

# **GenoCAD.org: a web site to design and verify synthetic genetic constructs derived from standard biological parts**

Yizhi Cai<sup>1</sup>, Brian Hartnett<sup>1</sup>, Claes Gustafsson<sup>2</sup> and Jean Peccoud<sup>1\*</sup>

1. Virginia Bioinformatics Institute, Virginia Tech, Blacksburg VA, USA

2. DNA2.0, Inc. Menlo Park CA, USA

\*email: peccoud@vt.edu

The sequence of artificial genetic constructs is composed of multiple functional fragments, or genetic parts. Biologists have deciphered structural rules that the design of genetic constructs needs to follow in order to ensure a successful completion of the gene expression process, but these rules have not been formalized, making it challenging for non-specialists to benefit from the recent progress in gene synthesis. We show that context-free grammars (CFG) can formalize these design principles. This approach provides a path to organizing libraries of genetic parts according to their biological functions, which correspond to the syntactic categories of the CFG. It also provides a framework for the systematic design and verification of new genetic constructs consistent with the design principles expressed in the CFG.

Gene synthesis technology now enables molecular biologists to assemble long DNA molecules that may include multiple genes and their regulatory sequences. We will refer to these molecules as “genetic constructs” or just “constructs”. As the throughput of construct manufacturing increases, the design of complex genetic constructs becomes the bottleneck of the process. It becomes easier to assemble complex DNA molecules than to design them. A natural way of designing complex constructs involves combining basic building blocks also known as “biological parts” or “genetic parts” [1-3]. These parts are small DNA fragments implementing specific biological functions. The mechanisms of gene expression require that certain structural constraints are met in order for a construct to be functional. Parts of different types need to be placed in a particular order and next to each other in order to ensure that coding sequences are properly transcribed and translated. Certain parts are functional only in a specific context whereas other parts have proved functional in organisms other than the one from which they originate. For instance promoters are often restricted to specific organisms or even cell types [4-6] whereas genes coding for proteins can often be expressed in multiple species[7]. The design of complex genetic constructs such as artificial gene networks [2, 8-13] therefore requires an intimate knowledge of gene expression mechanisms. It is interesting to observe that more than six years after the description of the first artificial gene networks [9, 10], this technology has yet to find biomedical applications. It is likely that most biologists who could use sophisticated genetic constructs to control the expression of their gene of interest do not have the expertise to design the construct they need.

One way to lower the barrier to entry into synthetic biology is to formalize the structural constraints associated with the use of standardized biological parts in a construct. We have used a class of formal languages called context-free grammars to represent the structure of previously published artificial gene networks. We have embedded this formalism into GenoCAD.org, a web

site that includes a wizard guiding users in the design of their constructs by combining standard biological parts. In addition, GenoCAD includes a parser capable of verifying the structural validity of synthetic DNA sequences that have been designed outside of GenoCAD.

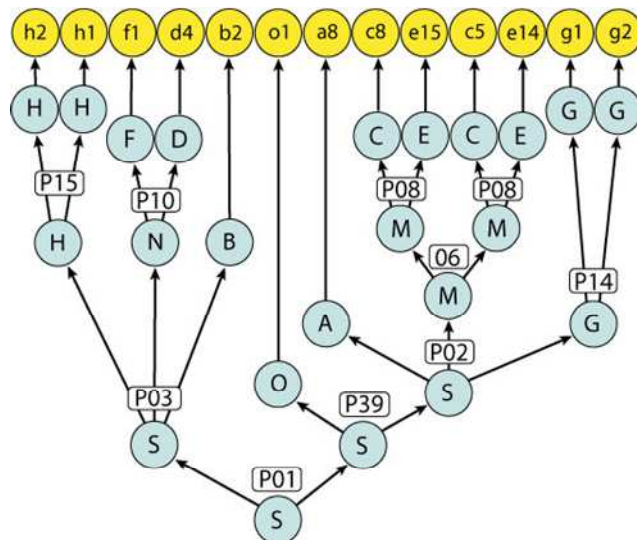
P01	$S \rightarrow SOS$	Start symbol (S), linker (O), start symbol (S)
P02	$S \rightarrow AMG$	Promoter (A), transcript (M), terminator (G)
P03	$S \rightarrow HNB$	Terminator rev (H), transcript rev (N), promoter rev(B)
P04	$S \rightarrow PMR$	T7 promoter (P), transcript (M), T7 terminator (R)
P05	$S \rightarrow TNQ$	T7 terminator rev (T), transcript rev (N), T7 promoter rev (Q)
P06	$M \rightarrow MM$	Transcript (M), transcript (M)
P07	$N \rightarrow NN$	Transcript rev (N), transcript rev (N)
P08	$M \rightarrow CE$	Ribosome binding site (C), gene (E)
P09	$M \rightarrow KIK$	Hammerhead (K), riboregulator (I), hammerhead(K)
P10	$N \rightarrow FD$	Gene rev (F), ribosome binding site rev (D)
P11	$N \rightarrow LJI$	Hammerhead rev (L), riboregulator rev (J), hammerhead rev (L)
P12	$E \rightarrow WUY$	Start codon (W), protein domain (U), stop codon(Y)
P13	$F \rightarrow ZVX$	Stop codon rev (Z), protein domain rev (V), start codon rev (X)
P14	$G \rightarrow GG$	Terminator (G), terminator (G)
P15	$H \rightarrow HH$	Terminator rev (H), terminator rev (H),
P16	$O \rightarrow OO$	Linker (O), linker (O),
P17	$Y \rightarrow YY$	Stop codon(Y), stop codon(Y)
P18	$Z \rightarrow ZZ$	Stop codon rev (Z), stop codon rev (Z)
P19	$U \rightarrow UU$	Protein domain (U), protein domain (U)
P20	$V \rightarrow VV$	Protein domain rev (V), protein domain rev (V)
P21	$A \rightarrow OA$	
P22	$B \rightarrow OB$	Linkers can be added next to some parts
É	É	
P0100, P0101É	$A \rightarrow a1 \mid a2 \mid \acute{E}$	
P0200, P0201É	$B \rightarrow b1 \mid b2 \mid \acute{E}$	All variables can be transformed into terminals.
É	É	
P2400, P2401É	$Z \rightarrow z1 \mid z2 \mid \acute{E}$	

