Defining Genes

A Computational Framework

Peter F. Stadler, Sonja J. Prohaska, Christian V. Forst, David C. Krakauer

Manuscript date Wed May 13 19:18:27 CEST 2009

Abstract The precise elucidation of the gene concept has become the subject of intense discussion in light of results from several, large high-throughput surveys of transcriptomes and proteomes. In previous work we proposed an approach to constructing gene concepts that combines genomic heritablity with elements of function. Here we introduce a definition of the gene within a computational framework of cellular interactions. The definition seeks to satisfy the practical requirements imposed by annotation, capture logical aspects of regulation, and encompass the evolutionary property of homology.

Keywords Gene Concept, Homology, Computation, I/O-Relations

P.F. Stadler
Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany
Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
E-mail: studla@bioinf.uni-leipzig.de

S.J. Prohaska
Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany E-mail: sonja@bioinf.uni-leipzig.de

C.V. Forst
University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX, 75390-9066, USA E-mail: christian.forst@utsouthwestern.edu

D.C. Krakauer
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
E-mail: krakauer@santafe.edu

1 Introduction

The concept of the *gene* has come under intense scrutiny in recent years. This is largely in response to the recognition that the "standard" model of genes as beads on a genomic DNA string is inconsistent with the findings of high-throughput transcriptomics, see for example, (Pearson, 2006; Pennisi, 2007). As a consequence, several modifications of the concepts of the gene have been explored, ranging from purely structural definitions in terms of groups of transcripts (Gerstein *et al.*, 2007), the consideration of transcripts themselves as the central operational units of the genome (Gingeras, 2007), to *functional* notions (Scherrer and Jost, 2007). In (Prohaska and Stadler, 2008) we suggest that a "useful" gene concept should satisfy several criteria:

- The gene concept combines structural and functional components.
  - The gene concept is based on a well-defined notion of function that is amenable to experimental measurement.
  - The gene has a well-defined structural representation at the genomic sequence.
- Genes are heritable (not to imply that all inheritance is embodied in genes). In particular, the concept must be compatible with a suitable notion of (phylogenetic) homology.
- The gene concept is embedded in a larger framework that views "gene expression" as a form of computation.
- Genes are "expressed" from the DNA, hence genes are associated with transcripts and/or further processing products.
- The gene concept relates genomic mutations in a unique way to changes in a gene product, and thereby allows for the explicit construction of genotype-phenotype maps.

In this short paper we introduce a framework that satisfies these requirements. We do *not* claim that this framework is unique or optimal. We view this as an exercise in deriving a concrete model using the road map outlined in Prohaska and Stadler (2008).

2 A Chemical/Computational Framework

The basis for our construction is an abstract computational model of regulation. We start with the observation that cellular processes can be described as chemical reactions. This includes the interconversions of metabolites, the interactions of regulators, the aggregation of supermolecular structures, and the transport of molecules. There will be no need to operate at the level of individual molecules. It may be much more practical to employ coarse grained representations. For instance, transcription could be viewed as an input/output (I/O)-relation that takes genomic DNA and a set of transcription factors as input, and results in a specific output transcript. In this way, we emphasize the emphasize the computational aspects of bulk chemical reactions.

More formally, each I/O-relation is a quadruple $(\chi,[x],[p],[y])$ which we write in the more suggestive form

$$\chi : [x] \times [p] \to [y], \tag{1}$$

where $[x]$ is a list of material components (inputs) transformed into a list of material outputs $[y]$ by means of a process χ that depends on a list $[p]$ of additional influences. We call $[x]$ the arguments and $[p]$ the parameters of χ. Equ.(1) is an abstract, and arbitrarily coarse-grained representation of a chemical reaction. In chemical notation, we could write it in the form,

$$x_1, x_2, \ldots x_m \overset{p_1,p_2,\ldots p_k}{\longrightarrow} y_1, y_2, \ldots y_n \tag{2}$$

