BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 2. 122–131
Advance Access published on 11 August 2010
doi:10.1093/bib/bbq060

Evolution of gene regulation—on the road towards computational inferences

Georg Fuellen

Submitted: 28th April 2010; Received (in revised form): 13th July 2010

Abstract

If fragments of DNA are transcribed (expressed), they deserve to be called (parts of) a gene. Whether transcription takes place depends on the 'gene regulatory network'. This network is defined as the complex interplay of the sequence, biochemical modifications and structure of the chromosomal DNA with the regulatory proteins/RNA (transcription factors, co-factors, regulating RNA and the transcriptional apparatus itself). Gene regulatory networks play a role in various stages of development as well as in the maintenance of the organism; in this review we will concentrate on the former. Their evolutionary reconstruction is daunting (to say the least), and bioinformatics tools are in their infancy. However, gain of understanding offers a reward beyond itself, since evolutionary considerations can enable discoveries in the first place, e.g. the computational identification of conserved transcription factor binding sites. We discuss the evolution of gene regulation in the context of the 'Genetic Theory of Morphological Evolution' as described by Carroll, identifying those parts of the theory that are relevant for bioinformatics, and their implications. We discuss the important question of how bioinformatics analysis results on the evolution of gene regulation may be validated. Finally, we briefly exemplify use of the UCSC genome browser, exploiting its pre-computed alignments to describe the evolution of gene regulation.

Keywords: *gene regulation; evolution; regulatory region; gene regulatory network*

BIOLOGICAL BACKGROUND

DNA and the network of regulators

Chromosomal DNA can be represented by a string of nucleotides. In a genome browser such as UCSC [1], it serves as the *x*-axis across which its features can be visualized, as in Figure 1. DNA includes transcribed parts (genes), which are often used as blueprints for proteins, and a large set of 'regulatory elements'. These elements, that is their sequence of nucleotides, their modification (such as methylation) and structural accessibility decide in part about the timing and the amount of successful transcription [2]. Successful transcription means that RNA is produced—the gene is expressed. The RNA is usually processed further. Transcription also depends on a multitude of other factors, which may be represented by a network of interacting proteins and RNAs. Among them are transcription factors, microRNAs and the proteins of the transcriptional apparatus, which produces the RNA, given the DNA template. This 'network of regulators' is dynamic in space and time. The concentration of the network components is subject to influences that are internal to the nucleus/cell (after all, transcription factors are transcribed themselves and they may also regulate their own transcription), external (environmental, driven by neighboring cells), and/or stochastic (due to effects that happen at random).

Regulatory elements and events

Since DNA can bend and may form loops in 3D space, the linear sequence of the regulatory elements, which are found before, within or after the gene, does not necessarily tell us much about their

Corresponding author. Georg Fuellen, Institute for Biostatistics and Informatics in Medicine and Ageing Research – IBIMA, University of Rostock, Medical Faculty, Ernst-Heydemann-Str. 8, 18057 Rostock, Germany, Tel: +49-381-494-7360; Fax: +49-381-494-7203; E-mail: fuellen@uni-rostock.de; fuellen@alum.mit.edu

Georg Fuellen is professor of Medical Bioinformatics and director of the Institute for Biostatistics and Informatics in Medicine and Ageing Research, Department of Medicine, Rostock.

