A Defense of Syntax-Based Gene Concepts in Postgenomics: Genes as Modular Subroutines in the Master Genomic Program

Tudor M. Baetu†‡

The purpose of this article is to update and defend syntax-based (conserved DNA-sequence motifs) gene concepts. I show how syntax-based concepts have been and can be extended to accommodate complex cases of genome expression, regulation, and processing. In response to difficult cases and causal parity objections, I argue that the syntax-based approach fleshes out a deflationary concept that defines genes in terms of sequences and organizational features of the genome that contribute to a phenotype.

1. Introduction. One of the consequences of the elucidation of the structure of DNA was the identification of genes with DNA sequences (Watson and Crick 1953). Shortly after, the central dogma (Crick 1958) allotted DNA the role of an archive containing “information” for the synthesis of proteins (the “blueprint” gene). The elucidation of the mechanisms of genome expression suggested that genes are characterized by a set of conserved sequence motifs required for the interaction with the transcriptional and translational machinery of the cell to generate gene products from a DNA template. The transcription unit and the open reading frame molecular concepts are often characterized in such a manner. For instance, the basic prokaryotic transcription unit (fig. 1) is typically characterized

†To contact the author, please write to: 4710 College Avenue, College Park, MD 20740; e-mail: tbaetu@hotmail.com.
‡I would like to thank Lindley Darden, Erika Milam, Eric Saidel, Pamela Henson, Joan Straumanis, Lane DesAutels, Christophe DiTeresi, and the DC History and Philosophy of Biology group for helpful discussion and comments on earlier drafts. This work is supported by Fonds de la recherche sur la société et la culture, Québec, Canada (grant 127231).
Figure 1. Prokaryotic transcription unit.

by two regulatory regions responsible for recruiting the RNA polymerase (enzyme synthesizing an RNA transcript from the DNA template), a transcribed coding sequence and a transcription termination sequence. Within the transcribed sequence, one or more open reading frames (sequences translated into protein gene products) are marked by a site responsible for recruiting ribosomes (the Shine-Delgarno box), a “begin translation” site (ATG codon), a succession of codons specifying the final peptide sequence, and an “end translation” site (termination codon).

According to syntax-based concepts, a region of the genome that is expressed as gene products is characterized by the presence of syntax-like conserved sequences providing instructions for transcription and translation; in contrast, a region of the genome that is not expressed does not contain such sequences. The assumption (amply verified by in vitro binding studies and the success of a variety of experimental techniques ranging from in vitro synthesis protocols to genetic engineering) is that if a given conserved sequence motif is present in a stretch of DNA, it will most likely serve as a binding site for some component of a mechanism of genome expression. Thus, if properly annotated and analyzed in the right order, the conserved syntax-like sequence motifs contained in the genome can help us predict how the genome will be regulated, what genomic sequences will be expressed and as what gene products.

Molecular biologists have relied on conserved sequences, ever since their discovery, as a means to define and discover genes, formulate hypotheses and make predictions about the regulation of genome expression and the function of gene products, guide genetic engineering, and design in vitro transcription and translation protocols. Sequence alignment algorithms such as Basic Local Alignment Search Tool (BLAST; Altschul et al. 1990) are used on a daily basis in laboratories around the world to identify putative genes based on the presence of syntax-like motifs typically required for transcription and translation. The establishment of bioinformatics as a new discipline aiming, among other things, to further develop
genome annotation protocols (Mandoiu and Zelikovsky 2008) suggests that such practices will continue in the near future.

Strangely enough, philosophers of biology do not partake of this enthusiasm. Griffiths and Stotz (2007) claim that the “nominal gene” defined by current annotation practices will soon be superseded by “postgenomic” concepts of the gene. In their view, the discovery of gene rearrangement, nested genes, alternative promoters, alternative splicing, trans-splicing, RNA editing, frameshifting, alternate stop codons, polyproteins, and various other complications due to regulatory and post-transcriptional/post-translational processing mechanisms posit insurmountable difficulties for a syntax-based approach. Mechanisms of genome expression regulation seem to determine which, when, and where coding sequences are transcribed and translated, while mechanisms of post-transcriptional/post-translational processing seem to control how the information contained in original DNA template is used. Given the shortcomings of current syntax-based concepts, there is a growing popular tendency to treat the genome as a set of sequences that can be “read” (transcribed and translated) and “processed” (RNA splicing and editing, post-translational modifications of peptides, etc.) in a variety of ways, depending on developmental (Griffiths and Neumann-Held 1999), cellular (Stotz 2011), and environmental contexts (Stotz 2006). According to this view, genes are best described as “things you can do with your genome” (Griffiths and Stotz 2006, 500).

