

Grammatical model of the regulation of gene expression

(bacterial σ^{70} promoters/transcription initiation/transformational grammar)

JULIO COLLADO-VIDES†

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Communicated by Boris Magasanik, June 26, 1992

ABSTRACT Based on a formal proof that justifies the search for generative grammars in the study of gene regulation, a linguistic formalization of an exhaustive data base of *Escherichia coli* σ^{70} promoters and their regulatory binding sites has been initiated. The grammar presented here generates all the arrays of the collection plus those that are predicted as consistent with the principles of regulation of σ^{70} promoters. "Systems of regulation," sets of regulatory sites that collaborate in a mechanism of regulation, are represented by means of syntactic categories. A small set of phrase structure rules restricted by an X-bar principle and by a hierarchical, c-command relation generates a representation of arrays of sites of regulation where the selection of the protein(s) identifying the system(s) of regulation occurs. Based on the features of the proteins, optional duplicated proximal and remote sites are generated by means of transformational rules. Consistency with the data, the predictions that the grammar generates, and important similarities and differences with some aspects of the generative theory of natural language are discussed.

The organized description of integrated control systems is one of the central aims in biology, and we need to find ways of systematizing the available information to make predictions easier. General rules of different properties of transcription initiation may begin to emerge (1-3). Once general principles are stipulated, the question is how to organize them into a theory capable of precise predictions. Such a theory may represent an important contribution for the computer scientist when dealing with the question of how to represent the expansion in data bases that the genome projects will catalyze in the near future.

I have recently obtained a result that formally justifies the use of grammars in the study of gene regulation (4). On the other hand, we have recently collected and analyzed the regulatory regions of an exhaustive set of *Escherichia coli* and *Salmonella typhimurium* promoters, which constitute the data set in the construction of the grammar here proposed (3). This grammar generates *all and only those* arrays that are consistent with the principles of the system of regulation of σ^{70} promoters.

Definitions and Antecedents

Operons and transcription units are referred to as units of genetic information (UGIs). Regulatory sites have been classified based on their position: "proximal," if their position enables direct contact of the regulatory protein with the RNA polymerase (i.e., sites that touch the domain between -65 and +20, the transcription start site being designated as +1) and "remote," when located elsewhere. A "system of regulation," initially defined as the site, or sites, of a UGI that bind the same regulatory protein (3), will be defined here as the set of regulatory sites that participate in a single mech-

anism of regulation. Let us consider the following partial derivation of the *lac* promoter.

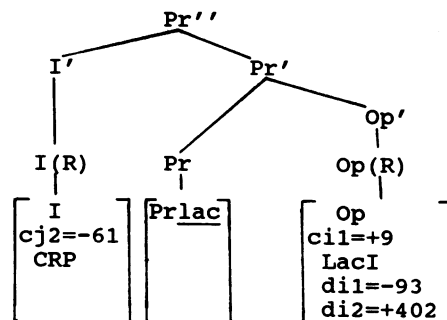


Diagram 1

This tree diagram represents rules of the form $X \rightarrow Y$, which read "rewrite X as Y " (5, 6). The set of nonterminal symbols (Pr'' , I' , Pr' , and Op') used, as well as their respective rules, are the subject of this paper and will become clear below. The complete sequences of promoter (Pr), operator (Op), and activator (I) sites are the smallest elements that can be used to stipulate productive substitutions determined by the criterion of regulatability. The grammar uses a dictionary where specific sequences of DNA corresponding to these *molecular categories* are listed with their respective pertinent properties of regulation, or *distinctive features* (unpublished data). These "words" are represented by matrices as illustrated in diagram 1. They contain features identifying the protein [i.e., cAMP receptor protein (CRP) or LacI] and features of positional information: the coordinates, cin ($n = 1, 2, \dots$), available to proximal sites, as well as distances of duplicated sites from the proximal referential (R) site. Insertional rules substitute I , Pr , and Op symbols by any word from the dictionary. When words with alternative c are selected at R sites, insertion randomly selects one c value, leaving the set of d values of available remote positions (i.e., $di1 = -93$ and $di2 = +402$ for the LacI Op site). These features are subsequently used to generate, by transformational rules, the final representation containing duplicated sites. The specificity of duplicated sites is identified by means of an index shared by an R site. In this way an *L1 linguistic representation* of UGIs reflecting the order in which categories occur in the DNA is generated.

