Data Integration and Mining for Synthetic Biology Design

September 2016

Goksel Misirli, and

Jennifer Hallinan, Matthew Pocock, Phillip Lord, James Alastair McLaughlin, Herbert Sauro, and Anil Wipat

Interdisciplinary Computing and Complex BioSystems (ICOS) research group

School of Computing Science

Engineering biological systems is challenging

C G A T → NNNNNNNNNN → 4^{10} ~ 1 million solutions

Order and layout

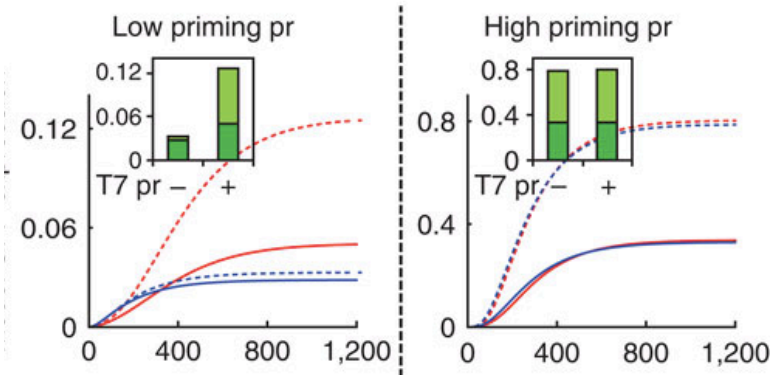
Regulatory elements

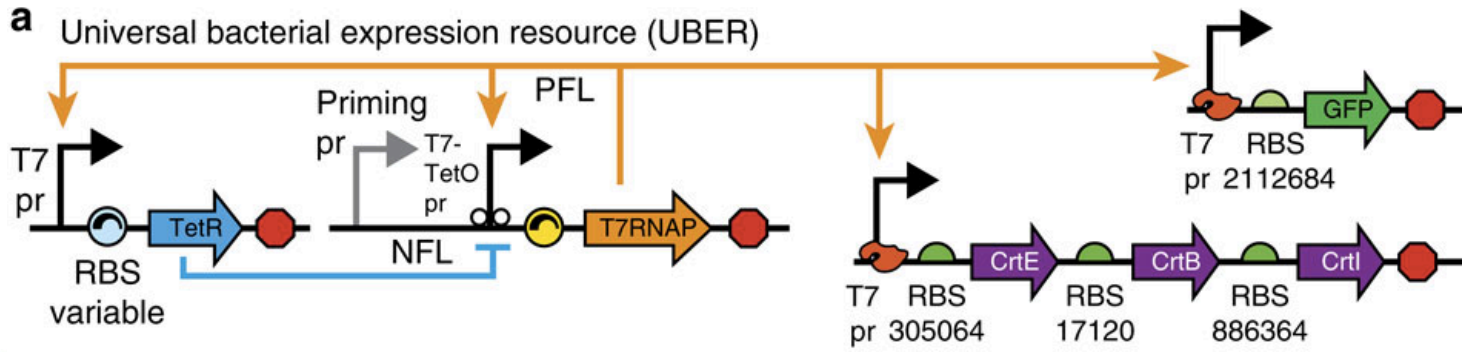
Molecular interactions

Strain/Host/Chassis

Biological Context

Experimental Conditions





Kushwaha and Salis, 2015

