An Overview of Cartwheel

C. Titus Brown

June 26, 2002

1 Introduction

Cartwheel is a toolkit being developed at Caltech that gives users the ability to do certain types of genomic sequence analyses through a Web site. It is the organizational core of a suite of programs that includes graphical interfaces, such as FamilyRelations, and database-backed Web sites, such as the Cartwheel analysis server Canal.

This short overview is meant for two types of people: biologists who want to know what Cartwheel does and why; and bioinformaticians who want to know what Cartwheel does and how.

1.1 What is Cartwheel good for?

Cartwheel was designed around the need to organize comparative sequence analyses of two or more eukaryotic genomic sequences, with the specific aim of finding cis-regulatory sequences. Since its inception, it has been expanding to become more of a general sequence analysis and annotation workbench system for labs. Our current efforts are focused on providing a fuller analysis pipeline and annotation system; in the longer term, we hope to add features that will help us locate and characterize binding sites for transcription factors.

The associated graphical user interface (GUI), FamilyRelations, is required for visual display of Cartwheel analyses. It allows users to view and interact with analyses and annotations on single sequences, sequence pairs, and sequence triples.
1.2 Why would I be interested in using Cartwheel?

There are several reasons you might be interested in using Cartwheel:

- You are a biologist doing directed sequencing of one or more genomes, and you would like to analyze your contigs and display the results graphically;
- You are a biologist with two or more orthologous genomic regions that you are searching for conserved non-coding regions;
- You are a bioinformatician looking for a simple reusable infrastructure for doing sequence analysis with 3rd-party tools.

1.3 Can you give me an example?

Sure! Suppose you’re a biologist working in mouse, and you are looking at the expression pattern of a terminal muscle differentiation gene (let’s call it “tmd”) that encodes a secreted protein. It turns out that tmd’s expression pattern is conserved across all vertebrates: that is, in chick and zebrafish, people have done *in situ* analyses that show that this gene is expressed in the same places as in mouse.

Because you’re interested in finding out what genomic element is responsible for tmd’s regulation, you go find contigs for chick, zebrafish, and mouse that contain the 50kb genomic region surrounding this gene. You upload the sequences to Cartwheel, create a three-way analysis, and use FamilyRelations to view the results. You find that there are five noncoding elements that are similar at 80% across several 50bp windows, and are all in the same order and orientation with respect to the transcription start site of tmd in the three organisms, suggesting that they are conserved, and hence functional. You then subclone each of the five elements into a reporter construct containing lacZ, create transgenic mice, and look for expression patterns that are muscle-specific.

At this point the remaining analysis is up to you! However, Cartwheel and FamilyRelations have helped you in several ways. First of all, they’ve organized your sequences on a server: you can return to the server and download them, or rename them, or even edit them. They’ve also provided a tool to analyze them in a simple way, and then given you a way to visualize the analyses and interact with them. All of this is done in an easy-to-use and
easy-to-learn environment that was designed with experimental molecular biologists in mind.

The above technique was used on a real live gene ($otx$) in $S.\ purpuratus$, the California purple sea urchin. It yielded a 60% success rate: 11 of 17 conserved elements tested drove expression positively.

1.4 How can I use Cartwheel?

Cartwheel runs on UNIX servers, and it requires fairly advanced computer skills to install. However, once installed, it can be used by many labs simultaneously; at Caltech we run a server that serves about 20 labs.

FamilyRelations works on all platforms that support Java 1.2; this is essentially all major operating systems except for Mac OS 9.

If you would like to try FamilyRelations and Cartwheel out, please contact us directly at cartwheel@caltech.edu. There is a tutorial for FamilyRelations at http://family.caltech.edu/ that you’re welcome to try out.

2 How does it all work?

Figure 1 is a diagram of how Cartwheel connects to various programs. Starting at the top, a number of analysis programs (such as seqcomp and BLAST) are coordinated by compute servers running the Cartwheel batchqueue code. These analyses are scheduled by and returned to a PostgreSQL database running the Cartwheel data-model. External programs (such as the canal Web interface or the Sea Urchin Genome Project) can interact with Cartwheel through several Application Programming Interfaces (APIs), or via Remote Procedure Call (RPC) mechanisms, or directly.

The canal Web interface also exports data directly to FamilyRelations, which allows FamilyRelations to display data from most of the analyses supported by Cartwheel.

Essentially, Cartwheel and FamilyRelations connect an analysis pipeline with a database-backed Web site and a graphical visualization tool. We would be happy to discuss adding specific analyses or specific functionality to it, and even happier to discuss how you can do it yourself!
Figure 1: A diagram of Cartwheel connectivity
2.1 Modifying and Redistributing FamilyRelations/Cartwheel

The source code for both FamilyRelations and Cartwheel is available under the GNU Public License (http://www.gnu.org). This means that everyone is free to use, modify, and redistribute the programs, as long as they do not restrict those rights for other people. For more information on what this license means see the Cartwheel Web site at http://cartwheel.caltech.edu/.

Cartwheel and FamilyRelations are © Copyright 2001, 2002 the California Institute of Technology.

2.2 Credits, Acknowledgements and a Brief History

Cartwheel was designed around the needs of two labs here at Caltech, the Davidson Lab and the Wold Lab, and around software produced as the result of a collaboration between these labs (SeqComp and FamilyRelations, both part of the FamilyJewels project: http://family.caltech.edu).

2.2.1 Authors

The primary perpetrator of Cartwheel thus far has been Titus Brown; Tristan De Buysscher developed the initial comparative sequence analysis tool (SeqComp) and (perhaps inadvertently) did most of the bug testing, too. Madeleine Price and Ramon Cendejas have also contributed to FamilyRelations in significant ways.

Both R. Andrew Cameron and Erich M. Schwarz contributed intellectually to the development and goals of this software, as did my and Tristan's advisors, Eric H. Davidson and Barbara J. Wold. They didn't write any code, though :).

2.2.2 History

The initial version of Cartwheel was developed between August and December of 2001 to provide a wrapper for SeqComp and BLAST analyses. The current version is a complete rewrite from April 2002; it was rewritten to take advantage of the Quixote and cucumber toolkits, which made it much more extensible than the old version. Development continues!