INTRODUCTION

Gene regulatory networks (GRNs) encode the developmental programs of animals. By representing both functional units – genes and regulatory regions – and the *interactions* between those units, GRN models provide a mechanism for integrating many types of data, including expression patterns, perturbation results, biochemical function, sequence data, and kinetic information.

A critical bottleneck in the construction of GRN models is the analysis of transcriptional regulatory regions and transcription factor binding sites within those regions. Regulatory regions can be thought of as hardwired logic functions, which connect upstream regulatory influences – transcription factors – to transcriptional effects. Specific sites within these regions direct transcription factor binding and represent immediate regulatory interactions; as such, they are important components of developmental GRNs. However, our ability to dissect regulatory regions into binding sites, and to generalize from specific binding sites to the entire genome, is technologically limited. Most experimental techniques are low-throughput; whole-genome experimental techniques are expensive and may not be of sufficient sensitivity to detect the sparse set of binding sites used in early development; and computational techniques for whole-genome analysis have gained little traction, largely due to high false-positive rates. Refining existing techniques and identifying new techniques is a critical component of GRN research.

One important use of GRNs is to understand the evolution of developmental novelty through GRN comparison. The vertebrate lineage contains a number of apparently novel developmental programs, including neural crest and placodes. There is some evidence that conserved regulatory modules from chordates and even the ancestral bilaterian have been co-opted for these functions, but we currently know very little about which components of the regulatory networks underlying vertebrate novelty are new and which have been re-used from already existing networks. Understanding the mechanisms by which developmental novelties arise is important for a basic understanding of evolution and development.

RESEARCH ACCOMPLISHMENTS

My research focuses on experimental and computational investigations of regulatory regions. During my graduate work in Dr. Eric Davidson's lab, I have continued the investigation of the *cyIIIa cis*-regulatory region and used a variety of experimental approaches to explore the binding sites contained within the region as well as its overall kinetic output. In particular, I have obtained nuclear protein concentrations for the five species of transcriptional activators that bind in this region; confirmed the identity of one of the factors through *in vitro* expression and binding assays; made individual mutations to binding sites and observed the results *in vivo*; and used perturbation analysis to help identify the only remaining unknown upstream regulator. I am currently working on integrating this information into a kinetic model of *cyIIIa* regulation. This research constitutes the final chapter of my thesis and is as yet unpublished.

I have also explored several computational approaches to regulatory region and binding site analysis.

I am the primary developer of the Cartwheel/FamilyRelations software package, a combination graphical interface and Web site for comparative sequence analysis and genome annotation **(Brown et al., Dev. Bio. 2002; Brown et al., BMC Bioinf. 2005)**. This software was originally developed to support the Endomesoderm Gene Regulatory Network effort and was used to find the regulatory regions of over 15 genes in the Network (**Davidson et al., Science, 2002; Davidson et al., Dev. Bio. 2002; Yuh et al., Dev. Bio. 2002)**.