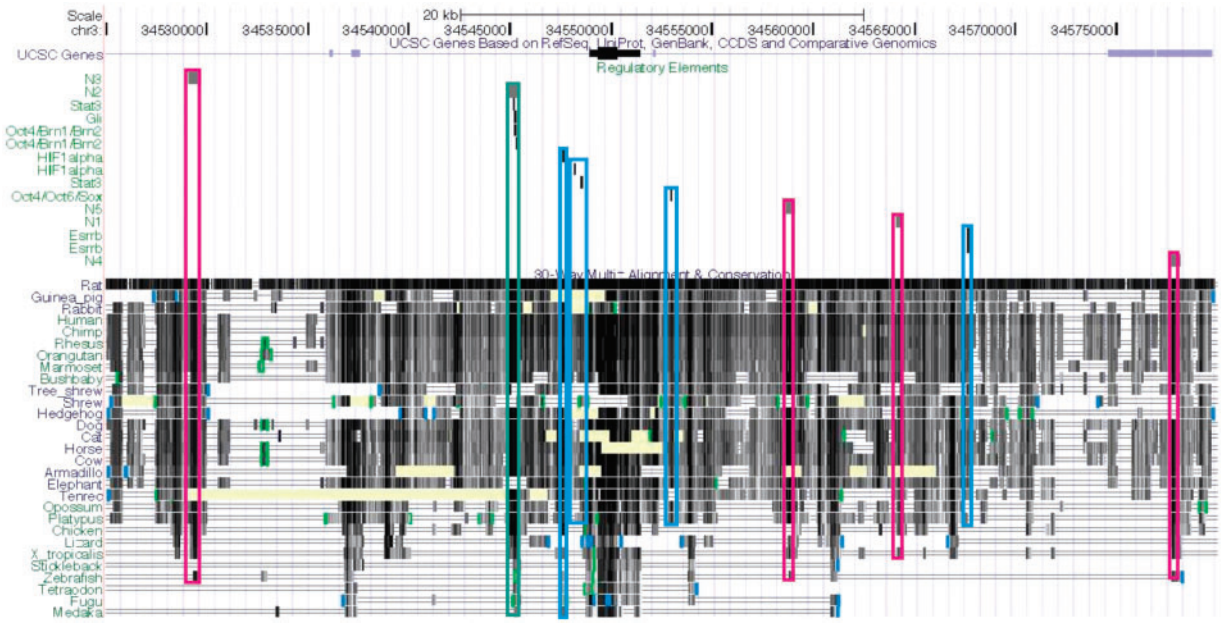

Figure 1: The Sox2 regulatory region, displayed using the UCSC genome browser [1]. The first two tracks display the scale bar and the chromosomal positions. The 'UCSC genes' track displays the Sox2 gene itself (in black), and parts of the Sox2 overlapping transcript [62] (in blue). Information on experimentally validated regulatory elements is displayed thereafter, using grey blocks and green text. The last tracks display alignment quality as grayscale density. UCSC convention is that yellow regions denote consecutive Ns (lack of sequence) and double lines denote unalignable bases. Red boxes mark binding sites involved in neural regulation, blue boxes mark binding sites involved pluripotency, and the green box marks the binding sites of the N2 region involved in both.

mutual interaction and their influence on regulation. Nevertheless, one can organize the elements into so-called modules, often termed CRMs, *cis*-regulatory modules. Furthermore, one distinguishes regulatory elements that are distal or proximal to the transcription start site. If transcription factors bind to (some of) these elements, transcription may be started or enhanced, reduced or silenced. Importantly, depending on the network of regulators that is active at a given time, the same elements may trigger a different, and sometimes opposite, effect. Thus, the typical regulatory region of a gene includes an array of regulatory elements, which may be enhancers or silencers. Closest to the transcription start site are the core and the proximal promoter, followed by distal elements. Here, the core promoter is the minimal portion of the promoter required to properly initiate transcription. The proximal promoter includes specific transcription factor binding sites (TFBSs) up to ~250-bp upstream of the transcription start site; these are also known as proximal elements. Distal elements are binding sites >250-bp upstream. Some components of the network of regulators form tight complexes called 'enhanceosomes', which bind to regulatory elements, in a competitive or cooperative fashion. A small change (e.g. the gain, loss, exchange or molecular modification of one component) may turn them into 'repressomes' [2]. The exact composition of the complexes depends on the concentration of their components in the nucleus [3]. The affinity of the complexes to the regulatory elements depends on their composition. In turn, affinity also influences complex composition, if some components bind to the DNA before the complex is assembled ([4], Box 3 and Figure 2 therein). Such variation in enhanceosome buildup, modification and binding starts to blurr the traditional distinction of tight and rigid 'enhancesome' binding sites on one hand and loose billboard-like sets of binding sites on the other hand. Moreover, bound complexes may move along the DNA, before contributing their effect to gene regulation [5]. Most likely, this movement is also influenced by a variety of factors. Even breakup of the DNA has been implicated in regulatory events [6, 7]. Once DNA is transcribed, the stability of the transcript depends on a multitude of factors. Finally, it may (or may not) be translated; these two aspects of 'gene regulation' are outside the scope of this review.

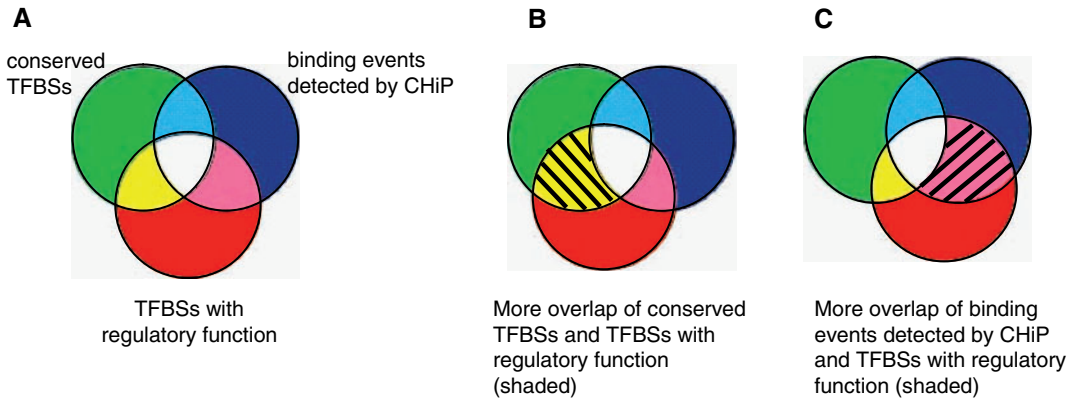

Figure 2: (**A**) The overlap between conserved transcription factor binding sites (TFBSs), binding events detected by ChIP, and TFBSs with regulatory function is not known with precision. In particular, the shaded overlap may be larger on the left- or on the right-hand side [panels (**B**) versus (**C**)]. In case (B), evolutionary analyses are more rewarding than in case (C).

