SoftwareX 6 (2017) 165–171

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

The Co-regulation Data Harvester: Automating gene annotation starting from a transcriptome database

Lev M. Tsypin, Aaron P. Turkewitz *

Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago IL, 60637, United States

ARTICLE INFO

Article history:
Received 16 March 2017
Received in revised form 29 June 2017
Accepted 30 June 2017

Keywords:
Bioinformatics
Evolution
Protists
Automation

ABSTRACT

Identifying co-regulated genes provides a useful approach for defining pathway-specific machinery in an organism. To be efficient, this approach relies on thorough genome annotation, a process much slower than genome sequencing *per se*. *Tetrahymena thermophila*, a unicellular eukaryote, has been a useful model organism and has a fully sequenced but sparsely annotated genome. One important resource for studying this organism has been an online transcriptomic database. We have developed an automated approach to gene annotation in the context of transcriptome data in *T. thermophila*, called the Co-regulation Data Harvester (CDH). Beginning with a gene of interest, the CDH identifies co-regulated genes by accessing the *Tetrahymena* transcriptome database. It then identifies their closely related genes (orthologs) in other organisms by using reciprocal BLAST searches. Finally, it collates the annotations of those orthologs' functions, which provides the user with information to help predict the cellular role of the initial query. 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.

Code metadata

Current code version	v1.2.0
Permanent link to code/repository used for this code version	https://bitbucket.org/ltsypin/cdhproject
Legal Code License	GNU GPL v3
Code versioning system used	git
Software code languages, tools, and services used	python
Compilation requirements, operating environments & dependencies	any system that can run python 2.7
If available Link to developer documentation/manual	http://ciliate.org/index.php/show/CDH
Support email for questions	coregulationdataharvester@gmail.com

Software metadata

Current software version	1.2.0
Permanent link to executables of this version	http://ciliate.org/index.php/show/CDH
Legal Software License	GNU GPL v3
Computing platforms/Operating Systems	OS X (10.6+) and Windows (x64) Vista or later
Installation requirements & dependencies	
If available, link to user manual — if formally published include a reference to the publication in the reference list	http://ciliate.org/index.php/show/CDH
Support email for questions	coregulationdataharvester@gmail.com

* Corresponding author.
E-mail address: apturkew@uchicago.edu (A.P. Turkewitz).

http://dx.doi.org/10.1016/j.softx.2017.06.006

1. Motivation and significance

Tetrahymena thermophila is a ciliate, one of the best-studied members of this large group of protists [1]. Its use as a model system led to the Nobel Prize-winning discoveries of telomerase and self-splicing RNA, as well as to other breakthroughs, including the isolation of dyneins and making the link between histone modification and transcriptional regulation [2–5]. These contributions to our understanding of important cellular pathways made use of genetic, cell biological, and biochemical approaches. More recently, genomic and transcriptomic data became available for *T. thermophila*, which have been used to infer functional gene networks [6–11]. The purpose of the software introduced in this paper, the CDH (Co-regulation Data Harvester), is to better integrate the genomic and transcriptomic data for *T. thermophila* with available data for all other organisms, in a way that permits researchers to infer functions for genes of interest.

The *T. thermophila* genome has been sequenced and assembled [6], and is available online on the *Tetrahymena* Genome Database (TGD) [7]. While the TGD collates the sequence data along with available gene annotations and descriptions, the genome overall remains incompletely annotated. An extensive transcriptomic database, the *Tetrahymena* Functional Genomics Database or *Tetra*FGD, is also available for *T. thermophila* [10]. These data were collected over a well-established range of culture conditions in which *T. thermophila* undergoes large physiological changes [8–11]. In addition to displaying individual expression profiles, the *Tetra*FGD can indicate the statistical strength of co-expression between any two genes, as calculated using the Context Likelihood of Relatedness (CLR) algorithm [9,12,13]. Co-expression in *T. thermophila*, as judged based on mRNA levels, can reveal functionally significant co-regulation. Gene regulation in this species appears to predominantly occur at the level of transcription [14], and so steady-state mRNA levels may explain the majority of steady-state protein levels, as reported in other systems [15]. In this report, we will refer to genes that are listed as co-expressed in the *Tetra*FGD as co-regulated.

A high-throughput analysis of *T. thermophila* gene expression profiles revealed that accurate gene networks can be inferred from co-regulation data [13], providing evidence that co-regulated genes tend to be functionally associated. This approach has also been used in bacteria, mammals, and apicomplexans [12,16,17]. Furthermore, there is experimental evidence in *T. thermophila* to support the conclusion that co-regulation corresponds to functional association: co-regulation data were used to successfully predict novel factors involved in the biosynthesis of a class of secretory vesicles, called mucocysts [18,19]. These results suggest that the *T. thermophila* transcriptome may be used to bioinformatically infer factors involved in an array of cellular pathways.