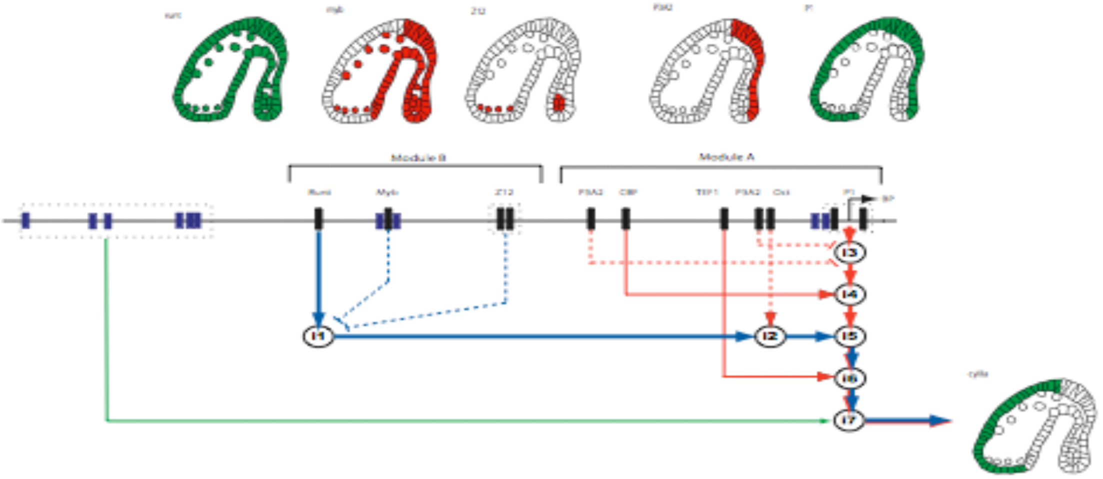
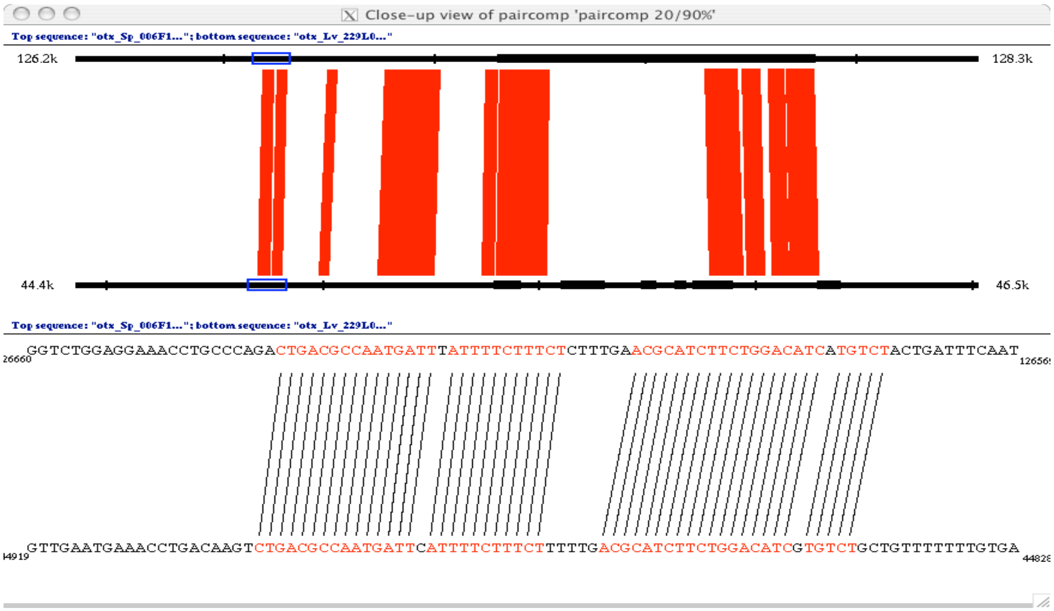


Figure 1: *cis*-Regulatory regions integrate diverse spatial and temporal inputs. The *cyIIIa* *cis*-regulatory region contains binding sites for 5 spatial regulators (shown above) and 4 temporal regulators. These inputs combine to drive *cyIIIa* transcription in the aboral ectoderm (shown on the lower right) throughout development.

FamilyRelations and Cartwheel are now being used by over 40 other labs, at Caltech and elsewhere, to find regulatory regions in animal and plant genomes (see publication list at http://family.caltech.edu/). Comparative sequence analysis to discover regulatory regions has become a commonly employed technique in animals, and the dynamics of *cis*-regulatory evolution revealed by our approach have helped us understand how genomes evolve.


Close-up view of paircomp 'paircomp 20/90%'

Top sequence: "otx_Sp_006F1..."; bottom sequence: "otx_Lv_229L0..."

Top sequence: "otx_Sp_006F1..."; bottom sequence: "otx_Lv_229L0..."

```
26660 GGTCTGGAGGAAACCTGCCCAGACTGACGCCAATGATTTATTTTCTTTCTCTTTGAACGCATCTTCTGGACATCATGTCTACTGATTTCAAT 12656
4919 GTTGAATGAAACCTGACAAGTCTGACGCCAATGATTCATTTTCTTTCTTTTTGACGCATCTTCTGGACATCGTGTCTGCTGTTTTTTTGTGA 4482
```

Figure 2: Strict conservation between homologous genomic regions suggests regulatory function. FamilyRelations can display alignments and comparisons generated by many different algorithms. Shown here is a dot-plot style comparison between genomic regions from two distinct species of sea urchin. The pattern of sequence conservation shown, with few gaps or mismatches, has been a good evolutionary signature of regulatory regions.

In a separate collaboration with Dr. Curtis G. Callan, a physicist at Princeton University, I investigated binding site mutation patterns and genome distribution statistics for the bacterial transcription factor CRP (**Brown and Callan, PNAS, 2004**). Our work demonstrates that many hundreds of the supposedly spurious CRP binding sites in the genome are evolutionarily conserved with respect to predicted CRP binding (but *not* necessarily sequence) between *E. coli* and *S. typhimurium.* This approach also identifies functionally important binding sites for ArcA in the facultative anaerobe *Shewanella oneidensis* (**Gralnick, Brown, and Newman, Mol. Micro. 2005**). Recent results indicate that this may be a general computational approach to validating binding sites in bacterial genomes (unpublished).

