Setting the standard in synthetic biology

Adam Arkin

Standards for characterization, manufacture and sharing of information about modular biological devices may lead to a more efficient, predictable and design-driven genetic engineering science.

Although genetic engineering—the technical ability to edit DNA—has led to impressive biotechnology applications, these generally require many years of work and trial-and-error experiments to implement[1]. A concerted effort among synthetic biologists and allied fields might increase efficiency by developing rigorous characterization and manufacturing protocols linked to formal sharing of information and material through registries of biological parts and standard 'datasheets' (**Box 1**). In this issue, Endy and colleagues[2] present a case study demonstrating a possible datasheet for a biological part. Although they focus on a particular composite part—a genetically encoded cell-cell communication device—the authors' broader assertion is that there is a science to be developed concerned with the proper packaging and characterization of 'modular' biological activities so that these may be efficiently assembled into applications. If successful, this science would yield the profound benefits seen in other engineering sciences but not yet realized in the biological engineering community.

Engineers are fond of standards. A good device standard defines sufficient information about discrete parts to allow the design of predictable complex composite systems. It also provides guidelines for the minimal characterization and manufacturing tolerances of new elements. If suitably designed, a standard can also lead to the abstraction of a composite element's behavior into a few key functions and requirements, thereby greatly simplifying the design and analysis of the engineered system. If the abstractions are chosen just so, they may form a complete mathematical framework for design, as Boolean logic does in electronic engineering.

Datasheets are an embodiment of such engineering standards. They contain a formal set of context-dependent, input-output behaviors, tolerances, requirements, physical interconnect 'form factors' (the mechanical requirement for physical incorporation of the device into a system) and other details about a particular part or subsystem. This compact form enables engineers to rapidly select from a vast list the parts that will meet their design requirements. Adherence to the set of standards ensures that each device and the systems made from them will work as advertised.

Endy and colleagues[2] characterize a cell-cell communication receiver to demonstrate what it might take to create such standards. First, the form factor of the device is stated to be compliant with the BioBrick standard with which

