Automated Verification and Modification of DNA Sequences regarding DNA Synthesis Constraints

Ernst Oberortner*, Jan-Fang Cheng, Nathan J. Hillson, Samuel Deutsch

*** eoberortner@lbl.gov**

JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY

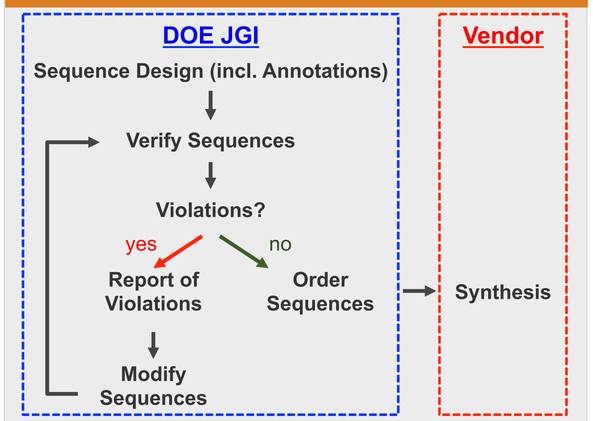
Abstract

Before *in silico* designed DNA sequences can be synthesized, the sequences must be verified and modified regarding DNA synthesis constraints, such as *repeats* or *GC content*.

At the DOE Joint Genome Institute (JGI), we developed the Sequence Polishing Library (SPL) to verify and – in case of violations – to modify DNA sequences in an automated manner. Modifications depend on the region of a DNA sequence that violates a constraint. For example, if a coding sequence contains repeats, then codon juggling can be performed. However, codon juggling cannot be performed if a violation occurs in a non-coding region, such as a promoter. We emphasize the use of standards that support meta-information about DNA sequences, such as via annotations.

In addition, SPL offers a yet simple but expressive language to specify DNA synthesis constraints. Our goal is to further develop and contribute this language to standardize the communication of DNA synthesis constraints.

Design Automation at the DOE JGI



Systematic Approach of Sequence "Polishing"

SPL automatically "**polishes**" DNA sequences in three steps:

Step I: *Verification*
Verify a sequence against a set of vendor-specific constraints, such as GC% or repeats.

Step II: *Report*
Comprehensive report of violations, such as location, length, or type of constraint.

Step III: *Modification*
Perform corrective actions on regions that violate constraints, such as through codon replacements, RBS re-calculation, or promoter/terminator switching.

STEP I
Verification of DNA Sequences against DNA Synthesis Constraints

Types of DNA Synthesis *Constraints*

- **Restriction Sites**
- **GC Content**
  - **Local ("Windowed") vs. Global**
- **Repeats**
  - **Local ("Windowed") vs. Global**
  - **Tandem vs. Interspersed**

```
(N)* ACGTGTCA       ACGTGTCA (N)*
(N)* ACGTGTCA (N)+ ACGTGTCA (N)*
```

  - **Direct vs. Inverted**

```
(N)* ACGTGTCA (N)* ACGTGTCA (N)*
(N)* ACGTGTCA (N)* TGACACGT (N)*
```

  - **Exact vs. Mutated**

```
(N)* ACGTGTCA (N)* ACGTGTCA (N)*
(N)* ACGTGTCA (N)* TGACACGC (N)*
```

- **Repeat Coverage**

Examples of DNA Synthesis Constraints

Description	Parameters
remove BsaI and AarI sites	BsaI: *GGTCTC*, AarI: *CACCTGC*
Local GC% (***X***bp window)	min% – max%
Global GC%	min% – max%
Homopolymers	A < ***X***, C < ***X***, G < ***X***, T < ***X***
Terminal repeats	No ***k***-mer in ***X***bp termini should be present elsewhere
Hairpins	No hairpin with stem greater than ***X***bp
Direct/inverted repeats	No exact repeat greater than ***X***bp within a ***Y***bp interval
Dimer repeats	No more than ***X*** consecutive Dimer repeats
Trimer repeats	No more than ***X*** consecutive Trimer repeats
Regions with high repeat content	Any sequence contains more that ***X*** % of bases that are part of repeats of ***Y*** bases or longer

SPL utilizes the **BBMap bioinformatics library** (http://sourceforge.net/projects/bbmap)

STEP II
Reporting Violations of DNA Synthesis Constraints

Requirement

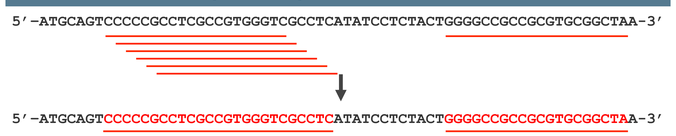
Consolidated Report
Comprehensive Format

Report: *Merging of Violations*

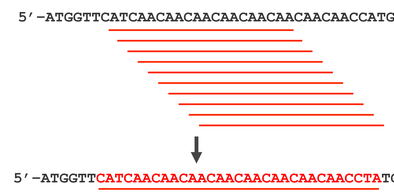
A Violation has three characteristics:
<Type, Location, Length>

Two violations can be merged if
(1) they are of the same type and
(2) they overlap

Example 1: Report of GC% Violations



Example 2: Report of repeating k-mers



STEP III
Correcting Sequences based on Modification Strategies

Characteristics of Modification Strategies

What constraint was violated?
Repeat, GC%

Where did the violation occur?
Coding Sequence, RBS, Promoter

How many codons should be replaced?
"Conservative" vs. "Liberal"

