Cell PRESS

Attributing functions to genes and gene products

Neil S. Greenspan

Wolstein Research Building, Room 5130, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7288, USA

A major focus of modern biochemical, biophysical and cell biological research is the attribution of function to elements of structure: gene products, genes and higher-order cellular structures. Misunderstandings and controversies can arise in connection with such assignments, in part because of the logical complexity inherent in the relating of structure to function and the failure to distinguish clearly among the different senses in which function can be imputed to elements of structure. I explore distinct ways in which functions are connected to structures and factors that contribute to the context-dependence of such associations so that the multiple senses of function can be made explicit.

The intricacies of mapping function onto structure

A dominant theme in molecular and cellular biology is the correlation of function with structure on a molecular level. It is now so commonplace to attribute functions to genes and gene products (protein or RNA), or portions thereof, that biomedical scientists might be forgiven for thinking that it is a simple, perhaps even trivial, process that proceeds unerringly once the relevant biochemical or genetic manipulations have been implemented and the right assays have been performed. The underlying thesis of this commentary is that the logic by which function is attributed to or associated with structure is neither necessarily nor generally simple and that the concepts of function and structure–function correlation embody subtleties that deserve rigorous examination.

The concept of function as applied to genes and gene products is often taken for granted in the sense that no definition is offered in a typical textbook of biochemistry or molecular biology. It is assumed that every reader will know what the term means. In fact, as is true of many biological concepts (e.g. gene or specificity) of great importance, multiple overlapping but non-identical meanings are in use (either implicitly or explicitly) for function in typical discussions of biological cells and molecules [1,2].

Consider an investigator who has purified a cellular protein to homogeneity and has prepared a buffered solution of that protein that permits detection of a particular catalytic activity. If this investigator then demonstrated a particular enzymatic activity in the presence of the appropriate substrate, it would likely be regarded as reasonable for the experimentalist to ascribe that activity to the protein in question and refer to it as the (or at least a) function of the protein. For the sake of convenience, the activities of the purified protein in a buffered solution will be referred to as biochemical function.

However, if another investigator had begun studying the function of the same protein by deleting or otherwise modifying the DNA sequence encoding its amino acid sequence, he/she would possibly observe alterations to the phenotype of either individual cells or of a whole organism. In this instance, it would likely be regarded as reasonable for the experimentalist to ascribe these altered activities to the gene and gene product (assuming that there was only one translation product encoded by the gene under consideration) in question and refer, in some manner, to them as the functions of the protein (henceforth referred to as the genetic function). The crucial point to note is that the functions associated with our hypothetical protein in these two different scenarios would not necessarily be identical [3–5].

In the first case, where we can observe the behavior of what I will call protein X in relatively pure form, we have the gene product with little to none of its *in vivo* biological and biochemical context. Function in this setting is interpreted as 'effects mediated directly by protein X'. By contrast, in the second case, we can observe the behavior of the entire system, whether cellular or organismal, within which protein X normally resides but without protein X *per se*. Function in this setting is interpreted as 'the differential behavior of the system without X versus the system with X'. So, in one case, we attribute function on the basis of studying X in isolation without its authentic biological context whereas in the other case we attribute function on the basis of studying the authentic biological context with and without X (Figure 1). If one accepts that what I have referred to as biochemical and genetic function are just the extremes of a continuum and possibly a multi-dimensional continuum, then the broader point is that functional attributions can be somewhat dependent on the methods underlying those attributions.

For example, deletion of *Src* in mice failed to yield the neurological or hematological phenotypes anticipated on the basis of observed c-Src kinase activity and expression levels in various cell types [6]. Instead, the most prominent phenotype of $Src^{-/-}$ mice was that they appeared to suffer from a defect in bone remodeling leading to osteopetrosis. Another, perhaps even more compelling, example is provided by the results obtained when the severity of experimental allergic encephalomyelitis (EAE) was evaluated in wild type mice and in mice manipulated in various ways to alter the concentration of interferon-gamma (IFN-γ) in relevant tissues. Given the earlier evidence that IFN-γ

Corresponding author: Greenspan, N.S. (neil.greenspan@case.edu).

