the BLASTs and analysis, as well
    as fill in any missing files,
    (3) overwrite only the analysis and fill in any
    missing files, or
    (4) sanitize database errors, or
    (5) run only the FGD/TGD search
    Your choice: 1

Send to Dropbox?
    (1) Yes, and also write new results locally.
    (2) Yes, but do not write new results locally.
    Remark: if you chose option (2) or (3) above,
    some files may still be synchronized between
    the Dropbox and local directories.
    (3) No, run everything locally.
    Your choice: 3

What kind of NCBI BLAST algorithm would you like to run?
    (1) BLASTp,
    (2) BLASTx, or
    (3) both?
    Your choice: 1

You may choose whether to look for homologs in all organisms
outside of the Ciliates, only within the Ciliates,
everywhere or using a custom entrez query:
    (1) BLAST outside the Ciliates
    (2) BLAST within the Ciliates
    (3) BLAST everywhere
    (4) Custom (please us the NCBI guidelines
        and instructions for formulating the entrez query)
    Your choice: 1
```

It is important to keep in mind that z-scores become meaningless when running a cross analysis, such as in this example. The z-scores are excluded from such reports.

3 Interpretation of the Data

The CDH returns what is called a 'comma separated values' or '.csv' file. To the computer, this is just a plain text file, where each value is delimited by a comma, tab, or some other systematic symbol. However, programs such as Microsoft Excel or Libreoffice Calc know how to break up .csv files into a human-readable spreadsheet.

The report given by the CDH can be considered in two parts: the known data collected from the *TetraFGD* and the TGD (Table 1), and the predictions from phrase analysis of the BLAST searches (Table 2).

TTHERM_ID:	Common Name:	Description:	z-score:
TTHERM_00216010	FTT18 (14-3-3 protein)	No description available	Queried gene
TTHERM_00160770	FTT49 (14-3-3 protein)	14-3-3 protein	7.34
TTHERM_00977740	None	hypothetical protein	6.09
TTHERM_00464950	None	Glycosyl hydrolases family 31 protein	5.88
TTHERM_00158520	HSP82 (Heat Shock Protein)	Hsp90 family member; 82 kDa heat shock-inducible protein; found in heat shock induced cytosolic hetero-oligomeric complexes with tubulin and hsp73; localizes to mature basal bodies and cortical cytoskeleton; involved in cortical patterning	5.84
TTHERM_00933400	None	hypothetical protein	5.77
TTHERM_00554600	None	hypothetical protein	5.77
TTHERM_00624710	None	EF hand family protein	5.44

Table 1: How the CDH reports publicly available data

The spreadsheet, as represented in Table 1, is ordered according to decreasing z-score (as listed in the fourth column), i.e., in order of decreasing strength of co-regulation. The first three columns correspond to the gene ID, its common name, and its description as available from the *TGD*. As shown in Table 2, the CDH gives four different annotation predictions. In the course of the reciprocal BLAST searches, the CDH separates the identified homologs into putative orthologs and potentially informative paralogs. The first prediction is the most commonly observed phrase among the putative orthologs' annotations. The second prediction is the most commonly observed phrase among the potentially informative paralogs' annotations. The third prediction is the most commonly observed phrase among both the orthologs and paralogs mixed together. The fourth prediction is the longest phrase that recurs among the orthologs and paralogs mixed together. Altogether, these four predictions allow for more subtle interpretations of the given gene's definition. In the specific example

of Table 2, BLAST searches were performed outside of the Ciliates. It is apparent from the annotation predictions that CTH3 is a Ciliate-specific cathepsin, and has potentially acquired new functions, while CTH4 is more likely to have conserved function with other clades. The different columns in CDH reports are summarized below:

TTHERM_ID:	Common Name:	BLASTp Ortholog Summary:	BLASTp Paralog Summary:	BLASTp Mixed Summary	BLASTp Mixed Longest Common Phrase
TTHERM_00321680	CTH3 (cathepsin 3)	hypothetical protein braflddraft114822	cathepsin d	cathepsin d	predicted cathepsin d isoform x1
TTHERM_00445920	CTH4 (cathepsin 4)	dipeptidyl peptidase 1	Paralogs found, but nothing informative	dipeptidyl peptidase 1	predicted dipeptidyl peptidase 1 isoform x

Table 2: Examples of CDH annotation predictions

TTHERM_ID: Here you will see the gene IDs of co-regulated genes. In addition to the standard codes, you may sometimes see identifiers from outdated annotations such as 38.m02218.