Equ.(1) can also represent transport "reactions", where input and output describe the same object(s) in different spatial locations or compartments, as well as other higher-level aggregate processes including replication, transcription, translation, or the production of biomass (if one chooses not to model such parts of the system in detail). In contrast to an implementation at the finest level, that of elementary chemical reactions, the I/O-relations are not required to satisfy conservation of mass or atom types. We are able, for instance, to ignore ubiquitous chemical species (such as H_2O, CO_2, or coenzyme A) and energy and redox currencies ATP and NADH, if we so choose. Our framework is consistent with, but may be more coarse-grained than, a full-fledged representation of all chemical reactions. This is a common coarse graining in Systems Biology models (Palsson, 2006). For our purposes, it will be convenient to model transcription and translation as I/O-relations that "produce" primary transcripts from a DNA template and a polypeptide from an RNA template. Equ.(1) may also include compartment/spatial information and thus can describe cellular processes of more than one cell or organism, including a complete microbial community or even entire ecologies with complex predator-prey dynamics. Note that some or all elements of the output list $[y]$ of χ will typically appear as inputs $[x]$ and/or parameters $[p]$ of other I/O-relations ξ.

A system Ξ of I/O-relations over a given domain of "objects" X has a natural interpretation as a model of *computations* on X (Berry and Boudol, 1992; Taylor, 1998). This gives us considerable freedom in implementing a model of cellular processes in the form of equ.(1) depending on: (1) the level of aggregation or abstraction beyond elementary chemical reactions; and (2) the effect that a parameter p must have on the outcome of χ to be considered relevant. For example, we may define p to be relevant to a particular I/O-relation χ if the absence of p makes the transformation χ impossible. Alternatively, we could consider p a relevant influence whenever it affects the reaction rate.

Before proceeding, a formal issue requires attention. Each process χ links a particular triple of input, output, and parameter lists. Hence, transformations utilizing the same input $[x]$ to produce different outputs $[y'] \neq [y]$ are necessarily two distinct reactions χ and χ'. Here, we admit only physical objects as elements of the input and output lists $[x]$ and $[y]$. The parameters $[p]$, on the other hand, may be either objects or physical quantities such as temperature or pH. The parameter list may be empty, $[p] = \emptyset$, e.g. in sponteneous chemical reactions or transport by diffusion.

If an object a appears both as an argument, $a \in [x]$, and as a parameter, $a \in p$, in the same I/O-relation χ, this implies an autocatalytic mechanism. The argument and the parameter are necessarily two different instantiations of the object type a. The simplistic distinction between arguments and parameters in the formalism is akin to the notions of *cis* and *trans* action in molecular biology. Note, however, that the concepts are not equivalent in all cases.

3 Information Metabolism

The crucial *assumption* in our exposition is that – given a suitable collection of I/O relationships – we can single out the transformations among informational molecules (i.e., heteropolymers that are capable of encoding information, such as RNA, DNA, and polypeptides) from the generic "reaction soup" of all I/O-relations. To this end, we identify those reactions in which both the input list and the output list $[y]$ contains informational molecules (DNA, RNA, or peptide) of a single type. Depending on whether the informational molecules in $[x]$ and $[y]$ are of the same type or not, χ represents either processing or one of several information transfer mechanisms (translation, transcription, reverse transcription, and replication). Each informational molecule has an

explicit representation as sequence of nucleotides or amino acids $x = (z_1 z_2 \dots z_n)$. Among the transformations of informational molecules, we single out those reactions that satisfy an additional property of *traceability*.

Definition 1. The I/O-relation χ is *traceable* if and only if for each informational molecule $y \in [y]$ in the output and each letter $y_k \in y$ we can uniquely determine whether

(a) y_k is an encoded letter, i.e., its identity can be traced to a single letter or an interval (e.g. a codon) on an input sequence $x \in [x]$; or
(b) y_k is not an encoded letter, in which case its identity is determined by the reaction χ.

