On the Concept of *cis*-Regulatory Information: From Sequence Motifs to Logic Functions

Dedicated to Grzegorz Rozenberg's 65^{th} Birthday

Ryan Tarpine*† Sorin Istrail*‡

Abstract

The regulatory genome is about the "system level organization of the core genomic regulatory apparatus, and how this is the locus of causality underlying the twin phenomena of animal development and animal evolution." [8] Information processing in the regulatory genome is done through regulatory states, defined as sets of transcription factors (sequence-specific DNA binding proteins which determine gene expression) that are expressed and active at the same time. The core information processing machinery consists of modular DNA sequence elements, called *cis*-modules, that interact with transcription factors. The *cis*-modules "read" the information contained in the regulatory state of the cell through transcription factor binding, "process" it, and directly or indirectly communicate with the basal transcription apparatus to determine gene expression. This endowment of each gene with the information-receiving capacity through their *cis*-regulatory modules, is essential for the response to every possible regulatory state to which it might be exposed during all phases of the life cycle and in all cell types.

We present here a set of challenges addressed by our CYRENE research project aimed at studying the *cis*-regulatory code of the regulatory genome. The CYRENE Project is devoted to (1) the construction of a database, the *cis*-Lexicon, containing comprehensive information across species about experimentally validated *cis*-regulatory modules; and (2) the software development of a next-generation genome browser, the *cis*-Browser, specialized for the regulatory genome. The presentation is anchored on three main computational challenges: the *Gene Naming Problem*, the *Consensus Sequence Bottleneck Problem*, and the *Logic Function Inference Problem*.

*Department of Computer Science and Center for Computational Molecular Biology, Brown University, 115 Waterman Street, Box 1910, Providence, RI 02912, USA

†ryan@cs.brown.edu

‡sorin@cs.brown.edu

1 Introduction

Gene expression is regulated largely by the binding of transcription factors to genomic sequence. Once bound, these factors interact with the transcription apparatus to activate or repress the gene. Unlike a restriction enzyme, which recognizes a single sequence or clearly defined set of sequences, a transcription factor binds to a family of similar sequences with varying strengths and effects. While the similarity between sequences bound by a single factor is usually obvious at a glance, there is as yet no reliable sequence-only method for determining functional binding sites. Even the seemingly conservative method of looking for exact sequence matches to known binding sites yields mostly false positives– since binding sites are small, usually less than 15 bases long, by chance alone millions of sequence matches are scattered throughout any genome. Very few of these have any influence on gene regulation. The activity of a putative binding site depends on more than the site sequence, and it is hoped that the site context (i.e., the surrounding sequence) contains the necessary information. To predict new binding sites, we must understand the biology of gene regulation and incorporate the missing sources of regulatory information into our model.

To this end, we have created a database for storing the location and function of all experimentally found and validated binding sites, the *cis*-Lexicon. The *cis*-Browser, a powerful visual environment for integrating many types of known genomic information and experimental data, together with the *cis*-Lexicon make up the core components of the CYRENE Project: a software suite, associated tools, and set of resources for regulatory genomics. Only by mining a large, high-quality collection of functional binding sites can we uncover the missing clues into what makes sites functional.

We have come across many fundamental issues as the CYRENE Project has developed. We crystalized three of these under the names the *Gene Naming Problem*, the *Consensus Sequence Bottleneck Problem*, and the *Logic Function Inference Problem*.

2 A Case Study

A gene with one of the best-understood *cis*-regulatory regions is *endo16* of the sea urchin *Strongylocentrotus purpuratus* [6]. A 2.3 kb sequence directly upstream of the transcription start site has been shown to contain binding sites for many transcription factors, clustered into six *cis*-regulatory modules[26], as seen in figure 1. Each module has its own independent function: Modules A and B drive expression in different periods of development, Modules DC, E, and F cause repression in certain regions of the embryo, and Module G is a general expression booster at all times. Each module has a similar organization: one or two binding sites for unique factors which don't bind anywhere else in the regulatory region, surrounded by several sites for factors which are found in other modules as well.

