

- 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.

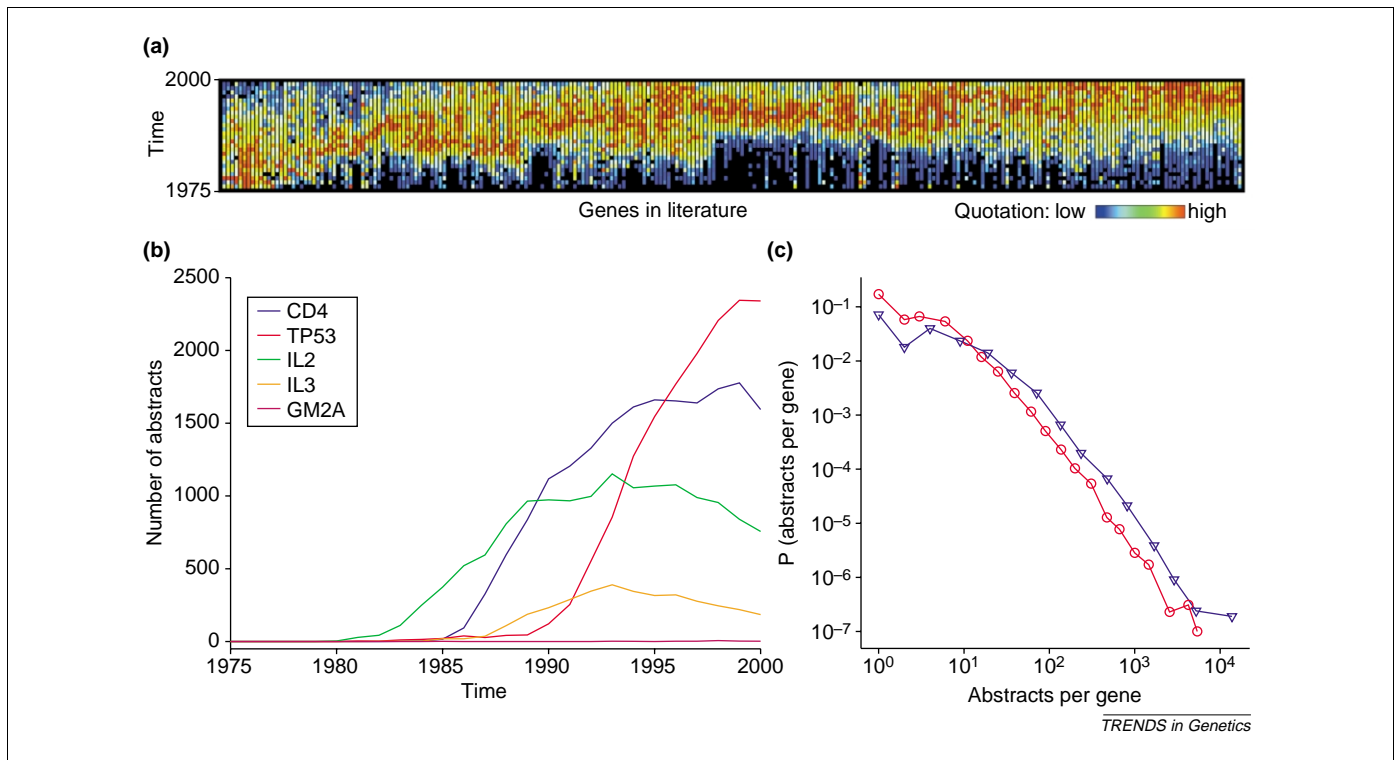
## Life cycles of successful genes

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).