The collection Υ of traceable I/O-relations involving informational molecules represents the "linear" part of information metabolism. We suggest that it is not only well-defined but encompasses important processing steps in the "life-history" of a transcript, including: primary transcription, splicing, translation, insertion editing, cleavage, intein extraction, chemical modification, poly-adenylation, etc. Thus we can interpret $\Upsilon \subset \Xi$ as the subsystem of *gene expression* in a cell, organism, or ecosystem.

In a traceable reaction, we can determine, for any collection of sequence intervals in the output $[y]$, the collection of all those sequence intervals in the input $[x]$ that gave rise to the encoded letters in the output. This inverse map χ^{-1} gives us a well-defined "footprint" of each stretch of output sequence on the input. The concatenation of such inverse maps is well-defined and allows each sequence position in an informational molecules z to be traced back to its genomic source, the *genomic footprint of z*. This construction will provide us with the structural part of our gene concept. Any letter x_k of x that is not contained in the genomic footprint $\Gamma(x)$ of x is identified as being inserted or appended at a particular stage in the production of z. The genomic "source" $\Gamma(x)$ can be one of the cell's genomes or, for instance, an intruding viral transcript.

In some cases a product x that appears in the linear part of the information metabolism may have an empty genomic footprint $\Gamma(x) = \emptyset$. This is the case e.g. for the so-called non-ribosomal peptides (NRPs), which are synthesized *de novo* without using the templating function of a messenger RNA (Walton, 2006).

Theoretically, the definition of the genomic footprint $\Gamma(x)$ requires detailed knowledge of the complete gene expression pathway leading to the production of x. In practise, however, $\Gamma(x)$ can be *approximated* by mapping the sequence of a biopolymer x to the underlying genome. Current procedures of genome annotation do this in a way that incorporates knowledge of the genetic code, splicing, end processing, editing, etc. In other words, given the data provided by proteomics and transcriptomics, computational procedures can already produce reasonable estimates of genomic footprints. These are used in current genome annotations and genome browser systems (Furey, 2006; Karolchik *et al.*, 2008). In line with the prevailing simplified model of the transcriptome and proteome, genome browsers restrict themselves to co-linear arrangements of footprints of a given product, thereby neglecting rearrangements, trans-splicing, and conceivably other "non-monotonous" processing mechanisms.

4 Function

Within the framework of I/O-relations, the *function* of an object $z \in X$ becomes a derived property. It appears natural to identify the function of z with those processes that it influences. (Structural proteins are captured by formulating pseudo-reactions that describe the formation and reorganization of supramolecular structures.) We insist that "being processed" (i.e., being an input for an I/O-relation), "being produced" (i.e., appearing as output of an I/O-relation), and "encoding information" is **not** in itself a function. The reason for this distinction is that we need to avoid the trivial notion that everyhing is functional just because it is present.

Formally, let us denote by $\mathrm{param}(\chi)$ the set of parameters of the I/O-relation χ. This leads us to

Definition 2. The *function* $\mathbf{Fct}(z)$ of an object $z \in X$ the set of I/O-relations

$$\mathbf{Fct}(z) := \{\chi | z \in \mathrm{param}(\chi)\} \tag{3}$$

While this looks somewhat contrived, it forms the basis of the official *Nomenclature of Enzymes* (see (NC-ICBMB and Webb, 1992) and the annual supplements at `http://www.chem.qmul.ac.uk/iubmb/enzyme/supplements/`), in which enzymes are named for the chemical reactions that they catalyse: An *alcohol dehydrogenase*, for instance is *per definitionem* an enzyme that catalyzes the dehydrogenation of alcohols. In our setting, this would be respresented by associating with z a set of I/O-relations $\mathbf{Fct}(z)$ that consists (largely) of chemical reactions χ describing the dehydrogenation of various alcohols. Also note that nothing precludes an object from having multiple disparate functions in this framework i.e., $\mathbf{Fct}(z)$ can be very large and contain several, semantically different, groups of I/O-relations.

