biology
letters

Biol. Lett. (2006) **2**, 431–434
doi:10.1098/rsbl.2006.0484
Published online 2 May 2006

Anecdotes, data and regulatory modules

James E. Balmer and Rune Blomhoff*

Institute of Basic Medical Sciences, School of Medicine, University of Oslo, 0316 Oslo, Norway
**Author and address for correspondence: PO Box 1046 Blindern, 0316 Oslo, Norway (rune.blomhoff@medisin.uio.no).*

Beginning in the late 1980s, Eric Davidson's group at Cal Tech developed a modularity hypothesis of developmental gene regulation, showing that in an expanding number of cases, particular aspects of development were governed by compact 'modules' of transcription factor binding sites (TFBSs), and that these modules were separable, complex and interconnected. Davidson made no attempt to further generalize the hypothesis, but others took up the idea, transported it out of development and extended it to a general rule of clustering. Despite such misbegotten origins, the 'extended' modularity hypothesis—that TFBSs in general tend to come in compact clusters—has been highly productive, yet it has never been challenged with a large, diverse and unbiased dataset to see how universal it actually is. The aim of the present paper is to do so. Applying human–mouse–rat phylogenetic footprinting to neighbourhoods of a diverse set of TFBSs, including both developmental and non-developmental signals, we find that the extended hypothesis holds in at least 93.5% of cases. Based on this particular sample, we found a mean module length of 609 nucleotides containing, on an average, 24.5 presumptive regulatory signals of length greater than 5 and averaging 8.5 nucleotides each.

Keywords: regulatory modules;
gene expression regulation; transcription factors

1. INTRODUCTION

The notion of 'regulatory modules' is firmly established in transcriptional systematics, denoting a class of basic regulators of gene expression. Regulatory modules are a *sine qua non* of many bioinformatics studies (Loots *et al.* 2000; Krivan & Wasserman 2001; Berman *et al.* 2002) and are so *given* in the literature-at-large as to require no explanation, reference or definition (e.g. Avise 2001). How they came to attain this canonical standing is rather surprising.

(a) *Example, conjecture, comparison, law*

During the late 1980s, Eric Davidson and collaborators were studying the spatial and temporal limits of gene expression during development. Their primary model was a developmentally regulated actin gene, *CyIIIa* from *Strongylocentrotus purpuratus*. They identified a stretch of regulatory DNA lying about 7 kb upstream of the *CyIIIa* transcription start site, which (together with a basal promoter) is necessary and sufficient for correct spatial and temporal expression. They were able to show that several transcription factors (TFs) with lineage restricted expression patterns interacted productively with that region, and after honing it down to a minimal effective length, they found that it could be subdivided into small, independent 'regulatory domains', each hosting binding sites for different sets of factors and handling different aspects of expression. This example led them to wonder about the 'modularity' of other developmental genes, and indeed, similar short, discrete regions were found to control developmental expression—in timing, location and amplitude—of *SM50*, a minor skeletal matrix protein, and *Endo16*, an endoderm marker protein. In each case, short clusters of binding sites for both unique and general factors were separated by longer stretches of DNA that appeared to be uninvolved in transcription. The group was drawn to conjecture—and later, to demonstrate through a masterful set of deletion studies—that the clusters indeed operated in a modular fashion, each adjudicating a particular regulatory decision. (See Kirchhamer & Davidson 1996; Davidson 2001 and references therein.)

As successive papers appeared, the group generalized these ideas, finally writing in a widely cited 1997 review, 'individual modules are always found to contain multiple TF target sites' and 'developmental *cis*-regulatory outputs never devolve entirely from *cis*-regulatory sites for a single species of factor' (Arnone & Davidson 1997). Importantly, however, they never generalized the modularity hypothesis beyond the realm of development (or of similarly delimited situations). And indeed, among all the examples they produced, few (e.g. from muscle) were even post-specificational, let alone non-developmental.