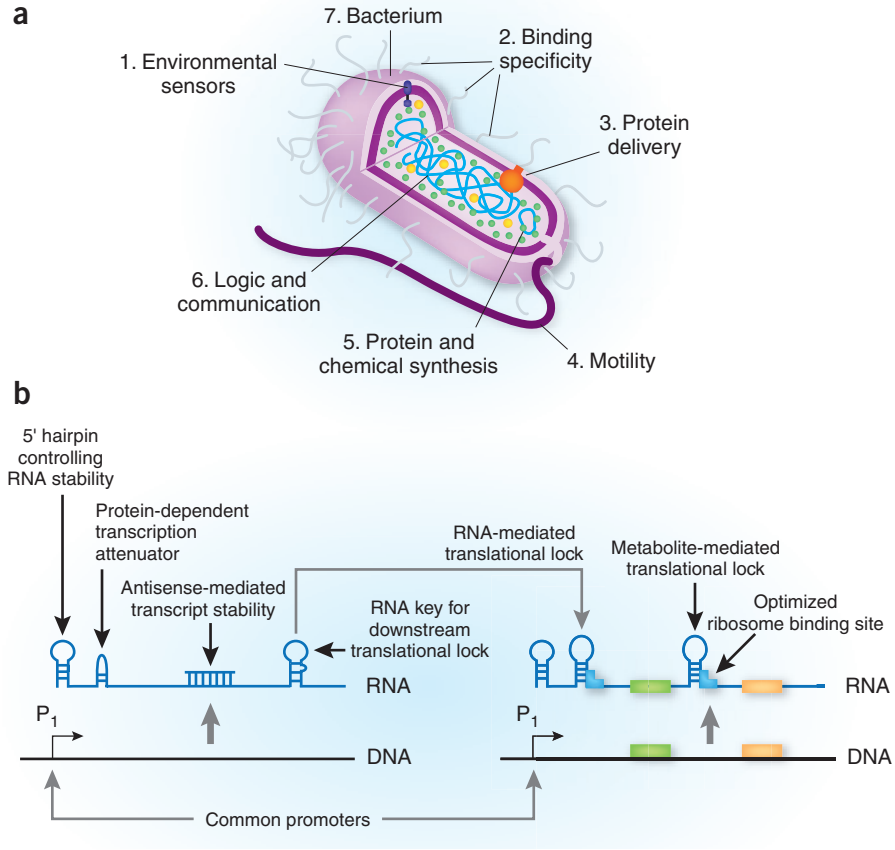

Figure 1 Examples of synthetic biological devices. (**a**) Different classes of a biological device. Tunable devices: 1 and 2. These devices perform functions whose particular features, such as specificity or affinity, may be changed with difficulty and only modestly. Once a datasheet is made for one member of a class, it can be applied to other members. Specific and/or complex devices: 3,4,5,7. These devices perform complex specialized functions used in one or a few variant forms in different applications. Their datasheets are necessarily non-standard and must be defined for each new device. Designable/scalable devices: 6. These devices are structured such that many new variants with new specificity or activity are easily designed (see **b**). Datasheets are reusable and become stable and mature quickly. (**b**) Gaining control over the central dogma with designable/scalable RNA devices. The diversity of known RNA structure–based mechanisms for regulating all aspects of gene expression and the emerging principles for altering their specificities suggest that it may be possible to develop designable, predictable and large families of parts to homogenize gene regulatory network design. The development of datasheet formats for each of the classes illustrated here would greatly aid the exchange of information about these parts and the assessment of their quality.

Adam Arkin is in the Department of Bioengineering, University of California Berkeley and the Physical Bioscience Division and the Virtual Institute of Microbial Stress and Survival at the Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Bldg. 977-257, Berkeley, California 94720, USA.

Endy is associated. BioBricks is a protocol for the relatively easy cloning and physical linking of biological parts together on a DNA strand. Parts adhering to this standard in the MIT Registry of Standard Biological Parts (http://parts.mit.edu/) follow a formal nomenclature. From its 'name', it may be inferred that the authors' part, BBa_F2620, is (i) compliant with the alpha release of the BioBricks repository (BBa); (ii) is of the 'F' or signaling class of function (there are thirteen such classes); and (iii) has an accession number of 2620.

Second, the authors define their device's input and output. The input is the chemical concentration of a particular homoserine lactone ($3OC_6HSL$), a type of small organic molecule produced in certain bacterial quorum-sensing systems[3]. The output is defined as PoPS (polymerases per second), or the number of RNA polymerases crossing a particular point in DNA per unit time. In this case, the output is from a specific promoter in the construct. The PoPS unit is also a standard of a sort, defining a 'common carrier' of transcriptional information (like current or voltage in electrical devices).

BBa_F2620 is a composite part made up of five other BioBrick components. These, in their order on the DNA strand, are: a TetR promoter, a particular ribosome binding site, the *luxR* gene (whose product is a transcription factor activated by the presence of certain HSLs), a transcriptional terminator and the LuxR-sensitive promoter. In cells lacking TetR, LuxR is constitutively expressed and will activate its promoter when the cell is exposed to certain HSLs. It is consistent with synthetic biology philosophy that one could characterize each of the subcomponents and from these predict the composite behavior of the overall device. It is equally consistent, as done by Endy and colleagues[2], to encapsulate all this internal function into a 'gray box' in which only the properly characterized input-output behavior is necessary for an engineer to know how to use this 'abstracted' composite component in a larger design. The authors cite many other groups that used this encapsulated device, BBa_F2620.

This philosophy (suitably hedged with caveats by the authors) raises the question of what proper characterization entails. Clearly, one should vary the inputs and measure the outputs. But should these be steady-state or dynamic measurements? What range of concentrations of input should be assayed? What methods will best quantify PoPS, and what precision is necessary? Should the response be measured during an exponential or stationary phase of growth? At 37 °C or 25 °C? In minimal media or rich? In *Escherichia coli* DH5α or MG1655? In a multicopy plasmid or integrated in the genome? Even for the obvious input-output behavior, the list of considerations goes on and is highly dependent on the final application. Synthetic biological devices reported in the past make somewhat arbitrary, if reasonable, choices about which of these to consider (see, e.g., ref. 4).

Endy and colleagues[2] also show that it might be important to measure device properties beyond those of the designed inputs and outputs. How responsive is the device to other members of the HSL family of signaling molecules? To what extent does operation of the device drain cellular resources and affect growth? What mechanisms of mutation inactivate the device, and how quickly do such mutants take over? The authors do an admirable job of implementing methods for measuring device performance for a wide array of these conditions, even using blunt instruments for measurement of expression such as green fluorescent protein. Their compact summary of results in their datasheet is impressive in its communication of these complex results, although there is much to be found in their supplementary information as well.

