CYRENE: A Database, Browser, and Library of Tools for Regulatory Genomics

Ryan Tarpine

May 12, 2008

Abstract

The CYRENE project seeks to address the fundamental problem of determining *de novo* the function of regulatory sequence by developing the cis-Lexicon, a database of known cis-regulatory modules, the cis-Browser, a next-generation regulatory genome browser, and a library of tools for assisting in the annotation pipeline. The cis-Lexicon will be a comprehensive catalog of experimentally-validated gene regulatory knowledge, designed to be a foundation and benchmark for future prediction algorithms. The cis-Browser is a high-speed integrative environment for viewing and annotating all types of genomic information. It is capable of displaying data from the cis-Lexicon, public online databases, BLAST hits, and precomputed comparative genomics analyses. To aid annotators' entry of information into the cis-Lexicon, we are developing high-throughput tools for finding relevant literature and assisting in the extraction of correct information. We suggest several algorithms to analyze the cis-regulatory data as the cis-Lexicon expands. The CYRENE project is being carried out in cooperation with Eric Davidson at the California Institute of Technology.

1 Background

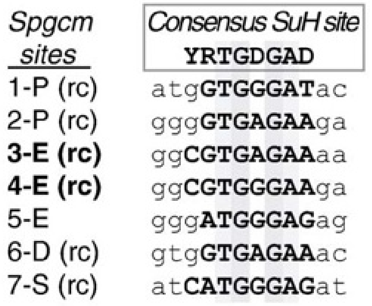
Most of the processes of life are carried out by molecules called *proteins* (and in some cases, *enzymes*). Every cell has inside itself a database of instructions for creating these molecules, called its *genome*. This database is stored not as bits like in a man-made computer, but as *DNA*, a sequence of small molecules called *nucleotides* (also known as *bases*). A chain of the four types of nucleotides (written A, T, C, and G) can be decoded to tell the cell how to make a specific protein. This code is simple and unambiguous. What is not so simple and thus far imperfectly understood is the code which tells when to make (*express*) the protein.

The cell has machinery which recognizes where a protein "recipe" (*gene*) begins and starts the first phase (*transcription*) of the decoding process, and the expression of a protein is largely determined by helping (*activating*) or hindering (*repressing*) this process. A class of proteins called *transcription factors* bind to particular substrings of DNA and interact with this machinery to regulate the transcription of (usually) nearby genes.

Case studies of the regulatory regions of several hundred genes has revealed certain facts about the arrangement of these binding sites.

1.1 Binding Sites

The first DNA-binding molecules that biologists discovered were "*restriction enzymes*," which cut DNA at a very specific sequence, such as Eco RI, which cuts every GAATTC it finds. Some restriction enzymes will cut a few nearly identical sites, such as Sal I, which cuts CCAGG and CCTGG. The concept of consensus sequence was invented to handle both of these types of enzymes. In addition to A, T, C, and G, extra letters were defined to represent this ambiguity of sites. Once W was made to stand for A or T, Sal I could be said to bind to CCWGG. This



Spgcm sites	*Consensus SuH site* **YRTGDGAD**
1-P (rc)	atg**GTGGGAT**ac
2-P (rc)	ggg**GTGAGAA**ga
3-E (rc)	gg**CGTGAGAA**aa
4-E (rc)	gg**CGTGGGAA**ga
5-E	ggg**ATGGGAG**ag
6-D (rc)	gtg**GTGAGAA**ac
7-S (rc)	at**CATGGGAG**at

Figure 1: Binding sites

representation of binding sites was concise yet accurate, because even the enzymes that bound to more than one sequence bound to all perfect matches and no sites with even a single base mismatched [19].

In contrast, a transcription factor generally binds to a variety of similar sequences. There are often differences between any two sites for the same factor. For example, Su(H) has seven known binding sites in spgcm, made up of six unique sequences (i.e., only one is found twice) (see figure 1)[16]. Unlike restriction enzymes, which are all-or-nothing when it comes to matching sites, the binding of transcription factors is affected to various degrees by changes in the site sequence. This enables fine-tuned "tweaking" of gene regulation–a few mismatches will still allow the factor to have its activating or repressive function, but to a lesser degree.