Module A has been described as a "hardwired logic processor", whose output

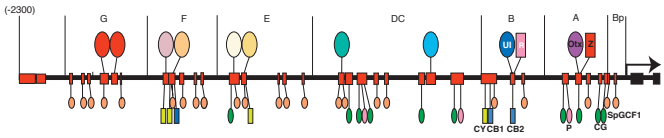


Figure 1: The structure of *endo16*'s regulatory region (from [25])

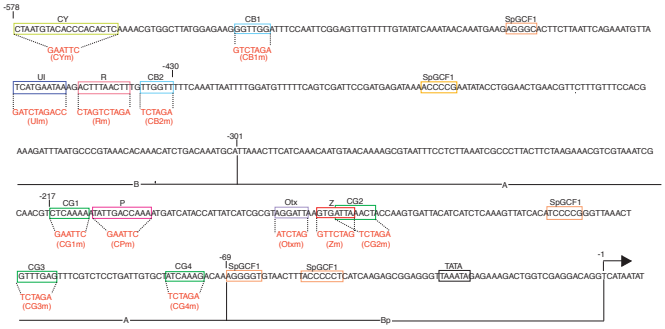


Figure 2: Quintessential diagram (from [25])

in terms of gene expression depends entirely on the binding of its inputs. The factor which binds uniquely to Module A, Otx, is the "driver" of the module, meaning that the change in the output of the module over time is a result of the change in the concentration of Otx [24]. The other factors which bind to Module A, while absolutely necessary for correct behavior, are ubiquitous and typically do not play a role in its output changing. The precise structure of a regulatory region is visualized in a "quintessential diagram", as seen in figure 2. We call such an image the quintessential diagram because it displays at a glance the most fundamental knowledge we have of the organization of a gene's regulatory region. When searching for papers with information relevant to our database, we found that no useful paper lacks this diagram.

The sites within Module A carry out a complex logic program. Sites CG_1 and P (highlighted green in figure 3) communicate the effects of Module B to the transcription apparatus–if either of them is disabled (by mutating the sequence), the gene expression is as if Module B were not present at all. Site Z (highlighted red) communicates the effects of Modules DC, E, and F. Sites CG_1, CG_2, and CG_3 (highlighted blue) together boost the final output of the module. If even one of them is disabled, then the boost is completely removed.

Module B was shown to have a similar complexity [25]. In fact, all *cis*-regulatory modules execute information processing of their inputs, and the sum total of the computations performed by the regulatory regions of individual genes, when considered within the complex network established by the dependencies among genes, forms a "genomic computer" [13].

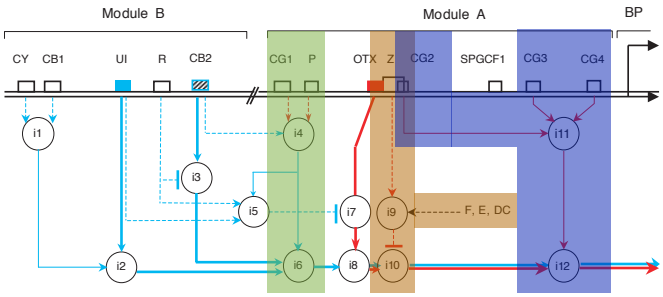


Figure 3: Computational logic model for Modules A and B of *endo16* (from [25])

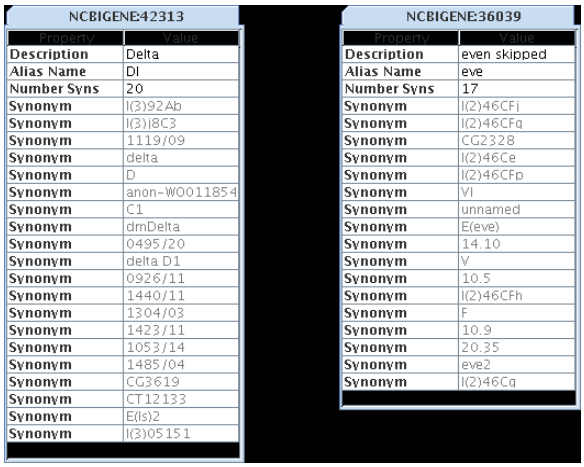
3 The Gene Naming Problem

> *"It all comes, I believe, from trying to define the indefinable."*
> – Charles Darwin [29]