Applying this notion of function, we identify the collection of *functional informational molecules* as those biopolymer sequences x for which $\mathbf{Fct}(x) \neq \emptyset$. Note, again, that in this way we pin function only to physical objects that *influence* transformations (of other objects). "Being transformed", on the other hand, by construction does not count as a function in itself. Furthermore, an informational molecule does not acquire a function, for example, by housing a cis-regulatory element that regulates its own subsequent processing step. Our model declares that the function of regulating/influencing subsequent processing step ξ is attributed to the parameter(s) of ξ, not to the argument of ξ. Intermediate

processing steps thereby will *not* typically have a function. For instance, if x is the mRNA coding for a protein p, we will often observe that $\mathbf{Fct}(x) = \emptyset$, while the translation product p of x typically will have some function as an enzyme, signal molecule, or structural protein, such that $\mathbf{Fct}(p) \neq \emptyset$, Fig. 1. Not all proteins are necessarily functional. The precursor p of one or more small hormone peptides $p'_1, \ldots, p'_r$, for instance, may not have any function alone (Dicou, 2008). In this case we have $\mathbf{Fct}(p'_i) \neq \emptyset$ for the hormone peptides p', but $\mathbf{Fct}(p) = \emptyset$ for the prohormone protein, see c and d in Fig. 1.

In practise, $\mathbf{Fct}(z)$ is dependent upon the experimental and computational methodologies employed to determine the processes (I/O-relations) that are dependent upon z. Improved measurements thus have the potential to change our representation of $\mathbf{Fct}(z)$.

We note, finally, that this simple notion of function brings with it a completely natural definition of *functional equivalence*: Two objects p and q are functionally equivalent if $\mathbf{Fct}(p) = \mathbf{Fct}(q)$.

5 Genes

We are now in a position to define a gene.

Definition 3. A *gene* on a given genome is the pair $(\Gamma(z), z)$ consisting of a **functional** informational molecule z and its genomic footprint $\Gamma(z)$, i.e., the collection of intervals on the genome that give rise to the encoded letters in the sequence z through a sequence of I/O-relations in Ξ.

In definition 3 we require that there is *at least one* sequence of I/O-relations linking $\Gamma(z)$ to the gene product z. Alternatively, we might want to include the specific sequence of I/O-relations in the definition of the gene. The distinction between these two alternative points of view is whether we would require that every gene has a unique way of being processed (each particular sequence of I/O-relations linking $\Gamma(z)$ to z), or whether we allow that a gene $(\Gamma(z), z)$ can be expressed in alternatives ways. At present, we lack sufficient evidence of alternative transcripts processed into the same functional "gene product", to decide which version is biologically more useful.

Definition 3 of course allows overlapping genes, and in particular, different genes with the same genomic footprint: if the same collection of genomic intervals gives rise to a different product (necessarily via a different processing cascade) we have two distinct genes. Thus, as in the proposal of (Scherrer and Jost, 2007), we label distinct (functional) splice variants as distinct genes. Similarly, if the same product z can be produced from different genomic footprints, we also speak of two distinct genes (an example are some pairs of paralogous microRNAs with identical mature products (Griffiths-Jones *et al.*, 2008)).

6 Functional Similarity and Homology

In taking an extreme "functional" point of view, one might want to interpret genes (in the above sense) as "the same" if products are functionally equivalent. As far as we can see, this choice leads to problems with notions of homology. We now discuss the connection of our gene definition with the homology concept.

