

domain is equally informative. 'A-type' and 'C-type' protein–protein interaction sites have been implicated in the transformation of prions across species by mediating binding of PrP^C to Protein X or PrP^{Sc}, respectively [2–7]. Of nine A-type residues, only one (position 215) is conserved in all vertebrate groups, whereas the remaining eight show variability patterns that generally correlate with vertebrate class groupings (Fig. 3b). Likewise, of all C-type residues (position 96–167), the first half are conserved with exception of *Fugu* and the other half are specific for each vertebrate group (Fig. 3b). Thus, PrP^{Sc}s could fail to convert PrP^Cs across vertebrate classes because of molecular incompatibility at specific contact sites, regardless of self-aggregation properties encoded at the N-terminal domain.

Together, the evidence presented here links discrete patterns of prion molecular evolution with important changes in their pathogenic properties. Particularly, changes at the N- and C-terminal domains could help explain why scrapie pathogenesis and transmission seem exclusive to mammals. Our discovery of a novel prion gene in fish, and possibly in a Urochordate, places the origin of prions in a common ancestor of all vertebrates, at least 550 million years ago. Moreover, our comparative analysis of PrP^C amino acid sequences reveals rapid rates of molecular evolution at the base of the vertebrate radiation without significant losses in protein structure, followed by a reduction in the effective substitution rates within each vertebrate class. The implications of these findings can now be tested experimentally.

Acknowledgements

E.R.-M. is a DAAD fellow. This work was supported by grants from the University of Konstanz AFF to E.M.-T. and from TSE, MWK, BW and FCI to C.A.O.S. We are grateful to the Joint Genome Institute, UK-HGMP Resource Center, Genoscope and the Whitehead Institute for Genome Research. This article is dedicated to our former colleague Joseph J.B. Jayabalan.

References

- Prusiner, S.B. (1991) Molecular biology of prion diseases. *Science* 252, 1515–1522
- Telling, G.C. *et al.* (1995) Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with another protein. *Cell* 6, 79–90
- Prusiner, S.B. (1998) Prions. *Proc. Natl Acad. Sci. U.S.A.* 95, 13363–13383
- Billeter, M. *et al.* (1997) Prion protein NMR structure and species barrier for prion diseases. *Proc. Natl Acad. Sci. U.S.A.* 94, 7281–7285
- Kocisko, D.A. *et al.* (1995) Species specificity in the cell-free conversion of prion protein to protease-resistant forms: a model for the scrapie species barrier. *Proc. Natl Acad. Sci. U.S.A.* 92, 3923–3927
- Telling, G.C. *et al.* (1994) Transmission of Creutzfeldt–Jakob disease from humans to transgenic mice expressing chimeric human–mouse prion protein. *Proc. Natl Acad. Sci. U.S.A.* 91, 9936–9940
- Horiuchi, M. *et al.* (2000) Interaction between heterologous forms of prion protein: binding, inhibition of conversion, and species barrier. *Proc. Natl Acad. Sci. U.S.A.* 97, 5836–5841
- Aguzzi, A. *et al.* (2001) Prions: health scare and biological challenge. *Nat. Rev. Mol. Cell Biol.* 2, 118–126
- Strumbo, B. *et al.* (2001) Molecular cloning of the cDNA coding for *Xenopus laevis* prion protein. *FEBS Lett.* 508, 170–174
- Donne, D.G. *et al.* (1997) Structure of the recombinant full length hamster prion protein PrP(29–231): the N terminus is highly flexible. *Proc. Natl Acad. Sci. U.S.A.* 94, 13452–13457
- Riek, R. *et al.* (1996) Prion protein NMR structure and familial human spongiform encephalopathies. *Proc. Natl Acad. Sci. U.S.A.* 95, 11667–11672
- Suzuki, T. *et al.* (2002) cDNA sequence and tissue expression of *Fugu rubripes* prion-like protein: a candidate for the teleost orthologue of tetrapod PrPs. *Biophys. Res. Commun.* 294, 912–917
- Murphy, R.M. (2002) Peptide aggregation in neurodegenerative disease. *Annu. Rev. Biomed. Eng.* 4, 155–174
- Viles, J.H. *et al.* (1999) Copper binding to the prion protein: structural implications four identical cooperative binding sites. *Proc. Natl Acad. Sci. U.S.A.* 96, 2042–2047
- Burns, C.S. *et al.* (2002) Molecular features of the copper binding sites in the octapeptide domain of the prion protein. *Biochemistry* 41, 3991–4001
- Gazit, E. (2002) Global analysis of tandem aromatic octapeptide repeats: the significance of the aromatic-glycine motif. *Bioinformatics* 18, 880–883
- Bamborough, P. *et al.* (1996) Prion protein structure and scrapie replication: theoretical, spectroscopic, and genetic investigation. *Cold Spring Harbor Symp. Quant. Biol.* 61, 495–509
- Gazit, E. (2002) A possible role for π -stacking in the self-assembly of amyloid fibrils. *FASEB J.* 16, 77–83
- Swofford, D.L. (2002) *PAUP**. *Phylogenetic Analysis Using Parsimony (and Other Methods)* (Ver. 4), Sinauer Associates
- Kumar, S. *et al.* (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17, 1244–1245
- Aparicio, S. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301–1310