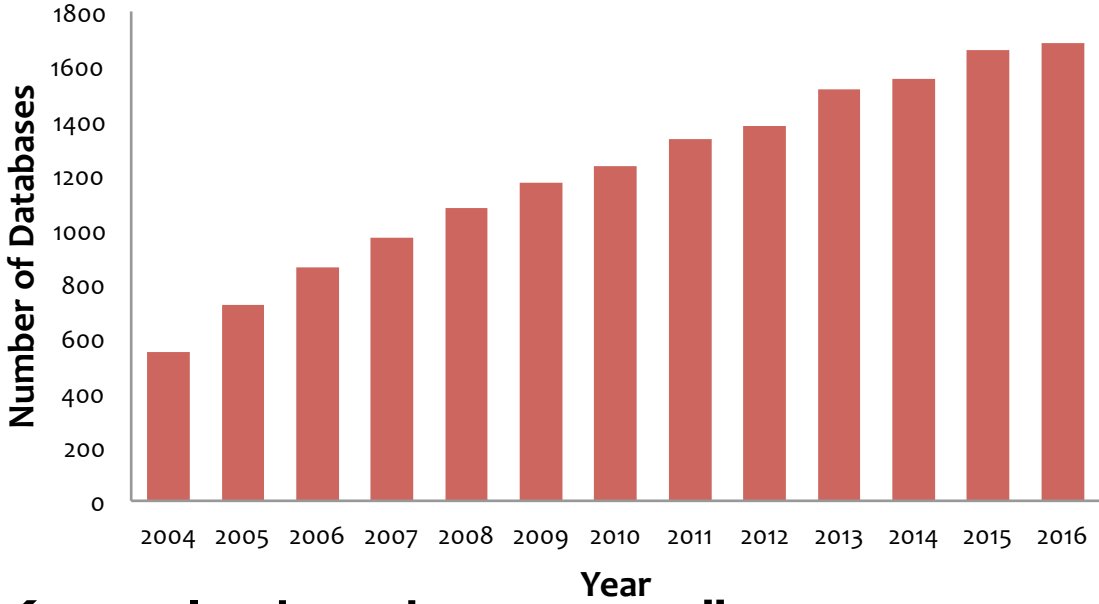
Newcastle University

I(C)$_2$S

FLOWERS

Data integration

* There is a large amount of information about model organisms such as *B. subtilis* and *E. coli*
* This information may be spread in
  * Different databases
  * Different formats
  * Different semantics
* This information can be Integrated and used to inform & constrain biological designs



"Knowledge is power"

Marx, Nature, 2013

Newcastle University

ICOS

FLOWERS

What is an ontology?

The representation of the entities of a domain is consistent and unambiguous

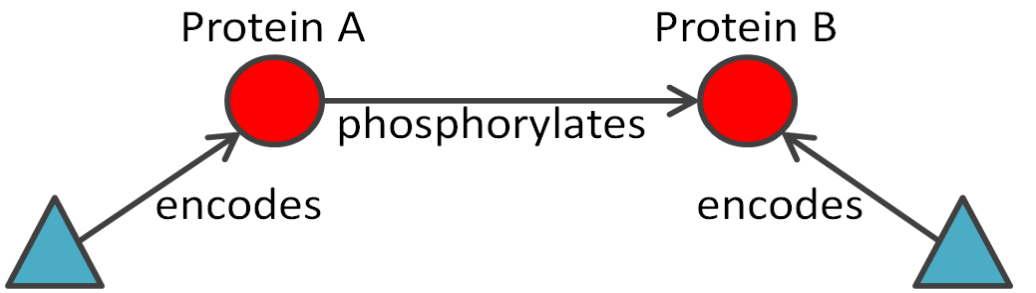
An abstract and simplified view of a domain being modelled

