Estándares de datos en SynBio

Martín Gutiérrez

November 8, 2024

Estándares

En esta sesión vamos a hablar de la representación de las partes para diseñar circuitos genéticos.

Habiendo ya visto gran parte de lo que se puede hacer en biología programable, es tiempo de revisar cómo es que se plasman los datos referentes a partes genéticas necesarias para construir circuitos/dispositivos y algunos recursos que están relacionados a esos datos.

Formatos

Expondré tres formatos que se usan mucho hoy en día, y nos centraremos en el último de ellos:

- FASTA/FASTQ
- Genbank
- SBOL

FASTA/FASTQ

FASTA/FASTQ son archivos de texto que principalmente describen una secuencia de nucleótidos/aminoácidos con anotaciones.

Ejemplo:

```
> guido meza
ACCTACAGATGGTGAATATCTCCCTGCGAGTGTTGTCTCGACCCAATGCTCAGGAGCTTC
CTAGCATGTACCAGCGCCTAGGGCTGGACTACGAGGAACGAGTGTTGCCGTCCATTGTCA
ACGAGGTGCTCAAGAGTGTGGTGGCCAAGTTCAATGCCTCACAGCTGATCACCCAGCGGG
CCCAG
```

Genbank

Genbank es otro formato que está basado en texto plano y que almacena una secuencia, pero adicionalmente anota características de la secuencia que lo hacen un formato más robusto y completo que FASTA/FASTQ.

Ejemplo: en el editor de texto

SBOL (I)

Synthetic Biology Open Language (SBOL) es un formato de almacenamiento de datos de partes biológicas que apunta a ser un estándar general dentro de la biología programable/sintética. Dicha representación de datos está basada en RDF (que es de base XML).

SBOL trata los datos con un enfoque jerárquico, orientado al intercambio a través de la internet y centrado en que una amplia gama de recursos puedan ser accedidos y utilizados tanto por humanos como por máquinas.

SBOL (II)

SBOL va en su versión 3.1.0 y el artículo y la especificación (del modelo de software/conceptual) completa se encuentran en `https://tinyurl.com/yjmh34h8`

Ejemplo de un archivo de SBOL: ver editor de texto.

Nota: SBOL gráficamente tiene la forma en la que han estado representado sus circuitos durante el semestre.

SBOL (III)

La ventaja de utilizar estos estándares radica en que muchos de los datos (en cualquiera de los formatos) ya está plasmado en repositorios de datos que almacenan las partes en sí, o bien referencias a ellas, de modo que sean accesibles inmediatamente y de una manera estándar a los investigadores que las necesiten.

En particular, SBOL ha cobrado mucha fuerza en el mundo de la biología sintética al posicionarse como formato por defecto en la comunidad, pero también por su amplia gama de herramientas y recursos asociados. Desventajas: la diversidad de versiones del estándar y poco avance hacia la última, incapacidad de funcionar nativamente como lenguaje de especificación para simulaciones de poblaciones celulares.

Hablaremos de unas pocos recursos y herramientas ligados a SBOL a continuación.

Los repositorios

iGEM Registry (`https://technology.igem.org/registry` y Synbiohub (`https://synbiohub.org`) son dos repositorios (vinculados) que concentran la mayor cantidad de partes genéticas (sintéticas) en formato SBOL (y Genbank y FASTA).