In order to analyze the notion of a function in greater detail, we assume that there is a distance function D that allows us to measure how different two I/O-relations χ' and χ'' are. In the analysis of chemical reaction networks, distance functions between reactions are typically based on a notion of differences Δ among underlying objects (Maggiora and Shanmugasundaram, 2004). Given a measure of dis-similarity of objects, one constructs a dis-similarity for input, output and parameter lists, which are finally combined into the desired measure D (Tohsato and Nishimura, 2007). We can use D to cluster the elements of $\mathbf{Fct}(z)$ into distinct functional classes and to construct a measure $\mathbf{D}$ of the functional differences between two objects. The functional distance D between I/O-relations can be extended naturally to sets of I/O-relations and hence implies a notion of functional distance $\mathscr{D}(x,y) = D(\mathbf{Fct}(x), \mathbf{Fct}(y))$, which is conceptually related to network distance (Forst *et al.*, 2006).

In the case of information molecules, we can think of the object distance Δ as an edit or alignment distance that measures a quantity of sequence similarity. In contrast, $\mathscr{D}$, which can be derived from Δ and the system of I/O-relations Ξ, measures functional dissimilarity. Homology-based gene annotation is based on the observation that Δ and $\mathscr{D}$ are correlated in practice. That similar sequences (of functional information molecules, or of their genomic footprints) often — but not always — give rise to products with similar functions. Deviations from this rule exist in nature and indicate either large changes in function that are acquired by closely related sequences (small Δ, large $\mathscr{D}$), or to horizontal replacement of a gene by a functionally equivalent one (small $\mathscr{D}$, large Δ). An example of the first type is the imprinting-related ncRNA *Xist* in Mammalia, which originated by pseudogenization of the *lnx3* transcript whose primary product is a PDZ-like ring finger protein (Duret *et al.*, 2006). An example of the latter type are several unrelated "clans" of serin proteases that share only the common catalytic triad Ser-His-Asp (Krem and Di Cera, 2001).

The concept of homology is also subject to intensive discussion, with several competing definitions, see e.g. (Laubichler, 2000; Brigandt and Griffiths, 2007). In the context of evolutionary and molecular biology, one requires that homologous characters are linked by common descent. In the strict "phylogenetic" definition, this is the only requirement. In our framework, phylogenetic homology of $(\Gamma(x), x)$ and $(\Gamma(y), y)$ is naturally established by the existence of a com-