uses

captures

"***An explicit and formal specification of a conceptualisation***"
Gruber

Shared understanding of a domain



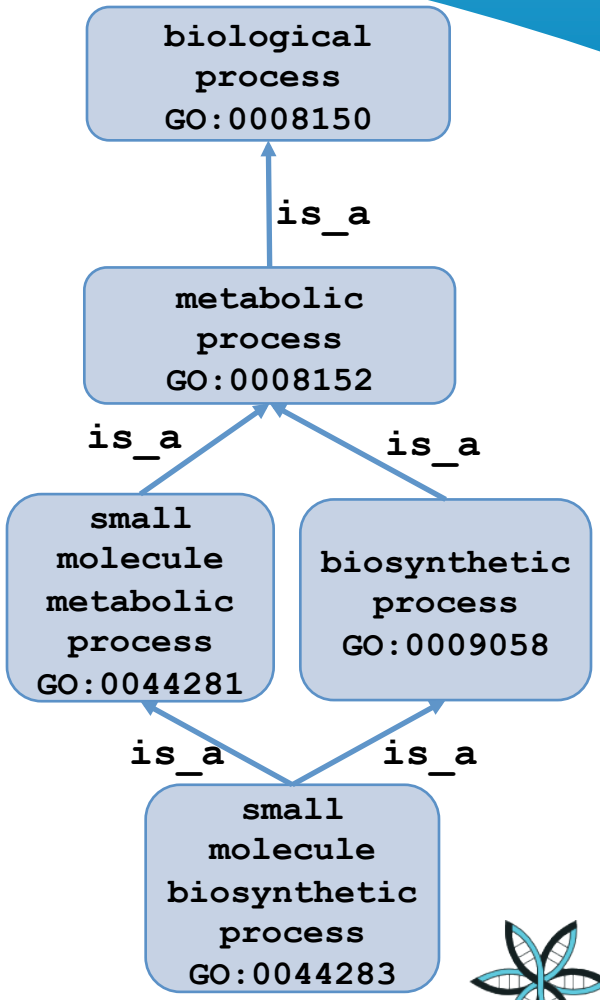
Newcastle University

I(C)$_2$S

FLOWERS

Semantic Web resources for synthetic biology

* Gene Ontology
* Sequence Ontology
* Systems Biology Ontology
* Synthetic Biology Open Language
* Standard Biological Parts Knowledgebase
* SBOL Stack
* Ontologies are needed to
  * Capture different relationships between biological parts
  * Facilitate data mining



Newcastle University

ICOS

FLOWERS

Synthetic Biology Ontology (SyBiOnt)

ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

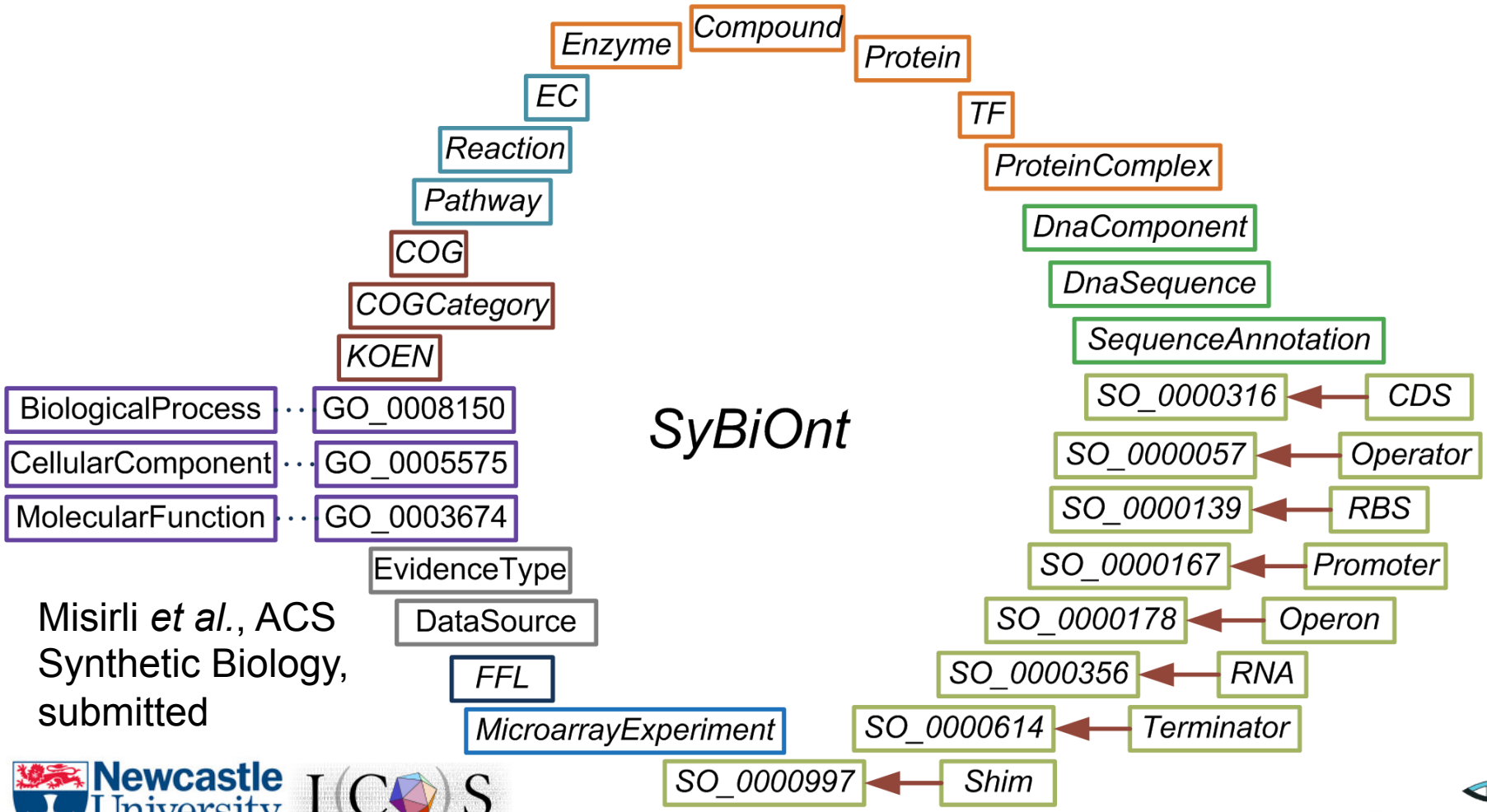
Data Integration and Mining for Synthetic Biology Design