The impression of a computer scientist after reading the literature is not just that gene regulation is extremely complicated, but there is also a plethora of unknown entities and relationships that still need to be elucidated. Considering this complexity *in silico* is a hard task. Thus, good news for bioinformatics is the few papers revealing simple rules. For example, there does not seem to be much interference of the regulatory effects of neighboring regulatory elements. Instead, effects seem to be additive [8]—an island of simplicity in an ocean of complexity. But whether such additivity holds in general remains to be seen. After all, complexes of transcription factors may still interfere with each other, if binding sites are sufficiently close.

The gene regulatory network and its evolution

The network of regulators on one hand and the regulatory elements on the DNA level on the other hand form the 'gene regulatory network'. In this terminology, evolution of gene regulation is concerned with the evolution of the gene regulatory network and its components. To begin with, there are studies of the evolution of transcription factors and co-factors (e.g. [9, 10]) and their interactions (e.g. [11, 12]), of regulatory RNA (e.g. [13]) and of the transcriptional apparatus itself (e.g. [14, 15]). Consideration of the evolutionary interplay (co-evolution) of the various components of gene regulation should increase the success rate of computational evolutionary reconstructions (see below). At least some of these components can be traced back to the roots of life [10], but details are the more nebulous the more we move back in time. In this review, we will concentrate on developmental gene regulatory networks. Davidson and Erwin [16] divide these into the following:

- 'network kernels' regulate general aspects of development (e.g. heart development) and are conserved across phyla,
- 'plug-ins and I/O switches' are concerned with developmental subcircuits (e.g. signaling) and are conserved within (sub)phyla, and
- 'differentiation gene batteries' execute the final developmental readout and are often only conserved in groups of closely related species.

As discussed below, good judgment in selecting the appropriate set of species is important for the success of computational inferences, because these different kinds of gene regulatory networks tend to be conserved for different sets of species.

The evolution of some regulatory elements on the chromosomal DNA can be traced back to the origin of the vertebrate lineage. In particular, there are still short conserved regulatory elements in lamprey [17], the earliest diverging extant vertebrate lineage. Evolution of many binding sites is due to mutations, insertions and deletions of nucleotides, and due to transposable elements [18]. Their volatility can lead to high turnover of binding sites in some cases, e.g. reducing conservation of Oct/Sox binding sites in rodents [18]. Binding (of a transcription factor), regulatory effect, and evolutionary conservation of the binding site are observations that may (or may not) co-occur (see Figure 2, and below). In case of human, the Encode pilot project [19] found no function in 40% of the conserved sequence regions, and

no conservation (across mammals) in 50% of the functional elements, see also [20].

Towards a theory of gene regulatory network evolution

The design and evaluation of computational analyses usually benefits from some 'theoretical' understanding. In this case, such understanding consists of general principles that are observed when we inspect gene regulatory networks in today's species, compare them, and try to come up with most parsimonious (or, most likely) explanations for our observations. For gene regulatory networks involved in development, Carroll [21] derived the following principles, which can motivate and guide computational analyses.

- 'Mosaic pleiotropy' and 'heterotopy' reflect that developmental regulators participate in a multitude of processes; they are promiscuous in time and space. Therefore, computational analyses must consider that data (e.g. on transcription factor binding and gene expression) and subsequent results/conclusions are dependent on space (tissue) and time. Generalizations along these two dimensions are often not permitted. However, we can study how this regulatory diversity may have evolved.

Bioinformatics developers have to be aware of the limitations that can be expected due to these two attributes of gene regulatory network evolution. Pleiotropy (for example, the co-option of regulatory elements for some novel biological phenomenon) may invalidate any conclusions of a straightforward evolutionary inference, but without background knowledge, it cannot easily be inferred.

- 'Ancestral genetic complexity', 'deep homology', 'functional equivalence of distant homologs' and 'infrequent toolkit gene duplication' are four principles, on which computational evolutionary analyses of (developmental) gene regulation rely. Basically, these principles reflect the existence of conserved 'network kernels' ([16], see above). More specifically, ancestral genetic complexity refers more or less directly to these kernels; Carroll writes about 'similar toolkits'. Without ancestral complexity, (computational) evolutionary inferences would stop early on when going back in time, because different entities of today would map to the same ancestral ones. Developmental processes such as heart formation are governed by deeply conserved homologous gene regulatory networks, enabling 'deep' computational inferences. Often, the 'toolkit proteins' are distant homologs that can nevertheless functionally substitute for one another, and they do not tend to be subject to duplication, possibly because developmental processes are rather sensitive to gene dosage. The latter two properties can simplify evolutionary analyses of function and regulation.

The four 'enabling' principles are at the same time challenging bioinformatics developers, in three ways. First, for homology to hold and to be useful, the right set of species must be selected. If the species are too closely related with respect to the regulatory process under study, no evolutionary steps can be inferred. Such information may of course still be useful; for example, high conservation of a regulatory network between rhesus monkey and human may be important for a pharmaceutical application. If the species are too far apart, homology becomes undetectable (if it exists at all) and evolutionary reconstruction cannot be successful. Most importantly though, species must also be selected based on the ready availability of reliable data. Second, integration of auxiliary data on the experimentally validated functional equivalence of homologous proteins can support statements on high evolutionary conservation of the role of these proteins in regulation. However, there is a lack of databases specialized on such data. Third, data on the duplication history of relevant genes can be of direct relevance to the process of evolutionary reconstruction. Such data are available at the NCBI Homologene database, and at the EBI Ensembl website.

