Chapter 22

Practical Computational Methods for Regulatory Genomics: A *cis*GRN-Lexicon and *cis*GRN-Browser for Gene Regulatory Networks

Sorin Istrail, Ryan Tarpine, Kyle Schutter, and Derek Aguiar

Abstract

The CYRENE Project focuses on the study of *cis*-regulatory genomics and gene regulatory networks (GRN) and has three components: a *cis*GRN-Lexicon, a *cis*GRN-Browser, and the Virtual Sea Urchin software system. The project has been done in collaboration with Eric Davidson and is deeply inspired by his experimental work in genomic regulatory systems and gene regulatory networks. The current CYRENE *cis*GRN-Lexicon contains the regulatory architecture of 200 transcription factors encoding genes and 100 other regulatory genes in eight species: human, mouse, fruit fly, sea urchin, nematode, rat, chicken, and zebrafish, with higher priority on the first five species. The only regulatory genes included in the *cis*GRN-Lexicon (*CYRENE genes*) are those whose regulatory architecture is validated by what we call the Davidson Criterion: *they contain functionally authenticated sites by site-specific mutagenesis, conducted in vivo, and followed by gene transfer and functional test.* This is recognized as the most stringent experimental validation criterion to date for such a genomic regulatory architecture. The CYRENE *cis*GRN-Browser is a full genome browser tailored for *cis*-regulatory annotation and investigation. It began as a branch of the Celera Genome Browser (available as open source at http://sourceforge.net/projects/celeragb/) and has been transformed to a genome browser fully devoted to regulatory genomics. Its access paradigm for genomic data is zoom-to-the-DNA-base in real time. A more recent component of the CYRENE project is the Virtual Sea Urchin system (VSU), an interactive visualization tool that provides a four-dimensional (spatial and temporal) map of the gene regulatory networks of the sea urchin embryo.

Key words: *cis*-regulatory architecture, gene regulatory networks, transcription factors, *cis*GRN-Lexicon, *cis*GRN-Browser, virtual sea urchin.

1. Introduction

When the GRN context is clear, we will use at times the shorthands "*cis*-Lexicon" and "*cis*-Browser". The *cis*-Lexicon and the *cis*-Browser are conceptually two separate entities, a database

I. Ladunga (ed.), *Computational Biology of Transcription Factor Binding*, Methods in Molecular Biology 674,
DOI 10.1007/978-1-60761-854-6_22, © Springer Science+Business Media, LLC 2010

and a visualization tool; however, from the user's point of view, they are intertwined into one integrated software environment for regulatory genomics.

The *cis*GRN-Lexicon is a database containing the regulatory architecture (the genomic regulatory region) of a set of transcription factor-encoding genes as well as of a number of other regulatory genes. This architecture is presented with full genomic structure known to date, including transcription factor binding site sequences, the organization into *cis*-regulatory modules (CRMs), and various other types of functional genomics (e.g., logic functions) annotations of the DNA regulatory region revealed by *cis*-regulatory analyses and systematic experimental perturbations of gene regulatory networks. The *cis*GRN-Lexicon annotations, accessible though the *cis*GRN-Browser, include the transcription factor binding site, the *trans* acting factor, the protein family to which the *trans* acting factor belongs, the *cis*-Regulatory Module (CRM) boundaries, the spatial and temporal functionality of the CRM, and the molecular function of the encoded protein. The *cis*GRN-Lexicon is embedded in and accessed through the *cis*GRN-Browser and is supported by various software libraries of tools for *cis*GRN-Lexicon annotators. One such system under development is CLOSE (*cis*-Lexicon Search Engine), a set of algorithmic strategies for literature extraction of *cis*-regulation articles to speed identification of new CYRENE genes and estimate the "dimension" of the CYRENE gene universe.

We describe the current state of the CYRENE *cis*GRN-Browser: its detailed architecture and its planned improvement in partnership with scientists from the Davidson Lab at the California Institute of Technology, where the *cis*-Browser has been in use for the past few years (18).

The Virtual Sea Urchin software system aims at giving a three-dimensional representation of the embryo's cellular anatomy stages in which gene expression is represented in time and within specific cell types. It will be integrated with the *cis*GRN-Browser and BioTapestry (19, 20) to present a *View from the GRNs.*

1.1. Algorithms for CRM Regulatory Architecture Prediction

The object of the *cis*-Lexicon is to create a data set that makes possible the prediction of *cis*-regulatory elements in DNA sequences of unknown function. To achieve a better prediction algorithm, the data set used must not be contaminated by low-quality data. Furthermore, the categories for annotation of a gene into the lexicon must be based on a relevant model of evolution and biological function.

Helpful surveys describing the state of the art of algorithms for prediction of sites, modules, and organization of regulatory regions are refs. 21–24. Many algorithms presented in the literature have aimed at addressing various aspects of the computational prediction problems related to regulatory genomics

(25–44, 19, 45–47). The present work aims at providing a database of *cis*-regulatory architectures, experimentally validated at the highest level as a basis for the design of the next generation of regulatory prediction algorithms.

1.2. CYRENE Genes and GRNs

The *cisGRN*-Lexicon presently contains the regulatory architecture of 200 transcription factors encoding genes and 100 other regulatory genes in eight species: human, mouse, fruit fly, sea urchin, nematode, rat, chicken, and zebrafish. The regulatory architecture of each of these CYRENE genes contains only functionally authenticated sites by site-specific mutagenesis, conducted in vivo, and followed by gene transfer and functional test. As the objective is to determine how genes are regulated in vivo, it follows that only in vivo *cis*-regulation studies should be admitted to the *cis*-Lexicon.

This database differs from other databases of gene regulation in several ways. It will be displayed in an interactive way that presents on one page the whole genome, a workspace for cis-regulatory analysis, and all the relevant gene functions. Many annotation categories and functions are unique to the *cis*-Lexicon. Experimental evidence required for admittance to the lexicon is stringently examined.

1.3. The Need for Integrated Cell Models

Combining GRN inference experiments (identification of regulatory genes (41, 48–50), perturbation experiments (51–52, 49, 53), *cis*-regulatory analysis (11, 13, 54, 55), etc.) with recent developments in systems biology imaging (56, 57) will make possible the construction of a full 4D spatiotemporal map of the sea urchin embryo. Such a map will fully describe the intra- and intercellular interactions of the GRN in the developing sea urchin embryo. The analysis and visualization tools needed to interpret these data seem to have lagged far behind experiments. Thus, we have been developing a natural visualization and analysis environment – the Virtual Sea Urchin (VSU) – that allows researchers to interrogate the developmental atlas at any time and at any position in the developmental process.

2. Materials

2.1. cisGRN-Browser: Software

The CYRENE *cis*-Browser was developed in the Eclipse IDE for Java Developers (http://www.eclipse.org). The foundation of the *cis*-Browser is the Celera Genome Browser (58), whose source code is available free on SourceForge.net (http://sourceforge.net/projects/celeragb/). The

cis-Lexicon is stored as an Apache Derby Database (http://db.apache.org/derby/).

2.2. Virtual Sea Urchin: Software

Our system is composed of OGRE (an open-source 3D graphics engine), a C++ visualization application, and data describing the GRN such as transcription factor binding site affinity and products, cell-signaling pathways, etc. Three-dimensional sea urchin embryonic models were created using Blender. Future directions will include the integration of the regulatory network simulator BioTapestry (19, 20) and the *cis*GRN-Browser Cyrene. Portions of the embryo viewing application were provided by OgreMax (http://www.ogremax.com).

3. Methods

3.1. Cis-Lexicon

3.1.1. Anatomy of the Lexicon

The clues given by biology for the rules of *cis*-regulation are copious; the difficulties lie in creating criteria to represent the clues in a meaningful way. We must thus develop a classification system that neither oversimplifies biology, so that categories lose their physical meaning, nor overcomplicates the issues by creating more categories than necessary. While biology is inherently resistant to precise definitions, categories are necessary for the sake of high-throughput data clustering. The many challenges in creating controlled vocabularies are discussed shortly. The vocabularies of the *cis*-Lexicon are based on two guiding principles: (1) vocabularies should represent phenomena in a way that fits their physical interaction and (2) vocabularies should facilitate comparison. Where possible, the vocabularies are externally linked in such a way that when the vocabulary is updated, so is the *cis*-Lexicon.

Cis-Lexicon annotations include the transcription factor binding site (TFBS), the function of the TFBS, the *trans* acting factor, the protein family to which the *trans* acting factor belongs, the *cis*-regulatory module (CRM), the spatial and temporal functionality of the CRM, and the molecular function of the encoded protein. These categories in the *cis*-Lexicon were developed for useful data clustering. When a gene could not be annotated accurately within the constraints of our chosen vocabulary, new categories were created in order to categorize the *cis*-regulatory architecture in a biologically relevant manner.

3.1.2. cis-Regulatory Ontology

Transcription factor binding sites (TFBS) usually span 6 to 8 bp, though sometimes many more. Every TFBS in the lexicon is annotated as performing one or more of the following functions (more functions may exist in the natural world, but this is the set

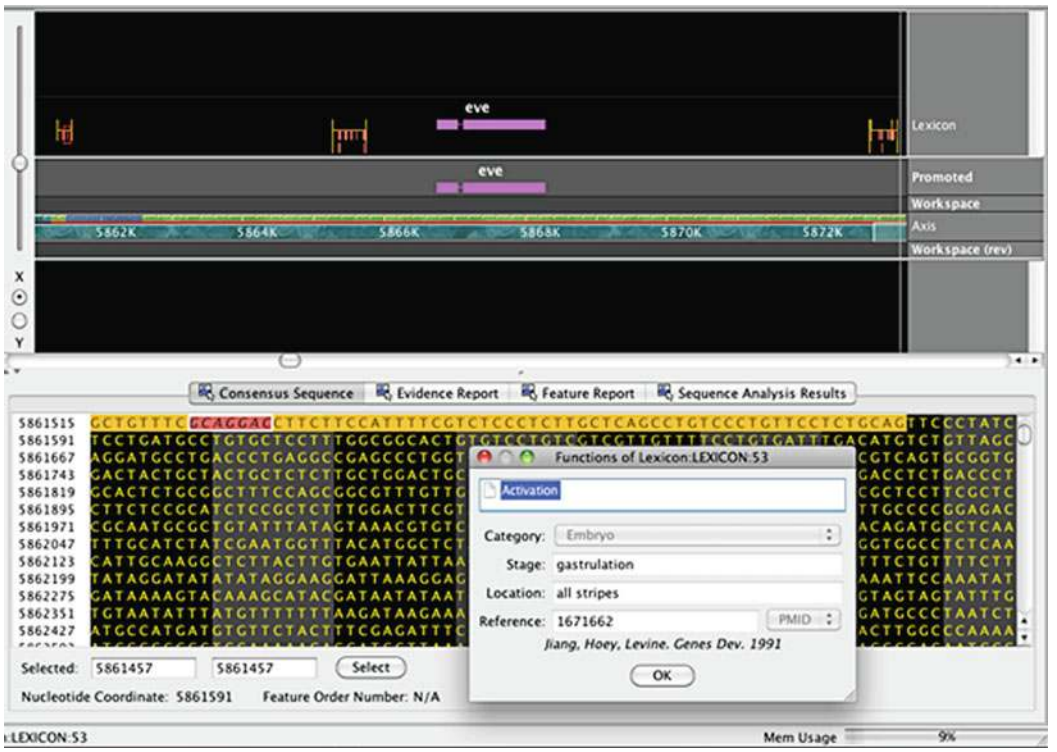

Fig. 22.1. Screenshot from the *cis*-Browser of the *Drosophila* transcription factor encoding gene, *eve*.

of all functions encountered so far) [*See* **Fig.** 22.1 for an example of *cis*-regulatory function in the *cis*Browser]:

- Repression – Indicates that mutating the TFBS increases gene expression or produces ectopic expression. Repressors may act "long range," when the repression effect may target more than one enhancer, or "short range," when repression affects only neighboring activators (59, 60). The function of repression applies in cases where the repressors interact with the basal transcription apparatus either directly or indirectly (61).
- Activation – Indicates that mutation decreases gene expression. An activator TFBS may act over a large genomic distance or short. See Latchman (62) for further discussion of some of the many ways a transcription factor can accomplish activation.
- Signal response – Indicates that the transcription factor has been shown to be activated by a ligand such as a hormone (phosphorylation is not included) (63).
- DNA looping – Indicates that the binding factor is involved in a protein–protein interaction with another binding factor some distance away that causes the DNA to form one or

more loops. This looping brings distant regulatory elements closer to the basal transcription apparatus (64).