DNA sequence alone is not sufficient to identify molecular categories since there is no available criterion to distinguish unambiguously Op from I sites. Knowing which protein binds may not help since the same regulatory protein can play the role of both an activator and a repressor (3). Therefore, the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: UGI, unit of genetic information (operon and transcription unit); CRP, cAMP receptor protein; Pr , promoter; Op , operator; I , activator; AP, asymmetry principle; R , referential.

†Present address: Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, A.P. 565-A, Cuernavaca, Morelos, 62271 México.

linguistic methodology presented here is restricted to the integration of "deciphered information" (i.e., UGIs of which we have a minimal regulatory understanding).

This grammar represents an illustration of an integrative methodology that may be extended in the future to the analysis of other regulatory data bases.

Regulatory Systems as Syntactic Categories

Grammars of natural language reflect the fact that languages have a structure. Words belong to lexical categories like noun, preposition, and verb. In addition, syntactic categories group several lexical categories. Descriptions of sentences by means of these categories enable linguists to stipulate properties of distribution, of syntax, and of meaning. Thus, for instance, the fact that "The boy living next door" can be substituted for "John" in a sentence is related to the fact that they are noun phrases.

Several observations point to the existence of syntactic structures within UGIs. The linear order of categories is not sufficient to obtain a general description of UGIs; in fact, the definition of an operon, a set of structural genes regulated by proximally located regulatory sites, implicitly uses the notion of a syntactic category (7). Furthermore, as shown in the following, molecular categories occur in groups even if located at a distance from one another.

The set of regulatory categories grouped within a syntactic category has to identify a substitutable unit, a "regulatory phrase." Thus, a system of regulation may include sites that bind different proteins. For instance, let us consider the regulation of *deoP2* by DeoR, CytR, and CRP. The DeoR sites define one system of regulation. In fact, these sites also regulate *deoP1* as an independent unit. On the other hand, the CytR and CRP sites together identify another system of regulation. Repression by CytR has been proposed to involve recognition of two CRP-bound molecules separated by a short fragment of DNA (8). In addition, the same arrangement of CytR and CRP sites can be found as a unit regulating other promoters (3). A different mechanism involving CRP and MalT proteins provides another example of a pattern of several interdigitated sites (9). These cases illustrate that regulatory mechanisms involving interactions between heterologous proteins bound to DNA determine the occurrence of arrangements of several sites that behave as a unit. It is reasonable to assume that the sites forming such units will have a limited number of possible relative positions restricted by the architecture of protein complexes.

Additionally, associated with the notion of syntactic categories is the idea of a unit that has a constant property even if the number of categories it contains can vary. That is to say, syntactic categories can be expanded, as in the case of noun phrases mentioned above. Arrays of regulatory sites show, in principle, a similar, restricted capacity of expansion. Thus, a pair of weak Op sites can be substituted by a single strong Op site that provides an equivalent degree of repression (10).

If systems of regulation can include binding sites for heterologous proteins, it must be possible to distinguish two independent different systems of regulation and a single heterologous system of regulation.

Fig. 1 groups the sites of promoters activated by CRP that are also subject to repression by another protein, where CRP activation is independent of repression. Fig. 2 groups the CRP sites of promoters regulated only by CRP activation. The sites occur mainly at -40, -60, or -70 (11). In these cases the CRP sites fall within the same positions as those in Fig. 1. Fig. 3 shows that CRP sites occur in much more variable positions in promoters that are also subject to activation by another protein. This distribution supports the idea of CRP activation mechanisms that depend on additional assistance by other proteins in order to facilitate the inter-

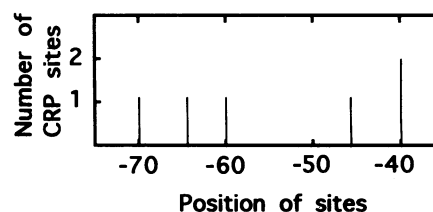


FIG. 1. Two systems of regulation: CRP activation and repression. The number of CRP sites from the collection in ref. 3 versus nucleotide position is illustrated. The transcription start site is designated as +1.

action of CRP with RNA polymerase (3). Thus, it is in principle possible to distinguish heterologous systems of regulation and several different simple systems of regulation. In summary, these cases illustrate that regulatory categories do not always occur independently. They are part of higher units, which stipulate their distribution.