- 'Modularity of *cis*-regulatory elements' and 'Vast regulatory networks' are principles that re-iterate the complexity of (developmental) regulation. Pleiotropy of the 'toolkit proteins' and their corresponding heterogeneous expression in time and space is made possible by a choice of regulatory elements, which a large gene regulatory network selects from. Vast networks are also a consequence of the pleiotropy of the downstream regulators which they have to control. To some degree, pleiotropy implies conservation 'in *trans*': the overall profile of the binding sites of a transcription factor

is conserved because a change would have a multitude of effects. In turn, changes 'in *cis*' tend to be non-catastrophic: they just affect the regulatory element of a single gene. Nevertheless, the debate on '*cis* versus *trans*' continues [22–24] and computational models of the evolution of regulation depend to some degree on its conclusion. Most likely the conclusion is a synthesis in the end, i.e. '*cis* and *trans*': Changes in the transcription factors themselves and in the regulatory regions of the genes probably play distinct yet overlapping roles in evolution.

For computational inferences, it would be most interesting and useful to explicitly consider the co-evolution of the regulators (transcription factors, etc) and their network on one hand, and of the regulatory elements on the other. For example, the duplication or mutational change of a transcription factor may go hand-in-hand with a partial re-organization of the regulatory regions of its target genes.

In Figure 3, the eight principles and their relevance for the computational inference of gene regulation are summarized, including some open questions and the bioinformatics challenges related to these principles. In summary, Carroll's observations imply that bioinformatics analyses of the evolution of gene regulation are not futile: there is something to be discovered due to ancient conservation, even though the complexity of the phenomena (and the high amount of volatility and noise) render the task difficult, especially if we want to reconstruct events that happened millions of years ago.

COMPUTATIONAL APPROACHES

Evolutionary bioinformatics and gene regulation

Tools and software for estimating, analyzing and/or visualizing the evolution of gene regulation are rare. In the next section we will describe the few approaches that we are aware of. Some aspects can be analyzed with standard tools, though. Using methods such as maximum likelihood, parsimony or Bayesian inference, sequence data are used to estimate the phylogeny of transcription factor families, transcriptional co-factors, some regulatory RNAs and the components of the transcriptional

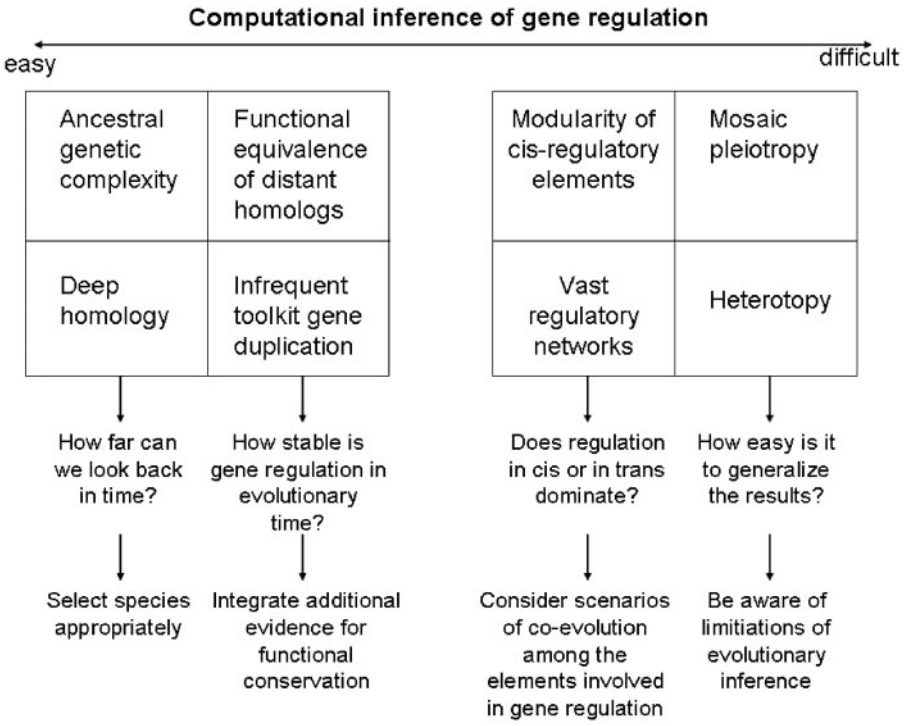

Figure 3: Carroll's eight principles and the computational inference of gene regulation. For example, such inference is eased by validity of Ancestral genetic complexity and Deep homology, and made difficult by validity of Mosaic pleiotropy and Heterotopy. At the bottom, some open questions and bioinformatics challenges pertaining to the principles are listed.