- Booster – Indicates that the TFBS does not increase gene expression on its own but can augment activation by other TFBSs.
- Input into AND logic – Indicates that the TFBS can activate gene expression only when two or more cooperating TFBSs are bound. Assigned to one of at least two TFBSs (14).
- Input into OR logic – Indicates that the TFBS can activate gene expression when either or both of two or more cooperating TFBSs are bound. Assigned to one of at least two TFBSs (14).
- Linker – Indicates that a TFBS is responsible for communicating between CRMs.
- Driver – Indicates that this TFBS is the primary determining factor of gene expression. The binding factor appears only in certain developmental situations and thus is the key input for directing gene expression. TFBSs that are not drivers usually bind ubiquitous factors (65).
- Communication with BTA (basal transcription apparatus).
- Insulator – Indicates that the TFBS causes *cis*-regulatory elements to be kept separate from one another. Insulators can separate the *cis*-regulatory elements of different genes as well as act as a barricade to keep active segments of DNA free of histones and remain active (66).

3.1.3. The trans Acting Factor

The transcription factor binding to the regulatory DNA is annotated as the gene name given in NCBI rather than the name given to the factor in the literature. For example, while Inagaki refers to human TF *c-Jun* (67), this is annotated in the *cis*-Lexicon as *JUN* for consistency. Each transcription factor in the lexicon is also assigned to a leaf of a transcription factor hierarchy adapted from TRANSFAC (68) (*see* **Figs.** 22.2 and 22.3)). More closely-related transcription factors may behave more similarly, so that when the data in the *cis*-Lexicon are clustered, patterns may be found by grouping transcription factors according to their evolutionary origins.

3.1.4. cis-Regulatory Modules

TFBSs occur in groups and each grouping usually directs gene expression in one temporal and spatial location. The CRM (1) includes the binding sites responsible for gene expression as well as the neighboring sequence established to enable the TBFSs to function correctly. Each CRM in the lexicon is annotated as functioning in a specific spatial and temporal location. There is currently no associated ontology for annotating this location in the lexicon. Exhaustively naming all locations and time points of

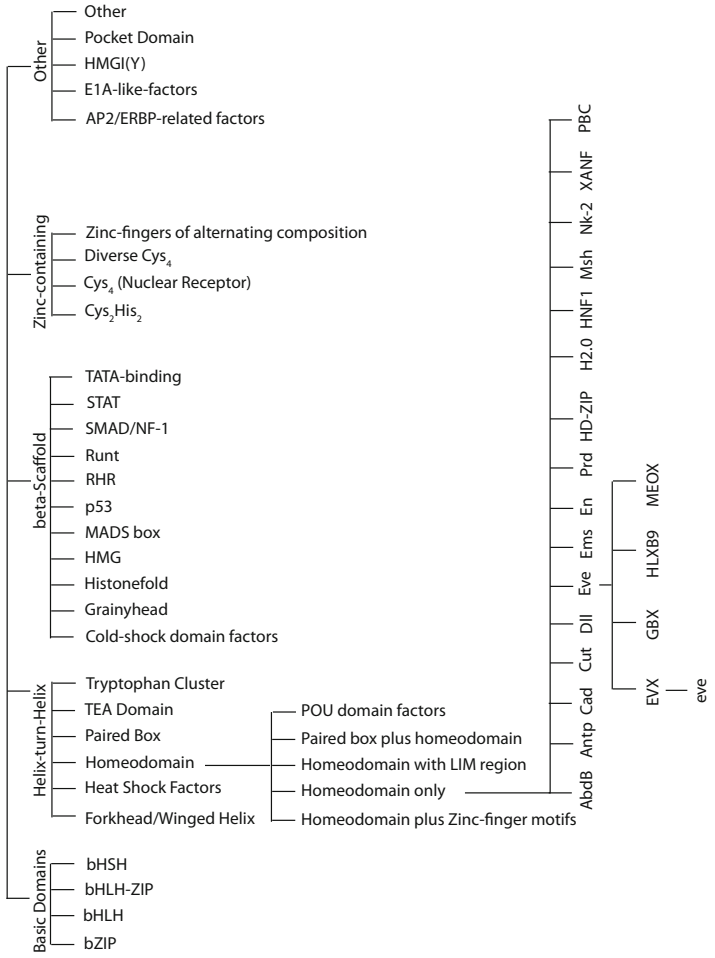

Fig. 22.2. Transcription factor hierarchy (homeodomain family expanded as an example).

an organism's development is beyond the scope of this project, though a controlled vocabulary of body parts and stages would be useful.

3.1.5. Gene Functions

Each gene whose *cis*-regulatory architecture is annotated in the lexicon is assigned to one of seven gene function categories. Many ontologies of gene functions already exist, such as the Gene Ontology (69) and Panther Classification System (70), but the *cis*-Lexicon Gene Functions ontology was created with the specific intent of grouping gene functions so that genes with similar *cis*-regulatory architecture are grouped together. The hypothesis is that housekeeping genes have *cis*-regulatory architecture distinct from transcription factor-encoding genes or signaling genes. GO annotations for each gene indicated in brackets show similarities and differences in gene function annotation between GO and the *cis*-Lexicon. *See* **Fig.** 22.4 for gene functions in the *cis*-Lexicon.

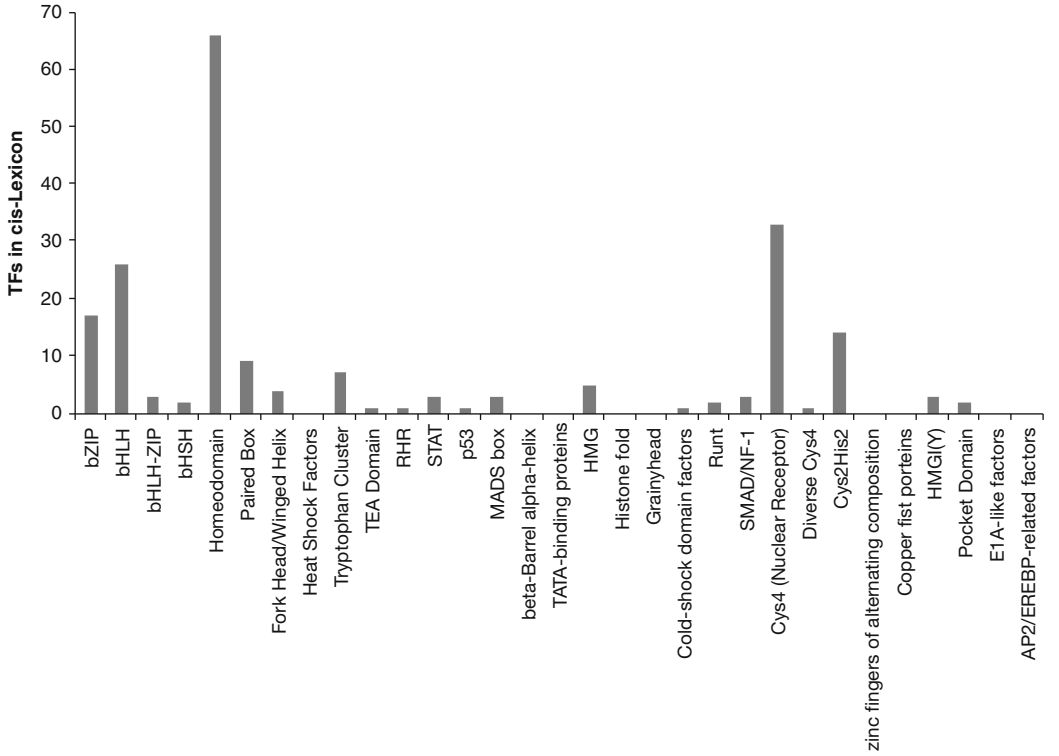

Fig. 22.3. Distribution of transcription factor encoding genes annotated in the *cis*-Lexicon categorized by transcription factor family (68).

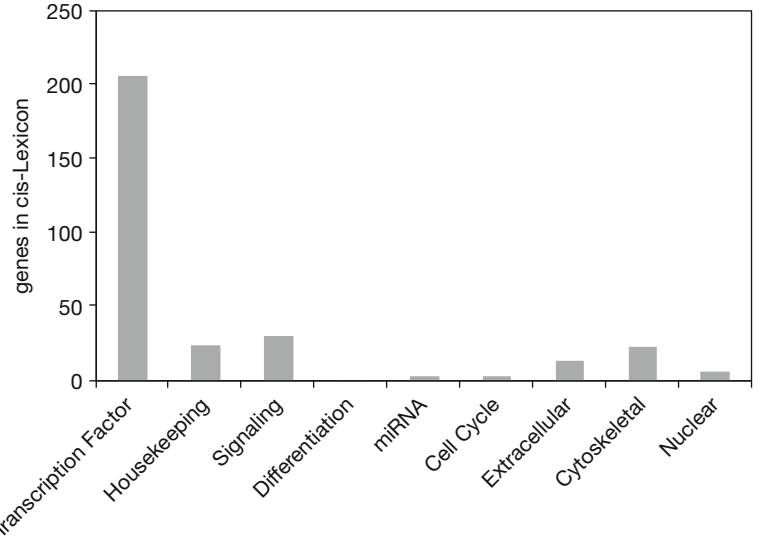

Fig. 22.4. Distribution of gene functions for which the *cis*-regulatory architecture has been annotated in the *cis*-Lexicon.