Nevertheless, the notion of 'regulatory modules' as discrete, compact arrangements of transcription factor binding sites (TFBSs) came to be seen as the normal arrangement of regulatory signals—in genes of all kinds and in all circumstances. What had been a carefully worded conjecture about compact, separable decision adjudicators networked together in a developmental or other closely specifiable context, became a general law about *typical* regulatory format. Numerous papers, including many of the standard references in evolution and computational transcription regulation (see Fickett & Wasserman 2000; Loots *et al.* 2000; Stone & Wray 2001; Elnitski *et al.* 2003) now rely on the notion of modularity—but one under which TFBSs are thrown together helter-skelter, without regard to functional connections or purposes. One can almost always trace a line of citations, directly or indirectly, back to one or another of Davidson's papers (cf. Wasserman & Fickett 1998; Bailey & Noble 2003; Ureta-Vidal *et al.* 2003); yet no one has ever investigated just how general modularity is. A thorough investigation would require a lengthy and complex analysis, but it is possible to quickly adduce one type of evidence that the hypothesis does, indeed, hold in general. One can simply test a diverse, unbiased set of TFBSs to see *what proportion* of them appear to reside within larger conserved blocks of non-coding DNA.

The electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2006.0484 or via http://www.journals.royalsoc.ac.uk.

Received 2 April 2006
Accepted 7 April 2006

Table 1. Various module parameters.

parameter	minimum	first quartile	mean	median	third quartile	maximum
module length	110	291.5	609	356	795	2734
percentage of nucleotides in blocks (%)	40.0	52.4	59.4	59.9	65.9	85.6
number of blocks ≧ 5 nucleotides	3	9	24.5	15.5	33.8	109
average block size per module	5.57	7.72	8.95	8.55	9.96	21.75
relative site position (non-truncated only)	8.9	44.4	58.4	56.7	80.2	97.2
centrality (balance point, conserved nucleotides)	0.375	0.503	0.530	0.537	0.558	0.697
centrality (non-truncated only)	0.375	0.485	0.519	0.529	0.558	0.595
dispersion (clustering)	1.07	1.77	2.94	2.43	3.72	9.37

In summary, the Davidson group formulated a hypothesis that the binding sites for TFs controlling individual decision points in developmental programmes would be densely clustered into distinct, separable units of non-coding space. The idea spread into the scientific literature at large, but with two crucial changes: the stricture of unitary functionality was lost and the connection with development (or differentiation) was ignored. The expanded modularity hypothesis had become the law.

(**b**) *A qualified dataset*

There is a considerable evidence that, at the very least, modularity represents a common regulatory architecture. There is nothing necessary about this, however, and all the evidence so far is anecdotal, deriving from elucidations of particular modular systems or from studies assuming modularity *ab initio*. As everybody knows, the plural of 'anecdote' is not 'data', and in order to have some confidence that the extended modularity hypothesis states a true regularity of nature, it might be useful to challenge the hypothesis with a large, unbiased and diverse set of TFBSs and try to create extended phylogenetic footprints around each. According to Goodman's corollary (Tagle *et al.* 1988), such footprints should delineate any larger regions of DNA that have been conserved in virtue of specific biologic roles. If clustering is the rule, then cross-species alignments will produce extended modular footprints (containing, perhaps, in addition to TF sites, various 'facilitator' or other structural components). Needless to say, one cannot be assured that all of the intramodular blocks have biological roles; but insofar as Goodman's insight is valid and applies to human–rodent comparisons, there is a high probability that most of them do.

The set of known binding sites for the transcriptional system driven by retinoic acid (RA) provides such a collection. This ancient system is involved in such processes as development, apoptosis, homeostasis, disease progression and disability. Its set of known target genes is so diverse as to appear arbitrary, and it has been studied in extraordinarily varied contexts: in clinical dermatology, oncology, nutrition, epidemiology, teratogenesis and development; as a differentiation agent; and as a model of inducible transcription. See Gronemeyer & Laudet 1995 for a review of the RA system within the context of nuclear receptors (of which it is one example), and Balmer & Blomhoff 2002 for a description and classification of RA-regulated genes.