apparatus itself (see above). One caveat is that these inferences just consider sequence evolution, ignoring the evolution of e.g. post-translational modifications of the regulators, of their variation due to alternative splicing, etc. Moreover, the inference error is usually the higher, the earlier the events one wishes to infer. The complex interplay of the regulators is even harder to trace back in time. There are a few automated approaches available to estimate the evolution of networks [25, 26], using a probabilistic inference framework. Inferring the evolution of regulatory elements is also very tricky. Here, most often we do not even have a good data set of elements in today's species to start with! In particular, the few databases of experimentally validated sites in metazoa/vertebrates (such as ORegAnno [27] and Pazar [28]) only cover a small fraction of what is known from the literature (they feature <10% of the sites curated for the case study below), which is only a very small fraction of what is there. Moreover, in contrast to the regulators, regulatory elements have low information content (binding sites feature a length of 4–20 bases, approximately), making their reliable *in silico* detection exceedingly difficult. Their experimental detection is also fraught with problems, because binding *in vivo* (with a subsequent regulatory effect!) is dependent on context, as described above. Sometimes, *in vitro* detection of the binding of a regulator (transcription factor) using an antibody (ChIP, see [29, 30]), often together with gene expression data documenting the up/downregulation of the regulated gene depending on the expression of the regulator, may provide a convincing story for a regulatory effect, but there is no proof. For example, the antibody may have picked up a protein interacting with the true regulator [31], and the expression data may just describe correlation, not causality. As discussed in the section on validation issues, ChIP-based TFBS data may be no better than computational predictions of conserved binding sites, if regulatory effect (and not binding) is being asked for.

Many TFBS prediction tools exploit libraries of known binding motifs *and* evolutionary conservation, and usually they infer sets of related sites (CRMs, *cis*-regulatory modules, see above). These modules are believed to be bound by sets of transcription factors corresponding to enhancesomes (see above). Methods for detecting CRMs have been reviewed recently in this journal [32]. Based on the simple idea that conservation goes with functional importance, 'phylogenetic profiling' [33] suggests that predicted binding sites are the more likely to be functional, the more conserved they are. This basic idea comes with some problems; for example, conservation may also correlate with distance to the TSS [34]. Nevertheless, phylogenetic profiling has been developed further. In the approach by Kheradpour *et al.* [35], occurrence of binding sites in multiple species is weighted by the length of the corresponding branches in the species tree. Other recent advances in TFBS/module prediction include PReMOd [34] and CompMoby [36, 37]. The latter does not use libraries of binding motifs; instead, it identifies subsequences (words) that are overrepresented. Detailed models of binding site evolution, going beyond conservation scores, are employed by some authors, to improve binding site prediction and alignment (CSMET [38], EMMA [39], Monkey [40], PhylCRM [41], eSimAnn [42]). Some binding site identification approaches integrate a large number of sources of evidence, including but going beyond sequence data, evolutionary conservation, and/or ChIP data [43–45]. Finally, Fredman *et al.* [46] review (Web) resources to identify and study conserved regulatory regions in metazoa, and a very recent interesting review is given by Vingron *et al.* [47].

Computational analysis of the evolution of gene regulation

As described, the ground on which to base inferences of the evolution of gene regulation is shaky, but such computational analysis is still a worthwhile effort. For once, insights into the genesis of complex structures are interesting *per se*. Moreover, they can be useful since evolutionary insight can improve our understanding of today's data, as demonstrated below in case of a gene involved in regulation of pluripotency. Finally, and most importantly, evolutionary analyses may result in by-products, which are predictions about entities or relationships in today's species based on the evolutionary analysis. The simplest example of this kind of thinking is phylogenetic profiling, yielding predicted TFBSs, as described above. As far as the author is aware, the only tool attempting to directly infer the evolution of gene regulation from the DNA perspective (that is, the gain (and loss) of regulatory elements and modules in phylogenetic history) is ReXSpecies, the first version of which was published in 2008 [48]. Given a gene to be analyzed, binding sites are predicted in its conserved upstream or downstream region, for

multiple species. Then, the most parsimonious scenario for the evolution of these sites is computed and visualized, using a standard phylogenetic tree of the species. Finally, as by-products, sets of predicted binding sites (modules) are identified and ranked according to a measure that highlights the 'most interesting' ones. These may be binding sites gained or lost together in subtrees of the species tree. Apart from ReXSpecies, if a user intents to analyze the evolution of the regulatory elements of a gene, she/he can inspect the corresponding genomic region in a genome browser such as UCSC, and investigate species-specific conservation tracks. An example for such an analysis is provided towards the end of the article.