0168-9525/03/\$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0168-9525(02)00032-X

Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites

M. Madan Babu and Sarah A. Teichmann

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

DNA-binding transcription factors regulate the expression of genes near to where they bind. These factors can be activators or repressors of transcription, or both. Thus, a fundamental question is what determines

whether a transcription factor acts as an activator or a repressor? Previous research into this question found that a protein's regulatory function is determined by one or more of the following factors: protein–protein contacts, position of the DNA-binding domain in the protein primary sequence, altered DNA structure, and

Corresponding author: M. Madan Babu (madanm@mrc-lmb.cam.ac.uk).

<http://tigs.trends.com>

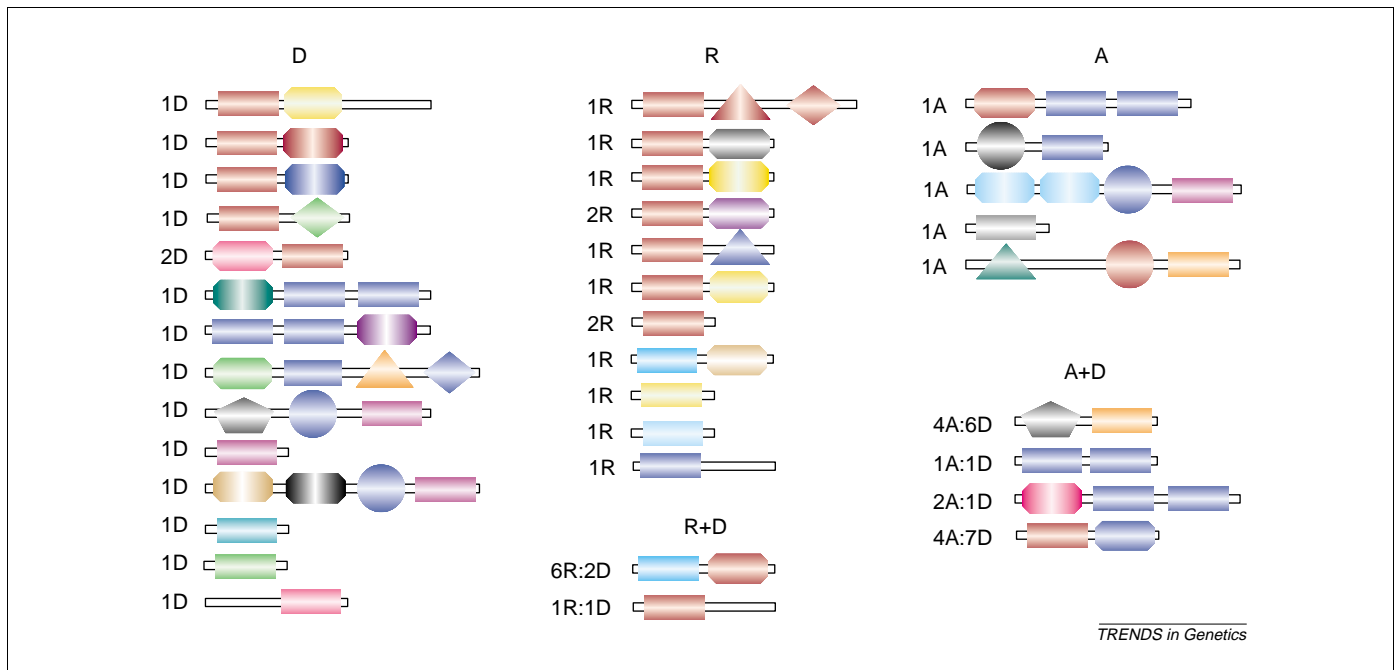


Fig. 1. The domain architectures for the 69 proteins are represented here. DNA-binding domains are shown as rectangles, with the different colours representing different families. The partner domains are shown as: octagon, small molecule binding domain; triangle, enzyme domains; circle, protein-interaction domain; pentagon, receiver domain; diamond, domains of unknown function; again colours represent different families. A, activator proteins; R, repressor proteins and D, dual regulators. The number of transcription factors of each domain architecture is shown to the right of the architecture. Of the 11 DNA-binding domain families, seven occur in both activators or repressors and dual regulators. Of the 28 partner domain families, 17 occur in transcription factors with different types of regulatory function. Six domain architectures are present in proteins that are activators or repressors as well as dual regulators. Thus, there is very little correlation between the regulatory function and the domain families of transcription factors. The full key to all the domains and families is given at http://www.mrc-lmb.cam.ac.uk/genomes/madann/ec_tf_bs/key.html.

the position of its binding site on the DNA relative to the transcription start site. Although there are many aspects specific to different transcription factors, in this work we demonstrate that, in general, in the prokaryote *Escherichia coli*, a transcription factor's protein family is not indicative of its regulatory function, but the position of its binding site on the DNA is.