It is easy to throw stones. For example, the BBa_F2620 data sheet states that the device is "qualitatively" compatible in different hosts. This seems strangely noncommittal for a datasheet. How would an engineer use this beyond knowing there was work to be done in strains of *E. coli* other than MG1655? Why weren't key device properties, such as turn-off time, measured? How constrained were the parameters of the model by the data for the primary input $3OC_6HSL$? Was the mathematical form of the model appropriate for the other HSLs tested? Did each bacterium in the population show identical behavior, or was there single-cell heterogeneity? Should a stochastic model have been used to describe the device? There are challenges here in model selection and parameter estimation known to plague every engineering science. Whereas aspects of this problem in synthetic biology have been considered by other laboratories[5–8], the characterization of BBa_F2620 sets up perhaps the first nearly complete standard to which future attempts can be compared. However, unlike many other engineering disciplines, biology does not yet possess a theory of what the minimal information about a biological part should be. But, in the words of Voltaire, "The perfect is the enemy of the good."

Endy and colleagues' results[2] also underscore conceptual issues with defining datasheets for biological parts. There is much uncertainty about what affects the behavior of biological circuitry and systems. For example, what precisely differs between the *E. coli* strains MG1655 and DH5α that causes differences in BBa_F2620 function? What untested cellular functions might this device perturb? Synthetic biologists might control for such issues by agreeing to use and characterize devices in a number of common 'chassis' organisms, but what happens when we try to put different devices into the same strain?

Even carefully designed device interfaces can yield unpredictable interactions. Imagine a PoPS-out device whose polymerases exit an open reading frame and a PoPS-in device that starts with a ribosome binding site followed by an open reading frame. The resulting multicistronic transcript formed by the composition could, for example, yield new RNA structures that affect both the expression of the upstream gene and the rate of polymerase read-through to the downstream gene. There are also likely to be parasitic and unpredictable interactions among components as well as with the host. It is possible, for example, that introduction of a particular device in a design will drain necessary common resources (such as ribosomes or transcription factors) from another device. There also may be unpredicted interactions among component and/or host molecules. Or a new device might place the cell in a stressed state that affects both growth and mutation rates of other devices.

In addition to the challenges posed by unpredictability, the other key challenge to standardization is the sheer heterogeneity of biological device types. There are elementary types of parts, such as DNA-binding protein domains[9], and extremely complex 'composite' parts, such as type III protein secretion systems[10]. Clearly, these represent very different categories of function whose properties require distinct types of experiments to characterize their behaviors, tolerances and compatibilities (**Fig. 1a** and **Box 1**). Even DNA-binding domains may belong to different protein families, each with a different set of key properties to measure (e.g., the ability to be fused with transcriptional activation domains).

It may seem as if there is such an infinite number of functions arrived at by evolution that there is no hope for standardizing their characterization. However, though large, the space of different types of elementary functions is finite and has been assembled in a limited number of ways to create the variety of organisms we see today. There are modules of function and evolvable structures of proteins and circuits that are shared, tuned and rewired across and within organisms to create new behaviors. This suggests that evolution has perhaps arrived at a tunable basis set of parts from which new complex organismal function can be rapidly evolved[11,12]. It seems we can exploit this for our own designs.

Some of these tunable functions, such as ribosome binding sites[13], riboswitches[14], eukaryotic protein interaction scaffolds[15] and zinc-finger

Box 1 What is in a datasheet?

Synthetic biology aims to create the standards, abstractions and protocols to make design and manufacturing of new biological function inexpensive, efficient, predictable and reliable. The standards seek to define modules of biological function both with regard to how parts are physically linked and which of their behaviors must be characterized to enable a designer to predict how they will function as a group. Datasheets are compact, prescribed formats for formally communicating this information. For every biological device, there will be certain information common to all devices (top) and data and protocols specific to the particular device in question (examples below).

Generic datasheet format for a biological device

Part-UID: an accession number for a registry
Name: a name compliant with a standard nomenclature
Brief description of function: a few paragraphs describing the device's key behaviors
Description of use and significance: a short narrative on the uses conceived for the device
Notes on usage: context dependencies, compatibilities, growth phase and media requirements, etc.
References: publications on the device and its use in larger systems
Authorship: information about the creators of the device
Declaration of intellectual property: information on patents and licenses associated with the device
Safety class: what sort of lab can use the device
Sequence: FASTA sequence
Packaging type: protocol for physical linkage (e.g., BioBrick version alpha)
Annotated sequence: the functions and other part-UIDs that make up the device
Data: any measurements on the device (see examples below)
  Property measured: one sentence or less
  Chassis ID: the host used for testing the system
  Vector ID: the location of the device in the host's genetic material
  Property description: what is measured and why (1–2 paragraphs)
  Protocol used: all information needed to understand the measurement
  Measurement data: data file including author information (may be different from above), the data format and the data

Possible cell-cell communication measurements (after Canton *et al*[2].)

