Databases and ontologies

**REDfly: a Regulatory Element Database for Drosophila**

Steven M. Gallo¹, Long Li², Zihua Hu¹ and Marc S. Halfon²,3,∗

¹Center for Computational Research, ²Department of Biochemistry and ³Center of Excellence in Bioinformatics and the Life Sciences, 140 Farber Hall, State University of New York at Buffalo, 3435 Main Street, Buffalo, NY 14214, USA

Received on September 29, 2005; revised and accepted on November 17, 2005
Advance Access publication November 22, 2005
Associate Editor: Chris Stoeckert

ABSTRACT

Summary: Bioinformatics studies of transcriptional regulation in the metazoa are significantly hindered by the absence of readily available data on large numbers of transcriptional cis-regulatory modules (CRMs). Even the richly annotated *Drosophila melanogaster* genome lacks extensive CRM information. We therefore present here a database of *Drosophila* CRMs curated from the literature complete with both DNA sequence and a searchable description of the gene expression pattern regulated by each CRM. This resource should greatly facilitate the development of computational approaches to CRM discovery as well as bioinformatics analyses of regulatory sequence properties and evolution.

Availability: [http://redfly.ccr.buffalo.edu](http://redfly.ccr.buffalo.edu)

Contact: mshalfon@buffalo.edu

INTRODUCTION

Despite their importance, the transcriptional cis-regulatory modules (CRMs) associated with the majority of genes in the higher eukaryotes are unknown, and few are yet included in genome annotations. This lack of a comprehensive collection of known CRMs presents a considerable roadblock to large-scale computational analyses of transcriptional regulatory sequences. Easy access to a compilation of CRM sequences would have considerable value for subsequent CRM discovery—for instance, by providing training data for supervised learning approaches—as well as for investigations into the nature and evolution of cis-regulatory elements.

Although *Drosophila melanogaster* has one of the most fully annotated metazoan genomes, fewer than 45 genes are annotated with documented CRMs ([http://flybase.bio.indiana.edu/annot/](http://flybase.bio.indiana.edu/annot/)). An additional collection of about 60 CRMs from about 24 different genes involved in early embryonic gene expression has also been developed (Lifanov et al., 2003; Schroeder et al., 2004). This collection has been used for a number of studies (Abnizova et al., 2005; Berman et al., 2004; Costas et al., 2003; Grad et al., 2004; Gupta and Liu, 2005; Lifanov et al., 2003; Papatsenko et al., 2002; Philippakis et al., 2005; Rajewsky et al., 2002; Schroeder et al., 2004; Zhou and Wong, 2004), but is limited by the fact that all of the CRMs are involved in regulating a similar pattern of gene expression and in binding a similar repertoire of transcription factors (TFs). Recently, Bergman et al. (2005) have compiled a comprehensive database of DNAse I footprints in *Drosophila*. However, the footprint data only detail TF binding sites, and not functional CRM sequences.

We introduce here a database of published *Drosophila* CRMs, REDfly (Regulatory Element Database for fly). REDfly currently contains over 600 CRMs along with their sequences and a description of the expression patterns for which they are responsible. The goal of REDfly is to provide a comprehensive source of sequence and expression pattern data for *Drosophila* CRMs.

OVERVIEW OF THE DATA

For the initial REDfly release we have focused on sequences that have been demonstrated to be sufficient to regulate gene expression, primarily through reporter gene assays in transgenic animals. Sequences necessary for expression, but not clearly sufficient—e.g. TF binding sites or sequences uncovered by small deletions—are not presently incorporated. Each record contains the DNA sequence of the CRM as well as coordinates mapped to the release 4 genomic sequence ([http://www.fruitfly.org/annot/release4.html](http://www.fruitfly.org/annot/release4.html)). We have also noted if the given CRM includes the associated gene’s promoter. A more detailed explanation of how sequences were chosen and mapped onto the genomic sequence are provided in the online User’s Guide.

REDfly currently contains an excess of 600 CRMs associated with more than 200 genes drawn from over 200 references. Curation of the database will continue both to add newly reported CRMs and to fill in previously reported CRMs; we estimate that better than two-thirds of the reported CRMs are presently included. Approximately 75% of the CRM sequences are <2500 bp in length and 50% <1125 bp. Approximately 25% of the included CRMs overlap other included CRMs, and 18% of the CRMs include their gene’s promoter. Greater than 75% of the CRMs regulate gene expression outside of the blastoderm embryo and are thus not included in the previous compilations of *Drosophila* regulatory elements (Lifanov et al., 2003; Schroeder et al., 2004). The total amount of (non-overlapping) CRM sequence in the database is slightly >1 Mb, or ~0.86% of the total *Drosophila* euchromatic genome, with sequences from each chromosome represented. The median distance between CRMs ranges from 23.4 kb on chromosome 2L to 275.4 kb on chromosome 4; the maximum distance in most cases is ~10% of the chromosome arm length. A more detailed bioinformatics analysis of the CRMs will be presented elsewhere.