To examine what determines whether a transcription factor is an activator, a repressor, or both, we extracted *Escherichia coli* transcription factors and their binding sites from the RegulonDB [1] database and from [2], forming a dataset of 71 proteins with experimentally verified regulatory information and binding-site data. Of these, 18 are activators, 20 are repressors and 33 are dual regulators. These 71 proteins regulate 529 genes through 1249 binding sites at 308 promoters; that is, approximately an eighth of all genes and a quarter of all transcription factors in *E. coli* [3].

Assignments of domains with a known 3D structure were available for 69 of the transcription factors in the SUPERFAMILY database [4]. In SUPERFAMILY, structural assignments are based on a sensitive multiple sequence comparison method [5] and the domain definitions and families of evolutionarily related domains are from the Structural Classification of Proteins (SCOP) database [6]. In SCOP, a combination of details of the structure, function and sequence are used to determine evolutionary relationships between the domains in SCOP superfamilies, which we call protein families here. In addition, domains were assigned from five families in the PFAM database [7].

<http://tigs.trends.com>

The different domain architectures in the set of 69 proteins are shown in Fig. 1. It is clear that most of the transcription factors are two-domain proteins, with a few three- and four-domain proteins as previously noted by others [8,9]. Each transcription factor consists of one or two DNA-binding domains (DBDs), and most have one or more partner domains. The DBDs come from 11 different families of known 3D structure, and the partner domains from 28 different families, including five PFAM domains. All together, these 71 proteins belong to 36 different domain architectures, where the proteins with the same domain architecture are probably direct duplicates.

Protein families are not indicative of a protein's regulatory function

A reasonable hypothesis would be to assume that activators, repressors and dual regulators are more closely related to the other proteins within each regulatory group than to proteins in other groups. This would mean that there would be protein families and domain architectures that were characteristic either of activators, repressors or dual regulators. Prag *et al.* [10] and Perez-Rueda *et al.* [11] reported evidence to support this by relating the position of helix-turn-helix motifs along the primary sequence to a protein's regulatory function. They used sequence comparison and profile methods to locate the helix-turn-helix motifs that represent DBDs, and found that the helix-turn-helix tends to be at the N terminus for repressors and at the C terminus for activators.