Göksel Mısırlı,[†] Jennifer Hallinan,[†,∥] Matthew Pocock,[†,‡] Phillip Lord,[†] James Alastair McLaughlin,[†] Herbert Sauro,[§] and Anil Wipat*[,†]

Open Access, available from ACS Synthetic Biology

w3id.org/synbio/ont

Synthetic Biology Ontology (SyBiOnt)



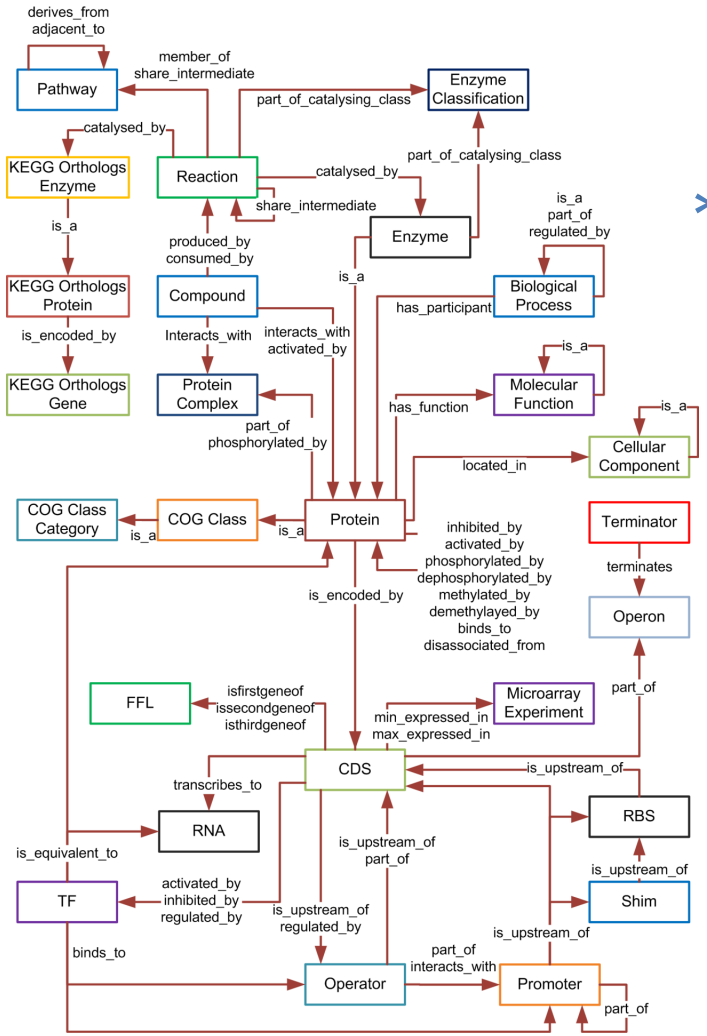
Misirli *et al.*, ACS Synthetic Biology, submitted