This raises the question, “Is this the end of syntax-based gene concepts?” According to these authors, the answer seems to be yes: “The ‘same’ DNA sequence potentially leads to countless different gene products; different sequences might code for identical products, and the need for a rare product asks for the assembly of a novel messenger RNA (mRNA) sequence. Hence the information for a product is not sufficiently encoded in the targeted DNA sequence but has to be supplemented through sequence information provided by elements outside the coding sequence, such as transcription, splicing, or editing factors” (Stotz 2006, 905). “The sequence of the DNA can . . . be compared to a sequence of letters without spaces or punctuation marks. . . . A different developmental system imposes a different scheme over the letters, that is, over the DNA sequence. It is therefore misleading to think of functional descriptions of DNA, such as ‘promoter region,’ as explicable solely in terms of structural descriptions of DNA, such as ‘sequence.’ . . . The gene is identified not with these DNA sequences alone but rather with the process in whose context these sequences take on a definite meaning” (Griffiths and Neumann-Held 1999, 661).

In this article I challenge this view. I argue that current syntax-based concepts can be extended to accommodate complex cases of genome expression regulation and processing. The syntax-based approach proposed
in this article has to its advantage three virtues—modularity, retrocom-
patibility, and the ability to explain why the genome is regulated, ex-
pressed, and processed in an orderly, predictive fashion—which rival ac-
counts cannot bolster. In response to difficult cases and causal parity ob-
jections, I argue that a syntax-based approach fleshes out a deflationary
concept that defines genes as genomic sequences and organizational fea-
tures of the genome contributing to (rather than determining) a pheno-
typic outcome.

The article is organized as follows: In section 2, I show how an updated
version of current syntax-based concepts can accommodate complications
due to mechanisms of genome expression regulation and processing. In
section 3, I elaborate a modular, three-level model of genome organization,
and, in section 4, I discuss its advantages. In section 5, I discuss the limits
of syntax-based approaches and possible avenues of improvement. Finally,
in section 6, I summarize my conclusions and arguments.

2. Expanding Current Syntax-Based Concepts. The critiques of current
syntax-based gene concepts are fueled by two main concerns:

1. Genome expression regulation (the role mechanisms of gene ex-
pression regulation play in specifying which DNA sequences are
expressed).
2. The breakdown of the DNA template–gene product sequence col-
linearity (due to posttranscriptional and posttranslational process-
ing).

The solution hinges on the fact that, for the most part, the mechanisms
of genome expression regulation/processing work in concert with con-
served DNA sequences. This suggests that one way to cope with the
complexities brought about by these two issues is to update and extend
available syntax-based concepts. A step in this direction is illustrated by
Gerstein and colleagues’ (2007, 671) attempt to define genes as “subrou-
tines in the genomic operating system.” The authors propose that the
structure of the genome should be described in very much the same way
that grammars are used to describe computer programs—with a precise
syntax of upstream regulation, exons, and introns.

Thus, instead of taking into account only the basic syntax of transcrip-
tion and translation underlying the transcription unit and the open reading
frame concepts, Gerstein et al. suggest an extended syntax including se-
quences recognized by transcription factors (e.g., activator/repressor bind-
ing sites associated with regulatory networks; Levine and Davidson 2005)
and sequences signaling splicing (Black 2003).