2. MATERIAL AND METHODS

Directly or indirectly, RA regulates at least 580 genes (Balmer & Blomhoff 2002, updated). The literature contains references to 220 genomic elements thought to be associated with this regulation—fully tested receptor (dimer) binding sites, unparsed 'active' regions, untested motifs and the like. Once inappropriate sites and orthologous redundancies were eliminated, 77 unique sites in non-coding, non-transcribed DNA remained. Each was from one of three target species: mouse (Mm), rat (Rn) or man (Hs). We located a native context of about 1000 bases for each site and used standard methods to find and align homologues. See supplementary table 1 and supplemental methods (electronic supplementary material) for more information.

3. RESULTS AND DISCUSSION

Given a TFBS from the literature, and an extended neighbourhood in the referenced species, we were almost always able to find extended homologous regions in the other two species. This is a strong presumptive evidence in favour of modules, suggesting that whether or not a particular *site* is conserved, neighbouring modular components may remain. In five cases—the sites in *Afp* (at −1961 relative to transcription start), *AQP1* (−2218), *Tgm2* (−1745) and *Thbd* both (−1531) and (−941)—we were unable to find homologous regions in at least one species. Nevertheless, all the three orthologues appear to exist, so insofar as three-way footprinting can determine, these five sites *may not* belong to modules. Notably, however, the authenticity of three of the sites has been questioned (reviewed in Balmer & Blomhoff 2002), so at most, between 2.5 and 6.5% of sites may lie outside regulatory modules. This is an acceptable exception rate, so the extended modularity hypothesis seems to be vindicated—insofar as any biological 'regularity' generalizes.

This finding not only validates the generalized modularity hypothesis, but it has bench-level implications for those studying gene expression. Namely, even if a *site* does not appear to be conserved among moderately similar species (such as primates and rodents), a quick check for modularity in a surrounding footprint can offer some comfort that the site is authentic. Conversely, given the notorious difficulty of establishing a definitive biological role for a putative TFBS, the lack of a surrounding footprint should serve as a caveat, inviting heightened scrutiny.

We truncated alignments that would otherwise extend into transcribed DNA. Among the 72