∗To whom correspondence should be addressed.
EXPRESSION PATTERN ANNOTATION AND SEARCHING

Each CRM has been annotated with a description of the expression pattern that it directs. REDfly uses the Drosophila anatomy ontology (http://obo.sourceforge.net/cgi-bin/detail.cgi?drosanat; Drysdale, 2001) for assigning expression patterns, which will enable high interoperability with other biological databases.

REDfly has two modes of searching for expression patterns. The ‘Expression Term’ search will search for records whose expression annotation matches the specified term. Alternatively, users can use the ‘Ontology’ search function, which will return records whose expression annotation matches the specified term or any of the descendents terms in the ontology hierarchy. For example, a search for ‘mesoderm’ using the Expression Term search will return only those CRMs whose annotation explicitly includes the word ‘mesoderm’. However, a similar Ontology search will also return mesodermal derivatives such as ‘embryonic somatic muscle’ and ‘cardioblast’. The Ontology search can be initiated either by entering an ontology term in the search box or by browsing the ontology tree in a pop-up window. This enables easy access to the terms and term hierarchy. A link is provided to the FlyBase gene expression report page (Drysdale et al., 2005), which provides a list of genes annotated in FlyBase with the current ontology term. Link-out is also provided to genes with similar expression patterns in the BDGP in situ hybridization database (Tomancak et al., 2002). As mappings between the anatomy ontologies of different organisms are developed, it should be possible to create links to similarly expressed genes in these organisms as well.

The REDfly expression pattern annotation is drawn from the textual descriptions given by authors. As these are provided in the literature in varying levels of detail and are typically not reported using the ontology terms, providing an exact annotation is not always straightforward. We have attempted to err on the side of more general rather than more restrictive assignments (e.g. ‘embryonic muscle system’ versus ‘abdominal dorsal acute muscle’, unless explicitly so annotated by the author). The Ontology search function therefore provides a way to identify CRMs that potentially drive similar spatial patterns of expression despite that expression having been described at different levels of detail in the literature.

The ability to search by expression pattern is a key feature of REDfly that promises to be highly useful for developing models for computational discovery of tissue-specific CRMs (e.g. Grad et al., 2004) and for investigating structural and organizational properties of CRMs (Erives and Levine, 2004; Senger et al., 2004). However, we note that the anatomy ontology does not at this time always provide a means to distinguish sub-tissue- or organ-level cell populations. Thus, e.g. two entries annotated as ‘wing disc’ may in fact refer to non-overlapping cell types within the disc. Users are therefore encouraged to consult original references for detailed descriptions of expression patterns.

GRAPHICAL DISPLAY, DOWNLOAD AND LINK-OUT

A number of options for graphical display, download and link-out to other databases have been provided. From the report page for any record, links are available to display the CRM in the UCSC genome browser (Kent et al., 2002) or in the Generic Genome Browser (Gbrowse; Stein et al., 2002). CRM sequences can be downloaded in multi-FASTA format, in the format for custom Gbrowse annotations, and in CSV or GFF3 format that includes additional field data. Links are available to the FlyBase report of the associated gene and to the PubMed citation of the primary reference. As noted above, for each expression term associated with a given CRM, it is also possible to link to a list of genes annotated as having the same expression pattern in both FlyBase and in the BDGP in situ hybridization database.

KNOWN LACUNAE AND FUTURE INCLUSIONS

A number of potentially important regulatory sequences have not yet been included in REDfly. These include CRMs inferred but not demonstrated to have specific activities based on deletion analysis, either from reporter gene assays or from genomic deletions, as well as silencer and boundary elements. REDfly is also currently limited to CRMs from D.melanogaster, despite the growing number of functionally tested sequences from other fly species. Future updates of REDfly will include such sequences along with a description of the evidence used to support their assignment as CRMs. We also hope to continue to upgrade the expression pattern search functions and the graphical display capabilities, and to improve cross-referencing with other databases.

ACKNOWLEDGEMENTS

We thank J. Leatherbarrow for assistance with literature curation, Q. Nguyen for programming assistance, H. Apitz, G. Mardon, J. Posakony and J. Wildonger for providing CRM sequences, and E. Wang and S. Sinha for comments on the manuscript and database. M.S.H. is supported by NIH grant HG002489.

Conflict of Interest: none declared.

REFERENCES