In fact, nothing prohibits us from extending this approach to any se-
quence-specific aspect of genome expression, including chromatin regu-
lation (matrix binding and nucleosome assembly; Turner 2001), RNA protein binding associated with translational regulation (Mazumder, Seshadri, and Fox 2003), and posttranslational modifications (glycosylation, phosphorylation, cleavage of polyproteins). While there are still many gaps to be filled, there is ample evidence that conserved DNA sequences play a necessary role in specifying sites within the genome (or its transcribed/translated counterparts) where various components of mechanisms of genome expression, regulation, and processing bind and initiate activities that lead to the expression of specific portions of the genome and specific modifications of transcribed/translated products. I propose therefore a more general syntax-based approach that goes beyond the immediate scope of Gerstein et al., who are concerned only with transcriptional regulation and RNA splicing.

3. The Gene as Modular Subroutine in the Master Genomic Program. Following the analogy developed by Gerstein et al. (2007), I argue that just as computer programs are organized into subroutines, that is, ready-made sets of instructions that can be accessed on demand in a variety of contexts, the genome is organized into gene modules. At each site where transcription is initiated, a modular “subroutine is run”; each of these subroutines counts as a gene. The genome behaves like a “master program,” relying on a set of specific syntax-like sequences specifying where in the genome and which transcription factors bind to “recruit” the transcriptional machinery. Each gene behaves like a modular subroutine because sequences within the transcribed DNA mark sites for further processing of the primary RNA transcript, leading to the synthesis of various final gene products.

The gist of the analogy is that the genome is organized as three nested levels of syntax-like DNA sequences:

1. The genomic level is the realm of transcription regulation (fig. 2, top). Regulatory sequences distributed at various sites throughout the genome play a role in specifying where in the genome and which transcription factors bind to enhance or repress transcription. The transcription regulatory regions are to be treated as structural features of the genome rather than belonging to a particular gene (e.g., the same regulatory sequences can be shared by several genes).

2. The gene level corresponds to the transcribed DNA (fig. 2, middle). For the most part, genes behave like independently processed modules because transcribed sequences are processed as dictated by the

1. Programs like the ExPASy Proteomics tools developed by the Swiss Institute of Bioinformatics allow for probabilistic predictions of posttranslational modifications.
Figure 2. Modular organization of the genome.
conserved sequence motifs contained within their boundaries alone. Genes are therefore characterized as modular sets of instructions for the genome expression machinery of the cell; these instructions are located within transcribed sequences known or expected to contribute to a phenotype and are processed independently of other such sets. Thus delimited, genes are shorter than standard transcription units but more extended than open reading frames. The transcription regulatory regions (promoters, enhancers) are part of the global organization of the genome (the “master genomic program”), whereas genes are rigorously delimited as DNA sequences contained between transcription start and stop sites or by homology with primary transcripts. Genes are not identical with open reading frames either, because they include the 5′ and 3′ untranslated regions (UTR), as well as introns and alternate reading frames.

3. The subgene level is the realm of translation, translational control, and posttranscriptional/translational processing (fig. 2, bottom). Sequences within the transcribed/translated DNA indicate sites for eventual RNA translation and further (alternate) processing of the RNA transcript/peptide, leading to the synthesis of one or more final gene products.

4. The “Gene Subroutine” in Practice. It is interesting to note the similarities between the approach proposed in this article and the genes as “things you can do with your genome” approach advocated by Griffiths and Stotz (2006). The activation of different transcription factors allow cells to use the genome in different ways. This accounts for cellular differentiation and the induction of genome expression in response to environmental cues. Each gene can also be used in a variety of ways, as dictated by its internal syntax-motifs. This accounts for cases of alternate posttranscriptional/translational processing, leading to the synthesis of distinct gene products from the same DNA template.

I also want to stress the differences. The approach advocated by Griffiths and Stotz downplays the specificity and organizational roles played by DNA conserved sequences. Their approach has the advantage of taking into account potential cases of genome regulation and processing that are not sequence specific. Unfortunately, no clues are given as to why there is order and regularity in the way the genome is processed. In contrast, a syntax-based approach has no trouble explaining the origin of this order and regularity: the genome is processed and expressed—either constitutively (e.g., housekeeping genes) or in response to regulatory stimuli (e.g., activation of transcription factors via signaling pathways)—as dictated by syntax-like motifs contained in its sequence.
Equally important, an extended syntax-based approach preserves a high degree of continuity with previous syntax-based gene concepts, both conceptually and by providing a substantial overlap in terms of the DNA sequences to which these concepts refer. This kind of retrocompatibility is important given that, in the vast majority of cases, whenever the term “gene” is mentioned in a scientific paper, the occurrence of this term refers either to a transcription unit, to a specific open reading frame, or, as is almost always the case, to a DNA sequence containing some of the proximal upstream regulatory sequences and at least one open reading frame.