> *"Scientists would rather share a toothbrush than use the same name for the same gene."*
> – Anonymous

As one tries to compile information from the literature into a database for automatic analysis, one of the issues that immediately crops up is that of naming. If information was inserted verbatim, then in order to find something one would need to use the exact terminology of the author. It would also be impossible to bring related information together automatically, unless large "translation" tables were compiled manually, which would suffer from problems with both completeness and precision. Completeness, because the addition of a new term in the database would require a new entry in the translation table, which could easily be left out by accident. Precision, since many terms are ambiguous when taken out of context even if they were completely obvious in the original setting. Errors in completeness lead to missed relationships, while error in precision lead to spurious ones. Therefore, it is clear that a standard set of terms must be developed that all of the descriptions given in the literature can be translated to at the time the information is added to the database. This does lead to some loss of information, as the translation will always be imperfect, but the set of terms can always be expanded if systematic problems are noticed. A prime example of such a controlled vocabulary is the Gene Ontology, which describes gene and gene product attributes[2]. In the course of developing the *cis*-Lexicon, we composed standard terms for describing regulatory function, such as activation, repression, and *cis*-regulatory module linkage, called the *Cis*-Regulatory Ontology. We are in the process of defining a similar vocabulary for describing the functional interactions among transcription factor binding sites.

In all of the above cases, the translation of terms to a standard vocabulary is relatively straightforward, because the properties described are clearly distinct



NCBIGENE42313

Property	Value
Description	Delta
Alias Name	Dl
Number Syns	20
Synonym	l(3)92Ab
Synonym	l(3)j8C3
Synonym	1119/09
Synonym	delta
Synonym	D
Synonym	anon-WO011854
Synonym	C1
Synonym	dmDelta
Synonym	0495/20
Synonym	delta D1
Synonym	0926/11
Synonym	1440/11
Synonym	1304/03
Synonym	1423/11
Synonym	1053/14
Synonym	1485/04
Synonym	CG3619
Synonym	CT12133
Synonym	E(ls)2
Synonym	l(3)05151

NCBIGENE36039

Property	Value
Description	even skipped
Alias Name	eve
Number Syns	17
Synonym	l(2)46CFi
Synonym	l(2)46CFq
Synonym	CG2328
Synonym	l(2)46Ce
Synonym	l(2)46CFp
Synonym	VI
Synonym	unnamed
Synonym	E(eve)
Synonym	14.10
Synonym	V
Synonym	10.5
Synonym	l(2)46CFh
Synonym	F
Synonym	10.9
Synonym	20.35
Synonym	eve2
Synonym	l(2)46Cq

Figure 4: Synonyms for *dl* and *eve*

and understood, regardless of the terminology–if one paper says a transcription factor "increases expression" and another says a factor "boosts expression", it is very clear that both phrases refer to our canonical term "activation". A much more difficult problem is the wide variety of names given to a single gene in the literature. A peek into GenBank for a well studied gene often yields more than 15 aliases–see figure 4 to see the synonyms for *dl* and *eve* of *D. melanogaster* displayed in the *cis*-Browser. To attempt to use tables of gene synonyms to determine which gene is really "meant" by a name can only lead to the troubles outlined in our discussion of translation tables earlier.

The *Gene Naming Problem* arises from the many techniques biologists use to study genes: examining a gene from the phenotype of a mutation, linkage map, or similarity to other known genes all yield different views and consequently different names. Given the genomic sequence, the obvious unifier would be the gene's locus, but this is not always known. In addition, even the definition of a gene and its locus is in flux[11].

When a gene is named according to its similarity with other known genes, different conventions lead to similar but separate names, such as *oct-3* versus *oct3*, which can both be found in the literature. One method which is less widespread is to prefix a homolog with an indicator of the new species, such as *spgcm* being a homolog of *gcm* in *D. melanogaster*. This method can be impaired by the difficulty in determining what qualifies as a new species – at least 26 definitions have been published [29].