The CDH was designed to facilitate genome-wide analyses of gene co-regulation. The CDH automatically mines existing data to find genes that are co-regulated with genes of interest. It then helps the user to deduce likely functions for the co-regulated genes, beginning with forward and reciprocal BLAST searches that identify orthologs in other model organisms. The CDH provides a systematic tool for gathering genomic information from public databases, and it can allow a researcher to quickly develop a hypothesis about the cellular pathways or structures in which a gene of interest may be acting, based upon the genes with which it is co-regulated.

2. Software description

2.1. Software architecture

The CDH was developed for Python 2.7, along with the following packages: `sys`, `os`, `platform`, `logging`, `re`, `dill`, `difflib`, `csv`, `pdb`, `shutil`, `xml`, `win32com.shell`, `requests`, `BeautifulSoup4`, and `Biopython` [20]. Executables for Windows (x64) and MacOS (10.6+) were made using the Pyinstaller library. The CDH gathers available data for a set of co-regulated genes from publicly available databases, which the user can exploit to predict possible gene functions (Fig. 1). The gathered information includes the co-regulation data from the *Tetra*FGD, and the gene names, sequences, and annotations from the TGD. The available annotations come from a combination of experimental results and inferences from homology [7]. The CDH can provide insights into gene functions based on the annotation of their respective orthologs, which are themselves identified by a series of forward and reciprocal BLAST searches *via* the National Center for Biotechnology Information (NCBI). These predicted annotations are generated by using the Ratcliff–Obershelp algorithm [21], as implemented in the python difflib library, to identify common phrases in the orthologs' annotations.

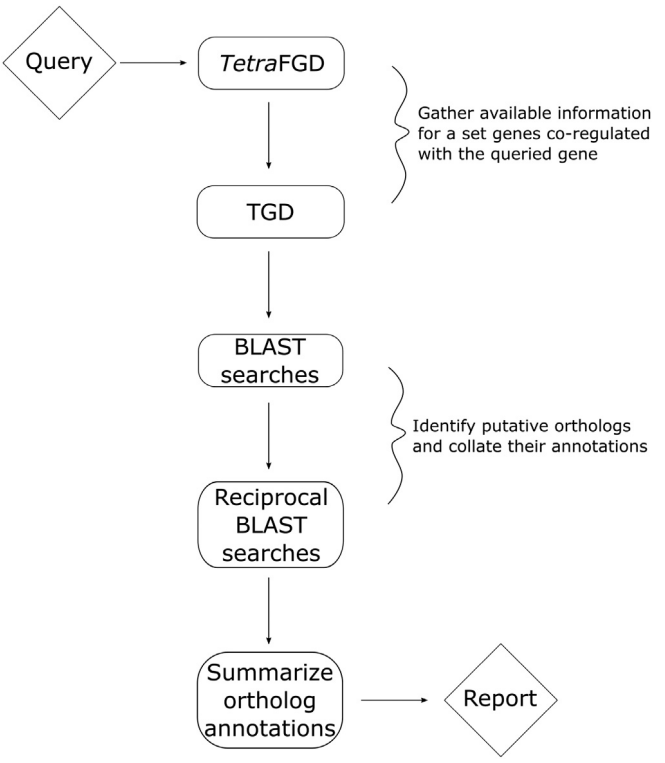

Fig. 1. CDH architecture. Beginning with a single *T. thermophila* gene as a query, the CDH identifies all genes that are co-regulated with it, via the TetraFGD. Next, the CDH uses the TGD to gather the annotation and sequence data for each gene in the co-regulated set. For each gene in the co-regulated set, the CDH then runs forward and reciprocal BLAST searches, through the NCBI and TGD, to identify likely orthologs. A phrase matching algorithm, based on the Ratcliff–Obershelp algorithm [21], as implemented by the python difflib library, is then used to summarize the annotations of the putative orthologs for each *T. thermophila* gene in the co-regulated set. These summaries, which provide predictions about the function (e.g., relevant biological pathway) of the *T. thermophila* gene query, are presented along with the other data gathered, in the final report.

2.2. Software functionalities

The basic functions of the CDH are to gather available co-regulation and annotation data, perform forward and reciprocal BLAST searches and predict gene annotations, and report this gathered information in a human-readable format. The CDH interface first asks the user to enter the ID for the gene(s) whose co-regulated