Common.Name: Here you will see the text that is available for the gene from the *TGD*.

Description: Here you will see the text that is available for the gene from the *TGD*. In the case that the gene is listed in the *TetraFGD* but not in the *TGD*, the analogous information from the *TetraFGD* will be displayed.

Gene.Ontology GO Terms that were collected from the *TGD*.

z-score: This will always display "Queried Gene" in the first row, since the first row corresponds to the query and a z-score is not applicable. All the other genes are ordered according to their z-score as per the *TetraFGD*.

BLAST(x/p).Ortholog.Summary: Most common phrase among the predicted orthologs for the gene in this row.

BLAST(x/p).Paralog.Summary: Most common phrase among the predicted informative paralogs for the gene in this row.

BLAST(x/p).Mixed.Summary: Most common phrase among both the predicted orthologs and informative paralogs for the gene in this row.

BLAST(x/p).Mixed.Common.Longest.Phrase: Longs phrase that is recurrent among both the predicted orthologs and informative paralogs.

4 Search Details, File Names, and Processing

4.1 Searching the *TetraFGD* and *TGD*

Normally, the search of the two databases is very straight forward. First the CDH opens goes onto the *TetraFGD*, gets all the co-regulation information, and then it goes to the *TGD* and gets all the descriptions and sequences. However, there are some complications to keep in mind. Sometimes, the *TetraFGD* and the *TGD* have slightly different cDNA and protein sequences listed for the same gene. By default, the CDH uses the sequences from the *TGD*. However, there are times when a gene is listed in the *TetraFGD*, but not in the *TGD*. In this case, the CDH will go back to the *TetraFGD* for the protein and cDNA sequence. Some genes have no cDNA or protein sequences available on either the *TetraFGD* or the *TGD*–these genes are excluded from further analysis by the CDH.

This whole process is narrated by the CDH, and saved in the log files, so you do not need to worry about missing any of this information. You will know how many genes were identified, you will be notified for each ten genes that had all their sequences assigned, you will be informed which genes were discarded. All of this narrative is generated in real time in the terminal.

4.2 BLAST details

Once the *TetraFGD* and *TGD* searches are done, the CDH accesses NCBI BLAST directly. The BLAST is run against the non-redundant "nr" database. It returns the top 50 hits, with an expect-value cutoff of 10.0. The parameter matrix used is BLOSUM62. These parameters are all, in principle, configurable, so if there is need to adjust them, it can be done by editing the source code. Every time that a BLAST profile is completed, the CDH displays a message saying what kind of BLAST was performed for which gene, and whether it was saved locally, to Dropbox, or to both.

The reciprocal BLASTs are performed via the *TGD* BLAST server. The homolog that is being reciprocally BLASTed is considered to be an ortholog if the e-value of the original *T. thermophila* gene is within two orders of magnitude of the top hit. If it is outside of this range, then the homolog may be classified as a potentially informative paralog, so long as its available annotation matches the annotation of the top hit. All other homologs are considered uninformative and are excluded from further analysis.

4.3 File naming

The CDH generates quite a few files while it is running: a file for the *TetraFGD/TGD* search; a file for each BLAST; a file for cleaned up homolog definitions; a file for the phrase analysis data; a file for the spreadsheet report. These file names are designed such that they are both machine- and human-readable. Here are the filename structures for all the files that the CDH produces.