Here, we use the domains of known 3D structure to identify the position and evolutionary relationships of DBDs and their partner domains. Seven out of the 11 DBD

families occur in both activators and repressors, or in dual regulators, as do 17 out of 28 partner domains. Thus, of the protein domain families present in transcription factors, 62% occur in activators and repressors. In addition, there are six domain architectures that appear in dual regulators as well as activators or repressors, as shown in Figure 1. In other words, a sixth of all domain architectures occur in dual regulators, and in some instances in an activator or repressor as well.

Ten of the 11 DBD families are helix-turn-helix families (the exception being the nucleic-acid binding domain), but the motif occurs in very different structural contexts, such that there is little or no evidence for evolutionary relationship beyond the families given here, which are SCOP superfamilies. The position of the DBDs is consistently N terminal for repressors as previously observed [10,11], whereas it is a mixture of N and C terminal in activators and dual regulators. Given that the 13 repressor domain architectures show that most of the repressors are either not directly or not at all related, the fact that the DBD is always N terminal is either a functional requirement or a coincidence, but not because the repressors are evolutionarily related by descent.

These observations show that the regulatory function of a protein does not depend upon the DBD type, partner family or domain architecture, and also that these cannot be used as a reliable measure to predict the regulatory functions of a protein.

Binding site position suggests regulatory function

To investigate whether the binding site on the DNA is informative as to regulatory function, we examined the distance from the transcription start site to the center of the TF binding site (obtained from RegulonDB) (Fig. 2). It is evident that all the activators have an upstream binding site (Fig. 2a) and the repressors have both upstream and downstream binding sites (Fig. 2b). For the dual regulators, the binding sites are equally populated upstream and downstream of the transcription start site (Fig. 2c). However, when we separated the data for dual regulators into activator binding sites (Fig. 2d) and repressor binding sites (Fig. 2e), the same pattern emerged as for the repressors and activators.

This observation suggests that the distance of the transcription factor binding site from the transcription start site is an important factor in determining regulatory function, as observed by Collado-Vides *et al.* from an analysis on a smaller dataset [12]. However, the regions upstream of the transcription start site are populated by both activators and repressors. We analysed binding-site positions by dividing the binding sites at a promoter into those upstream of the RNA polymerase binding site at -35, and other downstream sites. Transcription factors at the upstream sites generally do not prevent polymerase binding as such, although there are exceptions to this in DNA-bending proteins (e.g. FIS) that can influence polymerase binding and processing from upstream sites. Downstream sites are expected to hinder polymerase binding or processing.

These data show that in 91% of promoters, activators and activating dual regulators had only upstream binding