Newcastle University

ICOS

FLOWERS

SyBiOntKB knowledge base



* The knowledge base includes information about
  * Sequences, annotations
  * Metabolic pathways
  * Gene regulatory networks
  * Protein-protein interactions
  * Gene expression

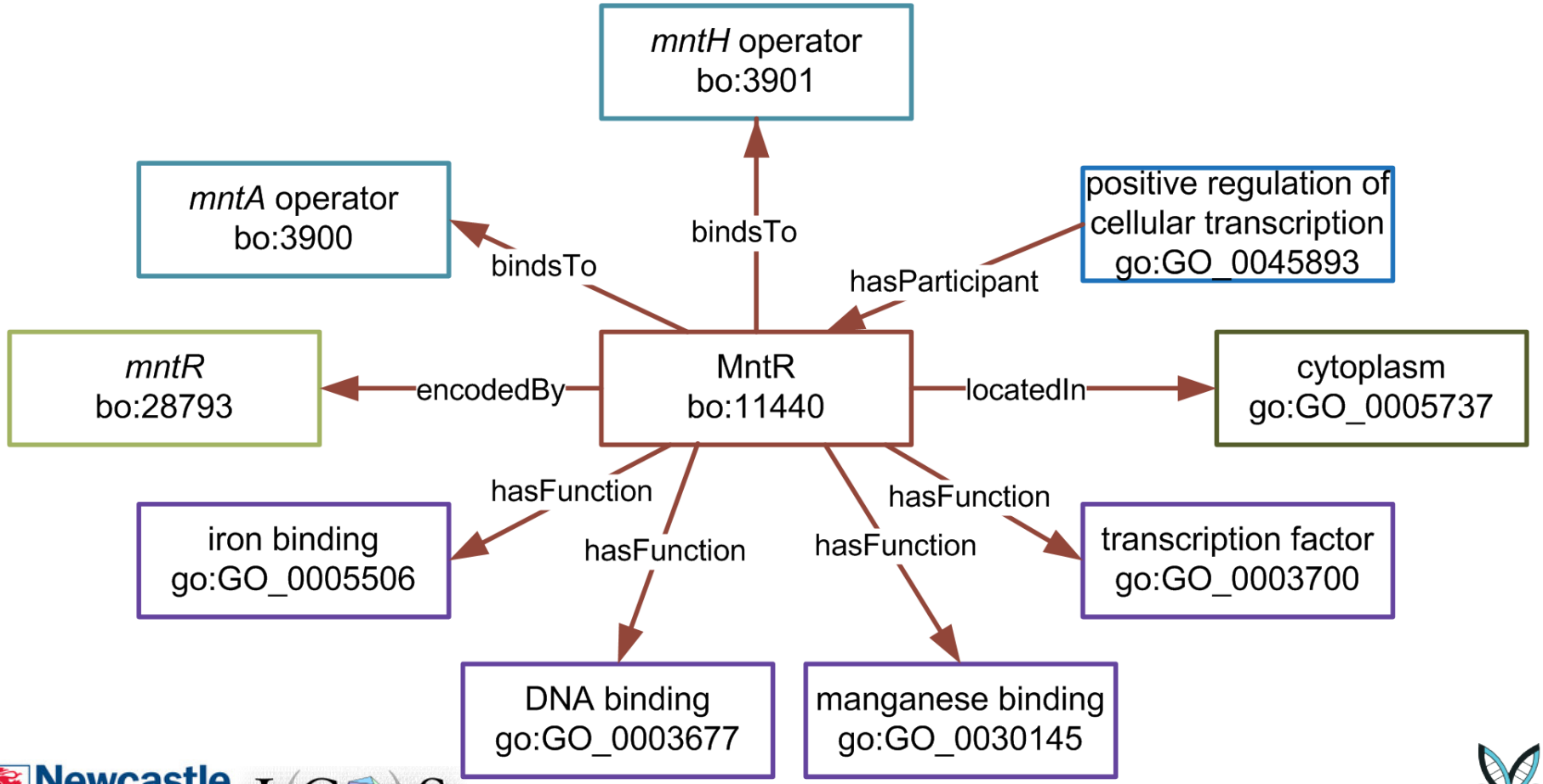
Misirli *et al.*, the Journal of Integrative Bioinformatics, 2013