**Fig. 1.** Life cycles of genes in MEDLINE abstracts from the past 26 years. (a) 250 human genes clustered along the horizontal axis according to their pattern of occurrence in the literature. Red areas indicate a peak in a gene's life cycle, black represents periods where a gene is not mentioned. (b) Characteristic life cycles of the genes CD4, p53, IL2, IL3 and GM2A. Annual frequencies of all genes were standardized to the year 2000 to account for the constant overall increase of articles per year. (c) Distribution of what we know about genes. Red circles, genes known for 10 years; blue triangles, genes known for 20 years. The  $x$ -axis represents the number ( $n$ ) of articles per gene (approximating the information known per gene). The probability  $P(n)$  of finding  $n$  articles about a given gene is plotted on the  $y$ -axis and decays as a power law,  $P(n) \approx n^{-\gamma}$ , appearing as a straight line on a log-log plot, where  $-\gamma$  is the slope of the line. The exponents for periods of different length,  $\gamma_{10 \text{ years}} = 1.6$  and  $\gamma_{20 \text{ years}} = 1.9$ , reveal that the extreme distribution remains although more knowledge has been accumulated during the longer period.

that have been known for 20 years. We find that the distribution of information for the genes decays as a power law function for both time periods (Fig. 1c); a few genes such as CD4 and p53 are frequently referred to and attract most attention, but the rest have had comparatively little published about them.

Power law distributions are the hallmark of systems in the critical state [3] (see Box 1 for definition). The scientific community as a small-world network is known to share important characteristics with these dynamic systems [4,5], where all members are permanently interacting and influencing each other. Given a certain complexity, the flow

of information within the scientific community can no longer be understood in terms of the behaviour of individuals; small changes can have effects out of proportion to their cause, leading, for instance, to the outstanding success of CD4. In other words, trends exist in the scientific community.

### The degree of protein interactions

These considerations raise the question of whether the frequency with which a gene is discussed in research abstracts reflects its importance in cellular functions. Or, is this extreme distribution of our scientific attention

#### Box 1. Small-world networks and the critical state

##### Small-world network

The small-world effect describes the finding that any two people in the world, chosen at random, are connected to one another by typically six intermediate acquaintances. Social networks of such topology allow for the rapid spreading of news, rumours, jokes or fashions. This also explains why diseases, transmitted from person to person, can result in global epidemics.

##### Critical state

Small-world networks meet the fundamental properties of complex systems, where the collective behaviour of a large number of interacting agents is not a simple combination of the behaviour of individuals. In physics, large dynamic systems have attracted great interest because of their tendency to organize into a poised state far out of equilibrium [a]. This critical state, sometimes called 'the edge of chaos', separates a frozen inactive state from a hot disordered state.

##### Domino effects

An important characteristic of systems in the critical state, studied on sand-pile models, is that a little perturbation (e.g. the addition of a single grain of sand) can lead to anything from an insignificant shift to an avalanche of unpredictable size. This unpredictability also applies to the dimension of epidemics, earthquakes or, as is the case here, the success of a gene within the scientific community. In molecular biology, power law properties have been discovered only recently in protein-interaction networks and metabolic pathways [b].

##### References

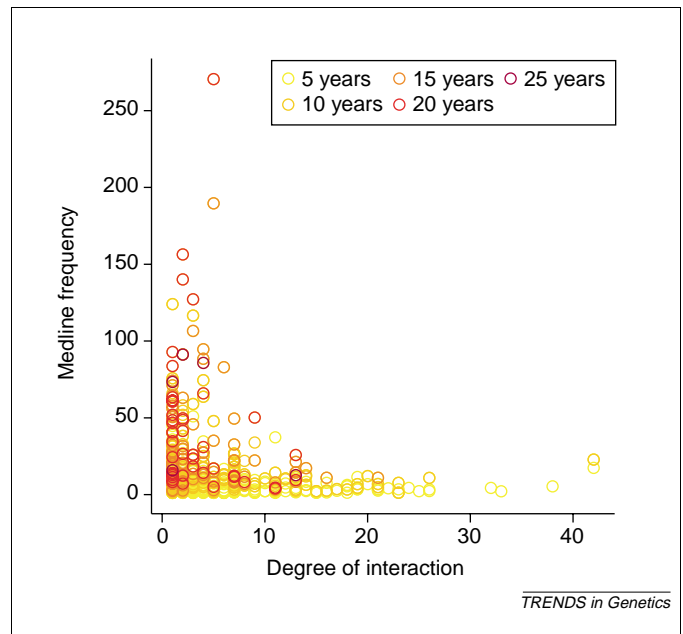
- a Bak, P. (1996) *How Nature Works: The Science of Self-Organized Criticality*, Copernicus
- b Jeong, H. et al. (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42