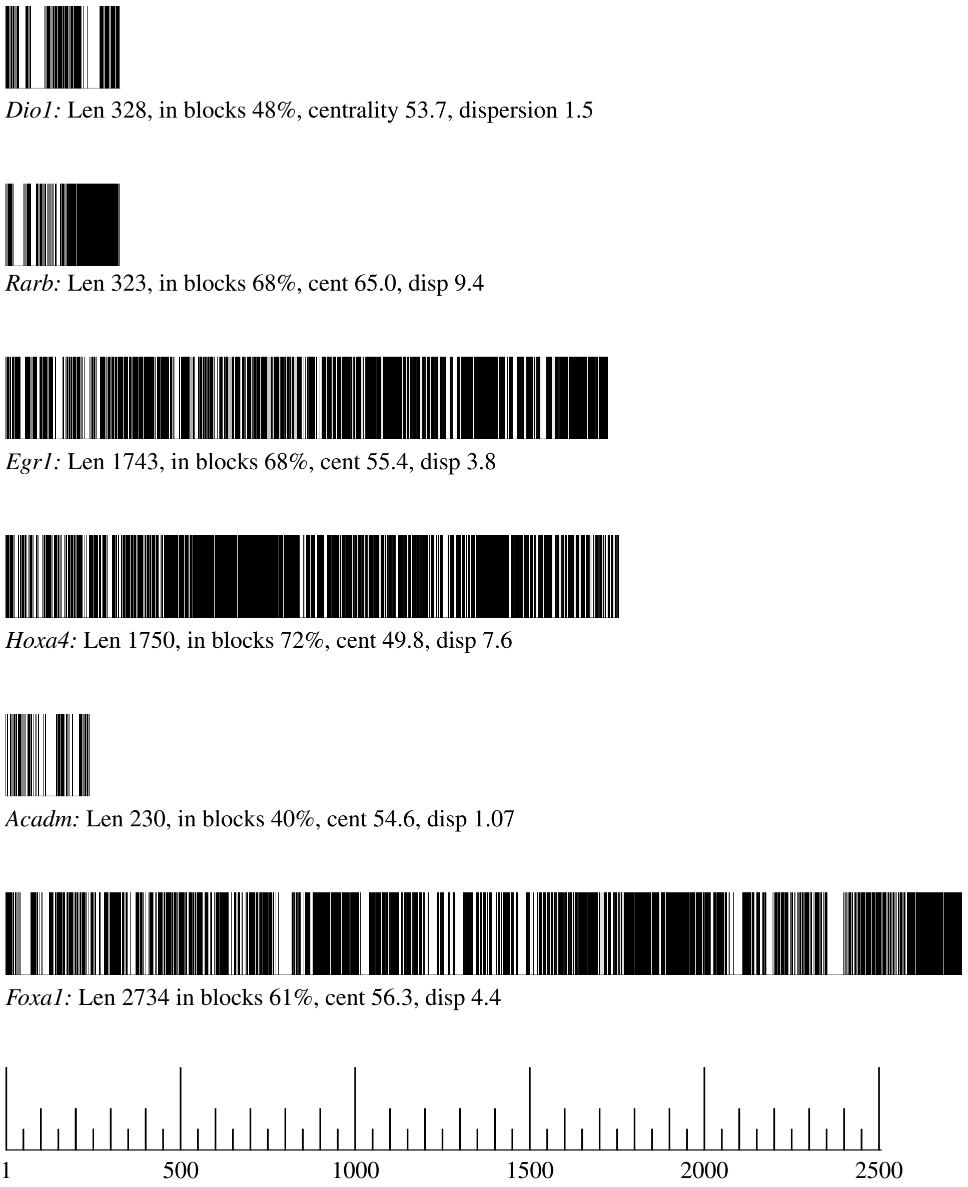

Figure 1. Barcode diagrams of typical regulatory modules. The basic characteristics of some typical modules. Identically conserved nucleotides are highlighted.

alignments, 58% were so truncated. The mean length of modules was 609 nucleotides with 59% identity in all three species; identity between mouse and rat was higher (86%) and identity between Mm and Hs (or Rn and Hs) averaged 63%. Considering only non-truncated modules, the mean length was 558.6 nucleotides with a maximum of 1750. Many of these numbers differ somewhat from analogous measures found in the literature (e.g. Levy *et al.* 2001; Frith *et al.* 2002; Suzuki *et al.* 2004), perhaps because our method starts with known TF sites rather than *ad hoc* searching rubrics.

The number of blocks of length five or greater—which most likely indicate TFBSs—varies between 3 and 109 per module (mean = 24.5). This does not correlate with module length ($\rho = 0.27$) beyond the obvious limits. Considering only blocks of length five or more, the mean length over the entire dataset is 8.95 nucleotides (as compared with the standard estimate of 6–20). See table 1 and supplementary table 2 (electronic supplementary material).

To get some notion of the distribution of identically conserved nucleotides, we computed measures of centrality and dispersion. First, we calculated the relative mid-point of each module, using the function $\mathfrak{S}(i) = 1$, if position i is identically conserved, $\mathfrak{S}(i) = 0$ otherwise. Marking the medians of these distributions, we found a 'balance point' for each module. These points tend towards the middle of the module, but are upstream-skewed.

To gauge dispersion (or clustering), we defined recursively, $\sigma(1) = \mathfrak{S}(1)$, and $\sigma(j+1) = \mathfrak{S}(j+1)[\sigma(j) + \mathfrak{S}(j+1)]$ for j ranging over positions in the module, and calculated $\Sigma_j(\sigma(j))/\max(j)$. This strongly rewards blocks for length without penalizing mismatches. Everything else being equal, modules with higher clustering are more compactly arranged, and on the basis of the present sample, higher compactness may be a mark of modules in developmental TFs (including toolkit TFs) and other elemental genes (such as *H1F0*). See figure 1, which gives an idea of what typical modules look like, showing examples with different lengths, centralities and degrees of dispersion.

