Analysis of the SBOL iGEM Data Set

Jeanet Mante, James Mclaughlin, Chris J. Myers
Department of Bioengineering, University of Utah, Salt Lake City, UT 84112

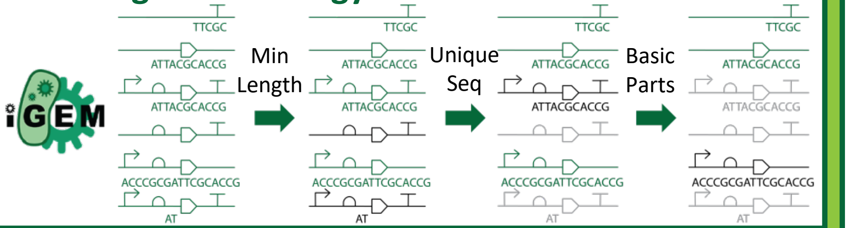
Introduction

- Use of the iGEM (international Genetically Engineered Machines) registry
- Converted to the Synthetic Biology Open Language(SBOL) data format [1]
- Preliminary analysis of the data to predict the size of a potential library and quantify current problems with the data

Results

- The goal is develop a library of parts (basic and composite) that are well annotated
- Suggest a library split into 2 types of data:
  1. **Innate**: the sequence and factors that relate solely to the sequence (e.g. the fluorescence of GFP)
  2. **Experience**: data related to the context (e.g. the strength of a promoter which relates to the organism in which it is used)
- This facilitates the reuse of parts and facilitates inter-organism
- As a step towards such a library, we analyzed the iGEM SBOL data from 2017
- This provided an estimate of the number of sequences that may be useful in future genetic engineering designs

Filtering Methodology



Filtering Results

SO Type (abbreviated)	Seq Length [min, av ± sd, max]	Min Length	Seq Count	Seq > Min Length	Unique Seq	Basic Parts
CDS	[0, 1140 ± 1430, 66880]	40	7,689	7,198	6,788	6,188
Chromosome	[0, 790 ± 1850, 8300]	6	73	13	13	10
Eng. Region	[0, 1730 ± 1560, 36200]	6	20,171	19,477	17,664	3,698
Transcript	[0, 220 ± 360, 3180]	6	595	556	538	485
oriT	[0, 960 ± 760, 3620]	6	41	39	39	20
Plasmid	[0, 4130 ± 3930, 49730]	40	609	526	484	398
Plas Vector	[0, 3790 ± 3000, 48170]	40	404	379	369	353
Polypeptide	[0, 590 ± 2450, 66190]	6	769	718	700	665
Primer	[0, 80 ± 320, 2510]	6	582	574	567	567
Promoter	[0, 380 ± 600, 7890]	6	3,106	2,965	2,770	2,493
Scar	[0, 70 ± 360, 2280]	6	40	26	25	24
RBS	[0, 70 ± 290, 4270]	6	525	494	454	448
Seq feature	[0, 800 ± 1730, 49730]	6	3,149	2,734	2,581	1,925
T7 Promoter	[0, 6220 ± 12790, 36940]	6	35	32	28	24
Tag	[0, 190 ± 520, 5850]	6	288	263	233	222
Terminator	[0, 150 ± 260, 3390]	6	388	381	335	329
Total			**38,464**	**36,375**	**33,588**	**17,851**

Snapgene Annotation

- Snapgene annotation of basic parts
  1. No known features contained
  2. Feature identified is the same as the type expected
  3. Features contain more and/or different feature types than expected

SO Type (abbreviated)	No Known	Single Feature	Wrong Feature(s)
CDS	4,145	973	1,070
Promoter	1,706	151	636
RBS	406	24	18
Terminator	239	53	37
Total	**6,496**	**1,201**	**1,761**

Further Work

- Manually exam and edit/remove the remaining basic parts to form a library
- Create a confidence metric which incorporates heuristics used for filtering
- Use analysis methods to create a pipeline for future library creation

Acknowledgements

This work is supported by the NSF via grant 1939892.

References

1. Galdzicki, M. et al. The synthetic biology open language (sbol) provides a community standard for communicating designs in synthetic biology. Nature Biotechnology 32, 6 (Jun 2014), 545–550