0968-0004/$ – see front matter doi:10.1016/j.tibs.2010.12.005 Trends in Biochemical Sciences, June 2011, Vol. 36, No. 6

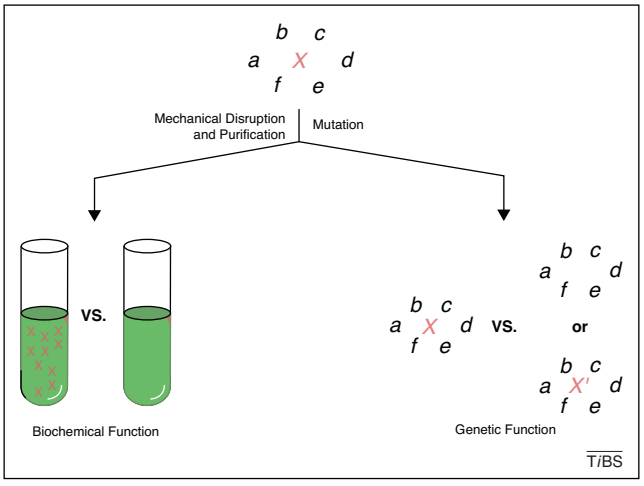

Figure 1. Illustration of substantive differences in alternative modes of function attribution for gene product X. The function of X can be assessed by purifying it to homogeneity and then determining what biochemical activities it exhibits in a simple buffer solution. This approach might be regarded as determining what we could call the biochemical function of gene product X. Alternatively, the function could be assessed by mutating the gene encoding gene product X and then determining relevant phenotypes of cells or organisms with the wild type and mutant genes. This approach might be regarded as determining what we could call the genetic function of gene product X. For biochemical function, the attribution of function relies on the behavior of authentic gene product X (in a simple buffer solution) in the absence of its normal biochemical and biological context and in comparison with buffer alone. In contrast, for the evaluation of genetic function, the attribution of function relies on the comparative behavior of cellular or organismal systems that possess or lack authentic wild type gene product X (either no gene product X or a mutant form of gene product X). The functions associated with gene product X by these contrasting experimental approaches might be overlapping but are not necessarily identical and might be challenging to relate to one another in some cases.

exhibited numerous activities that promoted inflammation, many experts in EAE anticipated that increasing concentrations of IFN-γ would worsen disease and decreasing concentrations of IFN-γ would greatly attenuate the disease process [7] In fact, in several studies, exacerbation of disease was seen in mice rendered IFN-γ-deficient (via antibodies or genetic manipulation) in comparison to wild type mice [8–11].

One further distinction worth noting is that between the function of a gene product (or gene) and the larger-scale process or processes to which it contributes. For example, a particular protein or RNA might participate in a noncovalent interaction or perform an enzymatic activity that serves to promote a higher-order effect, such as regulating blood pressure or mediating immunity to a particular pathogen. In ascribing a function to that gene product, it might help to minimize possible confusion if the functional attribution indicates how the activity of the gene product fits into a larger-scale pathway or network of interactions.

Alternative senses of function attributed to genes or gene products

Another complexity of interpreting the meaning of function relates to two (or more) different possible senses of the term. One sense is that the function of a gene product should be taken to refer to the biochemical activities that lead to outcomes favoring the survival and reproduction of the associated cell or organism. For example, according to this perspective a function of nuclear factor-kappa B (NF-κB) is to activate transcription of certain genes that contribute to evolutionarily adaptive inflammatory and immune responses [12]. In this sense, functions are basically the 'positive' contributions to cellular or organismal function.

Other interpretations of the meaning of function would be, somewhat in line with what is referred to above as a genetic definition, either the sum total of the effects on the cell or organism associated with a gene product or the net effect of a particular gene product on the cell or organism. Two possible comparative assessments are possible: (i) with reference to no such gene product or (ii) with reference to an allelic form (e.g. a putative wild type version) of the gene product. The sickle cell allele at the β-globin locus can be used to illustrate these points. The positive sense of function applied to the gene product encoded by the sickle cell β-hemoglobin allele would include oxygen transport and other related functions as well as the mediation of resistance to infection by *Plasmodiun falciparum*, which is observed when this allele is present in heterozygous form [13]. However, a more global view of the function (perhaps 'functional effects' would be preferred by many) of this allele and the corresponding gene product would include the effects associated with homozygosity at this locus (i.e. the often devastating cascade of pathophysiological consequences that follow from the vaso-occlusion and other abnormalities seen in sickle cell disease) [14].

Relational nature of molecular function

Of course, even the preceding discussion potentially oversimplifies. As is well recognized for many gene products, the effects attributable to a given protein (as defined by the amino acid sequences of the constituent polypeptide chains) or RNA molecule can vary in different cellular contexts, whether the latter term refers to different cell types in distinct tissues or organs or merely different stages of the same cell lineage. For example, activation of NF-kB, which is expressed in the majority of cell types, generates potentially different responses depending on the cell type and the range of stimuli received by the cell [12]. Although inter-disciplinary influence usually runs in the other direction, biomedical scientists would be well served by learning, as psychologists have, to be wary of falling prey to the 'fundamental attribution error' [15]. This fallacy involves automatically attributing the observed behavior of an entity to the intrinsic properties of that entity rather than considering the role of the particular circumstances pertaining to the entity during the period of relevant observation. For example, in studying the effects of administering recombinant IFN-γ on the severity of disease in adjuvant-induced arthritis in mice, the result depended upon when in relation to the disease-inducing immunization the cytokine was injected [16]. Conversely, it is possible to incorrectly attribute behavior (molecular or human) primarily to contextual factors when intrinsic properties are in fact dominant.