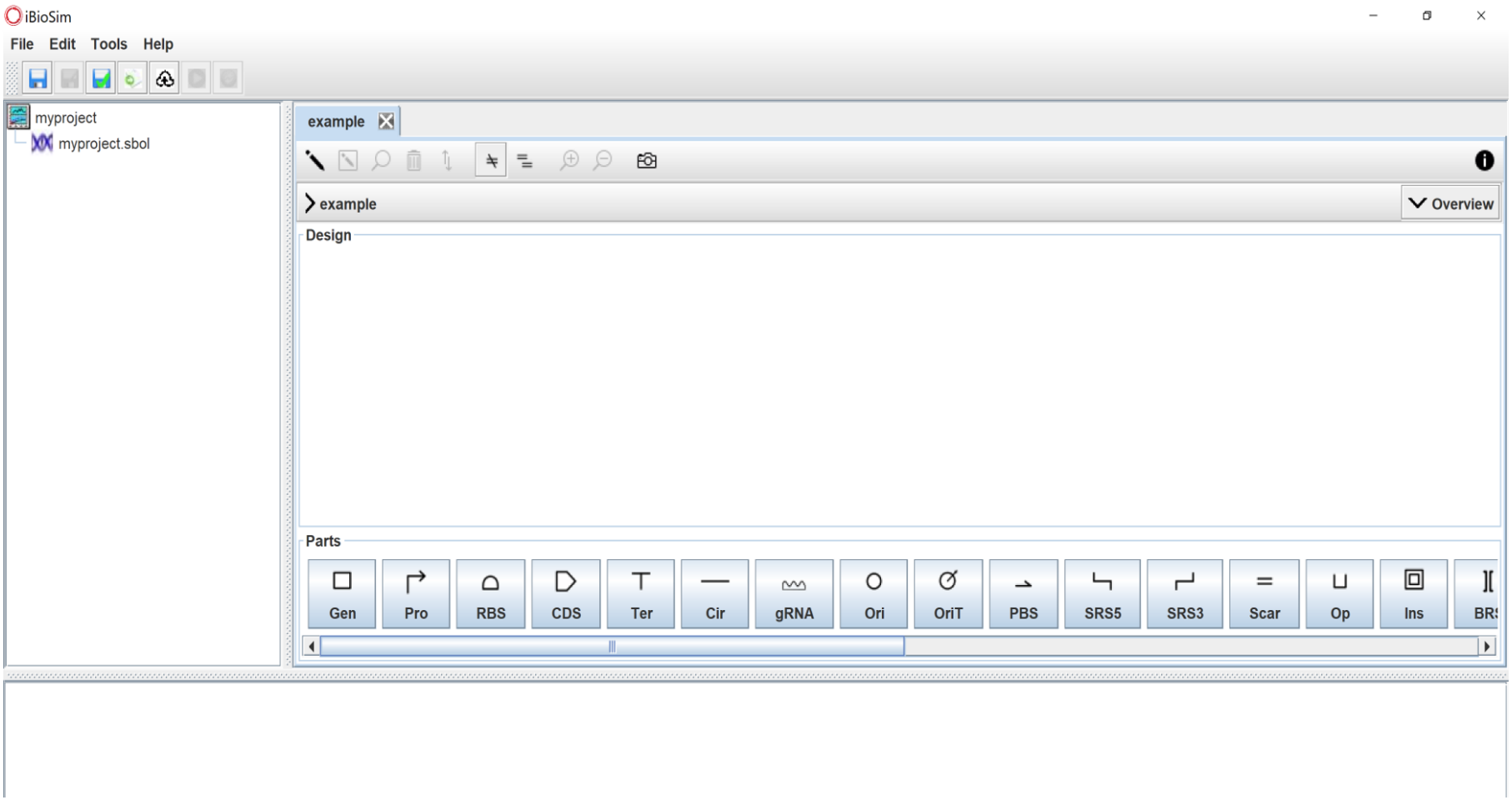
Estos repositorios se nutren de la competencia internacional de diseño y creación de circuitos sinntéticos de estudiantes, iGEM, y de los resultados obtenidos por investigadores en todo el mundo (a modo de GitHub), respectivamente.

iBioSim (v3) (I)