Occasionally, different transcription factors can bind to the exact same sequence. This is especially likely for factors in the same family which share an evolutionary history–the DNA binding motif may not have diverged much between them. In this case, when a sequence is found to be a functional site, for example by a gel shift assay or perturbation analysis, it is not clear what factor is actually binding there. It's also possible for two putative binding sites

to overlap, where it is not clear whether only one or both play a role in gene regulation. This can play havoc in determining the function of individual sites.

When biologists attempt to determine which factor is binding to a site by finding the molecular weights of proteins isolated by gel shifts or other similar techniques, often several weights are found. This can be caused by measuring different subunits of a single protein, alternative splicings of a single gene, different translation products from a single mRNA molecule, or even completely distinct factors. For example, when the structure of *endo16* was first mapped, 8 of the 14 inputs tentatively identified were found with more than one molecular weight[26]. Only by cloning the cDNAs encoding the factors can it be determined whether the multiple weights are from the same gene or not. The molecular weight can be thought of as a type of hash code for a protein: its value depends on the protein's sequence, but contains less information. If two proteins have a slightly different weight, they can still be from totally unrelated genes and their sequences can be very different. Analogously, even protein products from the same gene can have very different weights, if they differ in terms of splicing (entire exons added or subtracted which might not affect the final function a great deal). Biologists depend, however, on the assumption that proteins from different genes will have different weights. This may be true with very high probability, but it cannot be universally true, especially since our methods of measurement are always imperfect. This is similar to how unique identifiers are created and used in computer science: IDs are given by random number generators from a large enough range (e.g., 64-bit numbers) that the probability of a clash is almost zero.

4 The Consensus Sequence Bottleneck Problem

> *"[E]ssentially all predicted transcription-factor (TF) binding sites that are generated with models for the binding of individual TFs will have no functional role."*
>
> – Wasserman and Sandelin[23]

Given the wide variety of sequences that a transcription factor binds to, there is no straightforward yet accurate model. The earliest model, consensus sequence, while still preferred by most biologists, suffers from being forced to choose either selectivity (matching only known site sequences) or sensitivity (matching all known site sequences)[21]. As almost any pair of binding sites for a single factor differ in at least one base (see [19] for a study where out of seven binding sites for the same factor, only two sites have identical sequence and only three match the known consensus in all positions), the consensus sequence model allows for differences in two ways: (1) permitting a range of bases in certain positions through symbols such as Y (allowing C, T, or U); and (2) permitting a set number of global mismatches, such as allowing one or two bases which do not match the consensus at all. Both methods greatly increase the number of hits to random sequence: replacing an A with a R or V increases

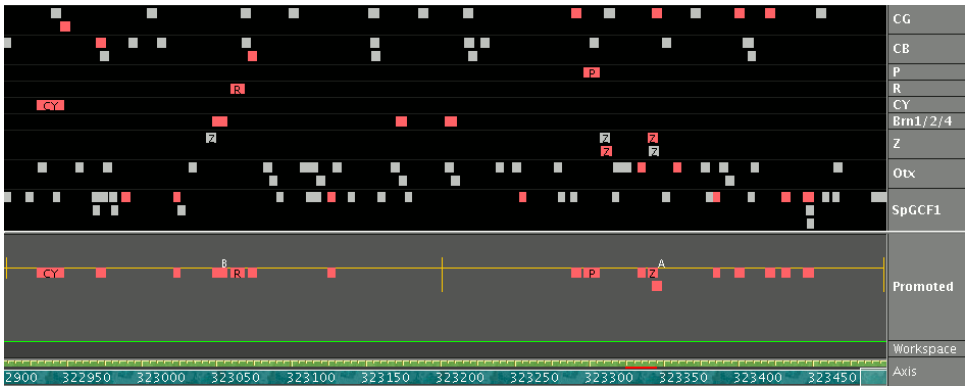


Figure 5: Matches for known binding site sequence in *endo16*

the probability of a match by chance by a factor of 2 or 3, respectively. Allowing for 1 or 2 mismatches anywhere in a consensus sequence increases it by a factor of up to 4 or 16.