- Steady-state and dynamic induction curves of the output promoter by different homoserine lactone inputs in different cells
- Reliability over time (mutational inactivation rate)
- Homogeneity of induction in members of the population
- Effect of induction on cellular growth rate
- Effect of growth phase on above function of the device
- Reference to DNA-Binding Protein Domain datasheet for element of current device, LuxR

Possible DNA-binding protein domain measurements

- Crystal structure
- Binding constants for different DNA sequences
- Toxicity of overexpression in different cells
- Stability in different cells
- Composability with different transcriptional activation domains

Possible therapeutic bacterium measurements

- Survival in different hosts
- Tissue localization in different hosts
- Cell-type targeting efficiency in different hosts
- Immune response of host to therapeutic organism
- Dose-efficacy curves of therapeutic organism
- Efficacy of safety measures
- Mutation rate

proteins[16] (whose functionality can be changed by engineering a few key sites), suggest that a careful choice of parts 'families' developed to support synthetic biological application would be very powerful. A family of parts is a set of devices derived from the same basic core structure in which each member has been slightly modified to vary a particular key property. Since members of the family are closely related they are likely to share physical mechanisms and therefore characterization protocols.

For example, it may be possible to find a small set of parts families that allow us to construct transcriptional-translational control circuits of any complexity at will. Designable RNA elements that control mRNA stability, transcription and translation have all been reported; there are examples of each that are sensitive to metabolites, proteins or other RNA in the cell[14]. Designing and characterizing large families of such RNA parts responsive to different inputs could be a large step in gaining control over the central dogma (**Fig. 1b**).

Such a basis set would allow each RNA logic gate to be transcribed from a separate copy of the same promoter, thus providing to each gate a homogeneous 'transcriptional power'. The physical basis for the functioning of these parts is largely governed by RNA-folding physics and in many cases may be designable by Watson-Crick base pairing. This part set may be as close as we can get to a scalable, physically homogeneous, computationally designable basis set of biological parts. The network effect gained by having many labs working simultaneously to standardize characterization and design of these parts is greatly facilitated by the use of datasheets and repositories like those proposed by Endy and colleagues[2]. Assessment of both data quality as well as efficacy of design and prediction tools are greatly enhanced by such central resources, as users of the main

biological databases and competitors in CASP (Critical Assesment of Techniques for Protein Structure Prediction) can attest.

Unavoidably, some devices will be nearly one-off characterizations. Complex multifactor systems such as type III secretion, flagellar biosynthesis or photosynthetic systems will require very specialized measurements for their characterization[10]. The existence of specialized parts is prevalent in other engineering systems. But the work of Endy and colleagues[2] and others in the community gives hope that there will be basis sets of parts that make scalable, predictable, reliable design of certain functions a reality for biological systems.

No standard, however mature, is set in stone. It must evolve with the development of a field and its technology. Some engineering fields have more formal and less mutable standards than others owing to the nature of their substrate and the uncertainties that plague their manufacture and deployment. Standards can be quite contentious things, especially when the principles of design and the predictability of manufacture are still in their infancy. Synthetic biology is in early gestation, although it is developing quickly. BBa_F2620 is built, for example, to comply with BioBricks version alpha, in which cutting and pasting together of parts is accomplished by particular restriction enzymes and ligation protocols. New protocols for efficient and automated cloning and assembly of synthetic biological parts are being continually developed. The ability to simply synthesize very large pieces of DNA quickly, cheaply and without error is rapidly improving, as are methods for integrating these large constructs into organisms. Whole viral and bacterial genomes have been constructed in one or a few lengths of synthetic DNA[17,18]. Further, our ability to measure the circuit behavior in cells, even at single-molecule resolution, is rapidly advancing. Thus, what constitutes satisfying standards for manufacturing and characterization is changing quickly as well. In the words attributed to Ken Olsen, the founder of Digital Equipment Corporation: "The nicest thing about standards is that there are so many of them to choose from."

Those of us with pressing practical or commercial applications of synthetic biology will certainly use whatever means necessary to create and optimize our systems and may feel that it is too early and burdensome to develop standards. But it is in our interests to contribute to this mission both because we are familiar with the practical need for and limitations of different proposed approaches and because we have the most to gain if the effort is successful. With their work, Endy and colleagues[2] have enunciated a challenge. However difficult and imperfect our standards may be, let's push this idea to its limits and see where it will take us.

