Analysis of the SBOL iGEM Data Set

Jeanet Mante
jv@mante.net
University of Utah
Salt Lake City, Utah

James Mclaughlin
j.a.mclaughlin@ncl.ac.uk
Newcastle University
Newcastle-upon-Tyne, UK

Chris J. Myers
myers@ece.utah.edu
University of Utah
Salt Lake City, Utah

1 INTRODUCTION

Synthetic biology is a movement to standardize genetic engineering and make it more repeatable. An important advancement was the development of standardized genetic parts known as *BioBricks*, which can be composed using restriction enzyme assembly [3]. The iGEM (international Genetically Engineered Machines) competition is an important synthetic biology outreach activity which is run by the iGEM foundation in keeping with their principles of the advancement of synthetic biology via education, competition, and development of an open and collaborative community. As part of the iGEM competition students submit records of any 'parts' they create to the iGEM registry (http://parts.igem.org/Main_Page). The iGEM registry was converted to the *Synthetic Biology Open Language* (SBOL) data format, a standard language for describing genetic designs, [2] and a preliminary analysis of the data was carried out to predict the size of a potential library as well as quantify current problems with the registry data set.

2 RESULTS

The ultimate goal of our analysis is to develop a library of parts (basic and composite) that are well annotated enough to be easily reused and computationally modelled. To do this, this paper proposes thinking of two separate types of data, the innate versus the experimental. The innate data of a part would be the sequence and factors that relate solely to the sequence (such as the fluorescence of GFP). Any data related to the context of the part is considered 'experience' data. For example, the strength of a promoter is experience data as it relates to the organism in which it is used. We suggest a library that contains core parts and their innate data and links in each 'experience' of part use and the data related to that via a provenance annotation. This library model facilitates the reuse of parts as it highlights the 'popularity' (a useful heuristic for confidence) of a part via the number of experiences it has. This also facilitates inter-organism work as it more clearly separates the data linked to a particular model organism and encourages the collection of experimental data (such as strain and growth medium) for every measured property of a part.

As a step towards the development of a library, we analyzed the iGEM SBOL Data set created in 2017. The initial analysis provides an estimate of the number of unique sequences that are useful in future genetic engineering designs.

A detailed analysis is shown in Table 1. This table is based on the sequential application of filters based on the type of part represented using the Sequence Ontology (SO) [1]. The first filter applied was the minimum length filter ("Sequences Over Minimum Length" column). The minimum length used is shown in the column "Minimum Length Parameter". Initially, the minimum length for many parts is set to 6 base pairs (bp) or the equivalent of 2 codons/amino acids (aa). However, for CDS 40 bp (about 13 amino acids) was used as the shortest human enzyme found in UniProt (Cytochrome P450 2A7) is 20aa (60bp). As plasmid and plasmid vectors generally contain a CDS their minimum length was also increased to 40bp. This simple filter removes almost 2,000 components which had no sequence associated with it or a very short sequence. The next filter that was applied looked at unique sequences per SO Type. It removed any exact sequence copies. However, as it worked by SO type the same sequence may, for example, still be repeated as both a terminator and CDS. This can be seen as 33,113 is the total number of unique sequences over a minimum length whereas by role the total is 33,588. Thus 475 sequences are repeated exactly but given different SO types. The final filter of which considers looking for basic parts. Components may be 'basic' or they may be 'composite'. Basic parts do not contain any sub-parts whilst compound parts have one or more sub-parts.

3 DISCUSSION

The initial analysis has indicated that there are probably fewer than **18,000** unique, non-composite sequences which might lead to well described parts with complete records. However, the filtering is not exact. For example, the exclusion of all composite parts is perhaps too strict as there are composite parts (such as the double terminator BBa_B0015) which are useful, and others such as Engineered Regions which would be expected to have sub-parts. As a rough pass, removing parts such as terminators that contain promoters as sub-parts and thus are likely to be mis-annotated, useless, or both, it is a simple and effective heuristic.

The initial filtering removed roughly 45% of the data. Whilst some of this may be too conservative and worth revisiting, there are also parts contained in the final 17,851

Table 1: iGEM Conversion. A sequential filtering was carried out to try and determine the 'useful unique entities' in the converted iGEM dataset. Analysis was carried out per SO type as expectations for different types are different (e.g. RBSs would be expected to be shorter than chromosomes). Sequence count provides a simple count of the number of sequences, sequences over the minimum length is the number of sequences with a length greater than that specified by the minimum length parameter, unique sequences over a minimum length takes the previous count but removes any duplicate sequences within the category. Finally, an analysis was carried out to filter out composite parts; NB: this assumption may be too stringent for types such as Engineered Region.

SO Type	iGEM Types	Minimum Length Parameter	Sequence Count	Sequences Over Minimum Length	Unique Sequences Over Minimum Length	Unique Over Minimum Length Basic Parts
CDS	Basic, Coding	40	7,689	7,198	6,788	6,188
Chromosome	Cell	6	73	13	13	10
Engineered Region	Composite, Device, Generator, Intermediate, Inverter, Measurement, Project, Reporter, Signalling, Translational_Unit	6	20,171	19,477	17,664	3,700
Mature Transcript Region	RNA	6	595	556	538	485
oriT	Conjugation	6	41	39	39	20
Plasmid	Plasmid	40	609	526	484	398
Plasmid Vector	Plasmid_Backbone	40	404	379	369	353
Polypeptide Domain	Protein_Domain	6	769	718	700	665
Primer	Primer	6	582	574	567	567
Promoter	Regulatory	6	3,106	2,965	2,770	2,495
Restriction Enzyme Assembly Scar	Scar	6	40	26	25	24
Ribosome Entry Site	RBS	6	525	494	454	448
Sequence feature	DNA, Other, Terminator	6	3,149	2,734	2,581	1,923
T7 RNA Polymerase Promoter	T7	6	35	32	28	24
Tag	Tag	6	288	263	233	222
Terminator	Terminator	6	388	381	335	329
Total			**38,464**	**36,375**	**33,588**	**17,851**

that still need to be filtered out to create the final part library. We suggest further work on refining part sets based on SO type considering sequence similarity clustering, automated sequence annotation, and machine learning based on the part descriptions provided. We hope to be ready to present the resulting library in the expanded paper arising from this abstract in January. Additionally, we note that whilst the iGEM registry is a large repository of synthetic biology information it does have several drawbacks: 1) The un-standardised nature of fields used, 2) the lack of part verification, 3) the lack of part removal, and 4) part duplication. We suggest that to create a useful library of parts from the iGEM library these four concerns must be addressed, either during a conversion process to SBOL or in the registry itself.

REFERENCES

[1] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. The sequence ontology: a tool for the unification of genome annotations. *Genome biology 6*, 5 (2005), R44.

[2] Galdzicki, M., Clancy, K. P., Oberortner, E., Pocock, M., Quinn, J. Y., Rodriguez, C. A., Roehner, N., Wilson, M. L., Adam, L., Anderson, J. C., and 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.

[3] Shetty, R. P., Endy, D., and Knight, T. F. Engineering biobrick vectors from biobrick parts. *Journal of Biological Engineering 2*, 1 (Apr 2008), 5.