It is still not clear how to best represent mathematically the sites that a given factor will bind to. A straightforward extension to the consensus sequence method is to allow a bounded number of mismatches. There is a clear tradeoff between sensitivity (ensuring all known sites are identified) and specificity (ensuring only binding sites are matched by the representation). [19] gives the example of six short sequences from the E. coli genome which differ enough that in order to identify all of them the model must be so generic that a match is made every 30 base pairs (bp).

A more powerful representation is the position-weight matrice (PWM). These matrices contain a weight for every possible base at each position, so a matrix for a 6 bp site would have 24 entries (4 rows and 6 columns). The score of a possible site is assigned by adding together the values from each column which correspond to the bases at each position. This assumes that each base contributes independently, which, while not true in fact, is a reasonable approximation [2]. The logarithm of the observed base frequencies has been shown to be proportional to the binding energy contribution of the bases [3], so there is clear biological significance to using these values as the weights of the matrix. Typically the matrices are tuned to handle biased genome composition (e.g., if a genome has very many As, then a position which is usually A is not necessarily significant) by dividing the observed frequencies by the background frequencies.

1.2 Motif-Finding

None of the existing methods of representing a binding site can predict which sites are functional. Unlike restriction enzymes, which have an effect by themselves, transcription factors only work by affecting the transcription machinery. Some factors do this directly, while others only communicate via intermediaries. Even a short sequence will contain what looks like many sites for many different transcription factors, and it is difficult to determine which actually determine gene expression.

One method to bypass this problem is to look at a set of genes that appear to be coregulated–i.e., they are expressed at the same time in the same location. It is very likely that the same transcription factor regulates these genes. Therefore most of the promoter regions of these genes should contain a binding site for that factor. By simply searching for a short sequence which is found to be overrepresented (i.e., more common than expected by chance), we should be able to find the binding site.

Unfortunately, the binding sites will probably not be identical. Some type of tolerance for mismatches must be added to the search algorithm, which complicates things considerably (otherwise a simple count of the number of occurrences of, e.g., every 8-mer would suffice). Some algorithms model the motif they are looking for combinatorially as a consensus string with a maximum number of mismatches [15], while others use a probabilistic or information-theoretic framework [13].

1.3 Phylogenetic Footprinting

Without being able to discern *de novo* the regulatory regions of genes, we know that they should be conserved between closely-related species. Like the protein-coding sequence, regulatory sequence has a functional purpose and most mutations to it will cause harm to an organism. Therefore few offspring who have any changes to the regulatory sequence will survive, in contrast to those who have changes to sequence outside the regulatory and coding regions, which should have no difficulty. Over generations, while a few minor changes occur within functional regions, large changes will accumulate in the rest. By examining the sequence of species at the right evolutionary distance, we should see clear conservation only where the sequence has a specific function. Since there are known methods for predicting protein-coding sequence, we can exclude that from our analysis and only look at the conserved patches of unknown function, which are likely to contain regulatory sequence. For the highest accuracy, several species can be compared simultaneously [4].

1.4 Evaluation

Existing motif algorithms perform reasonably well for yeast, but "significantly worse in higher organisms" [7]. Several evaluations of the many proposed methods have been attempted, but the use of real genomic promotor sequences is hampered by the simple fact that "no one knows the complete 'correct' answer" [20, 14]. For an overview of the algorithms and the models they are based on, see [7].

2 Beyond Sequence

We argue that sequence comparison alone is not sufficient to crack the regulatory code. Studies of the logic implemented by regulatory regions [24, 12] have demonstrated that individual binding sites rarely have a direct effect. The function of a site depends greatly on context–what other sites are nearby, the spacing and order between the sites, the location relative to the start of transcription, and most likely other conditions not yet known. All of these factors contribute to the regulatory information that determines when a gene is expressed. Regulation will never be fully understood until each of these elements are fully understood and incorporated into a model of regulatory information.

Few sites interact directly with the transcription machinery; many affect it indirectly through interactions with neighboring sites [8]. The site functions combine through various logic operations to yield the final effect on gene expression [11]. The inputs to cis-regulatory modules tend to be of two types: time-/space-varying and constant (i.e., ubiquitous) inputs. The time- and space-varying inputs appear at a glance to determine gene expression alone, but in reality they depend on their neighbors.

