Pura rarg Kni Bcl6 Sna Stat5a rara Foxa2 E74 Hb Nr4a1 e(spl) CI h TEF-1 (TEAD-1) Myf6 Myf6 Hoxb2 Nkx2-1 ac HOXD4 Cebpa Pit-1 (Pou1f1) Tll En-2 HoxA4 Foxa1 EN Ubx E2F2 Foxa3 IRF1 WT1 tin etv4 otx2 RORA rara gsc POU3F2 Pgr Kr TCN2 Ahr Gcm EGR-3 HSF1 bcd elk1 Nkx2-1 Fos POU4F1 HLHmgamma TCF7 Nupr1 Cdx-2 IRF8 gata6 Abd-A Mitfa PDX-1 Nkx2-1 Sox2 Gsb SMAD7 Nkx6-1 a-myb pax4 Pb NR0B2 (SHP) Tp53 Ankrd1 HNF1A Ddit3

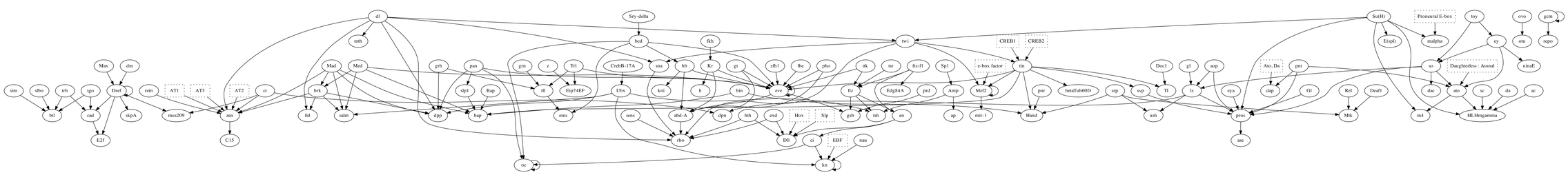
Report on the CYRENE Project: A *cis*-Lexicon containing the regulatory architecture of 586 regulatory genes experimentally validated using the "Davidson Criteria"

Ryan Tarpine, James Hart, Timothy Johnstone, Derek Aguiar, Sorin Istrail

Center for Computational Molecular Biology, Brown University

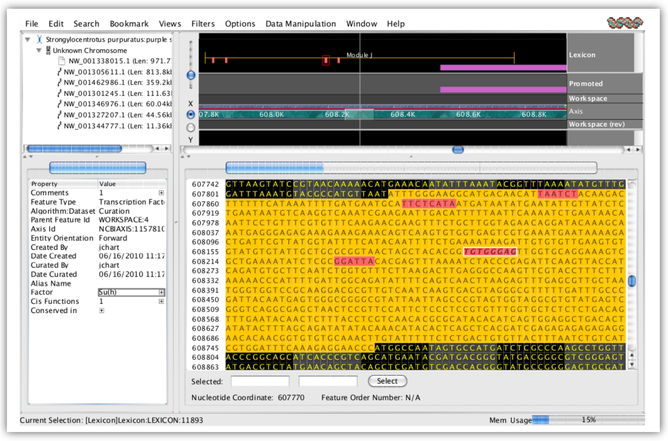
All correspondence including getting the cisGRN Browser to Istrail Lab, Center for Computational Molecular Biology and Department of Computer Science, Brown University,
sorin_istrail@brown.edu

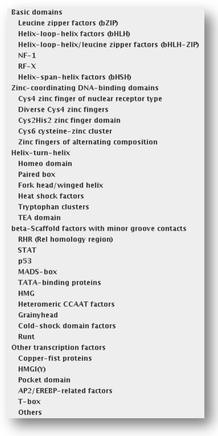
The CYRENE cis-Lexicon presently contains the regulatory architecture of 393 transcription-factor-encoding genes and 194 other regulatory genes in eight species: human, mouse, fruit fly, sea urchin, nematode, rat, chicken, and zebrafish, with a higher priority on the first five species. The regulatory architectures of each of these CYRENE genes are validated using the "Davidson Criteria:" sites must be shown to physically bind proteins and functionally confirmed by in-vivo disruption. The cis-Lexicon annotations include confirmed transcription factor binding sites, the *cis* Regulatory Module (CRM) boundaries, the spatial and temporal functionality of the CRM, and the molecular function and classification of the encoded protein. Included is an update on the CLOSE System (cis-Lexicon Ontology Search Engine) -- a set of algorithmic strategies for automated literature extraction of cis-regulation articles – that is used to speed up the identification of new CYRENE genes in the literature and to estimate the "completeness" of the CYRENE transcription factor universe. Here also we discuss the newly released CYRENE cisGRN-Browser, a full genome browser dedicated to cis-regulatory genomics. This work has been done jointly with Eric Davidson of Division of Biology at California Institute of Technology.