X-Bar Principle and C-Command

One of the most salient features of the collection is that at least one proximal site is always required in the constitution of a σ^{70} array of sites. Duplicated proximal and remote sites contribute to enhance repression or activation, but in principle are not equally obligatory (3). This distinction between *optional* and *obligatory* categories is also useful at the level of regulatory systems and promoter activity, in the sense that when a mutation impairs the regulatory protein, repressor-controlled systems become constitutively expressed, while activator-controlled systems become super-repressed. The absence of positive regulation impairs transcription activity more drastically than the absence of negative regulation.

To incorporate into the grammar these characteristics, two procedures will be introduced that restrict the types of derivations and that, additionally, introduce a notation for syntactic categories. Let us define the X-bar principle as follows:

(i) Every syntactic category is a *projection* of a molecular head category.

(ii) $X(n)$ immediately dominates $X(n-1)$, down to $X(0)$, where $X(0)$ is a *head category*.

A category X can have one $X(1)$ or successive $X(2), \dots, X(n)$ projections (12). These can also be written as X', X'' , etc.

Thus, syntactic categories are projections of either Pr, Op, or I categories. This principle implies that within any syntactic category there is at least one obligatory category, its *head*, plus other nonhead, optional molecular categories. For instance, the promoter is the head of a UGI since this is the only molecular category that occurs in any UGI—constitutive promoters are not regulated.

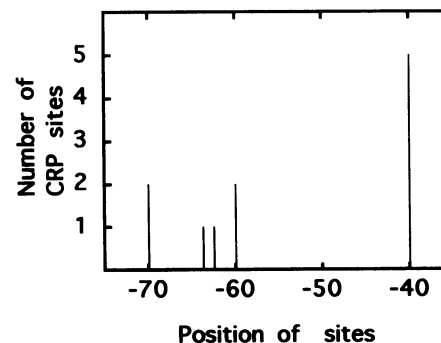


FIG. 2. One system of regulation: CRP activation. There is also a site at -105 in the *colE1* promoter, which has a proximal site centered at -62 (3).

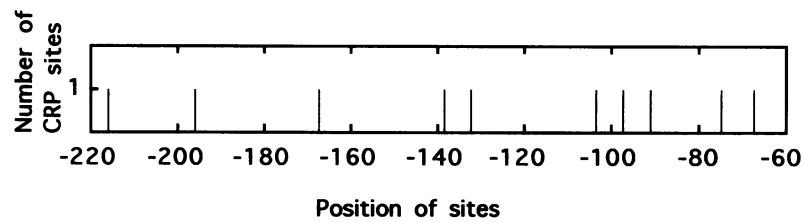
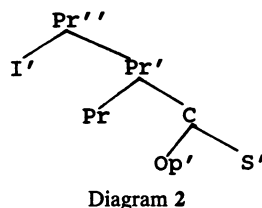


FIG. 3. Heterologous systems of regulation.

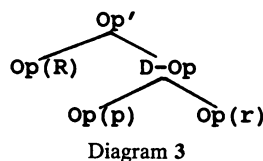
The second proposal is to define a hierarchical relation among categories such that obligatory categories within a derivation are hierarchically higher than optional ones. Thus, for instance, a site A can be defined as hierarchically higher than B if the number of intermediate symbols between A and the initial symbol in the derivation is smaller than that between B and the initial symbol. Thus, a principle can be proposed as follows:

For an I' to regulate a Pr, I' must be hierarchically higher than Pr. For an Op' to be able to affect a Pr, Pr must be hierarchically higher than Op' . This is a modified version of the asymmetry principle (AP) (7), with hierarchical relations involving syntactic categories. Certainly, the asymmetric behavior of mutants on which this principle is based is independent of the details (i.e., number of sites, of the positive and negative regulatory phrases). Restricted by this AP, and recalling that Pr is the head of UGIs, a simple grammar can be proposed as follows:

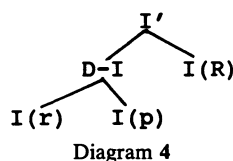


where Pr' and Pr'' are projections of Pr, the head of UGIs; S' is the domain of structural (S) genes; Op' and I' are regulatory phrases; and C is a category required to satisfy the AP.