Factor-factor interactions require that the sites be spaced at the correct distance. If a pair of sites are too close to each other, the two factors cannot bind simultaneously. If the sites are too far apart, the factors cannot reach each other. In *cyIIIa* it was even demonstrated that two copies of a group of sites without regard to spacing is less effective than one copy with the correct distribution [6]. There are also examples of known modules which need to be placed a minimum distance from the transcription start site [16].

The orientation of a site determines the orientation of the transcription factor that binds to it, which also affects the interactions it is capable of. Modules thousands of bases away from the gene interact with the transcription apparatus through looping, where the DNA molecule itself folds or forms loops to bring the two regions close together. In this case, orientation is unimportant [8]. When two neighboring sites interact, on the other hand, their orientations must be correct in relation to each other to ensure that the factors that bind them have their correct sides in contact. There are also motifs seen in several species where two sites for the same factor are arranged as an inverted pair (i.e., one site, a short spacer, and then another site in the opposite orientation) [16].

2.1 Function

It is not sufficient to merely identify the acting binding sites. In order to determine gene expression, we must also determine the sites' function. Certain transcription factors sometimes act positively and at other times negatively, depending on the context.

The overall function of any given module is the "combinatorial outcome" of the "unit operations" of its individual sites [8].

3 The Cis-Lexicon

Given that existing transcription factor databases [23, 17] do not contain sufficient information for cis-regulatory analysis, we have started the cis-Lexicon project. The cis-Lexicon is a database for all of the types of information described above, with an emphasis on site and module function and logic. Only regulatory elements that have been empirically tested and validated will be entered into the cis-Lexicon–no putative or predicted sites or function will be permitted, because this database is intended to be a foundation and benchmark for future algorithms. Data will be taken only from published papers.

3.1 Defining the Model

To develop the cis-Lexicon, we first had to design a data model: what data from each paper needs to be extracted, what the format should be, and what the relationships between elements are. We began by forming a team of three researchers: one computer scientist and two biologists. We studied the Strongylocentrotus purpuratus genes gcm and endo16 [16, 24] in detail.

With our new understanding and informal process, we then expanded our database to encompass nine additional genes: blimp1/krox, brachyury, cyIIIa, cyclophilin, gatae, nodal, otx, sm50, and wnt8.

3.2 Cis-Regulatory Ontology (CRO)

There exists a well known system of canonical names for gene function, the Gene Ontology (GO) [1]. GO terms allow researchers to give consistent descriptions of gene products that are amenable to computer processing. Standardized naming allows algorithms to use data from multiple sources. The GO project manages three separate controlled vocabularies, describing biological processes, cellular components and molecular functions in species-independent terms.

In the course of analyzing the genes above, we developed a cis-regulatory ontology (CRO), choosing canonical names for the various types of function transcription factor binding sites and cis-regulatory modules have. Our current list is: Spatial Control, Quantitative Control, Repression, Activation, Signal Response, DNA Looping, Booster, Input into AND logic, Input into OR logic, Linker, Driver, Insulation, and BTA Communication.

These terms are nonexclusive; in fact, many of them will often be used in combination. If a site is marked as being the Driver, then it probably also causes Activation of the gene. It should be classified as directing Spatial (ensuring the gene is expressed in the right place) and/or Quantitative (ensuring the gene is expressed in the correct amount) Control as well.

We initially established controlled vocabularies for the locations and times in which the cis-regulatory functions take place, but since these would have to be developed from scratch for each species and the terms would not allow cross-species searches, we ultimately decided not to continue their use. When new genes are added to the cis-Lexicon, for the time being the locations and times are entered using whatever description the authors of the papers used. Once the database grows, we will analyze these to determine canonical names automatically.

3.3 Streamlining the Process

With the experience of several months of analysis, the awareness that such time cannot be invested again, and the plan of hiring additional staff, we knew that we had to establish a formal process for adding genes to the cis-Lexicon. Only then could we expect annotators with less experience to quickly extract and add correct information to the database.