Finally, modularity is a very important methodological requirement (Callebaut and Rasskin-Gutman 2005). If a system is modular, it can be carved into functional modules that can be characterized and investigated on a largely independent basis (Lauffenburger 2000). A nested syntax-based account successfully conveys the notion that genome expression can be divided in three distinct stages: genome expression regulation achieved at the level of transcriptional control, gene expression proper achieved at the level of transcription, and gene sequence processing achieved at the level of posttranscriptional modifications, translation, and posttranslational modifications.

In practice, these three virtues of syntax-based gene concepts—modularity, retrocompatibility, and the ability to explain why the genome is regulated, expressed, and processed in an orderly, predictive fashion—are reflected as follows:

1. Reference continuity is ensured by providing a significant overlap with the transcription unit and open reading frame molecular gene concepts, as well as BLAST-generated and GenBank sequences. This means that molecular biologists can continue to identify genes via currently available molecular techniques and genome annotation protocols. Current syntax-based concepts are based on a solid and very successful experimental methodology and are not likely to be abandoned any time soon.

2. Since a gene is not a piece of genomic DNA but a transcribed DNA sequence, there is no restriction on overlapping or nested genes. Also, distinct genes are allowed to share promoters (operons, overlapping promoters) and distal regulatory elements (common enhancers). In contrast to traditional transcription units, alternative promoters are also accommodated.

3. Unlike traditional open reading frames, it is not required that protein end-products are generated. RNA products playing structural (ribosomal, transfer, small nuclear RNAs, ribozymes) and regulatory (microRNAs, RNA interference, and other noncoding RNAs) roles are also coded by genes. Also, by keeping the 3′ and 5′ untranslated
regions as part of the gene, certain mechanisms of translation regulation are accommodated as part of the gene subroutine; this allows DNA templates for RNA species that are not immediately processed for translation (such as developmental maternal factors) to count as genes.

4. Since the instructions contained in each gene module may prompt events such as splicing, perfect homology/collinearity between DNA templates and gene products is not required. Waters (1994, 78) proposed that molecular genes are sequences of DNA coding for homologous RNA or peptide product sequences. This concept was criticized on the grounds that splicing and RNA editing can generate significant divergences between the original DNA template and the final gene product sequence (Falk 1986, 2003; Portin 2002; Stotz, Bostancı, and Griffiths 2006).

5. More than one gene product can be associated with any given gene. The fact that the same gene can be involved in more than one phenotype was acknowledged a long time ago by classical geneticists (Morgan 1935). Alternative splicing, alternative reading frames, and polyproteins are newly discovered molecular mechanisms that further contribute to the generation of functional diversity.

6. Trans-splicing (splicing of exons from two different primary RNA transcripts) can be accommodated as a special case. I acknowledge that even if most gene subroutines are modular (i.e., once initiated, they are not influenced by other gene subroutines), trans-splicing is an example of nonmodular processing. However, there are a number of attenuating circumstances. First, the genes implicated in trans-splicing are unambiguously differentiated since each has its own well-defined transcription initiation and termination sites (e.g., Finta and Zaphiropoulos [2002] explicitly and unambiguously distinguish between the various cytochrome genes involved in trans-splicing). Second, trans-splicing is driven by conserved sequences required for splicing in general and therefore is an agreement with a syntax-based approach. And third, naturally occurring trans-splicing involves either transcripts originating from the same transcription unit (Caudévilla et al. 1998) or highly homologous units (genes sharing “a high degree of similarity,” such as duplicated genes; Finta and Zaphiropoulos 2002, 5882); we are thus dealing with a local, homology-driven cross-modular processing involving the same or very similar gene modules, not with a generalized breaking of modularity.

7. Finally, this concept is compatible with an important aspect of Griffiths and Stotz’s (2006; Stotz et al. 2006) postgenomic gene: each gene can be used in a variety of ways.