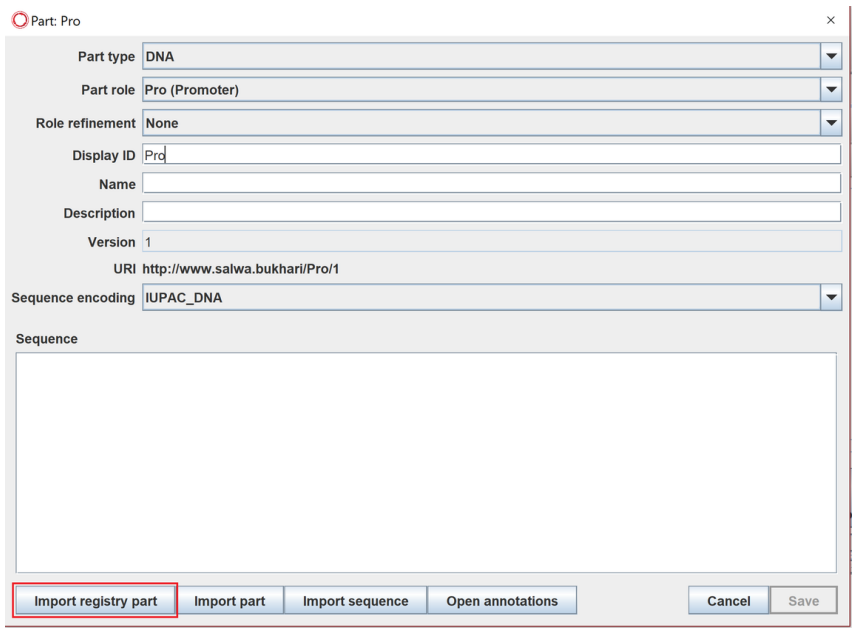
Es un simulador de comportamiento monocelular que trabaja con un diseñador de circuitos genéticos (sintéticos). Se conecta con los repositorios mencionados previamente para extraer las características de las partes necesarias en el diseño del circuito. Con ello, logra simular, en el tiempo, el comportamiento del circuito dentro de la célula en el tiempo.

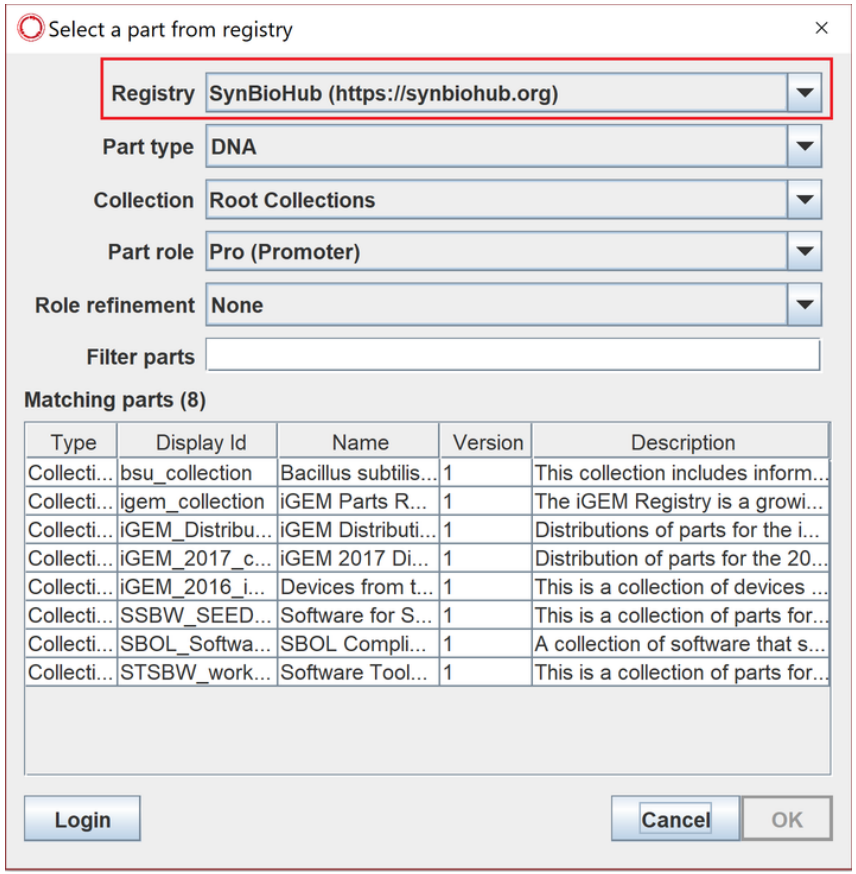
Dicho simulador usa SBOL (versión 3) y Systems Biology Markup Language (SBML) para configurar y efectuar la simulación del circuito. El artículo referente a iBioSim v3 se encuentra en `https://pubs.acs.org/doi/10.1021/acssynbio.8b00078`.

iBioSim (v3) (II)



iBioSim (v3) (III)





Cello (v2.1) (I)