In collaboration with Eric Davidson of the California Institute of Technology, we developed “The Annotator's Worksheet," which laid out a step-by-step process that annotators could follow to ensure they extracted all of the useful information without being distracted by the extra details of each paper that aren't of concern to the cis-Lexicon. Each step is described in an unambiguous manner such that anyone able to understand the results of a paper detailing cis-regulatory analysis can reliably follow the process. It does not require the annotator to understand all of the methods and details of a particular experiment.

We will be testing this process by evaluating the addition of the next fifty genes, which have been chosen from Chapter 2 of [8] as prime examples of cis-regulatory analysis. This chapter discusses the basic principles of cis-regulation and presents several genes to illustrate each point. Choosing the next genes from this chapter ensures that there will be plenty of information to find, and that the genes will come from many species and they will have been studied by different laboratories through different methods. It will allow us to thoroughly evaluate our annotation pipeline.

3.4 Literature Mining

Chapter 2 is only an initial source of new genes. To build the cis-Lexicon, we will need access to the thousands of papers with cis-regulatory analysis that are known to exist. Practically all recently-published papers are available via PubMed [9]. We are currently developing tools for finding through PubMed the papers containing the data we need.

We have started by collecting the keywords and titles from the papers we have already extracted data from. The next fifty genes are currently in the middle of this process. We will use these words to search for additional papers in PubMed. There is a trade-off between using too few or too general keywords, which will return mostly papers in related areas that don't contain cis-regulatory analysis; and using too many or too specific keywords, which will return only a few papers. We will use our initial set of known papers as a benchmark, attemping to ensure that all of the papers are returned in our search with a minimum amount of other papers. Once we have established a reliable method for finding papers, we can poll PubMed regularly for newly published material for our database.

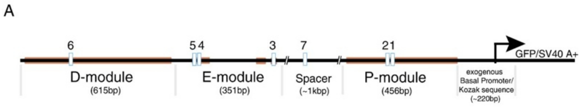


Figure 2: Cis-regulatory map

Even with the correct papers in hand, it is still difficult for the annotators to find the information they need. A fifteen-page document will contain only a few paragraphs with the results of the cis-regulatory analysis. This often isn't consolidated in one location; rather, it is spread among the details of the method, intermediate results, and other concerns. Therefore we have also begun developing tools for helping the annotators locate this information. Our first tool extracts all of the text from a PDF paper and searches it for key phrases which may signify useful information–activation, repression, mutation, and so on. By seeing a cluster of these terms in one area of the paper, an annotator can quickly zoom in to see whether the information he needs is there. It can also display all of the images in a document quickly so that the annotator can look for the "quintessential diagram," the one displaying a map of the cis-regulatory inputs (see figure 2). Without this diagram, the paper will not discuss cis-regulatory analysis, and so the annotator can discard the document as a false positive result of the paper-finding method (which needs to be sensitive more than selective in order to ensure the cis-Lexicon will be a comprehensive database).

4 The Cis-Browser

The cis-Browser is the other core component of the CYRENE project. The browser is partly a view of the cis-Lexicon, but more fully it is designed as a laboratory for the researcher to use throughout his experiments. Motivation for this is found in the disparity between the format of data that is published and the format of the researcher's own notes. The figures and graphs used in published papers display volumes of information in a small space, and help to