Furthermore, the same protein can pair with multiple different protein partners to form dimers or other oligomers that might mediate different and sometimes even directly opposed effects, such as facilitating or inhibiting transcription of particular target genes [17]. In fact, even the function of a given gene product can vary dramatically as a function of the intracellular concentration of that gene

product [18]. Furthermore, in the context of what has been termed squelching, an excess of the activating domains of the yeast transcription factor GAL4 (named for its role in galactose metabolism) recruited to certain DNA sequences can inhibit transcription of genes lacking the relevant cognate DNA sequences [19]. The suggested mechanism for this phenomenon is titration of one or more key components of the general transcriptional complex so that transcriptional activators binding to non-GAL4-binding sites lack these ancillary factors necessary for efficient transcription. This latter phenomenon illustrates the danger of thoughtlessly classifying molecules dichotomously, such as pro- or anti-activating for a particular cell type or pro- or anti-inflammatory (e.g. cytokines) with respect to tissue physiology and pathophysiology [7,20].

An additional layer of complexity is provided by the diverse functional consequences of different forms of post-translational modification and different sites of such modifications. For example, many of the transcriptional functions traditionally attributed to NF-kB are exhibited only if the molecule is post-translationally modified by a lysine methylase, nuclear receptor-binding SET domain-containing protein 1 (NSD1) [21].

A crucial aspect of molecular function, whether with respect to proteins, nucleic acids, other macromolecules or even small molecules, is that function, as normally understood, is generally not a completely intrinsic attribute of a molecule. Most function arises out of the interactions between molecules or between forms of energy and molecules [22]. Therefore, operationally, the functions that can be attributed to a given molecular species can depend on the presence or precise structures, configurations, or states of the molecules with which the species of interest interacts.

To illustrate the relational nature of molecular function, consider a cytokine, hormone or growth factor, X, that is bound by a single cell-surface receptor, XR. The functions associated with X are observed following engagement of sufficient numbers of XR molecules by a corresponding number of molecules of X. Now imagine that the gene encoding XR is deleted or mutated so that X no longer binds or no longer binds effectively with respect to the activation of signal transduction from XR. Although the structure of X is unchanged, the functional effects normally associated with X are now no longer observed in the presence of physiological concentrations of X. In other words, the functions of X are not attributes of X in isolation but are attributes of the relationship between X and XR in a particular cellular and molecular context.

For example, Laron-type dwarfism is due to homozygous or heterozygous mutations in the growth hormone receptor that render it less- or nonfunctional. The phenotype is clinically indistinguishable from growth hormone deficiency [23]. Similar parallels have been noted in the case of interleukin-4 (IL-4) and interleukin-4 receptor (IL-4R) deficiencies [24]. Of course, in cases involving multiple ligands for a given receptor or multiple receptors for a given ligand, the phenotypes for receptor deficiency and for ligand deficiency might differ. The crucial inference stemming from this point is that the function attributed to growth hormone or IL-4 is utterly dependent on the interaction with the receptor. In other words, the function actually arises out of the nature of the relationship between the hormone, growth factor or cytokine and the receptor or receptors with which it interacts.

Protein thermodynamics as a model for understanding genotype–phenotype relationships

There is an analogy (although imperfect) between the thermodynamic assessment of the contributions of individual amino acids to the energetics of protein–ligand interaction and the assessment of functional contributions of individual alleles to cellular or organismal phenotypes. In the context of protein–ligand binding, each position in the polypeptide amino acid sequence can be viewed as a locus and each amino acid that occurs at that position can be regarded as an alternative allele. Site-directed mutagenesis can then be used to determine which substitutions change ligand recognition, in what direction (better or worse) and to what extent in terms of affinity and the free energy change of complex formation, which in this domain of investigation is the functional or phenotypic endpoint of interest. What is perhaps uniquely advantageous about this arena of inquiry is that there is a highly validated and thoroughly quantitative theory, i.e. protein thermodynamics, to employ in dissecting the individual roles of various amino acids at various positions. No such comparable experimentally tractable and quantitative theoretical framework exists for definitively assessing the contributions of alleles to cell or organism function.