- *Cell cycle* – Genes involved in cell division that decide when to replicate DNA, when to divide the cell, etc., especially before cells terminally differentiate. Example: cyclins run in the *cis*-Lexicon: IL3 (*Homo sapiens*) (71) [extracellular space, cytokine activity, growth factor activity, etc.]; BCL2 (*H. sapiens*) (72) [apoptosis, etc.].
- *Cytoskeletal* – Genes involved in maintaining the structure of the cell, for example, actin and tubulin. In the *cis*-Lexicon: myh4 (*M. musculus*) (73) [actin binding, myosin filament, etc.; ASL (*G. gallus*) (74) [structural constituent of eye lens].
- *Differentiation* – Genes "expressed in the final stages of given developmental processes… They receive rather than generate developmental instructions" (2). In the *cis*-Lexicon: malpha (*Drosophila melanogaster*) (75) [cell fate specification, notch signaling pathway, sensory organ development].
- *Extracellular* – Genes whose product is released from the cell, such as hormones. These do not include membrane proteins, which could be categorized as housekeeping or signaling. In the *cis*-Lexicon: Col5A2 (*H. sapiens*) (76) [eye morphogenesis, skin development, extracellular matrix structural constituent, etc.]; SOD3 (*H. sapiens*) (77) [cytoplasm, extracellular region, zinc ion binding, etc.].
- *Housekeeping* – Genes continuously expressed that regulate processes inside the cell such as transcription apparatus, ribosomes, degradation proteins, and many enzymes. In the *cis*-Lexicon: btl (*D. melanogaster*) (78, 79) [endoderm development, glial cell migration, negative regulation of axon extension, etc.]; BACE1 (*Homo sapiens*) (80) [proteolysis, peptidase activity, etc.].
- *Transcription factor* – Genes whose product binds to DNA in a sequence-specific manner to affect gene expression. Does not include basal transcription factors. In the *cis*-Lexicon: twi (*D. melanogaster*) (81–84), [specific RNA polymerase II transcription factor activity, etc.]; Hoxa4 (*H. sapiens*) (85); Foxa2 (*Mus musculus*) (86, 87) [RNA polymerase II transcription factor activity, enhancer binding, etc].
- *miRNA* – Genes encoding micro RNAs. In the *cis*-Lexicon: DmiR-1 (*D. melanogaster*) (88) [cardiac cell differentiation, regulation of notch signaling].
- *Signaling* – Genes acting as part of a signaling pathway such as hormones, hormone receptors, and kinases. In the *cis*-Lexicon: IL4 (*M. musculus*) (89) [extracellular space, B cell activation, interleukin-4 receptor binding, etc.]; ins2 (*R. norvegicus*) (90) [cytoplasm, extracellular space, hormone activity, etc.].

3.1.6. Quintessential Diagram Problem

The categories used in the literature to classify the function of a TFBS are often described in simple terms that prevent the annotator from fully describing the function according to the *cis*-regulatory ontology (*see* **Section** 3.1.2). In the literature, a TFBS is generally declared an "activator" if deletion lowers output and a "repressor" if deletion increases output. Other more complex mechanisms may cause increased or decreased expression, such as DNA looping, communication with basal transcription apparatus, etc., but these are often unreported in the literature. Ideally, these more complex mechanisms would be known for each *trans* acting factor in the *cis*-Lexicon. Such biochemical clues would make possible effective data clustering, thus presenting clues for predicting *cis*-regulation. For example, DNA looping between two TFBSs cannot occur at less than a certain minimum distance, while Su(H), a transcription factor activated by signaling, may have a maximum distance from the transcription start site while still being able to direct gene expression of the sea urchin gene, *gcm* (9). Our lexicon is designed to handle more complex fields, but this information is not always available.

Perhaps more *cis*-regulatory information can be derived from the quantitative data obtained by mutating a TFBS (**Fig.** 22.5). Most literature containing data that meet the criteria of the *cis*-Lexicon contains a bar chart quantitatively describing gene expression as a result of mutating each of the TFBSs individually and in combination (18). Gene expression is thus a function of each of the inputs. Gene expression, the output of the function, is the combined effect of each of the individual inputs. This function is not simply the sum of the effect of each individual input; rather, the output depends on the interaction of the inputs. Thus, describing the gene expression requires a more complicated function than summing the effects of mutating each TFBS individually. A generalized mathematical function has been suggested, but applying the function to a broad range of mutational studies is difficult (15).

The *cis*-Lexicon currently does not handle the annotation of the quantitative data from literature. Knowing the relative impact of each TFBS on gene expression is important in properly describing *cis*-regulatory architecture. Thus a format for collecting these data that describe the biology effectively needs to be implemented from quantitative experimental data. Since gene expression can depend on which nucleotides are mutated within a TFBS, there is a relatively low certainty associated with these quantitative data, adding further complexity to the problem.

3.1.7. Examples in the Lexicon

See **Figs.** 22.6 and 22.7.

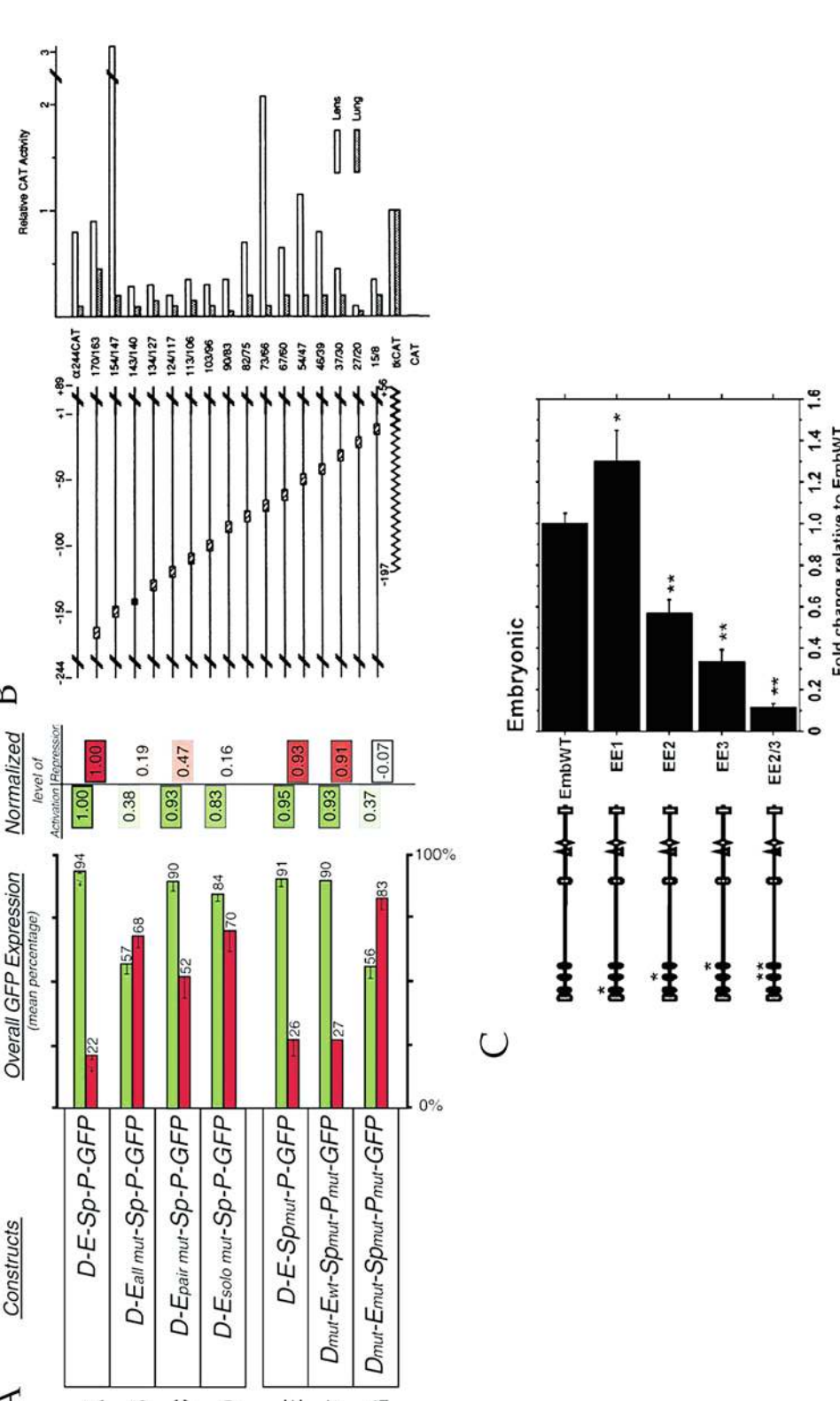

Fig. 22.5. Quintessential diagrams of *cis*-regulatory architecture. **a** sp-gcm in *Strongylocentrotus purpuratus* (9); **b** alphaA-crystallin in *Gallus gallus* (91); **c** *myh3 in Mus musculus* (92).

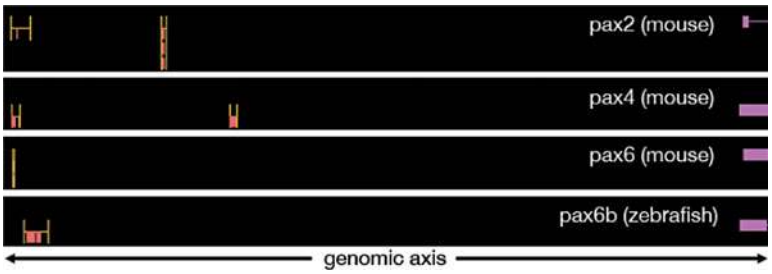

Fig. 22.6. Comparison of four pax genes in the *cis*-Lexicon (axis not to scale). The *purple* square represents the first exon, the *yellow* double-ended bar represents the CRM, and the *orange* blocks represent the TFBS.

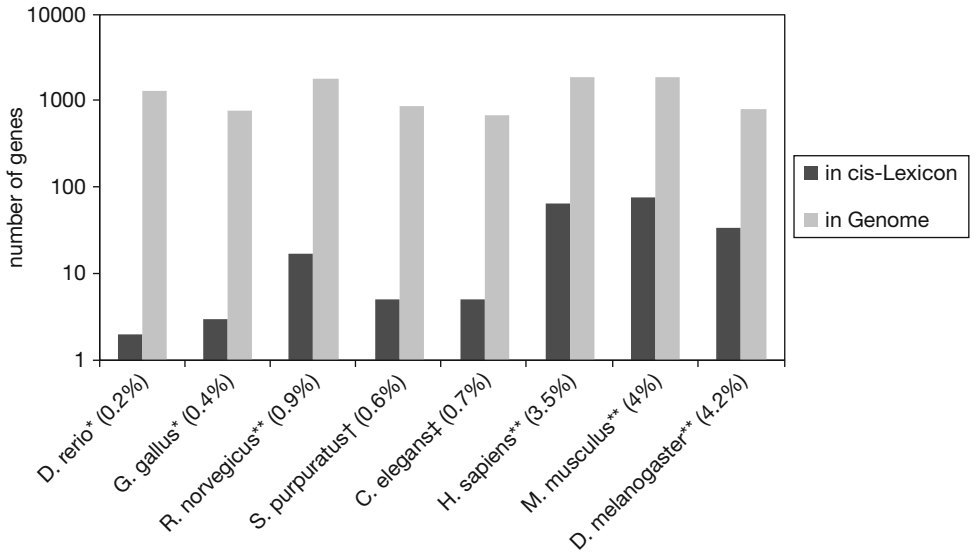

Fig. 22.7. Annotation progress for transcription factor encoding genes. The percent in parentheses is the percentage of transcription factor encoding genes in the corresponding genome that have been annotated in the *cis*-Lexicon. Genome-wide transcription factor encoding gene totals were taken from the following sources: * – DBD (93), ** – Panther (70), † – SpBASE (48), ‡ – (94).