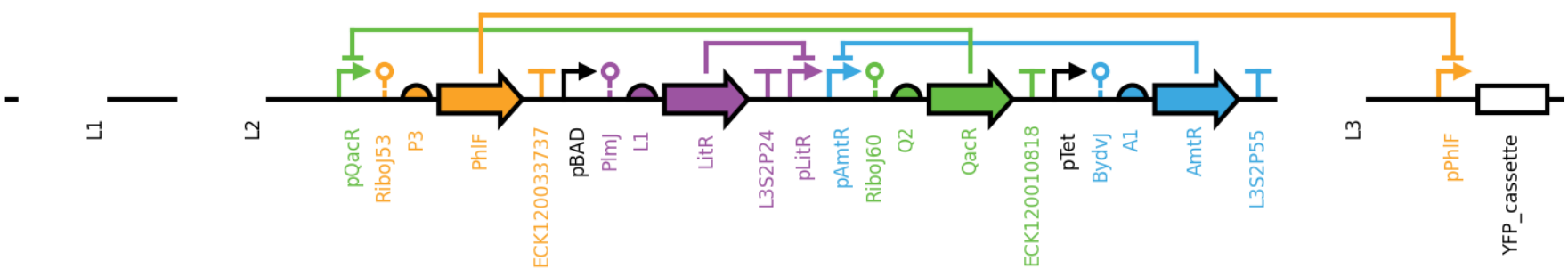
Cello es un software de "traducción" de lógica digital a circuito genético: se basa en una especificación de Verilog para definir el comportamiento booleano esperado del circuito (como si fuese un circuito digital). Con dicha especificación, Cello calcula (utilizando métodos metaheurísticos) posibles circuitos genéticos que ejecutan esa lógica.

Dicho software funciona en una versión de escritorio (`https://github.com/CIDARLAB/Cello-v2-1-Core`) y una versión online (`https://www.cellocad.org`). El output de Cello se expresa en formato SBOL. El paper de Cello v2.0 se encuentra en la siguiente referencia: `https://www.nature.com/articles/s41596-021-00675-2`

Cello (v2.1) (II)

```
/*
 Nand gate
*/
module nand_gate
(
 a,
 b,
 out
 );

   input a;
   input b;

   output out;

   assign out = ~(a & b);

endmodule // nand_gate
```