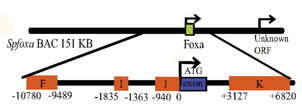


cis-Lexicon Connectivity Map (D. Melanogaster)

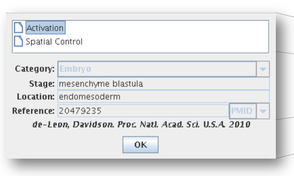
cisGRN Browser

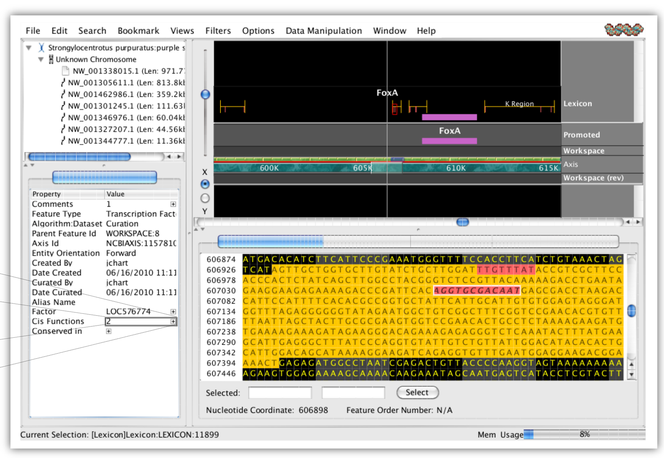




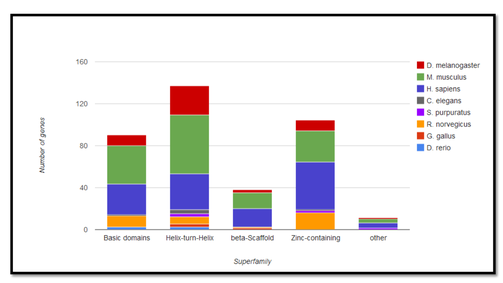


Davidson and de-Leon, 2010

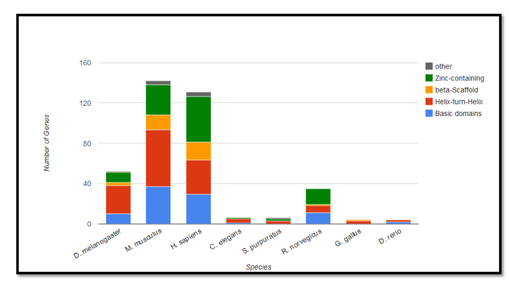




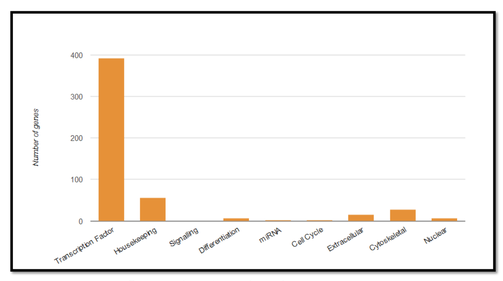
cis-Lexicon



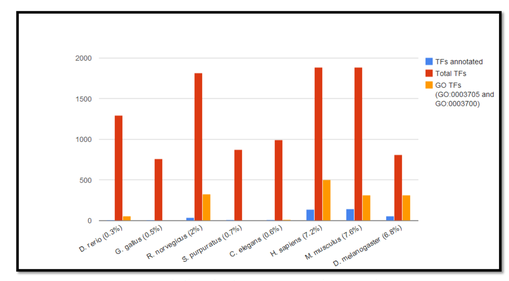
Distribution of cis-Lexicon transcription factors by TF superfamily



Distribution of cis-Lexicon transcription factors by Species



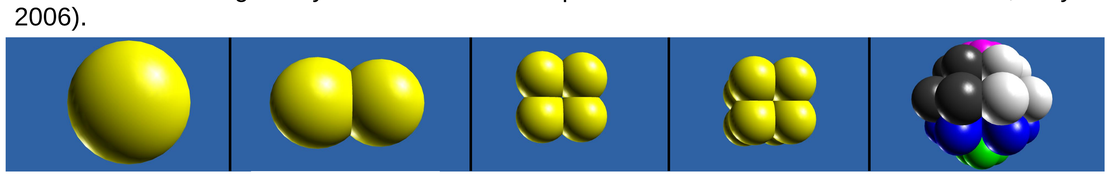
Cellular function of cis-Lexicon genes



Transcription factor coverage by species

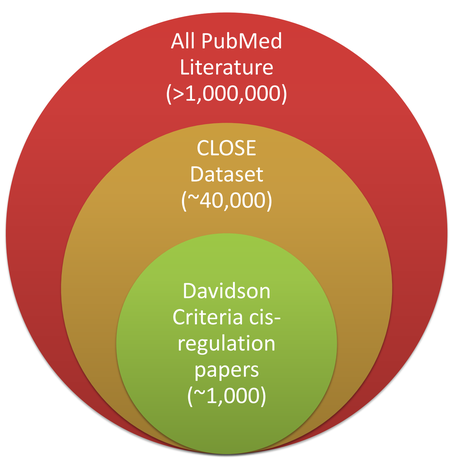
Virtual Sea Urchin

The Virtual Sea Urchin (VSU) uses spatial models and a graphics engine to simulate the 4-dimensional sea urchin embryo, allowing the researcher to probe the GRN at various levels of granularity -- from the multicellular embryo to the gene-regulatory network of an individual cell-type. The VSU currently provides models for the *S. purpuratus* embryo at 6h (shown), 10h, 15h, 20h, and 24h which were created by extrapolating cross sectional color coded tracings from photomicrographs to three dimensions (Eric H. Davidson. The Regulatory Genome: Gene Regulatory Networks In Development And Evolution. Academic Press, May 2006).



