

FamilyRelations: comparative sequence analysis of genomic data

C. Titus Brown
Computation and Neural
Systems
California Institute of
Technology
Pasadena, CA, 91125
titus@caltech.edu

Tristan De Buyscher
Computation and Neural
Systems
California Institute of
Technology
Pasadena, CA, 91125
tristan@caltech.edu

Eric Davidson
Biology
California Institute of
Technology
Pasadena, CA, 91125
davidson@mirsky.caltech.edu

ABSTRACT

The construction of detailed models of gene networks can be facilitated by many types of computational tools. Because gene regulatory interconnections are largely encoded in genomic DNA, tools to locate and characterize *cis*-regulatory regions are of particular importance for the rapid constructions of gene networks.

This need to locate *cis*-regulatory regions and the increasing availability of genomic sequence from closely related organisms has resulted in a pressing need for tools to aid in phylogenetic footprinting. We present a pair of tools, **seqcomp** and FamilyRelations, designed for comparative sequence analysis of orthologous non-coding regions in BAC-sized sequences of eukaryotes.

seqcomp is an implementation of a simple $N \times M$ fixed-width comparison algorithm: it exhaustively compares all possible contiguous windows of a given width for similarity above a particular threshold. Because even 15-bp exact matches are extremely unlikely to exist in two random sequences, this simple comparison technique is a powerful tool for phylogenetic footprinting; moreover, because no underlying statistical model of sequence change is assumed, **seqcomp** is species-agnostic.

FamilyRelations is a graphical user interface for exploration of **seqcomp** results, as well as other analyses. FamilyRelations is designed to be used by wet-bench biologists working on a desktop machine; it supports sequence feature display, pairwise display, dotplots, and motif searching.

FamilyRelations and **seqcomp** have been used in several biology labs at Caltech to locate and characterize putative *cis*-regulatory regions. Because the tools have no dependence on organism-specific genomic features, they can be used to analyze genomic data from many different species; so far, this has included *C. elegans*, *S. purpuratus*, *Mus musculus*, *D. melanogaster*, and *F. rubribes*.

Both tools are freely available and redistributable under the GNU Public License. **seqcomp** is written in C, and FamilyRelations is a cross-platform GUI written in Java. For more information, please visit <http://family.caltech.edu/>.