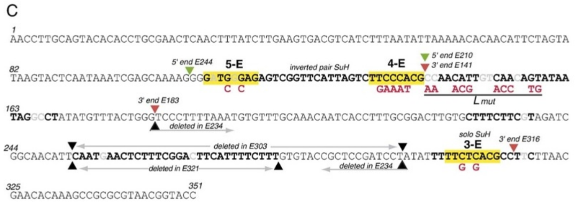

C
1
AACCTTGCAGTACACACCTGCGAACTCAACTTTATCTTGAAGTGACGTCATCTTTAATATTAAAAACACAACATTCTAGTA
82
5' end E244
5-E
inverted pair SuH
4-E
5' end E210
3' end E141
TAAGTACTCAATAAATCGAGCAAAAGGGGATGGGAGAGTCGGTTCATTAGTCTTCCCACGCCAACATTGTCAACAGTATAA
C C
GAAAT AA ACG ACC TG
L mut
163
3' end E183
TAGGCCTATATGTTTACTGGGTCCCTTTTAAATGTGTTTGCAAACAATCACCTTTGCGGACTTGTGCTTTCTTCGTAGATC
deleted in E234
244
deleted in E303
solo SuH
3-E
3' end E316
GGCAACATTCAATGAACTCTTTCGGACTTCATTTTCTTTGTGTACCGCTCCGATCCTATATTTTTCTCACGCCTTCTTAAC
deleted in E321
deleted in E234
G G
325
351
GAACACAAAGCCGCGCGTAACGGTACC

Figure 3: Cis-regulatory analysis

communicate many of the researcher's insights into the problem he is studying (see figure 3). As he is performing his experiments, however, this type of view is unavailable. A powerful, visual environment for integrating known genomic information and experimental data would allow him to better formulate new hypotheses for testing in the lab.

The cis-Browser is built upon the the Celera Genome Browser [21], an application developed by a team of programmers over several years with constant input from biologists and ultimately released as Open Source software. The cis-Browser itself is released under the LGPL license, permitting non-commercial modification and use of its source code.

4.1 Features

The Browser permits high-speed navigation of large genomic regions, fully supporting full chromosomes many megabases in length. It displays simultaneously both graphical and sequence views of the data, enabling a high-altitude view of the many features present in a genomic region while giving the precise residue sequence of a single feature. The graphical view can pan across wide regions instantly and allows the user to zoom to any level of detail, from the whole chromosome to the nucleotide level. New features can be added simply by selecting the nucleotides making one up and clicking a menu item.

Since we have been working primarily with the sea urchin developmental biology community, we added to the Browser the ability to download the sequence of any scaffold of the S. purpuratus genome and display every known and predicted gene (according to the Gnomon models at NCBI). The binding sites and cis-regulatory modules added to our database will be overlayed on this data.

4.2 Types of Data

The cis-Browser is capable of displaying many types of data. The simplest forms are the genomic features, such as transcripts and cis-regulatory modules. Cis-regulatory elements are supplemented with their functional and logic information not relevant for other structural elements on the genome.

The Browser is also capable of displaying BLAST hits. This makes it convenient to examine the context around the hits–the sequence and surrounding genes are downloaded and displayed automatically. In order to reduce the time spent looking at hits individually, the user can choose to concatenate all of the resulting scaffolds end to end and display them all at once. This permits visualizing all hits and their contexts simultaneously.

Solexa is a next-generation high-throughput sequencing technology. Unlike standard Sanger sequencing, Solexa sequencing yields much smaller reads (30 bp vs 400-600 bp) but with higher throughput (more total bases read in a given amount of time) and lower cost. While assembling a genome *de novo* is difficult with Solexa reads alone [22], it is possible to map reads to existing genomes. By mapping reads to an already-sequenced genome of the same organism, that assembly can be verified or corrected. By mapping reads to the genome of a different but related species, a restricted (but inexpensive) form of interspecies analysis becomes possible. The cis-Browser is capable of displaying tens of thousands of mapped Solexa reads for this type of analysis (described in detail later).

Currently in development is a more direct method of interspecies comparison–drawing lines connecting regions shared between two or more species. This can be used to display nucleotide-level comparisons, as in FamilyRelations [5] and Atavist (citation), or inter-module comparisons, highlighting binding site conservation whose lack of exact sequence conservation and possible change in order or orientation would elude simple sequence comparison.

5 Cis-Regulatory Meta Analysis

Once the cis-Lexicon contains several hundred genes, it will be ready for mathematical and computational analyses of the cis-regulatory code.

5.1 Interactive

There are countless factors that in theory could determine the function of a binding site, and rather than limit the use of the cis-Lexicon to predefined algorithms that examine a few possibilities, we will give researchers the ability to browse the database through a variety of searches that will permit them to manually look for patterns. These putative patterns can then be formalized and tested through automated analysis.