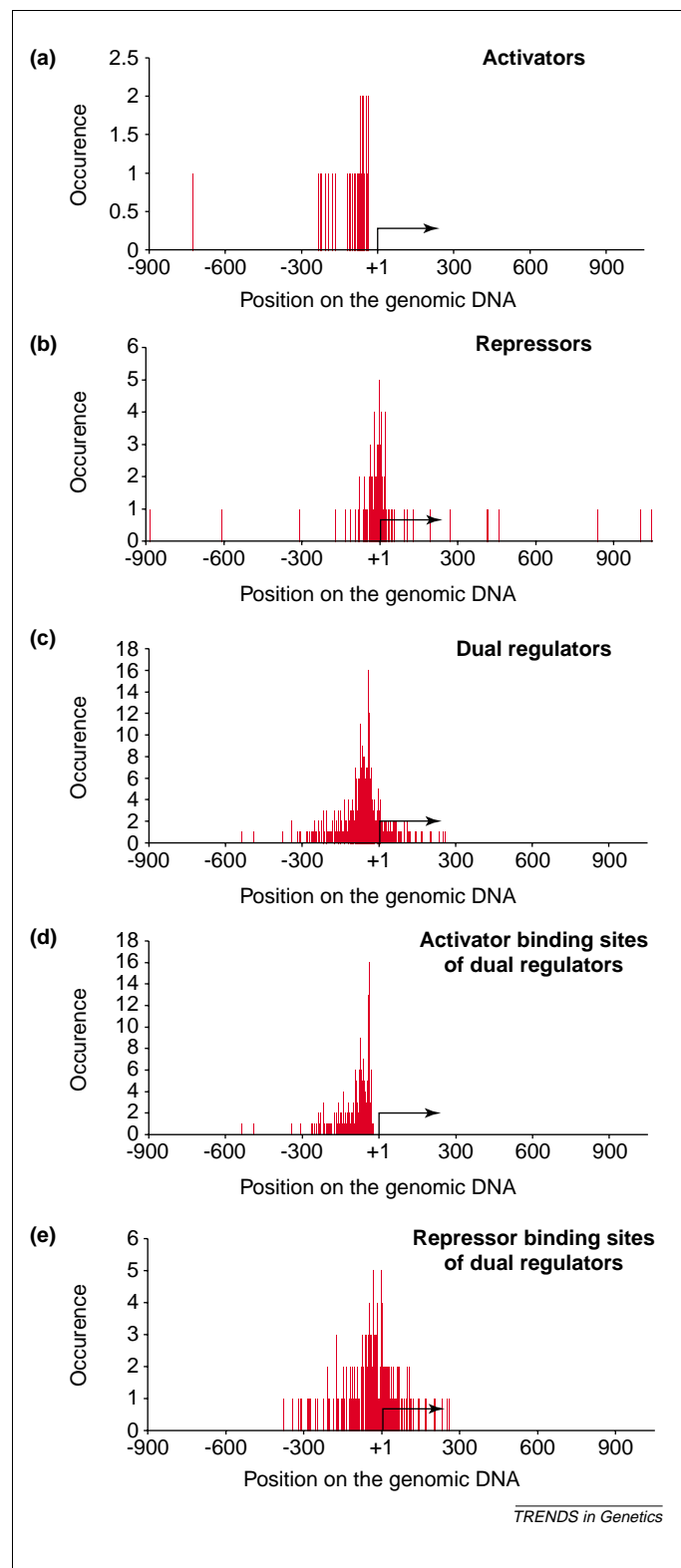


Fig. 2. The transcription start site is represented as a right-angled arrow in all the figures. (a) All of the activator binding sites occur upstream of the transcription start site. (b) One-hundred-and-twenty repressor binding sites occur both upstream and downstream of the transcription start site. (c) Six-hundred-and-three dual regulator binding sites have equal populations in both upstream and downstream regions. (d,e) When separated into activator binding sites only and repressor binding sites only, the dual regulators have the same pattern as in (a) and (b). For upstream repressor binding sites of both repressors and dual regulators, there is usually another repressor binding site after the start site. Also note that the repressor binding sites are seen beyond 300 bases after the transcription start site, which most binding site prediction programs do not consider.

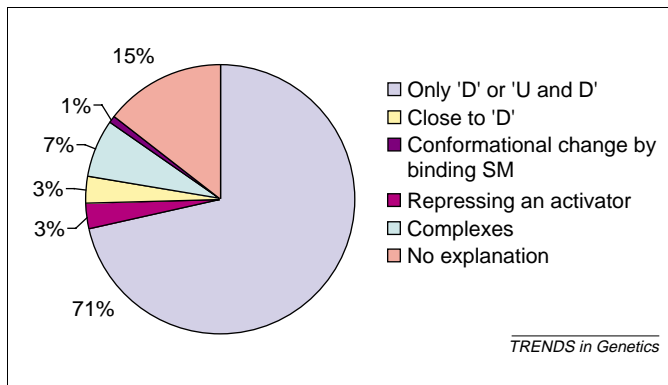


Fig. 3. Here we consider all repressor binding sites, including those of dual regulators. There are 46 repressors (including dual regulator repressors) regulating 228 promoters through 403 binding sites. Upstream or 'U' sites are defined as all sites before -35 , the most upstream RNA polymerase footprint. All other sites are downstream, 'D', sites.

sites, and in another 3%, the downstream sites are low affinity binding sites according to the literature. For the remaining 6%, there is no uniform explanation, but proteins such as MerR are included. MerR binds between the -35 and -10 sites on the opposite face of the DNA to the polymerase, and aids polymerase binding and transcription initiation by unwinding and altering DNA conformation [13].