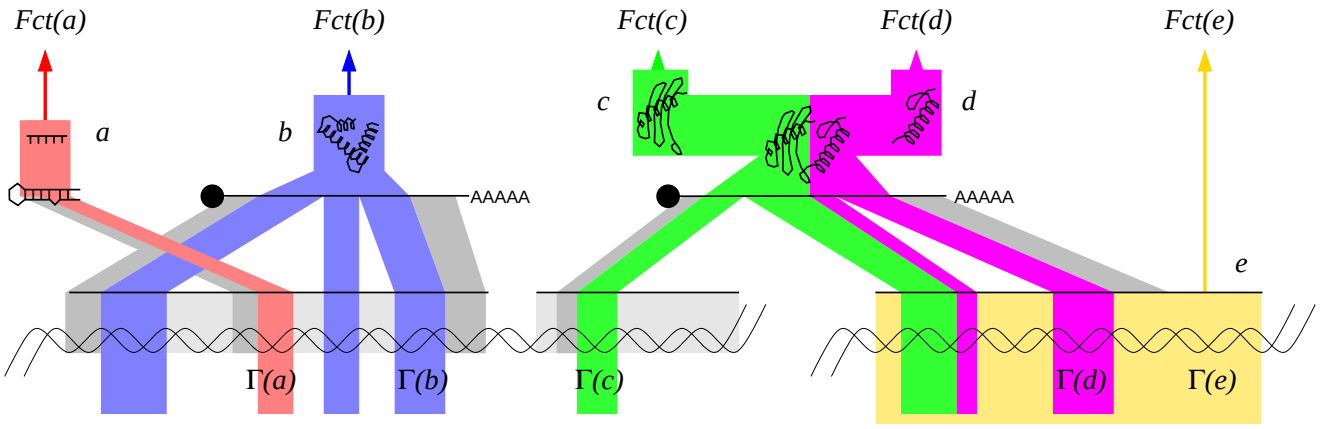

Fig. 1 Functional objects a to e and relationships with their genomic footprints $\Gamma(a)$ to $\Gamma(e)$.
A functional RNA molecule (e.g. a miRNA) with function $\mathbf{Fct}(a)$ is processed in two steps from an intronic sequence. Its image on the DNA is the genomic footprint $\Gamma(a)$. The genomic footprint $\Gamma(b)$ of the functional protein b is a discontinuous stretch of DNA corresponding to the coding sequence (CDS) including the start codon but excluding the stop codon. The mRNA includes UTRs that also map back to the DNA as well as parts without footprints on the DNA (the 5'-cap and the poly-A tail). The functional proteins c and d are obtained by cleavage of the (non-functional) precursor cd. The later is encoded by a trans-spliced mRNA. The footprint $\Gamma(c)$ is distributed over two DNA molecules. The primary transcript e has an additional function $\mathbf{Fct}(e)$ that is independent of its role as precursor of the mRNA of cd. As a consequence, $\Gamma(e)$ overlaps with both, $\Gamma(c)$ and $\Gamma(d)$. In all cases, the gene is the pair $(\Gamma(x),x)$ composed of the genomic footprint $\Gamma(x)$ and the resulting functional molecule x.

mon ancestor of the genomic footprints $\Gamma(x)$ and $\Gamma(y)$. In practice, evidence for homology of genes can be evaluated by comparative sequence analysis of $\Gamma(x)$ and $\Gamma(y)$, i.e., in terms of Δ, the same way as this done for protein, RNA, or DNA sequences. The only modification is a more precise recipe for delimiting the sequences that need to be compared.

The strictly functional notion that identifies functionally equivalent genes runs into an insurmountable problem because there is nothing to prevent it from identifying objects that do not share common descent. This would lead to a gene concept that is not compatible with (phylogenetic) homology.

Our framework also provides a starting point for formalizing notions of homology that postulate functional similarities in addition to common descent, e.g. via a suitable concept of homology for I/O-relations. We view this as a research agenda beyond the scope of this contribution.

7 Discussion

In this contribution we have introduced a formal framework that satisfies several of the intuitive requirements of a gene definition.

- We have emphasized a functional/computational notion that remains instantiated and identifiable in genomic material. We argue that both properties are necessary: Purely structural gene-definitions are useless at best and harmful at worst for annotation purposes because they tend to blur the information provided by transcriptomics and proteomics data. On the other hand, modern Molecular Biology can only work with a gene concept that is firmly rooted to sequence information and therefore annotatable at a genomic level.
- The gene concept includes genomic heritability and hence can be used to establish homology relationships over large phylogenetic distances.
- As far as we can tell, the concept is consistent with a fine-grained system of homology concepts that distinguish between sequence homology (of the genomic footprint), homology of the gene (in terms of both sequence and function), and concepts that also include homologies between intermediate processing products. This will be valuable when extending this approach to, e.g., Developmental Biology.
- Our framework suggests a definition of the *phenotype* as the collection of functions "performed" by an organism, $\Phi = \{\mathrm{Fct}(z)\}$. This phenotype is in turn evaluated by a complex *fitness function* $f(\Phi)$ to determine the viability and selective properties of the phenotype Φ. This provides compatibility with operational models of evolution such as Population Genetics.
- Our construction is consistent with the concept of genetic engineering, which is based on the assumption that genomic changes can be transferred (most of the time) in an unambigous way into gene products.
- There is a relatively simple relationship between "classical genes" and our concept: for instance, the genomic CDSs of functional proteins, as well as the genomic loci of mature microRNAs, tRNAs, rRNAs, are all by default (genomic footprints of) genes in our sense.