Imagining the set of all DNA sequences of a single length to be a coordinate space, these two methods of allowing for differences act to create a bubble whose contents are the sequences matched by the model. The more differences allowed, the larger the bubble. The assumption is that since the set of sequences bound by a transcription factor are evidently similar in some unclear way, they should be clustered in this coordinate space and we should be able to find a "bubble" which contains this cluster tightly–it should contain all the sequences in the set, and none outside it. In practice, however, this is not the case. Stormo showed that given just 6 sites in *E. coli*, in order to match all of them the "consensus" sequence must be so vague that a match is found by chance every 30 bp[21]. It's clear that at such levels of generality, the sets accepted by consensus sequence models begin to overlap as well–the distinction between binding sites for individual factors effectively disappears.

To demonstrate the futility of searching for individual sites using known binding site sequences, we examined the gene *endo16* from *S. purpuratus*. Two *cis*-regulatory modules of this gene have been studied, yielding 13 unique transcription factor inputs binding to 17 sites[24, 25]. We searched for additional sequences which look like binding sites for these same 13 factors within the two modules by searching for sequences like those we have recorded, except allowing for one single mismatch. For one of the inputs, *otx*, we knew of 4 other sites in other genes (*blimp1/krox*[16] and *otx* itself[28]), and we included those in our search, yielding 3 unique site sequences total for *otx*. For another input, *brn1/2/4*, we knew of one other site in *blimp1/krox*[16], so we included its sequence as well. The result of our search as visualized in the *cis*-Browser can be seen in figure 5. Exact matches are highlighted in red, and matches with one mismatch are drawn in gray.

The second major model of transcription factor binding sites is the position

weight matrix (PWM). By incorporating not only the bases known to appear at each position but also their probabilities of occurrence, a much more precise model is made. While the independence/additivity assumptions are imperfect, they are a good approximation of reality, especially for the simplicity of the model–while there are a few notable exceptions, most factors are described relatively well by a PWM [3]. The logarithm of the observed base frequencies has been shown to be proportional to the binding energy contribution of the bases [4], so there is clear biological significance to using these values as the weights of the matrix. But a matrix of coefficients is not sufficient to predict binding sites–there is still the question of the best cutoff score. Unlike a consensus sequence, a PWM assigns a *score* to every sequence, and a cutoff must be chosen as the minimum score to expect a sequence to be a functional binding site. Even with a cutoff chosen to best match the experimentally known sites, it is unclear how "good" a site is if it is barely over the limit, or how nonfunctional a sequence is if it is just under the cutoff. As true binding sites are not merely "on" or "off" but have an effect whose degree may depend on the strength of the bind, the concept of a cutoff may be incorrect.

4.1 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, either directly or indirectly. Therefore many 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), in principle we should be able to find such a 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 [18], while others use a probabilistic or information-theoretic framework [14].

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. We can exclude the protein-coding sequence from our analysis, either by using predicted gene models (e.g., [10]) or by transcriptome analysis (e.g., [20]), and only look the conserved patches of unknown function, which are likely to contain regulatory sequence (see, for example, [27, 17, 19, 1]). For the highest accuracy, several species can be compared simultaneously [5].

4.2 Evaluations of Motif-Finding

Existing motif algorithms perform reasonably well for yeast, but not for more complex 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" [22, 15]. For an overview of the algorithms and the models they are based on, see [7].

5 The Logic Function Inference Problem

> *"Development of Western Science is based on two great achievements—the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)."*
>
> – Albert Einstein

The observed expression of a gene in space and time is not determined by a single transcription factor binding site, but by a collection of sites whose effects are combined in nontrivial ways. Istrail & Davidson catalogued the known site interactions into 4 categories: transcriptional activation operators, BTA control operators, combinatorial logic operators, and external control operators[12]. To fully understand the regulation of any gene, we must (1) identify all binding sites, (2) understand the function of each site, and (3) understand the rules for combining their functions to infer the overall *cis*-regulatory output[8].