Virtual Sea Urchin's view of the *Strongylocentrotus purpuratus* embryo at 0, 1, 2, 3, and at 6 hours. VSU distinguishes cell type by color.

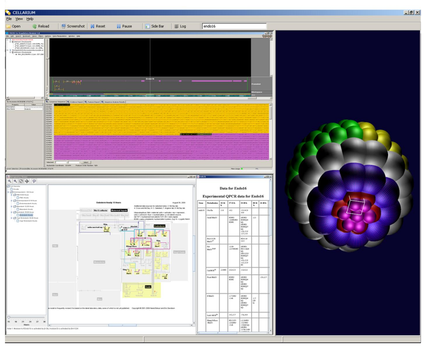
cis-Lexicon Ontology Search Engine (CLOSE)



The CLOSE algorithm combines human-curated knowledge of biological nomenclature with combinatorial optimization to home in on the few thousand papers that are relevant to the CYRENE Project out of the millions in PubMed. The CLOSE algorithm begins with a set of synonym lists, each carefully designed by biologists to capture the various ways that one concept can be described in the literature. Each list represents a particular aspect of cis-regulatory analysis that, when recognized in a title or abstract, would be evidence that the paper is relevant to the CYRENE Project. The CLOSE algorithm adapts itself to match as many known relevant papers as possible while minimizing the number of predictions that it makes, aiming to maximize both sensitivity and specificity. Within minutes, it determines a set of rules that match 95% of our known cis-regulatory papers while discarding 95% of our starting set—papers downloaded from journals which publish cis-regulatory analyses along with other biological research.

Future Direction: Cross-Platform Integration

The computational and data model for the VSU was recently completely rebuilt in Java using JOGL bindings to accommodate animation and integration with the cis-Browser. The development of an embryo can now be modeled using flat text files. The computational modeling of embryonic development will eventually feature realistic cell models and dynamics simulators. We also plan to combine the cis-regulatory sequence analysis capabilities of Cyrene and the network building, visualization, and simulation capabilities of BioTapestry with the temporal and spatial analysis of the 4D Virtual Sea Urchin to get a complete characterization of the S. purpuratus GRN.



HOXB4 nr1d1 Runx2 Pax6 mec-3 RELB Msx2 TFAP2c Ar NR0B2 (SHP) Pax2 otx GLI1 Mef2c cad IRF4 ndn IRF7 NFATc1 arntl Pit-1 HOXA10 E2F6 Srebf1 ppard GFI1B tal1 Tp63 Neurod1 Sp7 Bcl3 Nr4a3 Runx2 Ebf1 ase NR0B2 (SHP) Hif1a Ovol1 Hoxc8 Esr1 Rb1 Nr4a1 Myc nr1h4 Hmga1 hoxd9 SOX3 Gata1 nr5a1 (ad4bp, sf-1) pxr pokemon nr0b1 Prdm1 eve Hic1 Dfd EGR-1 rarb MYB (c-Myb) Hes1 REL Pou5f1 MYBl2 (b-myb) HNF-1a Hoxc8 Foxa2 ascl1 tsh Sox9 Gata4 Lyl1 Gbx2 C15 Smad6 Creb3 N3c1 Hif3a ikzf3 Ok2 chrebp Srebf1 Hmga1 Zeb1 Pou4f3 nr1h2 HNF1b tp73 Runx1 hes6 usf2 GATA1 car SREBF1 mxd1 hmx1 tbx20 neurog2 foxp3 coupt2 klf10 Nr4a1 Pf1a Ddit3 Hlh-6 ATF3 Sox10 Ebf1 Osr1 Snai1 Prox1 Nr4a1 Fos Foxf1a Foxl1 Jun Nkx3-2 CHREBP Mlpu1 (Zfh667) chrebp Id3 MYC CYP27B1 C11orf31 Foxq2p Foxa Sfpi1 EGR-1 NFKBIA dref

Dll E2F (dE2F) EDF-1 eve fGf4 FOS (c-fos) Fosl1 ftz gataE hand Hand-1 HNF1B Hoxa2 Hoxb2 hoxb2 Hoxb3 jing kn Krox20 lim-6 lz MafA Mafk mef2 Msx1 Myf5 Myf5 Myod1 MYOG nanog Nfe2 Nfe2l2 nfkb1 Nkx2-5 oc (otd) otp pax4 pax6b PDX-1 POU5F1 pros ptf1a Rb1 Rbl1 salm slp1 SMAD7 so Sox2 SP3 Srf STAT1 Stat3 STAT4 svp tcfap2a TFAP2a TFAP2c (AP-2gamma) TLX1(Hox11) vvl ybx1 zen Zfp106 gcm Zbtb7 Ahr Gcm EGR-3 ppard GFI1B tal1