The same hierarchical relation can be used within projections of regulatory categories. Let us distinguish regulatory categories by a subindex indicating if they are a referential (R) site, a proximal (p) or a remote (r) duplication. Thus, the regulatory phrase Op' , where $Op(R)$ is the head and $D-Op$ is an optional category grouping duplicated operators, can be derived as in diagram 3.



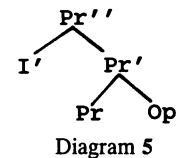
Similarly, activator phrases are generated by rules in diagram 4,



where subindexes have the same meaning as in diagram 3. Heterologous systems of regulation with positive and negative sites can be identified as a phrase with an R site and additional sites within a D^* category, where the asterisk indicates the presence of sites that bind different proteins.

In any regulatory phrase the referential site is hierarchically higher than the optional sites. The requirement for a referential proximal site derives from the X-bar principle. Furthermore, the precedence or left-to-right order of categories is such that if diagrams 3 and 4 are substituted in diagram 2, the sites occur as defined in the L1 linguistic representation of UGIs.

A *local hierarchical relation*, which could be independently evaluated within the set of sites of each system of regulation in diagrams 3 and 4, as well as within categories in diagram 2, irrespective of the details of I' and Op' , would be preferred. In fact, an additional benefit from a different hierarchical relation is to eliminate the C category and simplify the rules in diagram 2 to those in diagram 5,



where the domain of structural genes does not appear and therefore the need for the C category in diagram 2 also disappears. There are several reasons for this simplification. First, the relations that motivated the AP are independent of the nature and number of the structural genes transcribed by the promoter. Second, the separation of the range of transcription initiation from the domain of structural genes eliminates difficulties that rules in diagram 2 have in a grammatical model that produces derivations of complex UGIs with more than one promoter (diagram 5). Let us define the hierarchical notion of *c-command* that satisfies these requirements. Node A c-commands node B if and only if

- (i) A does not dominate B and B does not dominate A; and
- (ii) the first branching node dominating A also dominates B, where a node A *dominates* node B if and only if A is higher up in the tree than B and if one can trace a line from A to B going only downward (13).

Observe that I' c-commands Pr and Pr c-commands Op' in diagram 5. Similarly, the heads of regulatory phrases in diagrams 3 and 4 c-command the duplicated proximal and remote categories. These hierarchical relations are evaluated within their own domain of rules (diagrams 3-5, respectively). Thus the AP can now be reworded as follows:

For an I system to regulate a Pr, I' must c-command Pr. For an Op system to be able to affect a Pr, Pr must c-command Op' .

In summary, the X-bar principle distinguishes between obligatory and optional categories and enables us to identify UGIs as projections of Prs and referential sites as heads of regulatory phrases. A single hierarchical relation, c-command, provides a uniform way to make explicit the different degrees of obligatoriness of categories at different levels of description of UGIs. The specific meaning of such relation varies according to the categories involved.

If the proximal and remote positions of regulatory phrases can be occupied by one or several molecular categories, a unified grammatical description for a large number of UGIs

of the collection can be obtained by using the rules contained in diagrams 2–4.

These rules generate the possible positions that categories can occupy in any σ^{70} UGI. All the categories, except Pr, are optional; thus, a simple negatively regulated promoter will have no I', and a simple positively regulated promoter will have no Op'. But, once a regulatory projection occurs, the X-bar principle imposes the existence of its head, whereas all other categories are again optional.

It is clear that the grammar formed by the sets of rules used in diagrams 3–5 follows from the X-bar principle and the AP. In the following, we will discuss how the number and nature of the duplicated sites are generated in the grammar.