Thermodynamic contributions of individual amino acids change with single point mutations at other positions in the same polypeptide chain [25]. One of the early (and impressive) demonstrations of this point was performed by Lim and Sauer, who showed that amino acids different from the wild type residues could occupy the positions in the hydrophobic core (seven positions: 18, 36, 40, 47, 51, 57 and 65) of the N-terminal domain of phage lambda repressor without destabilizing the functional three-dimensional conformation of the protein [26]. In fact, they identified 30 distinct combinations of amino acids at the seven core positions that permitted the functionality of the protein, as assessed by the ability, like the wild type protein, to confer phage resistance when expressed at uninduced levels.

In assessing these 30 different amino acid sequences, Lim and Sauer found that certain amino acids at one position would permit functionality only when paired with a subset of the amino acid possibilities at another position in the core. So, for example, methionine (instead of the wild type valine) at position 36 was permissible (i.e. compatible with 'full function' as defined by the authors) if isoleucine, but not the wild type phenyalanine, occupied position 51. Alternatively, methionine could be accommodated at position 36 if non-wild type amino acids were found at positions 47 and 51 (isoleucine and methionine instead of wild type valine and phenylalanine) or 18, 47 and 51 (valine, isoleucine and leucine instead of wild type leucine, valine and phenylalanine).

The above results, in which interactions between any two amino acids exhibit flexibility when third party amino

acids are substituted, are analogous to findings relating to gene–gene interactions in fruit flies. In the pertinent studies, a mutation in one gene, *syntaxin-1A (Syx1A)*, altered which among 16 other genes enhanced or suppressed a behavioral phenotype (temperature-sensitive lack of coordination) interacted with one another and in what manner [27,28]. Similar flexibility in functional gene–gene interactions has recently been found in yeast [29].

Although thermodynamics provides an excellent framework for appreciating the context-dependence and interactivity involved in determining the functional contribution of any given structural element, there is an important difference between the thermodynamic contributions of individual amino acids and the biological contributions of genes and gene products. The energetic contribution of an amino acid can be reduced to a single number with units of kcal/mol, whereas the contributions of a gene product or gene will typically be multi-dimensional.

Gene, gene product, and mathematical functions

One way to think about the functions of genes and gene products that could be further developed relies on the explicit exploitation of mathematical concepts to a greater degree than heretofore. Perhaps the simplest notion, in this context, might be that the function of a protein is a value along some axis with the units of measurement specified, such as enzymatic activity in mass of substrate transformed into product per unit time or gene expression in number of copies of RNA transcript per unit time per fixed number of cells. A more satisfactory concept might be that the function of a gene product is a mathematical function of multiple variables. These variables might be the concentrations of other gene products, small molecule components of cells or extracellular fluid, and other environmental factors, such as temperature, ionic strength and pH, or even impinging electromagnetic radiation. In other words, the function of a gene product is the output resulting from multiple inputs. Because many gene products mediate two or more effects, it might be necessary to relate the function of such a gene product to a family of multivariable functions.

The functions of genes are frequently going to be even more complex to fully characterize than the functions of gene products, given that a single gene can give rise to multiple gene products. Thus, the mathematical description of the function of a gene that generates numerous gene products through alternative splicing of the mRNA and for which the resulting gene products are extensively subjected to post-translational modifications could be a large collection of functions corresponding to dozens or even hundreds of biochemically distinguishable gene products.

Although individuals employing computer-based models of cellular processes might wish to employ formal mathematical representations of gene product functions in their simulations, my intent with respect to most experimentalists is to encourage patterns of thought that more explicitly and more fully capture the complexities of gene product function than current default patterns.

Concluding remarks

There are multiple acceptable ways to map function to gene products and genes. Recognition of this reality can be valuable, both to avoid unnecessary misunderstandings and to recognize the limited range of application of any given structure–function correlation. Although it can be useful to pursue comprehensive compilation of gene product and gene functions in which all noncovalent interactions, enzymatic activities and biologically significant biophysical interactions (e.g. light absorption) are delineated, it should be kept in mind that in the general case, the contributions of proteins, RNAs and DNA sequences to the biological economy of a cell or organism and to the fitness of that cell or organism, are both context-dependent and subject to alteration mediated by new mutations.

Acknowledgements

I thank Toby Gibson for encouraging me to explore this topic, Derek Abbott and Brian Cobb for helpful insights, and the reviewers for their constructive comments. I declare no conflicts of interest pertaining to this article.