rather a result of the priorities within our society? For example, CD4, the most-frequently cited protein (Fig. 1b), is involved in HIV infection and of clear importance to our society, but from a purely biological point of view it has a similar role to other cell receptors. To address this question, we used the socially unbiased views from high-throughput experiments as a reference. In the past two years, large-scale methods have been used to generate global interaction networks of proteins. Based on yeast two-hybrid data, Barabasi and colleagues discovered the scale-free topology of these interaction networks, where a few proteins have many interactions, but most proteins have only a few interactions. There is a clear correlation between the number of a protein's interactions and its importance to the maintenance of cellular function [6]. Furthermore, the analysis of genomic data also reveals a strong selective constraint on genes with products that interact with many partners [7]. Thus, a significant indicator of a gene's importance to the cell is the degree of interaction of the encoded protein.

Therefore, we assessed this experimental measure of importance for those genes that are most frequently cited in MEDLINE and thus most important to the scientific community. Surprisingly, we find that there is no correlation between the degree of interaction [8] and the frequency of occurrence in MEDLINE for 380 yeast genes (Fig. 2). At the moment, there are no large-scale interaction data available for human, however, it is expected that this lack of correlation will be even more striking, because medical and sociological factors have an even stronger impact on human research. We believe that the discrepancy originates in the complex way information spreads within a small-world network such as the scientific community; a phenomenon that only becomes clear when considering the evolution of information over time.

#### Acknowledgements

We thank Ugo Bastolla for helpful discussion. This work was supported in part by the ORIEL and TEMBLOR EC projects.



**Fig. 2.** Lack of correlation between the degree of interaction of proteins and their frequency in the scientific literature. To ensure a high level of accuracy, protein interactions were only included if confirmed independently by at least two experimental methods: yeast two hybrid (Y2H), tandem affinity purification (TAP) and/or high-throughput mass spectrometric protein complex identification (HMS-PCI). Data from Mering *et al.* [8]. The length of time that genes are known to the scientific community (different colours) influences the correlation positively.

#### References

- Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28
- Dawkins, R. (1989) *The Selfish Gene*, 2nd edn, Oxford University Press
- Bak, P. *et al.* (1988) Self-organized criticality. *Phys. Rev.* 38, 364–374
- Milgram, S. (1967) The small world problem. *Psychol. Today* 2, 60–67
- Newman, M.E.J. (2001) The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA* 98, 404–409
- Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42
- Fraser, H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752
- Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403

0168-9525/03/\$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.  
PII: S0168-9525(02)00014-8

## The mouse *Ink4a/Arf* locus: a p53 pile-up at a tumour surveillance crossroads?

Robert G. Kelly

Department of Developmental Biology, Pasteur Institute, 28 Rue du Dr Roux, Paris, France

The *INK4a/ARF* (*CDKN2A*) gene is one of the most important defences against tumour development in mammalian cells [1]. The two products of this gene, p16<sup>INK4a</sup> and p14<sup>ARF</sup>, are encoded by separate 5' exons and alternative reading frames (Fig. 1b). p16<sup>INK4a</sup> and

p14<sup>ARF</sup> regulate cell division through Rb-dependent and p53-dependent mechanisms: p16<sup>INK4a</sup> inhibits the ability of cyclin D-dependent kinases to phosphorylate Rb; and p14<sup>ARF</sup> increases p53 stability by binding Mdm2 [1,2] (Fig. 1a). *INK4a/ARF* therefore acts at a crossroads in tumour surveillance, and is frequently lost or inactivated in human cancers.

Corresponding author: Robert G. Kelly (rkelly@pasteur.fr).