3.2. cis-Browser

3.2.1. Development of the cis-Browser

The CYRENE *cis*-Browser is a genome browser tailored for *cis*-regulatory annotation and investigation. It began as a branch of the Celera Genome Browser, which is available as open source at http://sourceforge.net/projects/celeragb/. The features of the original Celera Genome Browser centered on viewing and annotating gene transcripts, so many new capabilities were added to address our new focus.

First, support for *cis*-regulatory modules (CRMs) and transcription factor binding sites (TFBSs) was added. Each of these new genomic features possesses several unique properties and

associated information. Unlike gene transcripts, whose borders are determined solely by their exons, the boundaries of CRMs can extend beyond the known binding sites contained inside (e.g., if evidenced by sequence conservation). It is often known whether or not whole CRMs or individual binding sites are conserved across species, and this information can be added and viewed via the *cis*-Browser. Each TFBS has a specific factor (or, occasionally, a family of related factors) that binds there. The NCBI GeneID or name for this factor and its effect on gene expression can be annotated and viewed in the *cis*-Browser. Support for new types was added by creating new Java classes. For example, the class CuratedCRM was created to represent CRMs, and the existing class CuratedTranscript (from the Celera Genome Browser) was used as a reference, since transcripts and CRMs share key traits.

The focus of the *cis*-Browser is on annotations that are supplemental to known genes, rather than on discovering transcripts. Therefore, instead of requiring annotators to input the genes themselves, the capability was added to download genes directly from NCBI. Within the *cis*-Browser application, the user can search for genes (*see* **Fig.** 22.8) just as in the NCBI Entrez Gene web site. When a gene is selected from the results, the genomic sequence of the region is downloaded and all the gene's transcripts and exons are automatically displayed. Data are accessed via the NCBI Entrez Programming Utilities service (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html).

In the Celera Genome Browser, properties of genomic entities could be of two types: plain text (e.g., names) or a choice from a list of options (e.g., evidence type: *cis*-mutation, footprinting, etc). Properties could be nested, so that a single (parent) property could contain inside it several additional (child) properties. For *cis*-regulatory annotation, we required accurate recording of complex properties. First, we needed to support properties containing

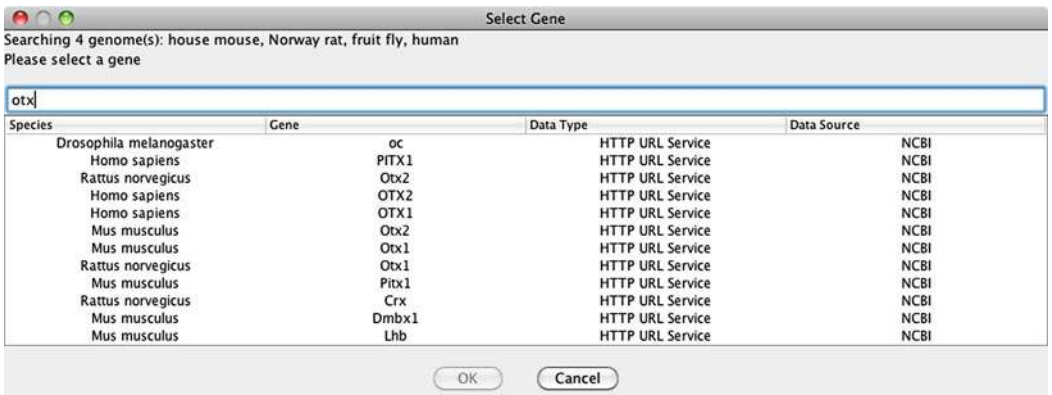

Fig. 22.8. Searching NCBI Entrez Gene within the *cis*-Browser.

multiple interdependent parts: for example, when annotating the factor that binds at a certain site, we must keep track of the factor name, its NCBI GeneID, and any synonyms mentioned in the literature, so that they do not fall out of sync. Second, we needed to support multiple values for a single property: multiple synonyms, multiple *cis*-regulatory functions, and conservation in multiple species.

For properties with multiple parts, we created rich dialog boxes ensuring that the user enters correct information. It would be tedious to ask the user to flip back and forth between the *cis*-Browser and the NCBI Entrez Gene web site to find GeneIDs for each binding factor, and it would be error-prone to make the user type in the factor names and GeneIDs, especially when the same factor binds at several sites for a single target gene. Therefore, the *cis*-Browser provides a special window for annotating the factor that binds to each site (*see* **Fig.** 22.9), so that the user can search the Entrez Gene site from within the browser; the search is automatically restricted to the species being annotated. If the same factor binds at multiple sites, for the second and later sites the user can select the gene from a menu rather than re-entering the information. There are similar windows for annotating conserved species (which searches NCBI for the correct scientific names and NCBI ID) and *cis*-regulatory functions (which ensures that the *cis*-regulatory ontology is followed in naming the regulatory

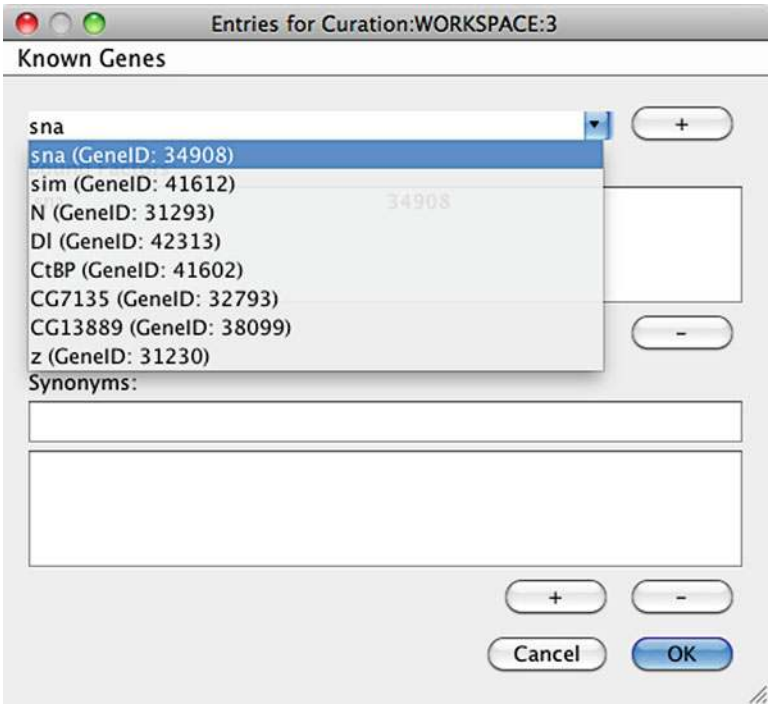

Fig. 22.9. Bound factor annotation with search results.

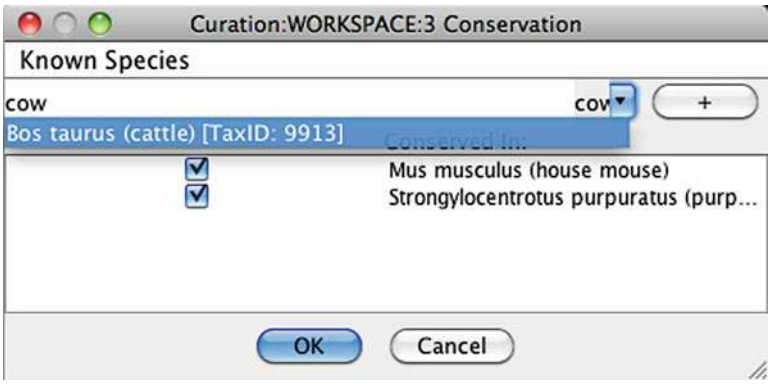

Fig. 22.10. Conserved species annotation with search results.

functions and verifies that PMIDs are typed correctly) – *see* **Figs.** 22.10 and 22.11.

Three different mechanisms were tested to support properties with multiple values; the first two are described in Notes section below. The problem was that in the Celera Genome Browser data model, only one property value can be assigned to a given name. In the GAME XML format supported by the browser, this is represented by the tag `<property name="property-name" value="property-value"/>`. It is not valid to give the same property more than one value; e.g., `<property name="synonym" value="gcm"/><property name="synonym" value="spgcm"/>`. We ultimately decided to represent a set of values for a single property as one property containing multiple child properties, one for each of the desired values, where each child has a unique name, for example `<property name="synonyms" value="gcm, spgcm"><property name="synonym1" value="gcm"/><property name="synonym2" value="spgcm"/></property>` (the child property names do not matter, since only their values are used). The value of the parent property is generally a human-readable summary of the contents, for convenience of display, but again this does not matter, since only the values of the children are used in computations and searches.

The Celera Genome Browser was one part of a three-tiered architecture and communicated with an application server to access a relational database back end. It supported loading genomic features from files, but this was meant to supplement the database (with, for example, output from bioinformatics tools), not replace it. Initially, our *cis*-Lexicon was simply a collection of these XML files. Searching the lexicon required the *cis*-Browser to open, read, and process every one of these files – and this was repeated for every individual search request.

Fig. 22.11. Regulatory function annotation.

At one point we created an ad hoc index as an interim measure to allow certain restricted queries (e.g., list all genes in the lexicon or ask whether a given gene is in the lexicon). Later, we implemented the *cis*-Lexicon as a true database using Apache Derby, an open-source relational database engine. Since Apache Derby is implemented entirely in Java, the *cis*-Browser remains entirely cross-platform. Derby can be run either in embedded mode, where the database is stored and accessed locally, or as a network client, where the database is stored remotely and accessed via a server. This allows the *cis*-Lexicon to be packaged with the *cis*-Browser, for ease of access, or to be stored in one central location, for ease of updating. Relational databases such as Derby support automatic indexing of specific fields in a database record. By indexing the NCBI GeneID field of gene records, for example, searching for genes by GeneID immediately becomes fast. Relational databases also support foreign keys, which ensure that

values intended to reference other entities are actually valid. For example, if the bound factor for a binding site is recorded as gene 373400, then the database either ensures that gene 373400 is present in the *cis*-Lexicon or rejects the annotation.

3.2.2. The CYRENE *cis*-Browser Interface

The *cis*-Browser interface has the same organization as the original Celera Genome Browser. The *cis*-Browser application window is split into four regions (clockwise from top left): the Outline View, the Annotation View, the Subview Container, and the Property Inspector View (*see* **Fig.** 22.12). The Outline View displays in a hierarchical tree format the species, chromosomes, and sequences loaded by the *cis*-Browser and ready for analysis. The Annotation View displays the locations of genomic features (e.g., transcripts, CRMs) on the sequence currently being examined. The Subview Container shows the user a set of views specific to the currently selected feature, and the Property Inspector View shows the properties of the selected feature in textual form.