5.1. The Scope and Intended Domain of Application of Syntax-Based Concepts. An immediate concern is that some forms of regulation and processing may not rely on conserved sequences. For example, it has been argued that some aspects of chromatin regulation involve sequence-independent mechanisms (Fox-Keller and Harel 2007; Stotz 2011). Extragenomic contributions to inheritance and development are another legitimate source of concern (Fox-Keller 2001). Finally, it has been argued that even if conserved sequences are useful in predicting certain biological outcomes, they do not suffice to explain these outcomes (Griffiths and Neumann-Held 1999; Fox-Keller 2000).

My answer to the above objections hinges on a deflationary view of what genes are and how they are defined. In contemporary scientific practice, gene concepts are not substitutes for explanation and therefore do not have to account for every single causal determinant of inheritance. Quite illustrative in this sense, the Human Genome Nomenclature Committee defines a gene as a “DNA segment that contributes to phenotype/function. In the absence of a demonstrated function a gene may be characterized by sequence, transcription, or homology” (Wain et al. 2002, 464). According to this definition, genes refer solely to genomic contributions to phenotypic outcomes. Nowhere is it stated that genes are unique causal determinants, that they contribute to all known instances of inheritance, or that they explain or suffice to explain why and how inherited phenotypes occur. I adopt a similar view in respect to syntax-based concepts: these concepts refer exclusively to genomic sequences and organizational features of the genome contributing to a pattern of genome expression and, if known, its associated phenotype. I am not claiming that genes or the genome are sufficient (they contribute—in conjunction with other factors—to phenotypic outcomes) or necessary (DNA may not contribute to all inheritance phenomena) causes of inheritance.

5.2. The Limits of Syntax-Based Concepts and Areas of Future Improvement. Although it is generally acknowledged that DNA sequence motifs play a crucial role in specifying binding sites for various non-DNA components (usually proteins) of mechanisms of genome regulation, expression, and processing, the predictions made in light of syntax-based concepts are notoriously probabilistic. Syntax-based concepts may fail to identify all and only DNA sequences contributing to a given phenotype. This may be in part due to the stochastic nature of molecular interactions and in part due to an incomplete and oversimplified understanding of sequence-specific binding.

DNA-protein interactions require that chemical moieties specified by
the nucleotide sequence of a conserved DNA motif are exposed at the right distance and position in respect to each other. Inasmuch as the DNA molecule is not subjected to any stress (torque, bends, super-/under-coiling), it can be predicted with a high degree of confidence that if a conserved DNA sequence motif is present, it can serve as a binding site for its protein partner and therefore contribute to some aspect of genome expression in a predictable way.

One complication that syntax-based concepts fail to take into account is that whenever a protein binds DNA, it creates bends and torques in the DNA double helix, thus altering to various degrees the spacing and position of the chemical moieties required for the binding of other proteins; such changes in spacing and positioning can be important enough to result in a significant enhancing or repression of the binding of these other proteins. Even if a DNA sequence will not completely change its “meaning” due to deformations of the DNA double helix, its specificity/affinity for a given protein target may increase or decrease (sometimes to biologically insignificant values), thus reducing the accuracy of the predictions about genome expression.

It is important, however, to realize that this shortcoming of syntax-based approaches does not preclude future remedies. For instance, nucleosomes (a key component of chromatin structure) are particularly troublesome because of the extreme deformations they induce in the DNA double helix. There is, however, evidence suggesting that the binding and assembly of the nucleosomes is to a large extent sequence-dependent. This suggests that a more detailed knowledge of the ways in which coiling around nucleosomes and other deformations affect the structure of the double-helix is likely to allow for a fine-tuning of syntax-based approaches.

6. Conclusion. I have argued that an expanded syntax-based concept can handle many cases of genome expression regulation and processing while providing a number of key advantages, such as retrocompatibility with molecular gene concepts and current genome annotation protocols, a step-by-step modular methodology for investigating inheritance phenomena, as well as the ability to account for the orderly fashion in which the genome is processed. In response to objections, I proposed a deflationary view of what genes are and how they are defined. Finally, I acknowledged the limits of syntax-based concepts and discussed possible avenues of improvement.

REFERENCES