This view is shared by many in the field and is a central thrust within the Synthetic Biology Engineering Research Center (SynBERC, http://www.synberc.org/), to which the authors of this paper, and I, belong. There may also be an opportunity for journals to foster this activity during a period when only a few specialize in the field. For example, a newly launched journal, *Synthetic Biology*, will be accepting datasheets in the spirit of Figure 1 of the paper[2] (and the examples in **Box 1** above). Authors will be requested to store experimental constructs in a public repository. (In the spirit of full disclosure, I am Editor-in-Chief of this journal.)

Such community repositories will yield the most benefit when synthetic-biology designs scale to systems requiring many interacting parts, thereby limiting the utility of even inspired tinkering to optimize function. Our planes and computer processors are made possible by sophisticated engineering programs that model characterized parts that are designed and manufactured to work together predictably. Although we cannot quite yet imagine what synthetic biological applications might require the numbers and quality of elements on which these advanced technological systems rely, it is economically and socially important that we improve the efficiency, reliability and predictability of our biological designs. Engineering cells for production of chemicals in a fermentor remains a key technical and economic challenge[1]. But there also exist critical applications beyond the bioreactor—in the environment, in agriculture and in medicine—for which it would be at least soothing to know that they could be engineered for dependable and safe function. Setting the standards—high standards—is a clear prerequisite.

1. Keasling, J.D. *ACS Chem. Biol.* **3**, 64–76 (2008).
2. Canton, B., Labno, A. & Endy, D. *Nat. Biotechnol.* **26**, 787–793 (2008).
3. Camilli, A. & Bassler, B.L. *Science* **311**, 1113–1116 (2006).
4. Anderson, J.C., Voigt, C.A. & Arkin, A.P. *Mol. Syst. Biol.* **3**, 133 (2007).
5. Batt, G., Yordanov, B., Weiss, R. & Belta, C. *Bioinformatics* **23**, 2415–2422 (2007).
6. Kim, P.M. & Tidor, B. *Genome Res.* **13**, 2391–2395 (2003).
7. Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S. & Elowitz, M.B. *Mol. Syst. Biol.* **3**, 143 (2007).
8. Arkin, A.P. & Fletcher, D.A. *Genome Biol.* **7**, 114 (2006).
9. Maerkl, S.J. & Quake, S.R. *Science* **315**, 233–237 (2007).
10. Temme, K. *et al. J. Mol. Biol.* **377**, 47–61 (2008).
11. Singh, A.H., Wolf, D.M., Wang, P. & Arkin, A.P. *Proc. Natl. Acad. Sci. USA* **105**, 7500–7505 (2008).
12. Wolf, D.M. & Arkin, A.P. *Curr. Opin. Microbiol.* **6**, 125–134 (2003).
13. Barrick, D. *et al. Nucleic Acids Res.* **22**, 1287–1295 (1994).
14. Win, M.N. & Smolke, C.D. *Biotechnol. Genet. Eng. Rev.* **24**, 311–346 (2007).
15. Dueber, J.E., Mirsky, E.A. & Lim, W.A. *Nat. Biotechnol.* **25**, 660–662 (2007).
16. Mandell, J.G. & Barbas, C.F. *Nucleic Acids Res.* **34**, W516–523 (2006).
17. Cello, J., Paul, A.V. & Wimmer, E. *Science* **297**, 1016–1018 (2002).
18. Gibson, D.G. *et al. Science* **319**, 1215–1220 (2008).

The long and short of carbon nanotube toxicity

Kostas Kostarelos

Toxicological and pharmacological studies suggest guidelines for the safe use of carbon nanotubes in medicine.

The unique physical, chemical and electronic properties of carbon nanotubes (CNTs) have generated much interest in their potential medical applications. Although most studies have assessed the pharmacological efficacy, stability and toxicity of CNTs *in vitro*[1], two recent reports, in the *Journal of Toxicological Sciences*[2] and *Nature Nanotechnology*[3], explore their carcinogenic risk *in vivo*. Notably, these studies reveal that CNTs delivered to the abdominal cavity of mice can induce a response resembling that associated with exposure to certain asbestos fibers. What is the significance of these findings for efforts to develop CNTs as delivery vehicles for therapeutic and diagnostic agents?

Carbon nanotubes are seamless cylindrical structures comprising single or multiple concentric graphene sheets. Applications of both single-walled nanotubes (SWNTs) and multi-walled nanotubes (MWNTs) have long been haunted by fears of toxicity because of their

Kostas Kostarelos is at the Nanomedicine Lab, Centre for Drug Delivery Research, The School of Pharmacy, University of London, 29-39 Brunswick Square, London, WC1N 1AX, UK.
e-mail: kostas.kostarelos@pharmacy.ac.uk