The Annotation View allows real-time zooming from a chromosome-wide view down to the individual nucleotide level. When the user clicks on a genomic feature, information specific to the feature is visible in the Subview Container and the Property Inspector View. The Annotation View displays genomic features in tiers (horizontal rows grouping features according to their source) so that information from multiple sources is not intermixed and confused; in **Fig.** 22.13, for example, mapped Solexa

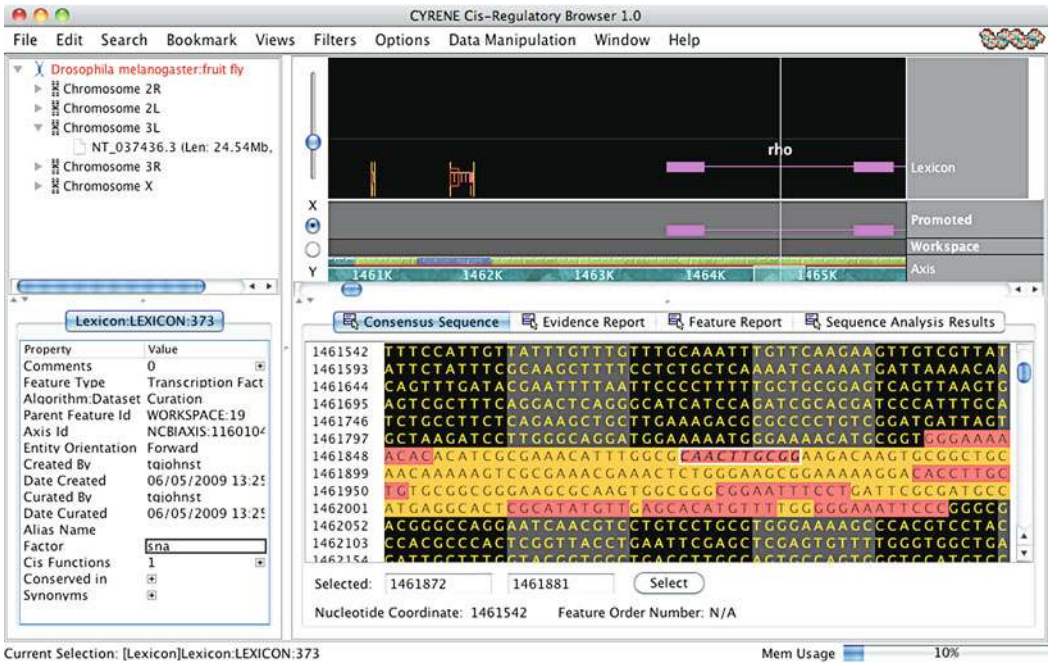

Fig. 22.12. The *cis*-Browser window.

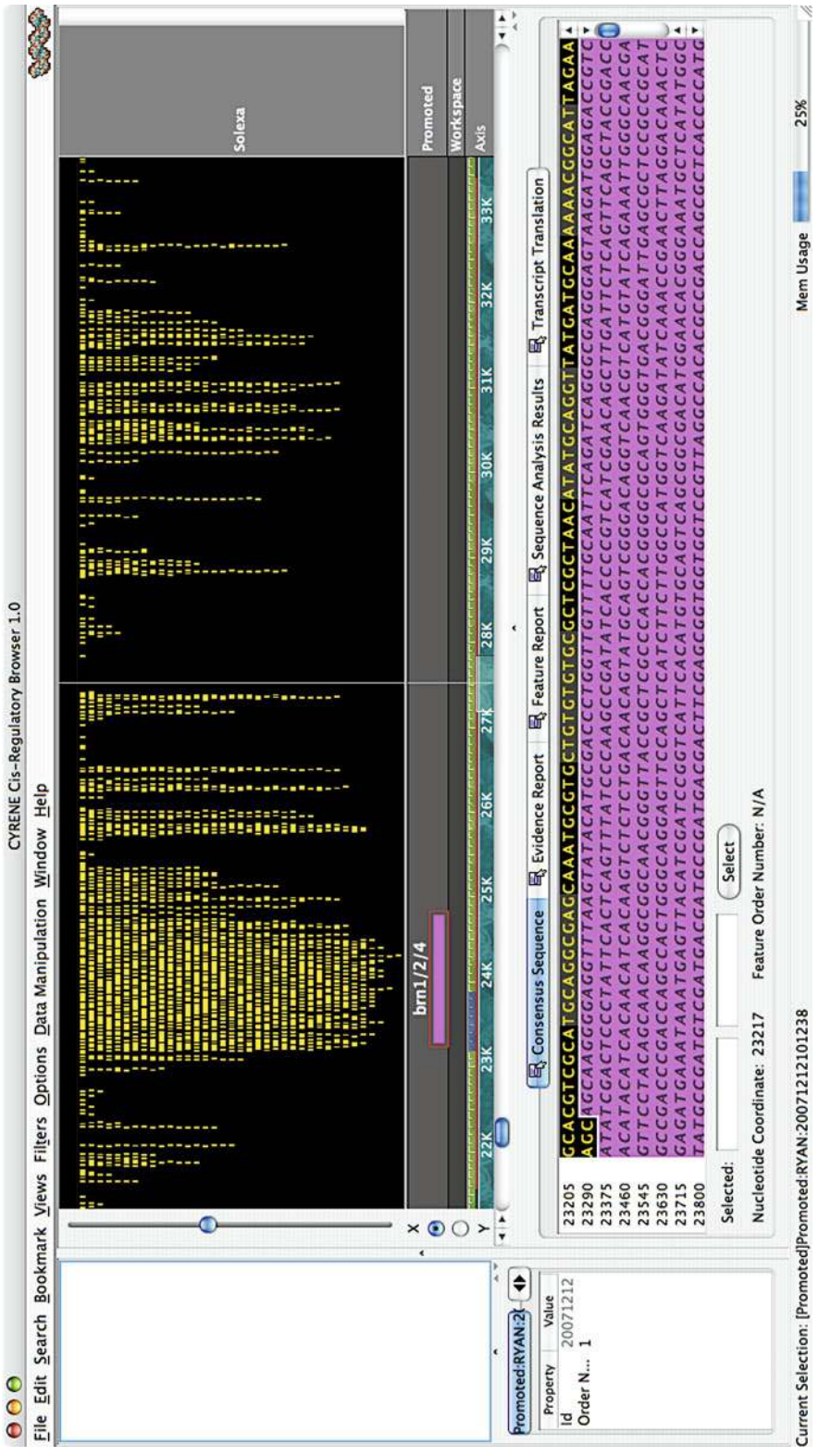

Fig. 22.13. The gene brn1/2/4 appears in the "Promoted" tier while mapped Solexa reads appear in the "Solexa" tier.

reads are grouped separately from genomic transcripts. One special tier in the Annotation View is the workspace, which is the tier that contains genomic features currently being edited. Features loaded from data sources such as XML files or the *cis*-Lexicon are considered immutable, so a copy must be made and placed in the workspace before it can be modified. New features added by the annotator, such as CRMs and binding sites, are always in the workspace.

Each type of genomic feature has particular traits that distinguish it from other types. For example, transcripts are translated into proteins and BLAST hits are the result of comparisons between different sequences. Therefore, viewing the translation of codons into amino acids is relevant only for transcripts, while examining the differences between the current sequence and a sequence that was searched against it is relevant only for BLAST hits. The Subview Container is the location for views such as these. When a feature is selected in the Annotation View, only the views relevant to that type of feature are shown in the Subview Container. Only one such view is shown at a time, to maximize the visible area; the rest are shown as tabs that the user may click on to switch to that view.

Every genomic feature has certain associated properties, such as name, NCBI accession number, or date of curation. The Property Inspector View displays these properties as a two-column table giving the name of each property on the left and the value on the right. Properties can be edited by double-clicking the current value. If the value is a simple string, then it can be edited in place; if it is more complex, such as binding factors, then a dialog box appears. Property changes affecting how the feature appears in the Annotation View are reflected in real time: when the name of a gene or CRM is modified, the new name appears immediately.

One subview (i.e., view appearing in the Subview Container) of critical importance is the Consensus Sequence View, which displays the sequence of the selected feature and the surrounding region and is also used to mark the location of new features. The user simply clicks and drags to select a sequence. Right-clicking shows a menu with options to create a transcript, CRM, or TFBS. The seqFinder feature quickly locates the exact coordinates of a sequence appearing in a published paper. Given a region of sequence to search within (e.g., a gene and its flanking sequence), the seqFinder lets the user type in only the minimum number of nucleotides to uniquely find the paper's sequence. For each letter the user types, the seqFinder tells whether the sequence typed so far is found more than once (i.e., multiple ambiguous matches, so more input is necessary), exactly once (i.e., a perfect match; no more typing needed), or never (i.e., a typo or possibly a true mismatch between the paper's sequence and the reference genome).

The user need only type a few letters from the beginning and then from the end to uniquely identify the entire sequence. When the start and end coordinates are known, the seqFinder automatically selects the sequence within the Consensus Sequence View. Typing the minimum sequence necessary lets the user locate the precise coordinates quickly yet accurately.

3.2.3. Annotation with the cis-Browser

To enter a genomic feature, the annotators first input the coordinates by locating them with the seqFinder. The relevant properties such as names, binding factors, *cis*-regulatory functions, and sequence conservation are set via the Property Inspector View. The Annotation View lets one do quick sanity checks – are the binding sites located upstream, downstream, or within introns of the regulated gene, as is usually the case? Are the CRMs of a reasonable size?

The annotators' work is saved as XML files in the GAME format, rather than directly input into the *cis*-Lexicon. This allows easy backup and sharing of past work and also prevents cluttering the database with half-finished or faulty annotations. A special software tool is required to move the annotations from these intermediate files into the *cis*-Lexicon; forcing the use of intermediate files and preventing unauthorized annotators from modifying the *cis*-Lexicon directly lets us keep the database at a strict high quality. An experienced annotator can verify the work of a trainee before it is entered into the *cis*-Lexicon.

3.3. Virtual Sea Urchin

We have worked closely with the Davidson laboratory at CalTech to produce a Virtual Sea Urchin prototype. The VSU uses spatial models and a graphics engine to simulate the four-dimensional sea urchin embryo, allowing the researcher to probe the GRN at levels of granularity from the multicellular embryo to the gene-regulatory network of an individual cell type. The embryo models were created by extrapolating to three dimensions cross-sectional color-coded tracings from photomicrographs (17).

The Virtual Sea Urchin currently provides models for the *Strongylocentrotus purpuratus* embryo at 6, 10, 15, 20, and 24 h. Cell types are defined by ambient and diffuse coloring as well as shape. Gene expression data are visible at a glance on an embryonic cell type using emission coloring (intensity of coloring is proportional to intensity of expression).