Keeping in mind that a central goal is discerning when a sequence that looks like a binding site really is functional and what that function is, there are many useful searches or filters on modules that will aid a user in examining sites in various contexts. Modules can be sorted by their position relative to the gene: is a module immediately adjacent to the transcription start site? Is it a long distance upstream? Is it in an intron? Since sites only cause an effect by interacting with other machinery, this property could be critical in determining the existence of a functional site.

Modules can also be filtered based on whether or not they have a certain transcription factor as an input (e.g., otx, gatae). Searching for pairs of sites (e.g., dorsal and twist) or larger groups within a given distance will also be possible (the distance limit has biological significance–transcription factors which bind nearby each other are likely to interact). Many other filters such as species and degree of inter-species module and/or site conservation will be available as well.

5.2 Automated

We have devised several initial automated algorithms to take advantage of the data we will have collected.

One of the main questions is: what sites often occur together? The effects of many transcription factors are mediated by other factors which bind nearby [8]. The software will allow the user to specify a specific transcription factor or run for all known factors. For a given transcription factor, the program will find all sites which occur within a maximum distance (as explained above) throughout all cis-regulatory modules where the factor is known to bind and function. The co-occurrence counts will be checked for statistical significance by comparing them to the number expected if the binding sites occurred independently.

We predict that many pairs will occur less than expected at random; this may be from transcription factors that are never present simultaneously (e.g., perhaps they are expressed at strictly different times and/or locations, and so will never be in the same module), but an interesting sub-case could be pairs where one factor actually represses the other.

Once common pairs of factors are known, these can be extended to larger complexes by running the algorithm on only the modules that contain these pairs and looking for a third or even fourth factor. This process can be tested by ensuring we catch known complexes like Dorsal, Twist, and Snail [10].

There are certain classes of groups of sites that should be found. The structure of DNA itself causes the proteins that bind to it to fall into two classes: major and minor groove binders. Most transcription factors are major groove binders. Factors binding to one groove often bend the DNA to give their neighbors easier access. We should be able to find a correlation between minor groove and major groove binders.

We can also use this analysis to find the factors that assist in signal transduction, the transformation of signals from molecules received from outside of the cell (ligands) to cellular state that influences gene regulation. The major mediator (the main transcription factor activated by the ligand) of many signaling pathways is known (e.g., Su(H) is the major mediator of N signaling in S. purpuratus; see papers cited in [16]). Since few factors affect transcription directly, there are often other factors that must bind nearby the mediator sites which work together. We should be able to determine which of these are mostly responsible for actual activation or repression of the genes involved.

Given the knowledge of which factors occur together, we can investigate whether the sequence of a binding site is affected by its neighbors. As discussed above, while there is the concept of a consensus binding site, few sites match this consensus. Besides allowing fine tuning of the degree of the effect of the bind, we conjecture that various sequences give the transcription factor some flexibility to interact with neighboring factors. If this is true, we should see that a given factor is biased toward specific "mismatches" from the consensus sequence when located near another site which it interacts with. We can test this by generating alignments of the binding site sequence under different contexts and calculating the Kullback-Leibler distance between the different distributions of each column. If this conjecture turns out to be true, then we will need separate position-weight matrices for each factor in different contexts.

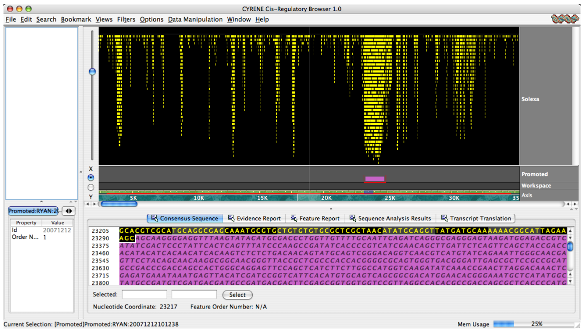
6 Beyond the Lexicon

Even without the cis-Lexicon, the cis-Browser is capable of assisting in cis-regulatory analysis.

6.1 Solexa Mapping

As described above, when comparing the genomes between two related species, functional sequence tends to be conserved much more than nonfunctional sequence. Solexa read mapping yields an inexpensive method to perform some interspecies analyses without waiting for an entire sequenced genome. Given a set of Solexa reads from one genome, only the reads from conserved regions will be mappable to the other sequence–nonconserved regions will have too many mismatches to be unambiguously mapped. A simple plot of the locations of mapped reads will display the regions of conserved sequence, which will often be functional regions.