References

1 Greenspan, N.S. (2002) Opinion – wishful thinking and semantic specificity. *Scientist* 16, 12
2 Greenspan, N.S. (2007) Conceptualizing immune responsiveness. *Nat. Immunol.* 8, 5–7
3 Greenspan, N.S. (1992) Epitopes, paratopes, and other topes: do immunologists know what they are talking about? *Bull. Inst. Pasteur* 90, 267–279
4 Chao, D.M. and Young, R.A. (1996) Activation without a vital ingredient. *Nature* 383, 119–120
5 Greenspan, N.S. (1998) Genomic logic, allelic inference, and the functional classification of genes. *Perspect. Biol. Med.* 41, 409–416
6 Soriano, P. *et al.* (1991) Targeted disruption of the c-src proto-oncogene leads to osteopetrosis in mice. *Cell* 64, 693–702
7 Steinman, L. (2007) A brief history of T(H)17, the first major revision in the T(H)1/T(H)2 hypothesis of T cell-mediated tissue damage. *Nat. Med.* 13, 139–145
8 Billiau, A. *et al.* (1988) Enhancement of experimental allergic encephalomyelitis in mice by antibodies against IFN-gamma. *J. Immunol.* 140, 1506–1510
9 Ferber, I.A. *et al.* (1996) Mice with a disrupted IFN-gamma gene are susceptible to the induction of experimental autoimmune encephalomyelitis (EAE). *J. Immunol.* 156, 5–7
10 Willenborg, D. *et al.* (1996) IFN-γ plays a critical down-regulatory role in the induction and effector phase of MOG-induced encephalomyelitis. *J. Immunol.* 157, 3223–3227
11 Krakowski, M. and Owens, T. (1996) Interferon-γ confers resistance to EAE. *Eur. J. Immunol.* 26, 1641–1646
12 Oeckinghaus, A. and Ghosh, S. (2009) The NF-kappaB family of transcription factors and its regulation. *Cold Spring Harb. Perspect. Biol.* 1, a000034
13 Allison, A.C. (1954) Protection afforded by sickle-cell trait against subtertian malarial infection. *Br. Med. J.* I, 290–294
14 Steinberg, M.H. (2006) Pathophysiologically based drug treatment of sickle cell disease. *Trends Pharmacol. Sci.* 27, 204–210
15 Ross, L. (1977) The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in Experimental Social Psychology* (Vol. 10) (Berkowitz, L., ed.), In pp. 174–214, Academic Press
16 Jacob, C.O. *et al.* (1989) Heterogeneous effects of IFN-gamma in adjuvant arthritis. *J. Immunol.* 142, 1500–1505
17 Gibson, T.J. (2009) Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.* 34, 471–482
18 Zheng, T.S. *et al.* (2000) Deficiency in caspase-9 or caspase-3 induces compensatory caspase activation. *Nat. Med.* 6, 1241–1247
19 Gill, G. and Ptashne, M. (1988) Negative effect of the transcriptional activator GAL4. *Nature* 334, 721–724

20 Gor, D.O., Rose, N.R. and Greenspan, N.S. (2003) T_H1-T_H2, a Procrustean paradigm. *Nat. Immunol.* 4, 503–505
21 Lu, T. *et al.* (2010) Regulation of NF-kappaB by NSD1/FBXL11-dependent reversible lysine methylation of p65. *Proc. Natl. Acad. Sci. U.S.A.* 107, 46–51
22 Klotz, I.M. (1997) *Ligand-Receptor Enegetics: A Guide for the Perplexed*, John Wiley & Sons
23 Laron, Z., Kowadlo-Silbergeld, A., Eshet, R. and Pertzelan, A. (1980) Growth hormone resistance. *Ann. Clin. Res.* 12, 269–277
24 Noben-Trauth, N. *et al.* (1997) An interleukin 4 (IL-4)-independent pathway for CD4+ T cell IL-4 production is revealed in IL-4 receptor-deficient mice. *Proc. Natl. Acad. Sci. U.S.A.* 94, 10838–10843
25 Di Cera, E. (1998) Site-specific analysis of mutational effects in proteins. *Adv. Protein Chem.* 51, 59–119
26 Lim, W.A. and Sauer, R.T. (1989) Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* 339, 31–36
27 Greenspan, R.J. (2009) Selection, gene interaction, and flexible gene networks. *Cold Spring Harb. Symp. Quant. Biol.* 74, 131–138
28 van Swinderen, B. and Greenspan, R.J. (2005) Flexibility in a gene network affecting a simple behavior in *Drosophila melanogaster*. *Genetics* 169, 2151–2163
29 Bandyopadhyay, S. *et al.* (2010) Rewiring of genetic networks in response to DNA damage. *Science* 330, 1385–1389