Defining Domain-Specific DNA Programming Languages Using GenoCAD 2.2 Grammar Editor

Jean Peccoud, Laura Adam, Mandy Wilson
Virginia Bioinformatics Institute
Virginia Tech
Synthetic DNA molecules as programs

Different Dialects
Plasmids for different purposes

Bacterial

Yeast

Mammalian

2013/07/24

BOSC 2013
Domain Specific Languages

No universal language
- C, SQL, HTML, Flash

Languages express design strategies
- Domain-specific
  - bacterial, yeast, mammalian
- Project-specific
  - gene therapy, synthetic biology
- Organization-specific
  - know-how, intellectual property

Empower end-users to develop their own DSL
- Gene expression in the chloroplast of microalgae

Chlamydomonas reinhardtii

Pieter Bruegel the Elder - The Tower of Babel
CHAPTER EIGHT

A STEP-BY-STEP INTRODUCTION TO RULE-BASED DESIGN OF SYNTHETIC GENETIC CONSTRUCTS USING GenoCAD

Mandy L. Wilson, Russell Hertzberg, Laura Adam, and Jean Peccoud

Meth. Enz. (2011) 498

GenoCAD for iGEM: a grammatical approach to the design of standard-compliant constructs

Yizhi Cai, Mandy L. Wilson and Jean Peccoud*

Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Washington St MC 0477, Blacksburg VA 24061, USA

Published online 18 February 2010

Received December 13, 2009; Revised January 28, 2010; Accepted February 1, 2010

Writing DNA with GenoCAD™

Michael J. Czar, Yizhi Cai and Jean Peccoud*

Virginia Bioinformatics Institute at Virginia Polytechnic and State University, Blacksburg, VA 24061, USA

Published online 8 May 2009

Received March 13, 2009; Revised April 21, 2009; Accepted April 22, 2009
Example of formal grammar

A grammar is a:

- Set of rules describing how to form sentences from a language’s vocabulary

```
R1: Sentence → Subject + Verb + Object
R2: Subject → NounPhrase
R3: Object → NounPhrase
R4: NounPhrase → NounPhrase + Modifier
R5: Modifier → PrepositionalPhrase
```
How to build a sentence?

Sentence

Subject  Verb  Object

R1: **Sentence** → Subject + Verb + Object
R2: **Subject** → NounPhrase
R3: **Object** → NounPhrase
R4: **NounPhrase** → NounPhrase + Modifier
R5: **Modifier** → PrepositionalPhrase
People are DNA’s way of making more DNA

~ Edward O. Wilson, 1975
## Identify and Define Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>5FLR / 3FLR</td>
<td>5’ / 3’ Flanking region for homologous recombination</td>
</tr>
<tr>
<td>SIS</td>
<td>Short Interval Sequences used to make polycistronic cassettes</td>
</tr>
<tr>
<td>STP</td>
<td>Stop codon</td>
</tr>
<tr>
<td>ATG</td>
<td>Start codon</td>
</tr>
<tr>
<td>GEN</td>
<td>Gene or protein domain. By convention does not include start and stop codons.</td>
</tr>
<tr>
<td>CDS</td>
<td>Open reading frame composed of several protein domains. Does not include start and stop codons.</td>
</tr>
<tr>
<td>TAG</td>
<td>Epitope tags. By convention does not include Start or Stop codons.</td>
</tr>
<tr>
<td>PBS</td>
<td>Sequence associated with the initiation of transcription and translation.</td>
</tr>
<tr>
<td>TCS</td>
<td>Targeted expression cassette. Expression cassette flanked with two adjacent genomic sequences for homologous recombination.</td>
</tr>
<tr>
<td>CAS</td>
<td>Expression cassette delimited by a promoter in 5’ and a transcription terminator in 3’.</td>
</tr>
</tbody>
</table>
Grammar Editor – Add/Edit Categories

Manage Grammar

Name: C. reinhardtii Chloroplast Grammar_v5
Description: This grammar captures rules to design expression vectors for the Chlamydomonas reinhardtii chloroplast. Authors: Sakiko Okumoto, Mandy L. Wilson, Jean Peccoud.
Starting Category: Start / Transcription unit (S)
Icon Set: main_icon_set

Categories

Rewritable Categories
- Start / Transcription unit (S)
- Cassette (CAS)
- Open reading frame (CDS)
- Gene (GEN)
- Terminator (TER)
- Targeted expression cassette (TCS)

Terminal Categories
- Opening reverse complement delimiter (\{)
- Closing reverse complement delimiter (\})
- Opening plasmid delimiter (\{)
- Closing plasmid delimiter (\})
- Promoter-RBS (PBS)
- Start codon (ATG)
- Vector (VEC)
- Short Interval Sequence (SIS)
- Stop codon (STP)
- Epitope tags (TAG)
- 5' Flanking Region (5FLR)
- 3' Flanking Region (3FLR)
- Orphaned Categories
- Opening chromosome delimiter (\{)

Category Detail

Letter: PBS
Description: Promoter-RBS
Genbank Qualifier: N/A
Icon: icon_c.png

Category Rules

Child Rules

There are no child rules for the selected category.