– Pragmatically, the currently available methods of data analysis e.g. in comparative genomics just need to be applied somewhat more carefully to selected data.
– This proposal does not require the introduction of an army of auxiliary concepts unlikely to be adopted by practicing biologists.

This approach, does require however that we forfeit properties that might be desirable for certain cases. For instance, the genomic footprints of genes are in general proper subsets of the footprints of transcripts that describes a more inclusive functional set. We reject however the notion that a gene comprises all region/regions of DNA required to produce a function. The reason for doing this is to keep the "sphere of influence" of a gene limited. In an all-inclusive view that includes everything necessary to unfold a given product, large parts of the cellular machinery (and their DNA loci) would become constituents of all genes. We find such an approach untenable because it does not lead to "genes" upon which one can perform meaningful, discriminatory experiments.

Acknowledgement

This work originated in the aftermath of the Workgroup on the *Complexity of the Gene Concept*, which took place at the Santa Fe Institute Mar 8-11, 2009 thanks to funding by the *James S. McDonnel Foundation in Robustness*. We are grateful for the intensive discussions on the topic with all the participants of this workshop. Sven Findeiß' comments on the manuscript are gratefully acknowledged. This work was funded in part by a grant from the Volkswagen Foundation on "Evolution of networks: robustness, complexity and adaptability" to PFS, and a grant on Innovation in Biological and Technological Systems from the David and Lucile Packard Foundation to DCK.

References

Berry G, Boudol G, 1992. The chemical abstract machine. Theor Comp Sci 96:217–248.

Brigandt I, Griffiths P, 2007. The importance of homology for biology and philosophy. Biology and Philosophy 22:633641.

Dicou E, 2008. Biologically active, non membrane-anchored precursors: an overview. FEBS J 275:1960–1975.

Duret L, Chureau C, Samain S, Weissenbach J, Avner P, 2006. The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312:1653–1655.

Forst CV, Flamm C, Hofacker IL, Stadler PF, 2006. Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. BMC Bioinf 7:67.

Furey TS, 2006. Comparison of human (and other) genome browsers. Hum Genomics 2:266–270.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M, 2007. What is a gene, post-ENCODE? history and updated definition. Genome Res 17:669–681.

Gingeras TR, 2007. Origin of phenotypes: Genes and transcripts. Genome Res 17:682–690.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ, 2008. `miRBase`: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158.

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ, 2008. The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res 36:D773–D779.

Krem MM, Di Cera E, 2001. Molecular markers of serine protease evolution. EMBO J 20:3036–3045.

Laubichler MD, 2000. Homology in development and the development of the homology concept. Integ Compar Biol 40:777–788.

Maggiora GM, Shanmugasundaram V, 2004. Molecular similarity measures. In: Bajorath J, editor, Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery, vol. 275 of *Methods in Molecular Biology*, (pp. 1–50). Heidelberg: Springer.

NC-ICBMB, Webb EC, editors, 1992. Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. San Diego: Academic Press.

Palsson BO, 2006. Systems Biology. Cambridge UK: Cambridge University Press.

Pearson H, 2006. Genetics: What is a gene? Nature 441:398401.

Pennisi E, 2007. DNA study forces rethink of what it means to be a gene. Science 316:15561557.

Prohaska SJ, Stadler PF, 2008. "Genes". Th Biosci 127:215–221.

Scherrer K, Jost J, 2007. The gene and the genon concept: A conceptual and information-theoretic analysis of genetic storage and expression in the light of modern molecular biology. Th Biosci 126:65–113.

Taylor RG, 1998. Models of Computation and Formal Languages. New York: Oxford University Press.

Tohsato Y, Nishimura Y, 2007. Metabolic pathway alignment based on similarity between chemical structures. IPSJ Digital Courier 3:736–745.

Walton JD, 2006. HC-toxin. Phytochemistry 67:1406–1413.