In cooperation with the Davidson lab, we mapped 11.6 million 25-bp Solexa reads from L. variegatus to Strongylocentrotus purpuratus with the software RMAP [18]. We configured the program to allow up to four mismatches. Since we knew the genes where the reads were taken from, we were able to save a great amount of time by only attempting to map the reads to the areas around the same genes in the other species. This reduced the amount of the genome we had to search to 4 million bases.



Property	Value
Id	20071212
Order N...	1

```
23205 GCACGTCGCATGCAGGCGAGCAAATGCGTGCTGTGTGTGCGCTCGCTAACATATGCAGGTTATGATGCAAAAAACGGCATTAGAA
23290 AGCAGCAAGGGGAGGTTTAAGTATACATGCGACCCTGGTTGTTTTGCAATTCAGATCAGGGCGAGGGAGTAAGATGGAGACCGTC
23375 ATATCGACTCCCTATTCACTCAGTTTATCCCAAGCCGATATCACCCCGTCATCGAACAGCTTGATTCTCAGTTCAGCTACCGACC
23460 ACATACATCACAACATCACAAGTCTCTCTGACAACAGTATGCAGTCGGGACAGGTCAACGTCATGTATCAGAAATTGGCAACGA
23545 GTTCCTACAGCAACAAGGCGGCAACGGGTTACCGCTCGCCCACCACGGGGCGCAGTGGGTGACGGATTGAGCGCTCCGCCGCAT
23630 GCCGACCCGACCAGCCACTGGGCAGGAGTTCCAGCTCATCTTCTTGGCCATGGTCAAGATATCAAACCGAACTTAGGACAAACTC
23715 GAGATGAAATAAATGAGTTACATCGATCCGGTCATTCACATGTGCAGTCAGCGGCGACATGGAACACGGGAAATGCTCATATGGC
23800 TATGCCGATGTCGATGACGATGCCGATGACGACTTCGAGCGGTGGTGGTCCGTTAGGCCACACGCCGACCAGCGCTCACCCCATG
```

Selected:

Nucleotide Coordinate: 23217 Feature Order Number: N/A

Current Selection: [Promoted]Promoted:RYAN:20071212101238

Figure 4: Solexa mapping

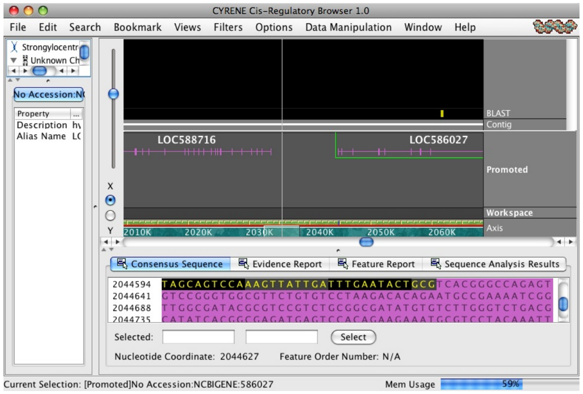
The mapped reads clearly show islands of definite conservation for many of the genes (see figure 4) . The Davidson lab has already reported that testing these patches has lead to the discovery of active cis-regulatory modules.

6.2 BLAST

We have also been collaborating with Gary Wessel and Mamiko Yajima to study insulator sequences. BLAST support was added to the cis-Browser in order to compute the location of possible insulator modules quickly and examine all of the returned instances. Many of the hits on the S. purpuratus genome occur on scaffolds of a very small size. Viewing them one by one is wasteful and inefficient. In order to bypass this problem we implemented support for concatenating all of the results end to end (without regard to order or chromosomal location, which is unknown). In this manner all of the hits and their surrounding genomic contexts (genes, etc) are visible simultaneously. See figure 5 for an example.