The VSU model of embryonic development will eventually be configurable, featuring realistic cell models and dynamics simulators. *In toto* imaging (56, 57) of the sea urchin embryo will enhance the model's accuracy and resolution, letting researchers probe the regulatory network activity per cell. We will ultimately combine the *cis*-regulatory sequence-analysis capabilities of CYRENE and the network building, visualization, and simulation capabilities of BioTapestry with the temporal and spatial

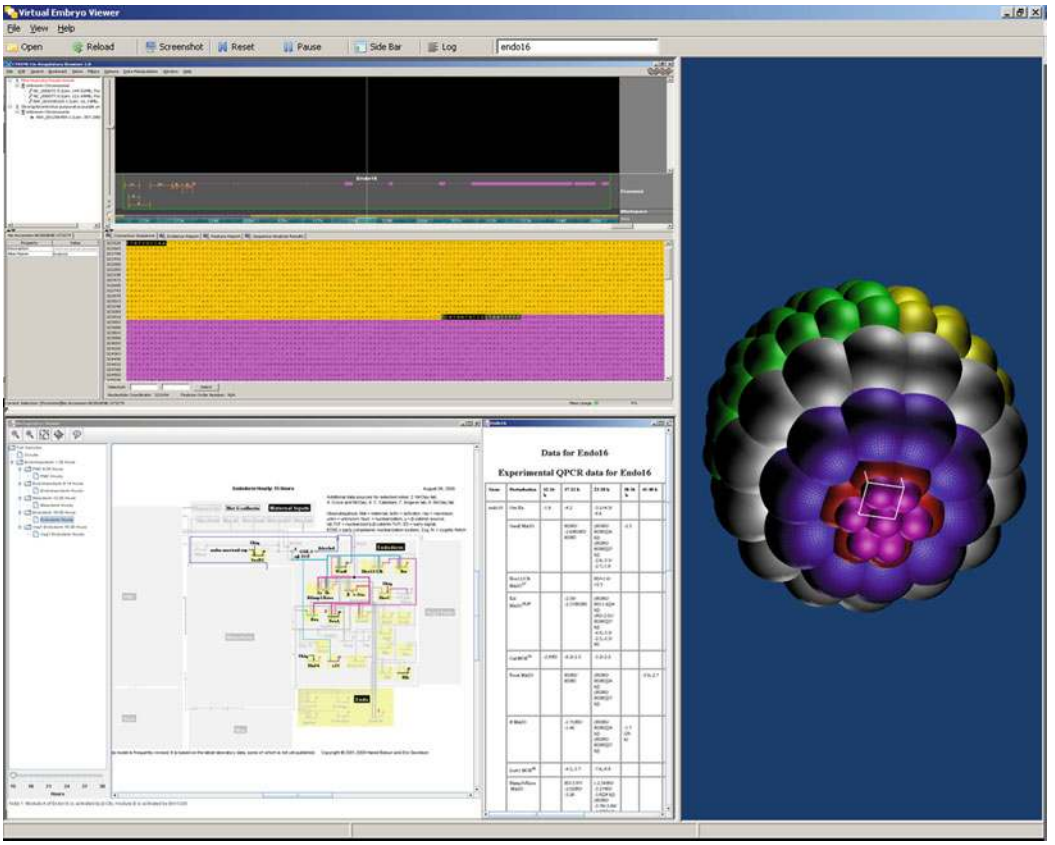

Fig. 22.14. Virtual Sea Urchin, BioTapestry, and Cyrene integration. In order to model the gene regulatory network of the developing sea urchin embryo completely, tools that specialize in analyzing different perspectives of the network must be integrated. BioTapestry operates at the network level, allowing users to manipulate and simulate network interactions. Cyrene operates at the DNA level, providing *cis*-regulatory module and binding site definition and other DNA sequence analysis. The Virtual Sea Urchin maps regulatory network information into space and time. The interoperability among these three views is a central component of future work.

analysis of the 4D Virtual Sea Urchin to yield a complete characterization of the GRN (*see* **Fig.** 22.14).

4. Notes

1. **Gene naming**
   Creating a controlled vocabulary starts with the name of the gene, as discussed previously (18), although there has been a large effort to establish common names across species (95). We would like to know the primary ortholog of some gene we have annotated to observe the similarities of *cis*-regulatory architecture across species. Our efforts toward

Table 22.1
Examples for gene name translation between mouse and *Drosophila*

Mouse	Drosophila	References
Beta-catenin	Armadillo	van Noort et al. (2002) (103)
Cux-1	Cut	Sharma et al. (2004) (104)
TLE-4	Groucho	Sharma et al. (2004) (104)
six3, six6	So	López-Ríos et al. (2003) (105)

such a gene name translation table are only in the nascent stages. Examples for gene name translation between mouse and *Drosophila* are shown in **Table** 22.1. A gene occurring in two species is more likely to have a similar *cis*-regulatory architecture if the two genes are related by evolution *and* have a conserved function. Sets of genes meeting these criteria we have termed *Davidson Orthologs*. The method usually employed for determining homology is by sequence rather than conserved function. *orthoMCL* (96) and *inparanoid* (97) are examples of this kind of ortholog table. While the latter definition of ortholog is more easily searched in a database of genes automatically, the definition is not as stringent.

2. **CRM boundaries**
TFBSs are usually well described and require no guesswork by the annotator, but CRMs are often not well defined in the literature. Some of this confusion stems from the lack of a precise definition of the function of a CRM; additionally, many research groups are not interested in finding the boundaries of CRMs and limit their scope to discovery of TFBSs. When annotating the *cis*-regulatory architecture of a gene, the annotator often must make certain assumptions about the boundaries of a CRM that can be classified as follows (examples are referenced): (1) CRMs are not discussed in the literature and the annotator defines the CRM by the minimal sequence that correctly directs gene expression, usually approximated to within 100 bp (96). (2) CRMs are not discussed in the literature, but a graph in the paper shows sequence similarity to the same gene in other species (98). The annotator defines the CRM as the sequence most conserved in other species. If a sequence remains highly similar over a great evolutionary distance, there must be selective pressure to conserve the sequence, and therefore the sequence probably plays an important role in the organism. (3) CRMs are not discussed in the literature and the

annotator defines the CRM by most extreme TFBSs determined to act in a specific location (99). That is, if three TFBSs are found to drive gene expression in a cell, the CRM annotation goes from the first nucleotide of the first TFBS to the last nucleotide of the last TFBS. Overall uncertainty and lack of consistency in CRM annotation reduce the quality of data on CRM boundaries.

3. **Caveats in Davidson criteria**
While the *cis*-Lexicon seeks to collect only the most reliable data, many uncertainties remain. Mutational studies show that a certain TFBS is important for correct gene expression, but the factor that binds to the site is not immediately certain. The sequence probably contains a transcription factor binding motif, so the factor that binds can often be guessed. In some annotations, the experimentalist has shown that a particular transcription factor binds to the TFBS in an assay (100) or that knockdown of the transcription factor also causes a change in gene expression. Such confirmations of *trans* acting factor are not always reported in the literature.

4. **Uncertainty in identifying the *tran*s acting factor**
Often in the literature the exact transcription factor binding to a TFBS is not known, but the transcription factor family to which the *trans* acting factor belongs is reported; Shen and Ingraham (101) report an E-Box *trans* activator. Sometimes the authors report that the *trans* acting factor could be one of several; for instance, Clark et al. (102) report that either an RAR/RAR or RAR/RXR dimer activates transcription. The TFBS cannot be compared to others in the *cis*-Lexicon since the *trans* acting factor is not known, and the TFBS loses its usefulness in the data set for clustering and prediction. The transcription factor hierarchy described above (**Fig.** 22.2) was added as an annotation tool to combat this problem. *Trans* acting factors can be clustered at different levels of the transcription factor hierarchy.

5. ***cis*-Lexicon search engine**
A great challenge in the annotation process has been finding literature relevant to building the *cis*-Lexicon. So far the literature has been located by PubMed or Google Scholar searches or by browsing references describing previously annotated genes (**Fig**. 22.15 shows journals cited in the *cis*-Lexicon). A formalized search process will rapidly uncover relevant literature; in addition, it will help determine the number of genes studied according to the *Davidson criteria* and give an estimate of *cis*-Lexicon completeness. When the *cis*-Lexicon is declared complete, searches will have to be performed continually to find new data. To accomplish these goals, the CLOSE Project (*cis*-Lexicon Ontology Search

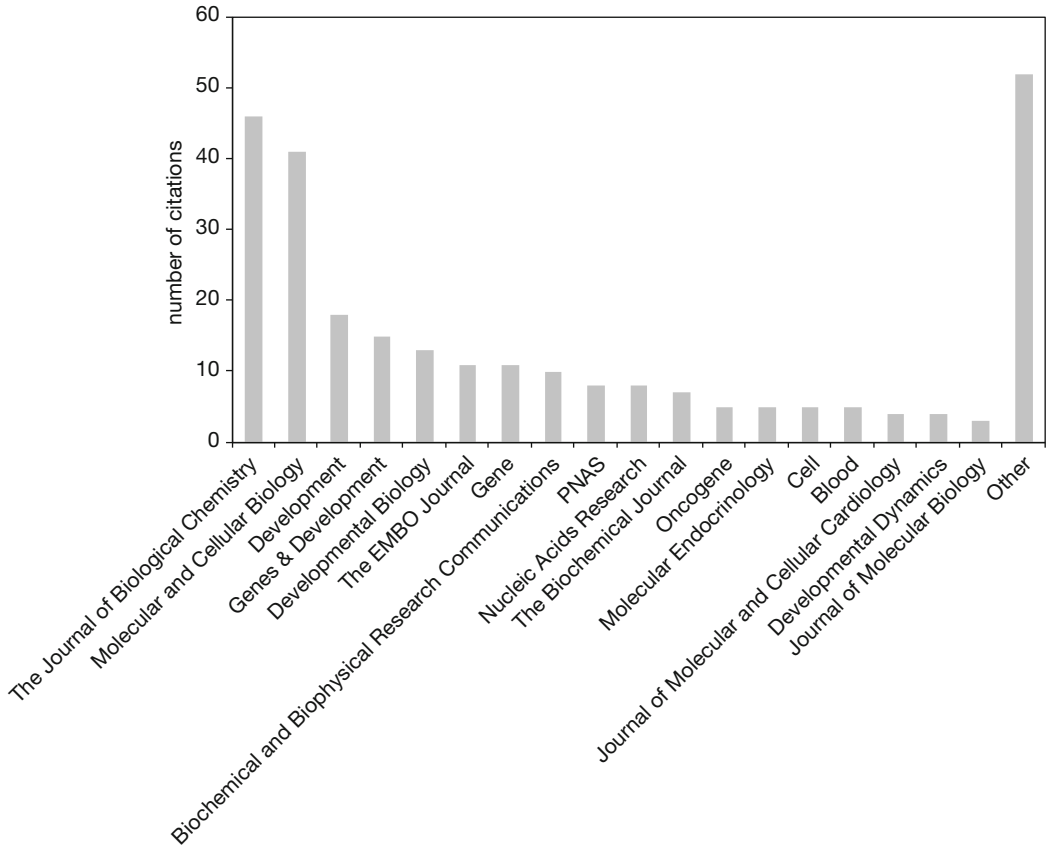

Fig. 22.15. Journals referenced in the *cis*-Lexicon.

Engine) aims to create a set of algorithmic strategies for literature extraction of *cis*-regulation articles.