```
Report: coreg_info_for_<TTHERM_IDs>_<clade>_<blastp/blastx/both>.<z-score>.csv
Log: <TTHERM_ID>.log
BLAST: <TTHERM_ID>_<clade>_<blastp/blastx/both>.XML
Web-search: coreg_list_for_<TTHERM_ID>.p
BLAST_definitions: homolog_dict_for_<TTHERM_ID>_<blastp/blastx>_<z-score>.p
BLAST summaries: best_phrase_dict_for_<TTHERM_ID>_<blastp/blastx>_<z-score>.p
```

The BLAST results are hidden from the user, because there is nothing to gain from messing with the intermediary data files. You can find them in a folder called CoregulationDataHarvester that is in the `/User/AppData/Local/CoregulationDataHarvester/` directory on Windows machines, and in the `/User/Library/CoregulationDataHarvester/` directory on Mac machines.

4.4 Annotation processing

The basic way that the CDH analyzes the homolog information, is by comparing all permutations of hit definitions pair-wise, using the Ratcliff-Obershelp algorithm, and counting common phrases.This is meant to emulate what a human would normally do: i.e., if the majority of annotations in a list of orthologs say "serine/threonine kinase", one would assume that that is a likely annotation for one's gene. In order to minimize uninformative results, the CDH also does a considerable amount of data cleanup. The CDH removes all the species and genera, as well as any gi-codes from all the hit definitions. It then removes all extraneous punctuation and capitalization. Finally, it does its best to discard useless phrases such as "hypothetical protein". Summarily, the CDH attempts to standardize the wildly varying hit definitions from different organisms so that they become machine readable. Some information, such as the species from in which the homologs are identified, is lost in this process.

The reciprocal BLAST searches sort the homologs into predicted orthologs and predicted informative paralogs. The phrase analysis is performed separately for the orthologs, informative paralogs, as well as for the orthologs and paralogs mixed together. Finally, in addition to reporting the most common shared phrase in each of these three groups, the CDH also reports the longest phrase that was shared among the orthologs and paralogs. Taken together, these four predictions allow the user to interpret things like whether the predicted annotations come from a ciliate-specific expansion, whether the existing annotation was inferred from orthologs or paralogs, etc.

5 Installation Considerations

Installation for Mac and Windows machines is different. For Windows, all you need to do is download the zipped folder and extract the executable file wherever you wish. On Mac, in addition to extracting the executable, you will also be likely asked to select what program to use to run the CDH. In this case, go to "Applications", then "Utilities", then select "Terminal".

The program requires an internet connection. The CDH also has a primitive integration with Dropbox for syncing files. In order to make use of this functionality, you need

to make sure that the desktop Dropbox application is installed in the default location on your computer and contains the folders `CoregulationDataHarvester/pickledData` and `CoregulationDataHarvester/BLASTresults`.

Since the binary executables are downloaded from the internet, you will probably need to give them initial permission to run. When the program launches, you will see a welcome message that informs you where your analysis will be saved. At this point, you are ready to start your search.

When you first launch the program you will see a message that depends on your operating system (Mac OS/Darwin, Linux, or Windows). The program automatically determines how to use directory trees on your OS, as well as where to save data files and analysis.

6 Contributing to the Project

Please send any feedback, questions, and suggestions to `coregulationdataharvester@gmail.com`.

6.1 Things/Thoughts in Progress

1. Updating/Keeping manual up to date
2. Testing use cases
3. Figure out why logs are stored in a different place on Macs
4. Developing official Dropbox API
5. Making "nearest neighbor" charts depicting homologs kept and tossed out during analysis
6. Integration with DAVID gene ontology database
7. Website hosting
8. Migrating data to SQL database
9. Automatic cleaning of old files

6.2 Known Bugs/Issues

1. Logs are saved in the top-level directory on Macs.