Newcastle University

ICOS

FLOWERS

An example network from SyBiOntKB



Newcastle University

I(C$_2$)S

FLOWERS

Testing the ontology with competency questions

* *Which parts can be used as inducible promoters?*

* **`Operator`**s have **`regulation type`** restrictions to indicate whether they are used positively or negatively in regulating gene expression. A **`Promoter`** with one **`Operator`** part that has the `Positive' **`regulation type`** restriction is an **`Inducible Promoter`**.

Newcastle University

ICO2S

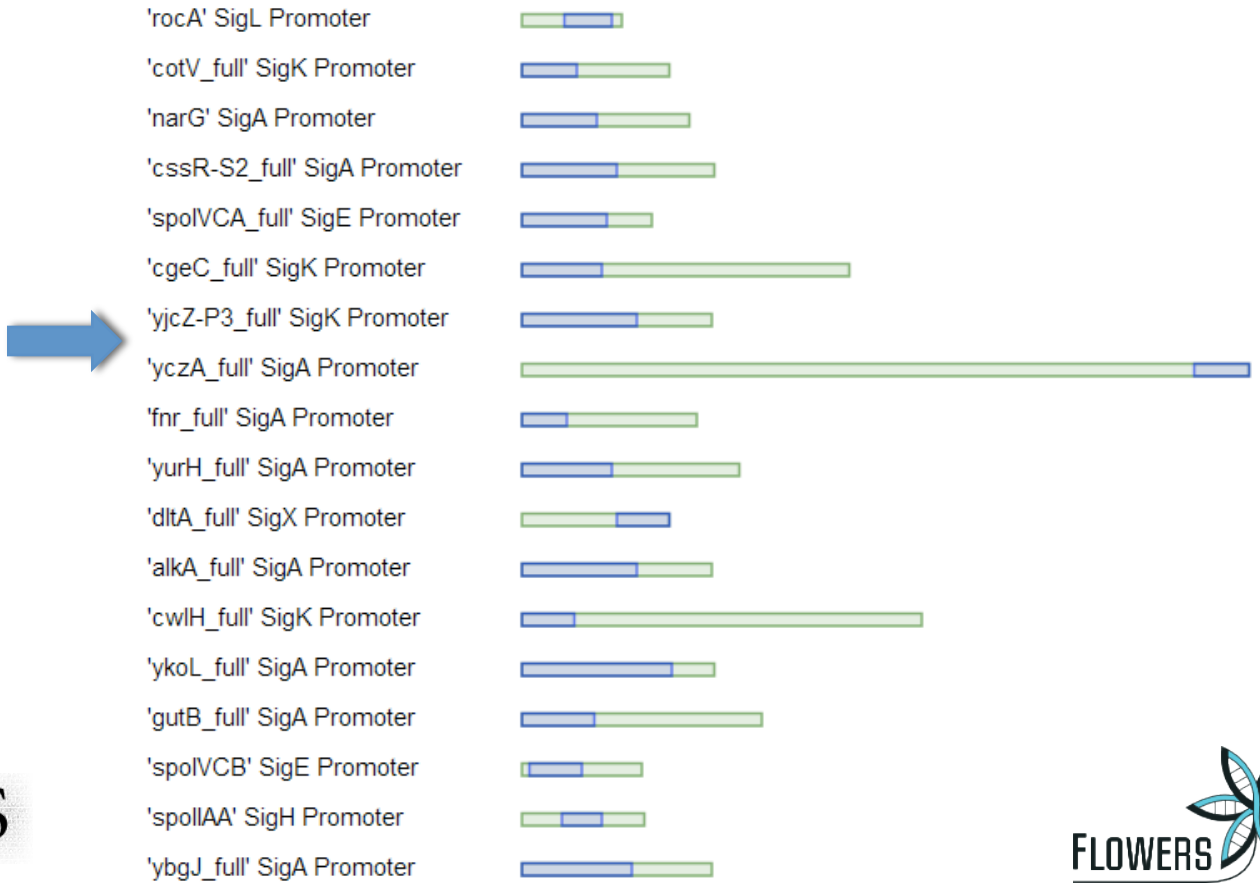
FLOWERS

InduciblePromoter query and the results

QUERY:

```
Promoter
and (has_part exactly 1 Operator)
and (has_part exactly 1 PositivelyRegulatedOperator)
```

RESULTS: In total, 51 promoters were classified



Newcastle University

ICOS

FLOWERS

Examples about the automated identification of biological parts

Part type	Count
Activator sites	222
Repressor sites	333
Inducible promoters	51
Repressible promoters	85
Inducible promoters with two inputs (AND/OR gates)	15
Repressible promoters with two inputs (NAND/NOR gates)	25

Part type	Count
SigA promoters	465
SigB promoters	67
Constitutive promoters	311
Repressor encoding CDSs	55
Activator encoding CDSs	44
Response regulator encoding CDSs	40
Kinase encoding CDSs	38

Newcastle University

I(CO)$_2$S

FLOWERS

More examples

* *Which pathways should be targeted for the over-production of ammonium?*