PubMed cannot perform unrestricted phrase searching of citations and abstracts; only phrases in the PubMed Index are found. If a phrase is not in the Index, then a PubMed search cannot return exact matches, even if it appears in citations or abstracts. Instead, the query is treated as a standard non-phrase search, yielding almost entirely irrelevant results (http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_450.html). This prevents the use of PubMed for advanced text-based queries. MeSH terms are not assigned consistently enough to make possible comprehensive searches by keyword alone (*see*, for example, Yuh et al. (13), which is not assigned a term for Base Sequence). Google does not yet offer an API to access Google Scholar, and the results of searching PubMed via standard search engines (Google, Yahoo, etc.) lack known important papers (unpublished

data). All of these point to the need for a specialized search engine to find the most relevant papers.

6. **Complex annotations in the *cis*-Browser**
   Three different mechanisms for making complex annotations with multiple values were tested. The first was to extend the GAME XML format with a new tag for each new type of data. This method was explored first because supporting multiple complex values appeared analogous to storing annotator comments, which was already supported by the Celera Genome Browser. According to the Genome Browser data model, comments are completely distinct from properties. In the XML format, each comment was stored in a `<comment>` tag containing custom author and date attributes. It therefore seemed reasonable to add a `<cisreg_function>` tag with function location and time attributes. Multiple `<cis_reg_function>` tags can be associated with a single genomic entity, allowing multiple functions to be annotated. This required significant work in modifying not only the XML parser itself but all of the loading code governing the interactions between the XML parser (and other possible storage implementations) and the browser. A more general solution was deemed necessary with the prospect of additional complex annotations, since it would be infeasible to make similar changes multiple times.
   The second mechanism was the addition of "facts," a new data type differing from properties in that multiple facts can be associated with the same name. Custom attributes could be replaced by child facts. This led to a straightforward conversion of `<cisreg_function>` tags to nested facts: `<cisreg_function function="booster" location="mesoderm" time="24 h"/>` became `<fact name="cisreg_function" value = "booster"><fact name = "location" value="mesoderm"/><fact name="time" value ="24 h"/></fact>`. New facts could be added to store interspecies conservation and bound factors without modifying the parser or loading code. However, the format was still not standard GAME XML, and therefore could not be read by the Celera Genome Browser (which, being open source, is still under development) or other applications using the GAME format.
   The desire to keep compatibility with the Genome Browser led us to consider the third option of using only nested properties, described in **Section** 3.

7. **Java XML parsing**
   Two significant issues arose in XML parsing with Java, the first concerning external DTD loading. XML data received

from the NCBI Entrez Programming Utilities service always refer to DTDs located on NCBI's servers. The XML language requires parsers always to access the DTD, and the extra time required to download the DTD whenever such an XML file is parsed led to difficult-to-trace slowdowns in the *cis*-Browser application. The correct solution was a custom org.xml.sax.EntityResolver implementation returning cached copies of NCBI files. Once given to an XMLReader object, it will utilize the cached copies rather than fetching the remote originals.

Second, computers with an old version of Java 6 may run out of memory when parsing large XML files, even when using efficient parsing methods (see http://bugs.sun.com/bugdatabase/view_bug.do?bug_id=6536111). The easiest solutions are to use Java 5 if available or to upgrade to a more recent version of Java 6.

8. **Virtual Sea Urchin embryo modeling**
   Producing the three-dimensional models for the Virtual Sea Urchin's 3D graphics engine is currently laborious; an embryology domain expert works with a 3D computer graphics and modeling software expert to create the embryonic models using a software suite such as Maya or Blender. These models are then exported and integrated into a format compatible with the VSU.

   We can streamline this process by animating the anatomical structure using a hierarchy of cells and cell types that completely categorize the embryo at all relevant time slices. This hierarchical tree representation, in which each tree level defines the embryo at a specific time and which is defined by the experimentalist, can be visualized by meiotically splitting cells into appropriate cell types. With this new embryo representation, experimentalists can interact directly with the VSU and easily define embryo development without needing outside expertise.

Acknowledgments

The support of the National Science Foundation under grant DBI 0645955 is acknowledged with gratitude. We would also like to acknowledge the tremendous impact on this work of our collaborator, Eric H. Davidson of the California Institute of Technology, who has guided every step of our efforts. This work would not have been possible without the contributions of three generations

of annotators, most notably Tim Johnstone, Jake Halpert, and David Moskowitz. (The first generation was David Moskowitz, Rohan Madamsetti, and Sanjay Trehan; the second generation was Tamar Melman, Mark Grabiner, and Kyle Schutter; the third generation is Tim Johnstone, Jake Halpert, Mei Cao, Kenneth Estrellas, Nicole Noronha, and Daniel Yang.) We would also like to thank Andy Ransick, Andy Cameron and Russell Turner for many discussions and valuable suggestions. Last but not least, many thanks go to Erin Klopfenstein for her outstanding work and many valuable contributions to the CYRENE Project.

References

1. Davidson, E.H. (2001) Genomic regulatory systems: In *Devel and evol,* Academic Press, San Diego, CA.
2. Davidson, E.H., and Erwin, D. (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311, 796–800.
3. Davidson, E.H. (1968) *Gene activity in early development.* Academic Press, New York, NY.
4. Sea Urchin Genome Consortium. (2006) The genome of the sea urchin *Strongylocentrotus purpuratus. Science* 314, 941–952.
5. Samanta, M.P., Tongprasit, W., Istrail, S. et al. (2006) The transcriptome of the sea urchin embryo. *Science* 314, 960–962.
6. Erwin, D.H., and Davidson, E.H. (2009) The evolution of hierarchical gene regulatory networks. *Nature Rev Gen* 10, 141–148.
7. Davidson, E.H., Rast, J.P., Oliveri, P. et al. (2002) A genomic regulatory network for development. *Science* 295, 1669–1678.
8. Britten, R.J., and Davidson, E.H. (1969) Gene regulation for higher cells: a theory. *Science* 165, 349–357.
9. Ransick, A., and Davidson, E. (2006) cis-regulatory processing of Notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. *Dev Biol* 297, 587–602.
10. Oliveri, P., Tu, Q., and Davidson, E.H. (2008) Global regulatory logic for specification of an embryonic cell lineage. *Proc Natl Acad Sci USA* 105, 5955–5962.
11. Yuh, C.H., and Davidson, E.H. (1996) Modular cis-regulatory organization of Endo16, a gut-specific gene of the sea urchin embryo. *Development,* 122, 1069–1082.
12. Yuh, C.H., Bolouri, H., and Davidson, E.H. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279, 1896–1902.
13. Yuh, C.H., Dorman, E.R., Howard, M.L. et al. (2004) An otx cis-regulatory module: a key node in the sea urchin endomesoderm gene regulatory network. *Dev Biol* 269, 536–551.
14. Istrail, S., De-Leon, S.-T., and Davidson, E. (2007) The regulatory genome and the computer. *Dev Biol* 310, 187–195.
15. Istrail, S., and Davidson, E. (2005) Logic functions of the genomic cis-regulatory code 2005. *Proc Natl Acad Sci USA* 102, 4954–4959.
16. Levine, M., and Davidson, E.H. (2005) Gene regulatory networks for development. *Proc Natl Acad Sci USA* 102, 4936–4942.
17. Davidson, E.H. (2006) *The regulatory genome: gene regulatory networks in development and.* Academic Press, San Diego, CA.
18. Tarpine, R., and Istrail, S. (2009) On the concept of Cis-regulatory information: from sequence motifs to logic functions. *Algorithmic Bioprocesses* In (Condon, A., Harel, D., Kok, J.N., Salomaa, A., and Winfree, E. Eds.) pp. 731–742 Springer-Verlag, Berlin Heidelberg.
19. Longabaugh, W.J.R., Davidson, E.H., and Bolouri, H. (2005) Computational representation of developmental genetic regulatory networks. *Dev Biol* 283, 1–16.
20. Longabaugh, W.J.R., Davidson, E.H., and Bolouri, H. (2009) Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. *Biochem Biophys Acta* 1789, 363–374.
21. Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23.
22. Wasserman, W.W., and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev Gen* 5, 276–287.
23. Sandelin, A. (2004) In silico prediction of *cis*-regulatory elements. Karolinska Institutet. Stockholm, Sweden, 4–130.

24. Tompa, M., Li, N., Bailey, T.L. et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnol* 23, 137–144.
25. Hannenhalli, S., and Levy, S. (2001) Promoter prediction in the human genome. *Bioinformatics* 17, S90–S96.
26. Hannenhalli, S., and Levy, S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res* 30, 1–8.
27. Hannenhalli, S., Putt, M.E., Gilmore, J.M. et al. (2006) Transcriptional genomics associates FOX transcription factors with human heart failure. *Circulation J Am Heart Assoc* 114, 1269–1276.
28. Singh, L.N., Wang, L.S., and Hannenhalli, S. (2007) TREMOR—a tool for retrieving transcriptional modules by incorporating motif covariance. *Nucleic Acids Res* 35, 7360–7371.
29. Markstein, M., Markstein, P., Markstein, V. et al. (2002) Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Dev Biol* 99, 763–768.
30. Linhart, C., Halperin, Y., and Shamir, R. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res* 18, 1180–1189.
31. Tompa, M. (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *7th International Conference Intelligent Systems for Molecular Biology*, 262–271.
32. Sinha, S., and Tompa, M. (2003) Performance comparison of algorithms for finding transcription factor binding sites. *Proceedings of the 3rd IEEE Symposium on Bioinformatics and Bioengineering*, 213.
33. Blanchette, M., Schwikowski, B., and Tompa, M. (2002) Algorithms for phylogenetic footprinting. *J Comput Biol* 9, 211–223.
34. Wasserman, W.W., and Fickett, J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 278, 167–181.
35. Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* 30, 4442–4451.
36. Keich, U., and Pevzner, P.A. (2002) Subtle motifs: defining the limits of motif finding algorithms. *Bioinformatics* 18, 1382–1390.
37. Ng, P., Nagarajan, N., Jones, N. et al. (2006) Apples to apples: improving the performance motif finders and their significance analysis in the Twilight Zone. *Bioinformatics* 22, e393–e401.
38. Badis, G., Berger, M., Philippakis, A. et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–1723.
39. Berger, M., Badis, G., Gehrke, A. et al. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133, 1266–1276.
40. Noyes, M.B., Christensen, R.G., Wakabayashi, A. et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* 133, 1277–1289.
41. Cameron, R.A., Rast, J.P., and Brown, C.T. (2004) Genomic resources for the study of sea urchin development. *Methods Cell Biol* 74, 733–757.
42. He, X., Ling, X., and Sinha, S. (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. *PLoS Comput Biol* 5, e100299.
43. Li, N., and Tompa, M. (2006) Analysis of computational approaches for motif discovery. *Algorithms Mol Biol* 1, 1–8.
44. Li, X., Zhong, S., and Wong, W.H. (2005) Reliable prediction of transcription factor binding sites by phylogenetic verification. *Proc Natl Acad Sci USA* 102, 16945–16950.
45. Papatsenko, D., and Levine, M. (2005) Quantitative analysis of binding motifs mediating diverse spatial readouts of the dorsal gradient in the *Drosophila* embryo. *Proc Natl Acad Sci USA* 102, 4966–4971.
46. Pilpel, Y., Sudarsana, P., and Church, G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet* 29, 153–159.
47. Zhu, Z., Pilpel, Y., and Churge, G.M. (2002) Computational identification of transcription factor binding sites *via* a transcription-factor-centric clustering (TFCC) algorithm. *J Mol Biol* 318, 71–81.
48. Howard-Ashby, M., Materna, S.C., Brown, C.T. et al. (2006) Identification and characterization of homeobox transcription factor genes in *Strongylocentrotus purpuratus*, and their expression in embryonic development. *Dev Biol* 300,74–89.
49. Oliveri, P., Carrick, D.M., and Davidson, E.H. (2002) A regulatory gene network that directs micromere specification in the sea urchin embryo. *Dev Biol* 246, 209–228.
50. Calestani, C., Rast, J.P., and Davidson, E.H. (2003) Isolation of pigment cell specific genes in the sea urchin embryo by differen-