Figure 3 shows that in 71% of promoters, repressors and repressing dual regulators had at least one downstream binding site. In addition, 3% of the repressor sites are within four nucleotides of the -35 site. For another 7%, the repressor acts as a complex or just in conjugation with other TFs, where its partner in the complex has a downstream binding site, and for another 3%, the repressor acts by inhibiting an activator from binding. There are also complex mechanisms as in the case of AraC, where two AraC monomers, each binding at different upstream sites, dimerize to induce looping out of DNA and prevent the polymerase from binding properly [14].

From this analysis, it is clear that the large majority of activators have only upstream binding sites. Most repressor binding sites are downstream, or upstream in conjunction with a downstream site. Thus, our summary of prokaryotic transcription factors provides further evidence that at a gross level, prokaryotic activators function by stabilizing the polymerase from upstream sites and repressors act by steric hindrance, blocking polymerase binding or processing. However, at a more detailed level, it is clear that there is a huge variety of different mechanisms for activation and repression, including DNA bending and unwinding [15,16], oligomerization [17,18], protein-protein contacts [19], position of the DNA binding domain on the protein primary sequence [10,11], altered DNA structure [13], and the position of the DNA binding site relative to the transcription start site [12,20,21]. The function of upstream repressor sites will vary for different transcription factors, but as they are generally not involved in the actual repression by steric hindrance, the role of most of these sites could be in recruitment of other repressors to downstream sites. This could occur either specifically by oligomerization [17] or cooperativity, or by simply increasing the local concentration

of the protein to increase its chances of binding to a functionally more important downstream site.

A third of the repressor binding sites occur after the transcription start site

Many transcription factor binding site prediction methods consider only intergenic regions and neglect coding regions because of the high number of false positives when these are taken into account. At the same time, regions downstream of the transcription start site contain 33% of the repressor binding sites in our analysis. This is evident from Fig. 2, which shows that repressor binding sites occur well beyond 100 bases after the transcription start site. Although it is not necessary to consider the entire coding region for binding site prediction, it is advisable to look within 300 bases downstream of the transcription start site.

Conclusions

In this work we demonstrate that activators, repressors and dual regulators in *E. coli* belong to many of the same protein families and even share some domain architectures. Therefore, a transcription factor's regulatory role is not determined by protein structure or evolutionary relationships, but to a large extent simply by the position of the transcription factor binding site: activators have essentially only upstream binding sites, whereas more than two thirds of repressors have at least one downstream binding site. Eleven activators and seven repressors share the same six domain architectures with 18 dual regulators (Fig. 1). This implies that 36 of the 71 transcription factors have evolved by duplication of an ancestral transcription factor, followed by a change in function through a shift in binding sites.

Data availability

The dataset used for this analysis is available in MySQL format at:

http://www.mrc-lmb.cam.ac.uk/genomes/madanm/ec_tf_bs/

Acknowledgements

We thank Julio Collado-Vides and Heladia Salgado for helping us with the information from RegulonDB, Julian Gough for the structural domain assignments from the SUPERFAMILY database, and Andrew Travers for useful discussions. We are grateful to the Medical Research Council, Cambridge Commonwealth Trust and Trinity College, Cambridge for financial support.

References

- Salgado, H. *et al.* (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* 29, 72–74
- Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68
- Perez-Rueda, E. and Collado-Vides, J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.* 28, 1838–1847
- Gough, J. and Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* 30, 268–272
- Karplus, K. *et al.* (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14, 846–856
- Murzin, A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540