Validation of computational analyses of the evolution of gene regulation

A crucial aspect of computational analyses is their validation. There is no 'time machine', so how do we know that any advances in understanding are valid? An indirect solution is to test 'by-products'. If they are valid, the evolutionary inferences are supposed to have some validity as well. One 'by-product' is predictions about gene expression levels of regulated genes. However, correlations between gene expression levels of regulators and regulated genes are not necessarily indicating causality. Another by-product, predicted binding sites, may be validated by ChIP data. However, ChIP data describe binding, not regulative effect, so validation suffers from the 'conserved versus binding' dilemma: Conservation of binding sites may be an equally good, worse, or even better indication of regulatory effect than physical binding as measured by ChIP [35, 49–51], as visualized in Figure 2. For example, using insect muscle genes for validation, Stark *et al.* [50] reported that in the regulatory regions of muscle genes, evolutionarily conserved predicted binding sites of muscle-specific transcription factors were as enriched as binding sites of the same transcription factors found by ChIP. Moreover, Cheng *et al.* [52] studied GATA1 binding sites and found that conservation correlates with functional activity. Most recently, Balmer and Blomhoff [53] conducted a case study of experimentally validated retinoic-acid-related nuclear receptor binding sites observing that a superficial analysis reveals a conservation rate of 58%. However, specific consideration of over-predictions (i.e. changing the status of some binding sites to unvalidated based on careful evaluation of experimental evidence) and of compensatory evolution (i.e. counting a specific case of binding site turnover as a case of binding site conservation; see their paper for a discussion) yields a conservation rate as high as 94%. Although Balmer and Blomhoff believe that their case study covers a representative set of binding sites, further investigations are definitely necessary.

INSPECTING THE EVOLUTION OF GENE REGULATION USING THE UCSC GENOME BROWSER

In the following, we will briefly investigate the evolution of the regulation of the mouse Sox2 gene involved in pluripotency [54]. We exemplify the use of pre-computed multi-species alignments available at the UCSC Genome browser [1], supplemented by information on regulatory regions and TFBSs obtained from the literature. The pre-computed multiple alignment used is the 30-way Multiz alignment, including 30 species ranging from fish to human. The regulatory elements are taken from the literature, as listed below. The alignments are visualized in gray scale in Figure 1, where black blocks correspond to high similarity, and grey blocks correspond to low similarity (see the figure legend for more information). In general, the amount of similarity is highest between mouse and rat, and moderate between mouse and other mammals, but traces of conserved non-coding elements can still be found in fish. (The UCSC 30-way Multiz alignment may have some false positive data; these are regions deemed conserved while in fact they are not, due to misalignments. However, it definitely features a large amount of false negatives, due to misalignment and, more importantly, due to missing data.)

More specifically, the N2 region involved in neural regulation [55] as well as in pluripotency (including validated Stat3 [56], Gli [57] and Oct4/Brn1/2 [58] binding sites) is conserved up to fish. The other regions involved in neural development [55] (N3, N4, N5) are also found in fish (N1 can be traced back to Xenopus frog). In contrast, the binding sites involved in pluripotency, around the downstream auto-regulatory Oct/Sox binding site [59], the downstream Esrrb binding sites [60] and the proximal Stat3 [56] and HIF1alpha [61] binding sites, are found conserved up to platypus, with the exception of the first HIF1alpha binding site.

In summary, the hypothesis emerges that neural regulation of Sox2 is as old or older than regulation implicated in pluripotency. These two overlapping sets of conserved regulatory elements exemplify 'mosaic pleiotropy', 'heterotopy', 'ancestral genetic complexity', 'deep homology' and 'modularity of *cis*-regulatory elements' as described by Carroll.

Given experimental and/or predicted binding sites for a gene of interest, the reader can use the UCSC genome browser to perform analyses similar to the Sox2 analysis above. Towards this end, the gene needs to be located and displayed in the genome browser. After zooming out to display its putative regulatory region, information on the binding sites has to be added as a custom track, and it can then be correlated to the tracks describing evolutionary conservation. In case of human, a pre-computed track of predicted binding sites called 'TFBS Conserved' is part of the set of 'Regulation' tracks and it can be displayed directly. Given information on binding sites and on conservation, the evolutionary age of specific regulatory elements can then be estimated.

As discussed above, an appropriate selection of species and integration of auxiliary data (e.g. on functional equivalence of homologous transcription factors) increase the chance of generating correct inferences. In the long term, a pipeline inferring both the evolution of regulatory elements and of the network of regulators (transcription factors, etc.) is desirable. A statistical (e.g. Bayesian) approach toward modeling such a complex evolutionary scenario of co-evolution may be the way to go forward. Currently, many of these tasks can only be done manually: Species selection relies on expert knowledge. Data integration is only supported in part (but gene trees from the Ensembl website may be integrated quite easily). Finally, the study of co-evolution of regulatory elements and the network of their regulators is just beginning.

Key Points

- Bioinformatics tools and software for investigating the evolution of gene regulation are still in their infancy. One reason is the complexity of gene regulation itself, not to mention the additional complexity of the evolutionary processes acting on it.
- At the time of writing, investigators can check genes of interest using genome browsers, and find out about evolutionary conservation in the putative regulatory regions. Using ReXSpecies [48], they can obtain visualizations of parsimony-based reconstructions of the evolution of (predicted) TFBSs.
- Validation of reconstructions is possible in part by investigating their by-products, e.g. TFBSs highlighted by investigating the evolution of binding across a species tree. However, experimental validation of predicted binding sites, e.g. by ChIP data, suffers from imprecise binding site location data, and from the 'conserved versus binding' dilemma: It is not easy to assess whether binding found by ChIP or evolutionary conservation is the better indicator of regulatory effect.
- A theoretical framework would be very helpful. Towards this end, Carroll's article [2I] may lay a foundation. From its analysis, we derive four recommendations: (i) Select an appropriate set of species. (ii) Integrate auxiliary data such as functional conservation. (iii) Attempt to model the complex process of gene regulation including co-evolution of regulators and regulatory regions. (iv) Be aware of the various limitations of inferring the evolution of gene regulation, due to pleiotropy, heterotopy, etc.