tial macroarray screening. *Development* 130, 4587–4596.
51. Imai, K.S., Levine, M., Satoh, N. et al. (2006) Regulatory blueprint for a chordate embryo. *Science* 312, 1183–1187.
52. Ransick, A., Rast, J.P., Minokawa, T. et al. (2002) New early zygotic regulators expressed in endomesoderm of sea urchin embryos discovered by differential array hybridization. *Dev Biol* 246, 132–147.
53. Stathopoulos, A., Van Drenth, M., Erives, A. et al. (2002) Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* 111, 687–701.
54. Revilla-i-Domingo, R., Minokawa, T., and Davidson, E.H. (2004) R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of SpDelta in micromeres. *Dev Biol* 274, 438–451.
55. Lickert, H., and Kemler, R. (2002) Functional analysis of cis-regulatory elements controlling initiation and maintenance of early Cdx1 gene expression in the mouse. *Dev Dyn* 225, 216–220.
56. Megason, S., and Fraser, S. (2003) Digitizing life at the level of the cell: high-performance laser-scanning microscopy and image analysis for in toto imaging of development. *Mech Dev* 120, 1407–1420.
57. Megason, S., and Fraser, S. (2007) Imaging in systems biology. *Cell* 130, 784–795.
58. Turner, R., Chaturvedi, K., Edwards, N. et al. (2001) Visualization challenges for a new cyberpharmaceutical computing paradigm, *Proceedings of the Symposium on Large-Data Visualization and Graphics*, San Diego, CA.
59. Gray, S., Szymanski, P., and Levine, M. (1994) Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes Dev* 8(15), 1829–1838.
60. Courey, A., and Jia, S. (2001) Transcriptional repression: the long and the short of it. *Genes Dev* 15, 2786–2796.
61. Nakao, T., and Ishizawa, A. (1994) Development of the spinal nerves in the mouse with special reference to innervation of the axial musculature. *Anat Embryol* 189, 115–138.
62. Latchman, D. (2008) *Eukaryotic transcription factors*. Fifth Edition, Academic Press, London.
63. Barolo, S., and Posakony, J. (2002) Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev* 16 , 1167–1181.
64. Zeller, R., Griffith, J., Moore, J. et al. (1995) A multimerizing transcription factor of sea urchin embryos capable of looping DNA. *Proc Natl Acad Sci USA* 92, 2989–2993.
65. Smith, J., and Davidson, E. (2008) A new method, using cis-regulatory control, for blocking embryonic gene expression. *Dev Biol* 318, 360–365.
66. West, A., Gaszner, M., and Felsenfeld, G. (2002) Insulators: many functions, many mechanisms. *Genes Dev* 16, 271–288.
67. Inagaki, N., Maekawa, T., Sudo, T. et al. (1992) c-Jun represses the human insulin promoter activity that depends on multiple cAMP response elements. *Proc Natl Acad Sci* 89, 1045–1049.
68. Matys, V., Kel-Margoulis, O., Fricke, E. et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34, D108–D110.
69. Ashburner, M., Ball, C., Blake, J. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25–29.
70. Mi, H., Guo, N., Kejariwal, A. et al. (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res* 35, D247–D252.
71. Gottschalk, L., Giannola, D., and Emerson, S. (1993) Molecular regulation of the human IL-3 gene: inducible T cell-restricted expression requires intact AP-1 and Elf-1 nuclear protein binding sites. *J Exp Med* 178, 1681–1692.
72. Regl, G., Kasper, M., Schnidar, H. et al. (2004) Activation of the BCL2 promoter in response to Hedgehog/GLI signal transduction is predominantly mediated by GLI2. *Cancer Res* 64, 7724–7731.
73. Wheeler, M., Snyder, E., Patterson, M. et al. (1999) An E-box within the MHC IIB gene is bound by MyoD and is required for gene expression in fast muscle. *Am J Physiol* 276, C1069–C1078.
74. Sekido, R., Murai, K., Funahashi, J. et al. (1994) The delta-crystallin enhancer-binding protein delta EF1 is a repressor of E2-box-mediated gene activation. *Mol Cell Bio* 14, 5692–5700.
75. Castro, B., Barolo, S., Bailey, A. et al. (2005) Lateral inhibition in proneural clusters: cis-regulatory logic and default repression by suppressor of hairless. *Development* 132, 3333–3344.
76. Penkov, D., Tanaka, S., Di Rocco, G. et al. (2000) Cooperative interactions between PBX, PREP, and HOX proteins modulate the activity of the alpha 2(V) colla-

gen (COL5A2) promoter. *J Biol Chem* 275, 16681–16689.
77. Zelko, I., Mueller, M., and Folz, R. (2008) Transcription factors sp1 and sp3 regulate expression of human extracellular superoxide dismutase in lung fibroblasts. *Am J Respir Cell Mol Biol* 39, 243–251.
78. Murphy, A., Lee, T., Andrews, C. et al. (1995) The breathless FGF receptor homolog, a downstream target of Drosophila C/EBP in the developmental control of cell migration. *Development* 121, 2255–2263.
79. Ohshiro, T., and Saigo, K. (1997) Transcriptional regulation of breathless FGF receptor gene by binding of TRACHEALESS/dARNT heterodimers to three central midline elements in *Drosophila* developing trachea. *Development* 124, 3975–3986.
80. Christensen, M., Zhou, W., Qing, H. et al. (2004). Transcriptional regulation of BACE1, the beta-amyloid precursor protein beta-secretase, by Sp1. *Mol Cell Biol* 24, 865–874.
81. Pan, D., Huang, J., and Courey, A. (1991) Functional analysis of the *Drosophila* twist promoter reveals a dorsal-binding ventral activator region. *Genes Dev* 5, 1892–1901.
82. Thisse, C., Perrin-Schmitt, F., Stoetzel, C. et al. (1991) Sequence-specific transactivation of the *Drosophila* twist gene by the dorsal gene product. *Cell* 65, 1191–1201.
83. Jiang, J., Kosman, D., Ip, Y. et al. (1991) The dorsal morphogen gradient regulates the mesoderm determinant twist in early *Drosophila* embryos. *Genes Dev* 5, 1881–1891.
84. Akimaru, H., Hou, D., and Ishii, S. (1997) *Drosophila* CBP is required for dorsal-dependent twist gene expression. *Nature Genet* 17, 211–214.
85. Doerksen, L., Bhattacharya, A., Kannan, P. et al. (1996) Functional interaction between a RARE and an AP-2 binding site in the regulation of the human HOX A4 gene promoter. *Nucleic Acids Res* 24, 2849–2856.
86. Sasaki, H., Hui, C., Nakafuku, M. et al. (1997) A binding site for Gli proteins is essential for HNF-3beta floor plate enhancer activity in transgenics and can respond to Shh in vitro. *Development* 124, 1313–1322.
87. Yoon, J., Kita, Y., Frank, D. et al. (2002) Gene expression profiling leads to identification of GLI1-binding elements in target genes and a role for multiple downstream pathways in GLI1-induced cell transformation. *J Biol Chem* 277, 5548–5555.
88. Sokol, N., and Ambros, V. (2005) Mesodermally expressed *Drosophila* microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes Dev* 19, 2343–2354.
89. Ho, I., Hodge, M., Rooney, J. et al. (1996) The proto-oncogene c-maf is responsible for tissue-specific expression of interleukin-4. *Cell* 85, 973–983.
90. Kajihara, M., Sone, H., Amemiya, M. et al. (2003) Mouse MafA, homologue of zebrafish somite Maf 1, contributes to the specific transcriptional activity through the insulin promoter. *Biochem Biophys Res Commun* 312, 831–842.
91. Matsuo, I., and Yasuda, K. (1992) The cooperative interaction between two motifs of an enhancer element of the chicken alpha A-crystallin gene, alpha CE1 and alpha CE2, confers lens-specific expression. *Nucleic Acids Res* 20, 3701–3712.
92. Belkin, D., Allen, D., and Leinwand, L. (2006) MyoD, Myf5, and the calcineurin pathway activate the developmental myosin heavy chain genes. *Dev Biol* 294, 541–553.
93. Wilson, D., Charoensawan, V., Kummerfeld, S., et al. (2008) DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* 36, D88–D92.
94. Haerty, W., Artieri, C., Khezri, N. et al. (2008) Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution. *BMC Genomics* 9, 399.
95. Bult, C., Eppig, J., Kadin, J. et al. (2008) The mouse genome database (MGD): mouse biology and model systems. *Nucleic Acids Res* 36, D724–D728.
96. Chen, F., Mackey, A., Stoeckert, C. et al. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34, D363–D368.
97. Berglund, A.-C., Sjölund, E., Ostlund, G. et al. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* 36(database issue), D263–D266.
98. Delporte, F., Pasque, V., Devos, N. et al. (2008) Expression of zebrafish pax6b in pancreas is regulated by two enhancers containing highly conserved cis-elements bound by PDX1, PBX and PREP factors. *BMC Dev Biol* 8, 53.
99. Warren, D., Simpkins, C., Cooper, M. et al. (2005) Modulating alloimmune responses with plasmapheresis and IVIG. *Curr Drug Targets Cardiovasc Haematol Disord* 5, 215–222.
100. Annicotte, J.-S., Fayard, E., Swift, G. et al. (2003) Pancreatic-duodenal homeobox 1 regulates expression of liver recep-

tor homolog 1 during pancreas development. *Mol Cell Biol* 23, 6713–6724.
101. Shen, J.-C., and Ingraham, H. (2002) Regulation of the orphan nuclear receptor steroidogenic factor 1 by Sox proteins. *Mol Endocrinol* (Baltimore, MD) 16, 529–540.
102. Clark, A., Wilson, M., London, N. et al. (1995) Identification and characterization of a functional retinoic acid/thyroid hormone-response element upstream of the human insulin gene enhancer. *Biochem J* 309, 863–870.
103. van Noort, M., van de Wetering, M., and Clevers, H. (2002) Identification of two novel regulated serines in the N terminus of beta-catenin. *Exp Cell Res* 276, 264–72.
104. Sharma, M., Fopma, A., Brantley, et al. (2004) Coexpression of Cux-1 and notch signaling pathway components during kidney development. *Dev Dyn* 231(4), 828–838.
105. López-Ríos, J., Tessmar, K., Loosli F. et al. (2003) Six3 and Six6 is moduated by members of the groucho family. *Development* 130, 185–195.