Table 1 A CFG generating the most common architectures of artificial gene networks.

Step	Production	String
1	P01	SS
2	P03	HNBS
	P02	HNBMAMG
3	P06	HNBMAMMG
	P10	HFDBAMMG
	P08	HFDBACEMG
4	P08	HFDBACEMG
	P15	HHFDBACECEG
5	P14	HHFDBACECEGG
	P21	HHFDBACECEGG
6		h02h01f01d04b02o01a08c08e15c05e14g01g02
7		tataa.....gcgttata

**B**

```
[tataaacgcagaaagccaccgcaaggtgagccagtgtga][gagagcgttcaccgacaacaacacaga
taaaacgaaaggocccagctcttgcactgagccttgcgttttatttgatgcctgg][ttaagc...cacca][
catcgaaacggtttctct][tcctttgcataccctgctgatgtgctcattataaccgcaaggattttat
gtcaaacccgcaagagataatttataccgcaagatgggttatctgtgcatg][ttatcaaaaaccatggg
tttgataa][ccatcgaaatggctgaaatgagctgtgacaattaatcatcggctcgtataatgtgtgg
aattgtgagcggataacaatttcacacagga][aggaacccggttatg][atgagca...ttacaa][agga
atttaaatg][atgct...aaataa][ccaggcatcaataaaacgaaaggctcagtcgaaagactggccc
tttgcgttttatctgtgtttgctggggaacgctctd][tcacaactggctcaactcgggtggcctttct
gcgtttata]
```



**Figure 1.** The successive applications of productions starting from  $S$  provide a framework to guide the design of genetic constructs (B). The verification of an existing DNA sequence requires the use of a lexical analyzer to identify the parts composing the sequence. The symbolic description of the sequence provided by the lexical analyzer can be parsed using an LR algorithm.

## References

1. Voigt, C.A., *Genetic parts to program bacteria*. Curr Opin Biotechnol, 2006. **17**(5): p. 548-57.
2. Heinemann, M. and S. Panke, *Synthetic biology - putting engineering into biology*. Bioinformatics, 2006.
3. Benner, S.A. and A.M. Sismour, *Synthetic biology*. Nature Reviews Genetics, 2005. **6**(7): p. 533-543.
4. Cavin Perier, R., T. Junier, and P. Bucher, *The Eukaryotic Promoter Database EPD*. Nucleic Acids Res, 1998. **26**(1): p. 353-7.
5. Munch, R., et al., *PRODORIC: prokaryotic database of gene regulation*. Nucleic Acids Res, 2003. **31**(1): p. 266-9.
6. Zhu, J. and M.Q. Zhang, *SCPD: a promoter database of the yeast Saccharomyces cerevisiae*. Bioinformatics, 1999. **15**(7-8): p. 607-11.
7. Goeddel, D.V., et al., *Direct expression in Escherichia coli of a DNA sequence coding for human growth hormone*. Nature, 1979. **281**(5732): p. 544-8.
8. Guido, N.J., et al., *A bottom-up approach to gene regulation*. Nature, 2006. **439**(7078): p. 856-860.
9. Gardner, T.S., C.R. Cantor, and J.J. Collins, *Construction of a genetic toggle switch in Escherichia coli*. Nature, 2000. **403**(6767): p. 339-342.
10. Elowitz, M.B. and S. Leibler, *A synthetic oscillatory network of transcriptional regulators*. Nature, 2000. **403**(6767): p. 335-338.
11. Guet, C.C., et al., *Combinatorial synthesis of genetic networks*. Science, 2002. **296**(5572): p. 1466-1470.
12. Chin, J.W., *Programming and engineering biological networks*. Curr Opin Struct Biol, 2006. **16**(4): p. 551-6.
13. McDaniel, R. and R. Weiss, *Advances in synthetic biology: on the path from prototypes to applications*. Curr Opin Biotechnol., 2005. **16**(4): p. 476-483.