Acknowledgements

Nitesh Singh and Stephan Struckmann helped with some of the figures.

FUNDING

Funding by the German Research Foundation (DFG), SPP 1356 'Pluripotency and Cellular Reprogramming' (FU583/2-1), and by the German Ministry for Education and Research (BMBF), 'Generation of pluri- and multipotent stem cells' (01GN0901) is gratefully acknowledged.

References

1. Kent W, Sugnet C, Furey T, *et al*. The human genome browser at UCSC. *Genome Res* 2002;**12**:996.
2. Maston G, Evans S, Green M. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006;**7**:29–59.
3. Pan Y, Tsai C, Ma B, *et al*. How do transcription factors select specific binding sites in the genome? *Nat Struct Mol Biol* 2009;**16**(11):1118–20.
4. Pan Y, Tsai C, Ma B, *et al*. Mechanisms of transcription factor selectivity. *Trends Genet* 2010;**26**(2):75–83.
5. Jeziorska D, Jordan K, Vance K. A systems biology approach to understanding cis-regulatory module function. *Sem Cell Dev Biol* 2009;**20**(7):856–62.
6. Larsen B, Rampalli S, Burns L, *et al*. Caspase 3/caspase-activated DNase promote cell differentiation by inducing DNA strand breaks. *Proc Natl Acad Sci USA* 2010;**107**: 4230–5.
7. Ju B, Lunyak V, Perissi V, *et al*. A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. *Science* 2006;**312**:1798–802.
8. Visel A, Akiyama J, Shoukry M, *et al*. Functional autonomy of distant-acting human enhancers. *Genomics* 2009;**93**(6): 509–13.
9. Degnan B, Vervoort M, Larroux C, *et al*. Early evolution of metazoan transcription factors. *Curr Opin Genet Dev* 2009; **19**(6):591–99.
10. Aravind L, Anantharaman V, Balaji S, *et al*. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* 2005;**29**(2):231–62.

11. Amoutzias G, Robertson D, Van de Peer Y, *et al*. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci* 2008;**33**(5):220–29.
12. Hinman V, Davidson E. Evolutionary plasticity of developmental gene regulatory network architecture. *Proc Natl Acad Sci USA* 2007;**104**(49):19404–09.
13. Piriyapongsa J, Mariño-Ramírez L, Jordan I. Origin and evolution of human microRNAs from transposable elements. *Genetics* 2007;**176**:1323–37.
14. Best A, Morrison H, McArthur A, *et al*. Evolution of eukaryotic transcription: insights from the genome of Giardia lamblia. *Genome Res* 2004;**14**(8):1537–47.
15. Bourbon H. Comparative genomics supports a deep evolutionary origin for the large, four-module transcriptional mediator complex. *Nucleic Acids Res* 2008;**36**(12): 3993–4008.
16. Davidson E, Erwin D. Gene regulatory networks and the evolution of animal body plans. *Science* 2006;**311**: 796–800.
17. McEwen G, Goode D, Parker H, *et al*. Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet* 2009;**5**(12):e1000762.
18. Bourque G, Leong B, Vega V, *et al*. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 2008;**18**(11):1752–62.
19. King D, Taylor J, Zhang Y, *et al*. Finding cis-regulatory elements using comparative genomics: Some lessons from ENCODE data. *Genome Res* 2007;**17**:775–86.
20. Alonso M, Pernaute B, Crespo M, *et al*. Understanding the regulatory genome. *Int J Dev Biol* 2008;**53**(8–10): 1367–78.
21. Carroll S. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 2008; **134**(1):25–36.
22. Lynch V, Wagner G. Resurrecting the role of transcription factor change in developmental evolution. *Evolution* 2008; **62**(9):2131–54.
23. Ettwiller L, Budd A, Spitz F, *et al*. Analysis of mammalian gene batteries reveals both stable ancestral cores and highly dynamic regulatory sequences. *Genome Biol* 2008;**9**(12): R172.
24. Wagner G, Lynch V. The gene regulatory logic of transcription factor evolution. *Trends Ecol Evolution* 2008;**23**(7): 377–85.
25. Dutkowski J, Tiuryn J. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics* 2007;**23**(13):i149–58.
26. Gibson T, Goldberg D. Reverse engineering the evolution of protein interaction networks. *Pac Symp Biocomput* 2008; **2009**:190–202.
27. Montgomery S, Griffith O, Sleumer M, *et al*. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* 2006;**22**(5):637–40.
28. Portales-Casamar E, Arenillas D, Lim J, *et al*. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res* 2009;**37**:D54–60.
29. Boyer L, Lee T, Cole M, *et al*. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 2005; **122**(6):947–56.
30. Loh Y, Wu Q, Chew J, *et al*. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* 2006;**38**(4): 431–40.
31. Sharov A, Ko M. Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Res* 2009; **16**(5):261–73.
32. van Loo P, Marynen P. Computational methods for the detection of cis-regulatory modules. *Brief Bioinform* 2009; **10**(5):509–24.
33. Pellegrini M, Marcotte E, Thompson M, *et al*. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999;**96**(8): 4285–8.
34. Blanchette M, Bataille A, Chen X, *et al*. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* 2006;**16**(5):656–68.
35. Kheradpour P, Stark A, Roy S, *et al*. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res* 2007;**17**(12):1919–31.
36. Grskovic M, Chaivorapol C, Gaspar-Maia A, *et al*. Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells. *PLoS Genetics* 2007;**3**(8):e145.
37. Chaivorapol C, Melton C, Wei G, *et al*. CompMoby: comparative MobyDick for detection of cis-regulatory motifs. *BMC Bioinformatics* 2008;**9**:455.
38. Ray P, Shringarpure S, Kolar M, *et al*. CSMET: comparative genomic motif detection via multi-resolution phylogenetic shadowing. *PLoS Comput Biol* 2008;**4**(6): e1000090.
39. He X, Ling X, Sinha S. Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. *PLoS Comput Biol* 2009;**5**(3):e1000299.
40. Moses A, Chiang D, Pollard D, *et al*. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* 2004;**5**(12):R98.
41. Warner J, Philippakis A, Jaeger S, *et al*. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat Methods* 2008;**5**(4):347–53.
42. Bais A, Grossmann S, Vingron M. Incorporating evolution of transcription factor binding sites into annotated alignments. *J Biosci* 2007;**32**(5):841–50.
43. Ambesi-Impiombato A, Bansal M, Liò P, *et al*. Computational framework for the prediction of transcription factor binding sites by multiple data integration. *BMC Neurosci* 2006;**7**:S8.
44. Lähdesmäki H, Rust A, Shmulevich I. Probabilistic inference of transcription factor binding from multiple data sources. *PloS One* 2008;**3**(3):e1820.
45. Fu W, Ray P, Xing E. DISCOVER: a feature-based discriminative method for motif search in complex genomes. *Bioinformatics* 2009;**25**(12):i321–29.
46. Fredman D, Engström P, Lenhard B. Web-based tools and approaches to study long-range gene regulation in Metazoa. *Brief Funct Genomic Proteomic* 2009;**8**(4):231–42.
47. Vingron M, Brazma A, Coulson R, *et al*. Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol* 2009;**10**(1):202.