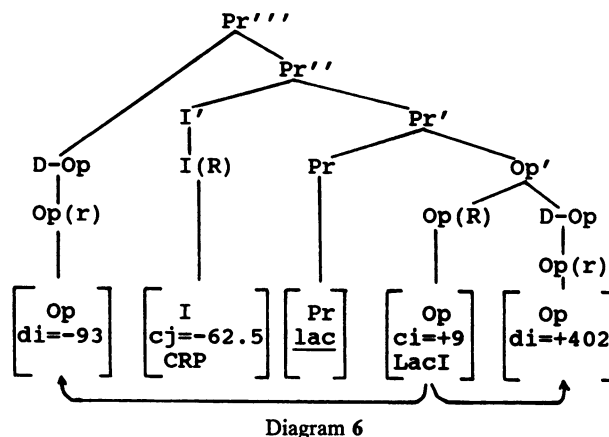
Duplicated Sites and Transformational Rules

Predictions generated by the grammar will result, in one way or another, from extrapolations of the data. When analyzing duplications of operator sites it was observed that a certain pattern of sites correlates with each protein. For instance, genes regulated by ArgR use duplicated sites irrespective of promoters being complex or simple, whereas LexA correlates in most of the cases with single sites (3). We will assume that the number of homologous sites is dictated by the protein.

In such a type of grammar, the protein has to be selected in the derivational process *before* the set of possible duplications is generated. This is achieved by a set of *transformational movement rules* and two sets of insertional rules. The grammar generates, first, a “primitive” or “protein” derivation, where only referential sites occur, as shown in diagram 1. Let us call this the P level. Insertional rules at this level identify the regulatory protein(s) of the UGI, as well as the feasibility of duplicated sites and their number and position. In the case of promoters subject to more than one system of regulation, or subject to heterologous systems, insertional rules will specify more than one protein in the respective referential site. Based on the features of referential sites, duplicated sites are generated. This process is described by means of optional transformational rules, which move the distance features of duplicated sites, initially contained within the features of referential sites, to the D categories that accept such values. The D categories are those shown in diagrams 3 and 4, plus a site for remote upstream Op sites that is proposed below. Let us call the final derivation the D level (for derived) with all sites with their inserted words or matrices of features. All the features have at least one value associated with them, except the one for protein–DNA specificity, which will contain information on the protein domain but not on the sequence of DNA. A second set of insertional rules selects, among the DNA sequences that match the protein specificity, one for each regulatory category. Incidentally, observe that the insertion of promoters at the P level is justified assuming that there is a correlation between features of promoters (i.e., affinity) and systems of regulation. Otherwise, insertion of promoters would occur at the D level.

As an illustration, let us complete the derivation of the *lac* promoter. After the P level shown in diagram 1, the D level is generated by the transformational rules, represented in diagram 6. To complete the range of positions observed in the collection, we added a D-Op(r) category for the remote upstream operators. This D-Op node is generated from a third Pr projection. The only argument for this configuration is that so far we have only used binary nodes.

Transformational rules generate in a combinatorial way as many derivations as the features permit. In fact, the array in diagram 1 is also acceptable since transformations are optional. Using this process, the derivations for all the arrangements of the data base can be obtained.



Consistency with Data and Predictions

The transformational grammar here described generates the collection of σ^{70} promoters by means of a reduced number of rules, those indicated in diagrams 3–5, plus the upstream D-Op generated from Pr''', together with transformational rules that move positional features from the referential site to other positions.

This grammar makes explicit in the derivation, by means of transformational rules, the fact that several sites can identify a system of regulation without necessarily being adjacent. This could not be achieved with rules of the type used in diagrams 3–5.

The representation in the grammar of remote Op sites implies that downstream positions have more in common with the referential position than the remote upstream positions. In fact, the downstream position occurs in exceptional cases in the absence of any proximal site, as in the *aroP* and *purR* promoters. No such cases exist with remote upstream positions (3). The formalization of this restriction is simply made by stipulating that unusual cases are generated at the P level in a proximal position and can only move to a position that is c-commanded by the referential site. The distinction in the grammar between the usual and exceptional cases corresponds precisely with the classification of the data set based on principles of the mechanism of transcription and regulation (3).