The dataset we chose for this study is made up of binding sites from a truly ancient and ubiquitous transcriptional control system. It differentiated out of the metazoan nuclear receptor superfamily members as early as the Porfiran divergence and has been found in all 'later' species investigated. According to the standard theory, and given the nearly ubiquitous availability of both receptors and ligand in nearly all cell types and life stages, one expects TFBSs to have accrued more or less at random throughout the various genomes over time. It is not surprising, then, that the system is involved in the regulation of a seemingly endless variety of target genes. Accordingly, our starting dataset contains genes of many types, from early toolkit genes (e.g. several *Hox* paralogs) to ATP-binding cassette transporters (*Abcc2*), intercellular adhesion molecules, regulators of glucose metabolism, etc.

This diversity suggests that our findings are probably representative of TFBSs overall. However, it does not allow us to meaningfully identify potential subclasses of modules, which might be associated with particular classes of TFs (such as tissue-, stage- or stimulus-specific factors) or with particular classes of genes (developmental, metabolic, cell-type-specific, etc.). These very interesting possibilities must await further study.

In conclusion, the extended modularity hypothesis appears to hold in general and we may comfortably continue to rely on it—with all its implications for transcription, evolution and bioinformatics.

This work was made possible through the generous support of the Norwegian Cancer Society, the Research Council of Norway, and the Throne Holst Foundation.

Arnone, M. I. & Davidson, E. H. 1997 The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851–1864.

Avise, J. C. 2001 Evolving genomic metaphors: a new look at the language of DNA. *Science* **294**, 86–87. (doi:10.1126/science.294.5540.86)

Bailey, T. L. & Noble, W. S. 2003 Searching for statistically significant regulatory modules. *Bioinformatics* **19**(Suppl. 2), II16–II25.

Balmer, J. E. & Blomhoff, R. 2002 Gene expression regulation by retinoic acid. *J. Lipid Res.* **43**, 1773–1808. (doi:10.1194/jlr.R100015-JLR200)

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. & Eisen, M. B. 2002 Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99**, 757–762. (doi:10.1073/pnas.231608898)

Davidson, E. H. 2001 *Genomic regulatory systems: development and evolution*. San Diego, CA: Academic Press.

Elnitski, L., Hardison, R. C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M. J., Schwartz, S., Miller, W. & Chiaromonte, F. 2003 Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13**, 64–72. (doi:10.1101/gr.817703)

Fickett, J. W. & Wasserman, W. W. 2000 Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* **11**, 19–24. (doi:10.1016/S0958-1669(99)00049-X)

Frith, M. C., Spouge, J. L., Hansen, U. & Weng, Z. 2002 Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.* **30**, 3214–3224. (doi:10.1093/nar/gkf438)

Gronemeyer, H. & Laudet, V. 1995 Transcription factors 3: nuclear receptors. *Protein Profile* **2**, 1173–1308.

Kirchhamer, C. V. & Davidson, E. H. 1996 Spatial and temporal information processing in the sea urchin embryo: modular and intramodular organization of the *CyIIIa* gene *cis*-regulatory system. *Development* **122**, 333–348.

Krivan, W. & Wasserman, W. W. 2001 A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11**, 1559–1566. (doi:10.1101/gr.180601)

Levy, S., Hannenhalli, S. & Workman, C. 2001 Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17**, 871–877. (doi:10.1093/bioinformatics/17.10.871)

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M. & Frazer, K. A. 2000 Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140. (doi:10.1126/science.288.5463.136)

Stone, J. R. & Wray, G. A. 2001 Rapid evolution of *cis*-regulatory sequences via local point mutations. *Mol. Biol. Evol.* **18**, 1764–1770.

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K. & Sugano, S. 2004 Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. *Genome Res.* **14**, 1711–1718. (doi:10.1101/gr.2435604)

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. & Jones, R. T. 1988 Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatis*): nucleotide and amino acid sequences, developmental regulation, and phylogenetic footprints. *J. Mol. Biol.* **203**, 439–455. (doi:10.1016/0022-2836(88)90011-3)

Ureta-Vidal, A., Ettwiller, L. & Birney, E. 2003 Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**, 251–262. (doi:10.1038/nrg1043)

Wasserman, W. W. & Fickett, J. W. 1998 Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **A278**, 167–181. (doi:10.1006/jmbi.1998.1700)