48. Struckmann S, Araúzo-Bravo M, Schöler H, *et al*. ReXSpecies–a tool for the analysis of the evolution of gene regulation across species. *BMC Evol Biol* 2008;**8**:111.
49. Ward L, Bussemaker H. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* 2008;**24**(13):i165–71.
50. Stark A, Lin M, Kheradpour P, *et al*. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 2007;**450**(7167):219–32.
51. Meireles-Filho A, Stark A. Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Curr Opin Genet Dev* 2009;**19**(6):565–70.
52. Cheng Y, King D, Dore L, *et al*. Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res* 2008;**18**(12):1896–905.
53. Balmer J, Blomhoff R. Evolution of transcription factor binding sites in mammalian gene regulatory regions: handling counterintuitive results. *J Mol Evol* 2009;**68**(6):6544–664.
54. Do J, Schöler H. Regulatory circuits underlying pluripotency and reprogramming. *Trends Pharmacol Sci* 2009;**30**(6): 296–302.
55. Kamachi Y, Iwafuchi M, Okuda Y, *et al*. Evolution of non-coding regulatory sequences involved in the developmental process: reflection of differential employment of paralogous genes as highlighted by Sox2 and group B1 Sox genes. *Proc Jpn Acad Ser B Phys Biol Sci* 2009;**85**(2): 55–68.
56. Foshay K, Gallicano G. Regulation of Sox2 by STAT3 initiates commitment to the neural precursor cell fate. *Stem Cells Dev* 2008;**17**(2):269–78.
57. Takanaga H, Tsuchida-Straeten N, Nishide K, *et al*. Gli2 is a novel regulator of sox2 expression in telencephalic neuroepithelial cells. *Stem Cells* 2009;**27**(1):164–74.
58. Catena R, Tiveron C, Ronchi A, *et al*. Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. *J Mol Chem* 2004;**279**(40):41846–57.
59. Tomioka M, Nishimoto M, Miyagi SK, *et al*. Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. *Nucleic Acids Res* 2002;**30**(14): 3202–13.
60. Feng B, Jiang J, Kraus P, *et al*. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* 2009;**11**(2):197–203.
61. Moreno-Manzano V, Rodríguez-Jiménez F, Aceña-Bonilla J, *et al*. FM19G11, a new hypoxia-inducible factor (HIF) modulator, affects stem cell differentiation status. *J Mol Chem* 2010;**285**(2):1333–42.
62. Amaral P, Neyt C, Wilkins S, *et al*. Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. *RNA* 2009;**15**(11):2013–27.