Finally, I have contributed extensively to several other genomics and computational projects. In particular, I built the data architecture and Web site for the Sea Urchin Genome Project site, http://sugp.caltech.edu/. I also developed the computational strategy used to locate and annotate over 400 transcription factors and more than 300 zinc finger genes in the then-unassembled *S. purpuratus* sea urchin genome (**Howard et al., in preparation; Materna et al., in preparation)**. In these projects, I have worked closely with experimental biologists, physicists, and mathematicians to adapt existing techniques and to develop new techniques for sequence analysis. The results have contributed extensively to experimental GRN research and have also advanced our understanding of the evolution of binding sites and regulatory regions. I hope to continue collaborating with a broad spectrum of scientists in my future research.

RESEARCH GOALS

Short-term Research Plans

My short-term research goals are to investigate the structure and evolution of specific developmental regulatory regions in vertebrates, using a combined experimental and bioinformatics approach in chick.

a. Build a library of important vertebrate regulatory regions. As part of a collaboration with Dr. Marianne Bronner-Fraser's lab to define the neural crest GRN, I will soon be engaging in the systematic experimental discovery of regulatory regions for genes important to neural crest specification. I will continue this effort, with special attention to regulatory regions for transcription factors involved in neural crest and placode development. This alone will be an important contribution to the understanding of vertebrate development: fewer than 20 regulatory regions involved in early vertebrate development are currently known.

b. Compare *cis*-regulatory elements among vertebrates and between vertebrates and chordates. By comparing known regulatory elements between different vertebrates and chordates, I can determine which upstream influences are conserved and which are novel. This, in turn, will allow me to make and test predictions about specific signaling and transcription factor involvement in neural crest specification.

c. Extend current computational approaches to whole-genome approaches, especially binding site analysis. While comparative sequence analysis works very well to identify vertebrate regulatory regions, we are lacking sensitive and specific tools to dissect regulatory regions. Binding sites, in particular, are underdetermined by current approaches. I will extend already existing approaches that work well in bacterial and other animal genomes – in particular, position-weight matrix extension, from my previous CRP and arcA work, and binding site positional correlation, which has been used extensively in *Drosophila* work – to the study of the larger, more complex vertebrate genomes.

Long-term Research Plans

In the long term, I will continue using a combined integrative and comparative approach to investigate the dynamics of gene regulatory network evolution, with special attention to the evolution of vertebrate novelty and the relation of vertebrates to other deuterostomes.

a. Dynamics of *cis*-regulatory evolution

We know very little about how *cis*-regulatory regions evolve. Most animal regulatory regions exhibit surprising degrees of nucleotide conservation across large evolutionary distances; recent results from echinoderms suggest that this may be due to the suppression of certain kinds of evolutionary events (insertion-deletions) rather than a requirement of strict sequence conservation. However, from starfish-sea urchin GRN comparisons, we also know that regulatory sequence can change dramatically while upstream regulatory influences remain identical. Precisely how this process works is still a mystery. Chordates offer an excellent opportunity to study how regulatory regions evolve in concert with changing regulatory roles, due to the myriad of subtle developmental changes and extensive genome duplication within the chordates, and the large number of sequenced and in-process genomes at varying evolutionary distances. A stacked comparison of regulatory regions from multiple animals around developmentally important genes, together with the presence or absence of specific regulatory influences, will help us understand the structure and evolution of GRNs.

b. Evolution of vertebrate novelty and comparison with chordate and other invertebrate deuterostomes.

The existing sequences of primitive chordates and the planned sequencing of the *Amphioxus* and sea lamprey genomes offer an exciting window into the dynamics of evolutionary change. We know from sequence conservation that most regulatory molecules and many regulatory regions evolved prior to the divergence of the vertebrates. By comparing protein expression patterns across species and performing cross-species *cis*-regulatory reporter experiments, we will gain insight into the genesis of the vertebrate developmental program.

c. Develop new and expanded computational approaches for GRN modeling and genomic analysis.

As our base of information on vertebrate development widens, it will inform the development of new computational tools, both for analyzing genomic data and integrating experimental data into gene regulatory network models. The vast amount of information upon which even the current limited models of development rest is already difficult to digest, and visualization and model-building tools will be need to help biologists build larger models. The wealth of sequence data is overwhelming, and we need to build automated tools for information extraction and interfaces for interacting with the data. Such tools will depend critically upon the types of questions being asked, the experimental approaches being used, and most especially the specific organisms under investigation; they cannot be built other than in close collaboration with experimental work.