Parent Rules

Code: PRCT Rule: → Edit Delete
Code: MRUL Rule: → Edit Delete
### Reserved Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ and ]</td>
<td>Negative orientation delimiters</td>
</tr>
<tr>
<td>( and )</td>
<td>Plasmid delimiters</td>
</tr>
<tr>
<td>{ and }</td>
<td>Chromosome delimiters</td>
</tr>
</tbody>
</table>

[Diagram showing delimiters and their use in biological systems]
Define rewriting rules

<table>
<thead>
<tr>
<th>Code</th>
<th>Rule</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAS</td>
<td>S -&gt; TCS</td>
<td>This rule is used to design only one expression cassette</td>
</tr>
<tr>
<td>1PLAS</td>
<td>S -&gt; ( VEC TCS )</td>
<td>This rule is used to specify the expression cassette along with the vector where it is inserted. The output is the entire plasmid sequence.</td>
</tr>
<tr>
<td>2PLAS</td>
<td>S -&gt; ( VEC TCS ) ( VEC TCS )</td>
<td>This rule is for designs that involve two plasmids.</td>
</tr>
<tr>
<td>TGS</td>
<td>TCS -&gt; 5FLR CAS 3FLR</td>
<td>Specifies the flanking regions for homologous recombination.</td>
</tr>
<tr>
<td>PRCT</td>
<td>CAS-&gt; PBS CDS TER</td>
<td>A gene expression cassette is composed of a promoter, open reading frame, and a transcription terminator.</td>
</tr>
<tr>
<td>2CAS</td>
<td>CAS -&gt; CAS CAS</td>
<td>This rule makes it possible to have more than one expression cassette on a construct.</td>
</tr>
<tr>
<td>rCAS</td>
<td>CAS -&gt; [ CAS ]</td>
<td>This rule is used to specify that the cassette is coded on the negative strand.</td>
</tr>
<tr>
<td>2CDS</td>
<td>CDS -&gt; CDS SIS CDS</td>
<td>This rule makes it possible to design polycistronic constructs.</td>
</tr>
<tr>
<td>SGEN</td>
<td>CDS -&gt; ATG GEN STP</td>
<td>The open reading frame is composed of a single gene flanked by a start and stop codon.</td>
</tr>
<tr>
<td>TGEN</td>
<td>GEN → GEN TAG</td>
<td>This rule is used to add a tag to a coding sequence. It can be used iteratively to add more than one tag.</td>
</tr>
<tr>
<td>2GEN</td>
<td>GEN-&gt; GEN GEN</td>
<td>This rule can be used to fuse two coding sequences that are not tags.</td>
</tr>
</tbody>
</table>
A gene expression cassette is composed of a promoter, open reading frame, and a transcription terminator.
Rewritable: used on the left side
Terminal: used only on the right side
Orphan: not used
Add parts in libraries
Grammar Summary

Name: C. reinhardtii Chloroplast Grammar_v5
Description: This grammar captures rules to design expression vectors for the Chlamydomonas reinhardtii chloroplast. Authors: Sakiko Okumoto, Mandy L. Wilson, Jean Pecou.
Icon Set: main_icon_set
Supports Attributes?: No
# Categories: 20
# Libraries: 1
# Parts: 53
# Rules: 14
Rule-based design of DNA sequences

Part: BBa_J13004
Designed by Jeff Tabor Group: iGEM_UTAustin (2005-06-28)

poly-cistronic CFP/YFP expression cassette
A single mRNA containing CFP and YFP under the strong RBS B0034 is driven by the tet repressible promoter, R0040.

Sequence and Features

Construct

Promoter
Cistron
Cistron
Cistron
Terminator

BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 20 2007, pages 2760-2767
doi:10.1093/bioinformatics/btm446

Systems biology

A syntactic model to design and verify synthetic genetic constructs derived from standard biological parts
Yizhi Cai¹, Brian Hartnett¹, Claes Gustafsson² and Jean Peccoud¹.*

2013/07/24
Point-and-click design tool
A dynamic language

You can change the model ...

► Add or Change a Rule in a Grammar
► Delete a Rule from a Grammar
► Remove a Part from a library
► Change a Part’s sequence
► Change a Part’s category

... but how does this affect the dependent designs?
Available Designs
Different design statuses

- **Valid** – the sequence could be decomposed into its parts, and the parts’ categories make up a grammar-sanctioned framework.

- **Needs validation** – either grammar, part, or library has changed, and the sequence has not been validated since

- **Under construction** – design is unfinished, so cannot be compiled.

- **Out of Date** – although design is still valid with respect to grammars and libraries, the parts have changed.

- **Invalid** – the sequence cannot be resolved.
Graphical Language Icons

**Icon sets:** visual language – color – structure
Hello World!": return 0;" // my first program

#include <iostream>

int main ()
{
    return 0;
}
Never odd or even.
No lemon, no melon.

Richard Wheeler (Zephyris) 2005; Schematic representation of the insertion of the bacteriophage lambda
Custom languages vs. Standardization

DSL in a standardized world
  ► GenBank
  ► Sequence Ontology
  ► Synthetic Biology Open Language

Need to map custom concepts to standards
  ► Data exchange
  ► Graphical representation
  ► Sequence annotation

Can we standardize customization?
Workshop Synthetic Biology and Language Engineering

► Bring together language designers and synthetic biologists with the goal of analyzing the different programming paradigms that have been or could be explored to write these biological programs more effectively

► Part of:
  o International Conference on Software Language Engineering (SLE)
  o International Conference on Generative Programming: Concepts & Experiences (GPCE)
  o Systems, Programming, Languages and Applications: Software for Humanity (SPLASH)

http://planet-sl.org/sble-at-sle2013/
Acknowledgements

VBI SynBio Group
► M. Wilson
► D. Ball
► M. Lux
► L. Adam
► C. Overend

GenoCAD Alumni
► Yizhi Patrick Cai (JHU)
► Mike Czar (Carillon)

GenoCAD Collaborators
► VBI: S. Hoops, J. Lewis
► SBOL: H. Sauro, C. Myers, D. Densmore, C. Rodriguez, M. Galdzicki and many more
► Language: Eric Van Wyck

Find GenoCAD
► www.genocad.org
► www.facebook.com/genocad/
► @genocad
► Google +
► LinkedIn