What codon should replace the current codon?
Relative Synonymous Codon Usage (RSCU)

Approach of Correcting Sequences

Frame of Violation → Amino Acid → Alternative Codons
Pichia stipitis CBS 6054

ATG	TTT	GCT	GGC
M	F	A	G
Methionine	Phenylalanine	Alanine	Glycine
ATG(1.00)	TTC(0.67) TTT(0.33)	GCT(0.43) GCC(0.29) GCA(0.22) GCG(0.06)	GGT(0.40) GGA(0.31) GGC(0.21) GGG(0.08)

Modification Strategies

Least Different	ATG	TTC	GCC	GGA
Mostly Used	ATG	TTC	GCT	GGT
Weighted	ATG	TTC	GCC	GGT
Random	ATG	TTT	GCG	GGC

Sequence Polishing Library (SPL)

SPL is a **computational tool** for the *in silico* verification and modification of DNA sequences against DNA synthesis constraints.

Input:
- **DNA sequences:** FASTA, CSV
  - **Preferably:** Annotated sequences (GenBank, SBOL)
- Set of **DNA Synthesis Constraints**
- **Codon Translation and Usage Tables**
  - Genetic Code (Default: ***Standard***)
  - Relative Synonymous Codon Usage (***RSCU***) of organism
- **Modification Strategy** and its parameters

Output:
- **Report** of violations
- **Log** of Modifications
- **Polished Sequences** (FASTA, CSV, GenBank, SBOL)

Input
DNA Synthesis Constraints

Language-based Approach

GC%			
Global GC	min [%]	max [%]	
Local GC	min [%]	max [%]	window size
Terminal GC	min [%]	max [%]	terminal size

REPEAT				
nucleotide	max. repeats			
k	direct/ inverted	space	window size	#mutations

RepeatCoverage	max [%]	window size

Input
Format of Codon Usage Tables

Codon Usage Database:
http://www.kazusa.or.jp/codon/

```
[triplet] [amino acid] [fraction] [frequency]([number])

UUU F 0.43 19.5 ( 37436)  UCU S 0.27 23.4 ( 44892)  UAU Y 0.41 14.5 ( 27806)  UGU C 0.69  7.0 ( 13408)
UUC F 0.57 25.7 ( 49198)  UCC S 0.17 14.6 ( 28000)  UAC Y 0.59 20.8 ( 39805)  UGC C 0.31  3.1 (  5992)
UUA L 0.15 14.6 ( 28010)  UCA S 0.16 13.9 ( 26655)  UAA * 0.35  0.7 (  1365)  UGA * 0.21  0.4 (   806)
UUG L 0.47 46.2 ( 88588)  UCG S 0.15 13.2 ( 25217)  UAG * 0.44  0.9 (  1686)  UGG W 1.00 10.3 ( 19688)

CUU L 0.16 15.9 ( 30449)  CCU P 0.35 15.2 ( 29151)  CAU H 0.54 11.4 ( 21784)  CGU R 0.15  6.5 ( 12542)
CUC L 0.11 11.1 ( 21307)  CCC P 0.18  7.7 ( 14787)  CAC H 0.46  9.9 ( 18903)  CGC R 0.05  2.0 (  3845)
CUA L 0.06  5.8 ( 11168)  CCA P 0.38 16.9 ( 32351)  CAA Q 0.52 19.7 ( 37858)  CGA R 0.07  3.2 (  6071)
CUG L 0.05  5.2 (  9929)  CCG P 0.10  4.3 (  8150)  CAG Q 0.48 18.5 ( 35412)  CGG R 0.04  1.9 (  3588)

AUU I 0.40 26.2 ( 50146)  ACU T 0.36 20.6 ( 39555)  AAU N 0.44 24.8 ( 47635)  AGU S 0.14 12.2 ( 23303)
AUC I 0.41 26.9 ( 51631)  ACC T 0.27 15.4 ( 29474)  AAC N 0.56 31.8 ( 61018)  AGC S 0.11  9.8 ( 18705)
AUA I 0.19 12.7 ( 24378)  ACA T 0.25 14.2 ( 27263)  AAA K 0.38 25.9 ( 49591)  AGA R 0.57 24.2 ( 46375)
AUG M 1.00 17.9 ( 34228)  ACG T 0.12  7.1 ( 13526)  AAG K 0.62 42.6 ( 81597)  AGG R 0.11  4.5 (  8679)

GUU V 0.32 19.7 ( 37769)  GCU A 0.43 25.5 ( 48823)  GAU D 0.52 30.5 ( 58542)  GGU G 0.40 21.1 ( 40515)
GUC V 0.25 15.3 ( 29428)  GCC A 0.29 17.2 ( 32893)  GAC D 0.48 28.6 ( 54883)  GGC G 0.21 11.0 ( 21126)
GUA V 0.22 13.4 ( 25711)  GCA A 0.22 13.1 ( 25097)  GAA E 0.67 44.8 ( 85900)  GGA G 0.31 16.4 ( 31396)
GUG V 0.21 13.0 ( 24861)  GCG A 0.06  3.7 (  7073)  GAG E 0.33 21.8 ( 41878)  GGG G 0.08  4.5 (  8542)
```

Pichia stipitis CBS 6054

Acknowledgements

The authors thank the following JGI members and collaborators for their contributions to the development of the SPL:
Xianwei John Meng, Lisa Simirenko, Brian Bushnell, Beat Christen, Adam Clore and Thomas Hofmeister.

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.