6.3 Dr. Gideon Koren

We have also begun work with Dr. Gideon Koren and Katja Odening in the rabbit genome. They are trying to determine the gene causing a phenotype of inter-



```
2044594  TAGCAGTCCAAAGTTATTGATTTGAATACTGCGTCACGGGCCAGAGT
2044641  GTCCGGGTGGCGTTCTGTGTCCTAAGACACAGAATGCCGAAAATCGG
2044688  TTGGCGATACGCGTCCGTCTGCGGCGATATGTGTCTTGGGTCTGACG
2044735  CATATCACGGCGAGATGAGTCCACAGAAGAAATGCGTCCTACAAATT
```

Selected:

Nucleotide Coordinate: 2044627   Feature Order Number: N/A

Current Selection: [Promoted]No Accession:NCBIGENE:586027

Figure 5: BLAST hits

est, and they have evidence that the gene is regulated by certain hormones. The sequence for binding sites mediating the effects of these hormones (hormone response elements) is known, but the rabbit genome has not been sequenced yet. We will map Solexa reads from the rabbit genome which are suspected to be from the causal gene to genes in a related organism such as human or mouse which contain hormone response elements.

References

[1] M. Ashburner, CA Ball, JA Blake, D. Botstein, H. Butler, JM Cherry, AP Davis, K. Dolinski, SS Dwight, JT Eppig, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.

[2] P.V. Benos, M.L. Bulyk, and G.D. Stormo. Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Research*, 30(20):4442–4451, 2002.

[3] OG Berg and PH von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–50, 1987.

[4] M. Blanchette and M. Tompa. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research*, 12(5):739, 2002.

[5] C.T. Brown, Y. Xie, E.H. Davidson, and R.A. Cameron. Paircomp, FamilyRelationsII and Cartwheel: tools for interspecific sequence comparison. *feedback*, 2005.

[6] T. Brown. *Tackling the regulatory genome.* PhD thesis, California Institute of Technology.

[7] M.K. Das and H.K. Dai. A survey of DNA motif finding algorithms. *feedback*, 2007.

[8] E.H. Davidson. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution.* Academic Press, 2006.

[9] National Center for Biotechnology Information. Pubmed home. `http://www.ncbi.nlm.nih.gov/pubmed/`.

[10] Y T Ip, R E Park, D Kosman, K Yazdanbakhsh, and M Levine. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. *Genes Dev.*, 6(8):1518–1530, 1992.

[11] S. Istrail and E.H. Davidson. Gene Regulatory Networks Special Feature: Logic functions of the genomic cis-regulatory code. *Proceedings of the National Academy of Sciences*, 102(14):4954, 2005.

[12] CV Kirchhamer and E.H. Davidson. Spatial and temporal information processing in the sea urchin embryo: modular and intramodular organization of the CyIIIa gene cis-regulatory system. *Development*, 122(1):333–348, 1996.

[13] CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[14] N. Li and M. Tompa. Analysis of computational approaches for motif discovery. *Algorithms for Molecular Biology*, 1(8), 2006.

[15] P.A. Pevzner and S.H. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 8:269–278, 2000.

[16] A. Ransick and E.H. Davidson. cis-regulatory processing of Notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. *Developmental Biology*, 297(2):587–602, 2006.

[17] A. Sandelin, W. Alkema, P. Engström, W.W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):W235–W241, 2004.

[18] A. Smith. Rmap: A program for mapping solexa reads. `http://rulai.cshl.edu/rmap/`.

[19] G.D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[20] M. Tompa, N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23:137–144, 2005.

[21] RJ Turner, K. Chaturvedi, NJ Edwards, D. Fasulo, AL Halpern, DH Huson, O. Kohlbacher, JR Miller, K. Reinert, KA Remington, et al. Visualization challenges for a new cyberpharmaceutical computingparadigm. *Parallel and Large-Data Visualization and Graphics, 2001. Proceedings. IEEE 2001 Symposium on*, pages 7–145, 2001.

[22] N. Whiteford, N. Haslam, G. Weber, A. Prügel-Bennett, J.W. Essex, P.L. Roach, M. Bradley, and C. Neylon. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, 33(19):e171, 2005.

[23] E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241.

[24] CH Yuh, H. Bolouri, and EH Davidson. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, 128(5):617–629, 2001.