- 7 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280
- 8 Morett, E. and Segovia, L. (1993) The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains. *J. Bacteriol.* 175, 6067–6074
- 9 Aravind, L. and Koonin, E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* 27, 4658–4670
- 10 Prag, G. *et al.* (1997) Structural principles of prokaryotic gene regulatory proteins and the evolution of repressors and gene activators. *Mol. Microbiol.* 26, 619–620
- 11 Perez-Rueda, E. *et al.* (1998) Genomic position analyses and the transcription machinery. *J. Mol. Biol.* 275, 165–170
- 12 Collado-Vides, J. *et al.* (1991) Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol. Rev.* 55, 371–394
- 13 Dai, X. and Rothman-Denes, L.B. (1999) DNA structure and transcription. *Curr. Opin. Microbiol.* 2, 126–130
- 14 Lobell, R.B. and Schleif, R.F. (1990) DNA looping and unlooping by AraC protein. *Science* 250, 528–532
- 15 Travers, A. and Muskhelishvili, G. (1998) DNA microloops and microdomains: a general mechanism for transcription activation by torsional transmission. *J. Mol. Biol.* 279, 1027–1043
- 16 Pemberton, I.K. *et al.* (2002) FIS modulates the kinetics of successive interactions of RNA polymerase with the core and upstream regions of the *tyrT* promoter. *J. Mol. Biol.* 318, 651–663
- 17 Rhee, K.Y. *et al.* (1996) Leucine-responsive regulatory protein–DNA interactions in the leader region of the *ilvGMEDA* operon of *Escherichia coli*. *J. Biol. Chem.* 271, 26499–26507
- 18 Rojo, F. (1999) Repression of transcription initiation in bacteria. *J. Bacteriol.* 181, 2987–2991
- 19 Hochschild, A. and Dove, S.L. (1998) Protein–protein contacts that activate and repress prokaryotic transcription. *Cell* 92, 597–600
- 20 Rhodius, V.A. and Busby, S.J. (1998) Positive activation of gene expression. *Curr. Opin. Microbiol.* 1, 152–159
- 21 Muller-Hill, B. (1998) Some repressors of bacterial transcription. *Curr. Opin. Microbiol.* 1, 145–151

0168-9525/03/\$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0168-9525(02)00039-2

Letters

Life cycles of successful genes

Robert Hoffmann and Alfonso Valencia

National Center of Biotechnology, CNB-CSIC, Cantoblanco, Madrid M-28049, Spain

By exploring time-series data from MEDLINE abstracts, we observe that only a few genes have been quoted with increasing frequency during the past 25 years. This is probably the result of selective pressure by the scientific community. Over the years, this selection has produced an extreme power law distribution of the information available for individual genes. Interestingly, those genes that are successfully selected are not necessarily the most important genes to the cell. To stress the implication of this finding we show that there is no correlation between a gene's impact in the scientific literature and its centrality in protein-interaction networks.

In the past 25 years, a tremendous effort by the biomedical research community has led to more than 10 million publications available in the PubMed (MEDLINE) database. In this study, we focus on a previously undervalued property of this outstanding repository: data from PubMed is time-resolved, because every article has a date of publication included. Thus, the evolution of scientific theories, terms and even gene names can be studied.

We have computed annual quotation frequencies for individual genes by tracing their names, symbols and synonyms in abstracts since 1975 [1]. We generated time series for 180 000 genes from human, mouse, *Drosophila*, yeast, zebrafish and *Escherichia coli*. Figure 1a shows the distribution of 250 of the human genes most referred to during the past 26 years. New gene discoveries are seen at different points in time, but subsequent reference to a gene

after its first description is clearly not random, and diverse patterns, or 'life cycles', can be distinguished.

Life cycles of successful genes

Characteristic life cycles of four genes are shown in Fig. 1b. These correspond to typical patterns found in 4532 genes that have appeared in the literature for at least 15 years. The glycolipid transporter GM2A, for example, is representative of the most frequent pattern, one that is shared by about 4200 genes. These have never attracted enough interest to become very important, and they exhibit a rather dull life cycle. Interleukin 3 (IL3) represents genes that have survived significant ups and downs in the collective scientific interest, but never boomed. The tumour suppressor gene p53, however, corresponds to a minor group that shows an exceptional increase of interest over time. These observations demonstrate how gene names have to overcome the selective mechanism of the scientific community to stand out from the rest [2]. The interest of the community in a specific gene, and thus its scientific impact, depends not only on a gene's molecular role, but also on the social needs within the scientific community, illustrated by the exceptional interest in genes such as CD4 and p53 which are involved in HIV infection and tumour development.

What we know about individual genes

The number of articles that mention a gene in a certain time period is a rough estimation of the information available for the gene. Considering this, we examined 8176 genes that have been known for 10 years and 2130 genes

Corresponding author: Robert Hoffmann (hoffmann@cnb.uam.es).