As discussed above, step 1 alone is an especially difficult problem, since knowledge of transcription factors is largely biased toward one category: the "drivers". The drivers are those transcription factors whose expression varies in space and time, and thus determine the expression of the genes they regulate. Other factors are more or less ubiquitous–they do not vary, but instead are always present, which makes their presence harder to detect. Only thorough studies such as perturbation analysis can expose them.

Perturbation analysis is currently the only method for the determining the functional interactions among modules and sites. Biologists take the known

regulatory sequence of a gene and modify it to determine the function of each part. When several modules are known to exist, a high level understanding of the function of each module and how they interact together can be found by observing the expression caused by different subsets of the full array of modules. For example, to discover the function of the modules of *endo16*, Yuh and Davidson tried each of the six modules individually and in many different combinations (A, B, DC, E, F, FA, EA, DCA, FB, EB, DCB, etc)[26].

When putative sites are identified, biologists can tease out the complex hidden logic underlying even the simplest regulatory region by mutating both individual sites and groups of sites, each in independent experiments. This "regulatory archaeology" is performed one experiment ("dig") at a time–after each experiment, possible models are postulated, and the next perturbation is chosen based on the results. When the *S. purpuratus* gene *spgcm* was studied by Ransick and Davidson ([19]), they identified three putative binding sites for Su(H) to investigate. They recognized that two of the sites were arranged in a manner similar to a pattern familiar from studies in *Drosophila* and mouse, so they treated the two as a single feature. To determine the precise function of the sites, they observed the expression resulting from mutating all three, only the pair, and only the third single site. Even with only two elements to examine, they found a non-trivial logic behind their function: having either element yielded nearly the full correct activation, while one element in particular (the solo site, not the pair) was clearly necessary for correct repression (mutating the pair resulted in impaired repression but did not remove it entirely).

There are a two obvious strategies for choosing the next construct when in the midst of a series of experiments: pick the perturbation whose results can rule out the greatest number of putative models (i.e., binary search); or pick the perturbation whose expected results would confirm the experimenter's current hunch (i.e., an expert heuristic). The better the algorithm, the fewer the experiments that need to be performed to uncover the regulatory logic–and that means less time and less cost.

The intermediate results of perturbation analysis is visualized as a "quintessential graph": with one row for each reporter construct tested, bars are drawn to show the expression and/or repression of each construct relative to the control, which contains the full regulatory region sequence.

It can be difficult even to test whether a model is correct or not–the measured output, gene expression, is not a simple value. It varies from organism to organism, and it cannot be characterized as a single number, but at best as a distribution with variance. This variance can cause the ranges of expression for perturbation constructs to overlap even when they are significantly different. More experiments would yield tighter distributions, but the cost can sometimes not be afforded. At times, biologists must choose the next construct to test and move on without achieving *statistically* significant results, although the results they achieved might have been significant to them in view of their domain knowledge.

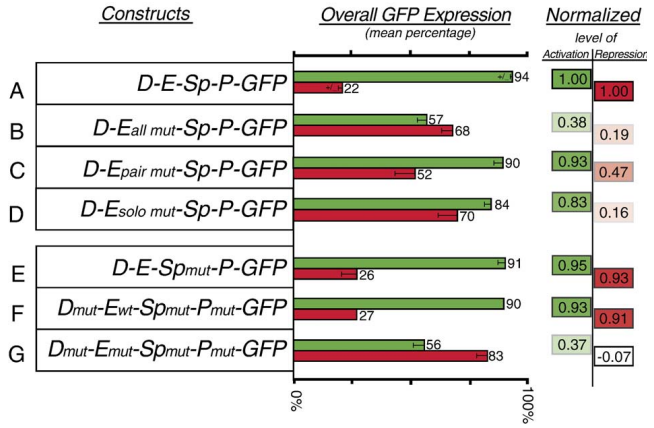


Figure 6: Quintessential graph (from [19])

6 Conclusion

The concept of *cis*-regulatory information abstracts the wealth of types of data required for designing algorithms for computationally predicting the most likely organization of a piece of regulatory DNA into *cis*-regulatory modules. In this paper we presented three problems that highlight the computational difficulties encountered in extracting such information. Two other sources of *cis*-regulatory information of fundamental importance are: (1) gene regulatory networks architecture (e.g., the endomesoderm network [9]) , and (2) genomic data from species at the right evolutionary distance that preserve only the *cis*-regulatory modules (e.g., *S. purpuratus* and *L. variegatus*). The challenge of extracting *cis*-regulatory information is much amplified by the relatively scarce data sets and the depth of the analysis required to unveil such rich information sources.