'Ammonium' is **produced by Reaction**s that are **member of** 'Arginine and proline metabolism' and 'Purine metabolism' **Pathway**s.

* *Which parts can be used to upregulate the production of ammonium?*

The Compound 'Ammonia' with the **accession** of 'C00014' is **produced by** the **Reaction** 'RN:R00131', which **consumes** the **Compound** 'Carbamide' (C00086). 'Carbamide' is **produced by** a **Reaction** that is **catalysed by** an **Enzyme**, which is a subclass of a **Protein encoded by** the *argI* CDS with the **accession** of 'BSU40320'.

* How can the SpoOA protein, the master regulator of sporulation, be phosphorylated to trigger sporulation?

'SpoOA' is **phosphorylated by** the 'KinC' and 'SpoOB' **Protein**s. The 'SpoOB' **Protein** is **phosphorylated by** 'SpoOF' **Protein** which is further **phosphorylated by** the 'KinA' and 'KinB' **Protein**s.

Newcastle University

FLOWERS

Summary

* SyBiOnt to capture complex biological data for synthetic biology computationally
* Facilitates semantic reasoning
* Complementary to existing standards and ontologies
* Open access and available at **w3id.org/synbio/ont**
* Extendible with new terms

Newcastle University

I(C$_2$)S

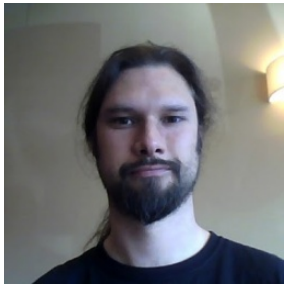
FLOWERS

Thanks

goksel.misirli@ncl.ac.uk
anil.wipat@ncl.ac.uk

Prof. Anil Wipat

Dr Jennifer Hallinan



Dr Matthew Pocock

Dr Phillip Lord

James McLaughlin

Prof. Herbert Sauro

Newcastle University
ICOS
NSF
EPSRC
Engineering and Physical Sciences Research Council
combine
FLOWERS