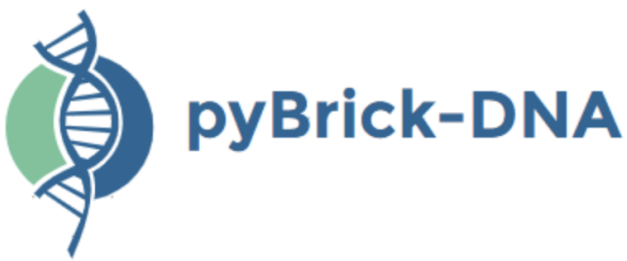
Design Option 1



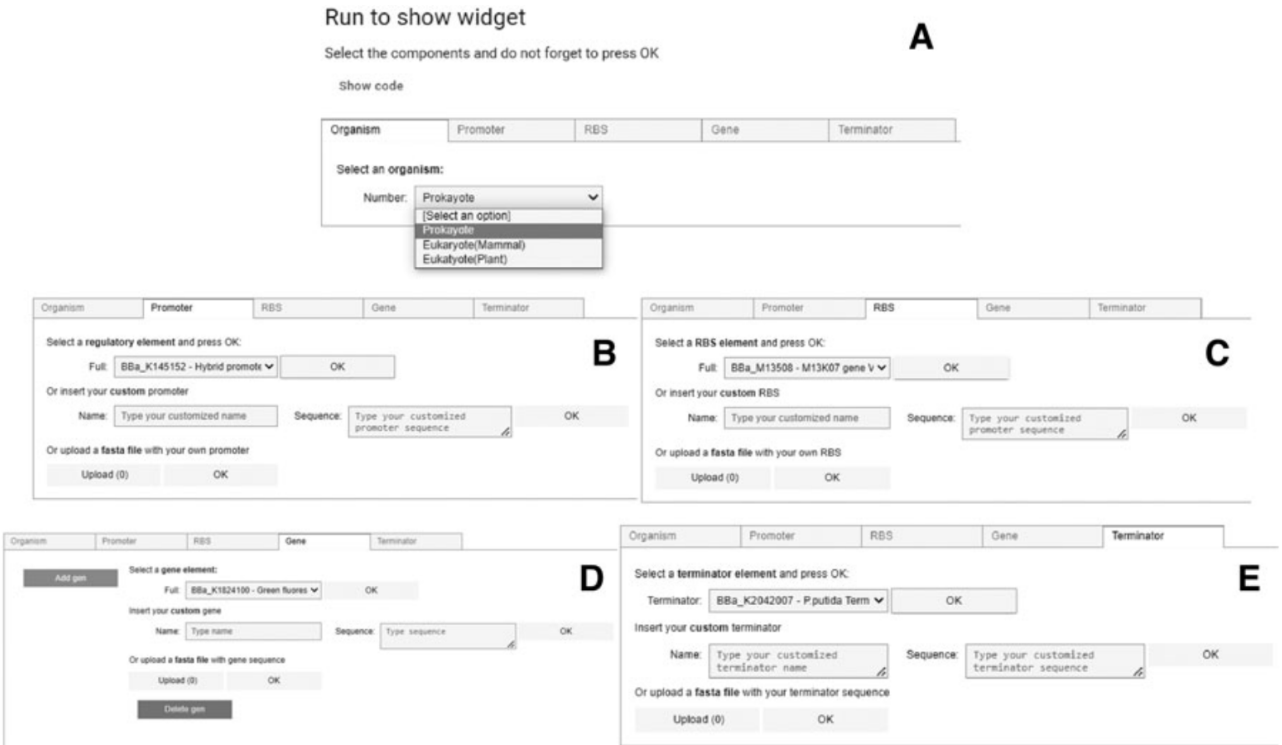
pyBrick-DNA (I)

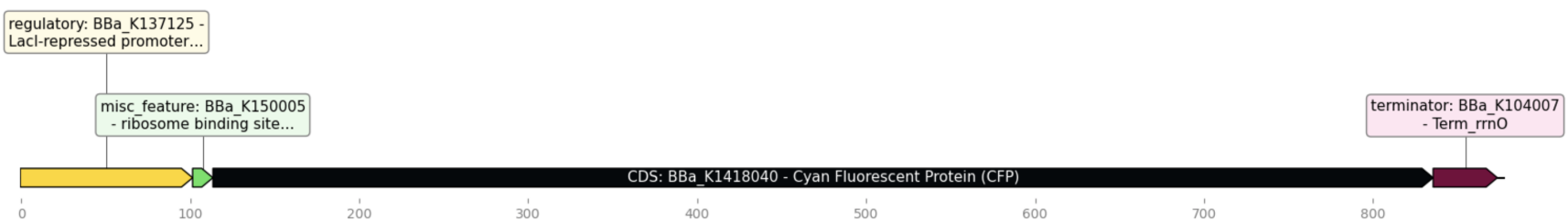
pyBrick-DNA (`https://www.liebertpub.com/doi/10.1089/cmb.2023.0008`, `https://github.com/gladyscavero/pyBrick-DNA`) es un notebook desarrollado en Google Colab (Python) mediante la colaboración entre UTEC (Perú) y la UDP para ensamblaje de circuitos sintéticos y sistemas de CRISPR en plantas.

Este software usa como output el formato Genbank.



pyBrick-DNA (II)





Conclusión

Los estándares de datos ayudan a organizar el conocimiento que se tiene sobre elementos necesarios en la construcción de circuitos/dispositivos genéticos. La caracterización y su formato inciden en la cantidad de público objetivo, facilitando el acceso a los datos.

Sin perjuicio de lo anterior, hay un activo desarrollo de herramientas que facilitan aún más el acceso y manejo de los datos. Una consecuencia del uso de estas herramienta es también abrir el campo de diseño y desarrollo a investigadores que no solamente son biólogos (llevando a estos investigadores hasta el punto de proponer una secuencia para su implementación en laboratorio).

Próximas clases

Nos centraremos en la simulación AbM y AI in-vivo.

Nos vemos!!!