Acknowledgments

We thank Eric Davidson for many inspiring discussions and other contributions to this project. This research was supported by National Science Foundation Award 0645955.

References

[1] Gabriele Amore and Eric H Davidson. cis-regulatory control of cyclophilin, a member of the ets-dri skeletogenic gene battery in the sea urchin embryo. *Developmental Biology*, 293(2):555–64, May 2006. PMID: 16574094.

[2] 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.

[3] 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.

[4] 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.

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

[6] Sea Urchin Genome Sequencing Consortium, Erica Sodergren, George M. Weinstock, Eric H Davidson, R. Andrew Cameron, Richard A. Gibbs, Robert C. Angerer, Lynne M. Angerer, Maria Ina Arnone, David R. Burgess, and et al. The genome of the sea urchin strongylocentrotus purpuratus. *Science*, 314(5801):941–952, Nov 2006.

[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] EH Davidson, JP Rast, P Oliveri, A Ransick, C Calestani, CH Yuh, T Minokawa, G Amore, V Hinman, C Arenas-Mena, and et al. A genomic regulatory network for development. *Science*, 295(5560):1678, 1669, Mar 2002.

[10] Christine Elsik, Aaron Mackey, Justin Reese, Natalia Milshina, David Roos, and George Weinstock. Creating a honey bee consensus gene set. *Genome Biology*, 8(1):R13, 2007.

[11] Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korbel, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-encode? history and updated definition. *Genome Res.*, 17(6):669–681, Jun 2007.

[12] 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.

[13] Sorin Istrail, Smadar Ben-Tabou De-Leon, and Eric H. Davidson. The regulatory genome and the computer. *Developmental Biology*, 310(2):187–195, Oct 2007.

[14] 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.

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

[16] Carolina B Livi and Eric H Davidson. Regulation of spblimp1/krox1a, an alternatively transcribed isoform expressed in midgut and hindgut of the sea urchin gastrula. *Gene Expression Patterns: GEP*, 7(1-2):1–7, Jan 2007. PMID: 16798107.

[17] Takuya Minokawa, Athula H Wikramanayake, and Eric H Davidson. cis-regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. *Developmental Biology*, 288(2):545–58, Dec 2005. PMID: 16289024.

[18] 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.

[19] 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.

[20] Manoj P. Samanta, Waraporn Tongprasit, Sorin Istrail, R. Andrew Cameron, Qiang Tu, Eric H. Davidson, and Viktor Stolc. The transcriptome of the sea urchin embryo. *Science*, 314(5801):960–962, Nov 2006.

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

[22] 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.

[23] Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr 2004.

[24] C. H. Yuh, H. Bolouri, and E. H. Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279(5358):1896–1902, March 1998.

[25] 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.

[26] CH Yuh and EH Davidson. Modular cis-regulatory organization of endo16, a gut-specific gene of the sea urchin embryo. *Development*, 122(4):1069–1082, Apr 1996.

[27] Chiou-Hwa Yuh, C Titus Brown, Carolina B Livi, Lee Rowen, Peter J C Clarke, and Eric H Davidson. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Developmental Biology*, 246(1):148–61, Jun 2002. PMID: 12027440.

[28] Chiou-Hwa Yuh, Elizabeth R Dorman, Meredith L Howard, and Eric H Davidson. An otx cis-regulatory module: a key node in the sea urchin endomesoderm gene regulatory network. *Developmental Biology*, 269(2):536–51, May 2004. PMID: 15110718.

[29] C. Zimmer. What is a species? *Scientific American Magazine*, 298(6):72–79, 2008.