The grammar also predicts that certain patterns will not occur (i.e., activator phrases cannot be interrupted by a remote Op site that is part of an independent negative system of regulation).

This grammar is an explicit set of instructions that generates many different arrays using the same dictionary as that of the data set. In that sense it identifies all and only those arrangements that are consistent with the principles of regulation of the σ^{70} system of transcription. This grammar is therefore testable, as well as expandable, if the analysis of emerging data requires it.

Discussion

The grammatical reconstruction of UGIs has produced some similarities with the study of natural language—i.e., the distinction between acoustic phenomena and a phonetic description parallels the distinction between “raw” sequences of DNA and a “deciphered representation” with regulatory information. The notion of c-command, here used, has been a central notion in the grammar of principles and parameters of Government and Binding (6, 14). This conceptual similarity either reflects some biological content, or it reflects a property common to the method of analysis. It is certainly quite natural to make use of the notion of c-command, once it is assumed that all nodes are binary and that the

notation of categories is restricted by an X-bar theory. However, the utility of these assumptions in the description of gene regulation is due to the adequate description of sets of sites by means of syntactic categories and hierarchical relations. Thus, it is possible that these common properties may illustrate the conservative character of evolution throughout two very remotely related discrete biological systems: DNA and natural language.

It may be useful to emphasize that DNA is frequently described as a language, but its analysis with linguistic methodology is sparse (15, 16). Formal approaches to the study of the regulation of gene expression are also limited (2, 4, 17, 18).

Once general principles can be stipulated, the question is how to convert such principles into an organized theory capable of precise predictions. This is the goal that the grammar proposed here begins to achieve. It should be emphasized that the importance of this work depends on the potential applications of this integrative methodology more than on the details of the model here proposed. The imminent results of the genome projects will provide food for thought for this type of integrative methodologies.

I want to acknowledge fruitful discussions with Dr. Boris Magasanik. This work was supported by U.S. Public Health Service Fogarty International Research Fellowship FO5-TWO4437.

1. Ptashne, M. (1986) *A Genetic Switch: Gene Control and Phage Lambda* (Cell Press, Cambridge, MA).
2. Savageau, M. (1991) *New Biol.* **3**, 190–197.
3. Collado-Vides, J., Magasanik, B. & Gralla, J. D. (1991) *Microbiol. Rev.* **55**, 371–394.
4. Collado-Vides, J. (1991) *Comput. Appl. Biosci.* **7**, 321–326.
5. Hopcroft, J. E. & Ullman, J. D. (1979) *Introduction to Automata Theory, Languages and Computation* (Addison-Wesley, Reading, MA).
6. Haegeman, L. (1991) *Introduction to Government and Binding* (Blackwell, Oxford, U.K.).
7. Collado-Vides, J. (1991) *J. Theor. Biol.* **148**, 401–429.
8. Pedersen, H., Sogaard-Andersen, L., Holst, B. & Valentin-Hansen, P. (1991) *J. Biol. Chem.* **266**, 17804–17808.
9. Raibaud, O., Vidal-Ingigliardi, D. & Richet, E. (1989) *J. Mol. Biol.* **205**, 471–485.
10. Brent, R. & Ptashne, M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4204–4208.
11. Gaston, K., Bell, A., Kolb, A., Buc, H. & Busby, S. (1990) *Cell* **62**, 733–743.
12. Jackendoff, R. (1977) *X-Syntax: A Study of Phrase-Structure* (MIT Press, Cambridge, MA).
13. Reinhardt, T. (1983) *Anaphora and Semantic Interpretation* (Univ. of Chicago Press, Chicago).
14. Chomsky, N. (1981) *Lectures on Government and Binding* (Foris, Dordrecht, The Netherlands).
15. Head, T. (1987) *Bull. Math. Biol.* **49**, 737–759.
16. Searls, D. B. (1988) *Proceedings of the 7th National Conference on Artificial Intelligence*, pp. 386–391.
17. Koton, P. (1983) M.S. thesis (Massachusetts Institute of Technology, Cambridge, MA).
18. Thomas, R. (1991) *J. Theor. Biol.* **153**, 1–23.