Pathway Synthesis using the Act Ontology

Saurabh Srivastava*
CS, UC Berkeley
saurabhs@cs.berkeley.edu

Jonathan Kotker
EECS, UC Berkeley
jkotker@eecs.berkeley.edu

Stephi Hamilton
Bioengineering, UC Berkeley
stephi@berkeley.edu

Paul Ruan
CS, UC Berkeley
paul.ruan@gmail.com

Jeff Tsui
CS, UC Berkeley
tsui.jeff@gmail.com

J. Christopher Anderson
Bioengineering, UC Berkeley
jcanderson@berkeley.edu

Rastislav Bodik
CS, UC Berkeley
bodik@cs.berkeley.edu

Sanjit A. Seshia
EECS, UC Berkeley
sseshia@eecs.berkeley.edu

ABSTRACT

We describe here the Act Ontology, a formalism for uniformly describing biochemical function, and its use in building an enzymatic pathway synthesizer. A formal description of biochemical function allows us to reason about it, and for the particular case of enzymes, this function allows us to build a synthesizer tool that given a target chemical can automatically infer the most likely pathway that leads to it. The pathway can include known as well as hypothetical enzymes with predicted function.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and genetics*

General Terms

Algorithms, Design, Standardization

Keywords

Biochemical formalism, enzymatic metabolic paths

1. INTRODUCTION

Synthetic biology is at the verge of a virtual overflow in sequence characterization and their availability for use in rationally designed genetic function. These designed genetic constructs, when inserted in a chassis such as *E. coli* (a bacteria) or *S. cerevisiae* (a yeast), impart desired function to the organism. While there will be a definite overabundance in the characterization of genetic material, there are as yet no formalisms in place to uniformly capture all that functional information to be used by computational tools.

Without a formal way of encapsulating that information the characterization data will remain computer-inaccessible and only available in human-readable tables and data sheets. Thus, design methodology will remain outside the purview of computational tools.

Towards remedying this situation, we propose the Act Ontology, which is a formal, uniform, and expressive mechanism for encoding biochemical function. We are developing the theoretical framework for specifying function, as well as a repository based on that formalism that will store biomolecular function.

Our current focus is on encoding enzyme function so that we can build a tool for automatically suggesting novel metabolic pathways to unnatural chemicals. The pathway synthesizer tool constructs pathways not only based on naturally-known reactions (for which there are many pre-existing tools), but also reactions that are inferred as plausible based on reaction operators. These operators are derived from abstracting from natural reactions and form an abstraction hierarchy that we intelligently traverse to derive pathways that have a high likelihood of success.

2. ACT ONTOLOGY

Act is a formalism for describing the molecular function of species. A specie is any entity that participates in a biochemical reaction. A genetic feature is a specific specie that encodes for and functions either a protein, a RNA, or in its DNA form itself. The central concept of Act is that of a family, akin to the Gene Ontology (GO [1]) concept of a family. In contrast to GO, however, the defining feature of Act families is not their location in the hierarchy of families, but the functional traits corresponding to a specie. In fact, in Act we do not even pre-specify the hierarchy, which can be inferred from containment of functional traits. As such, Act does not simply label families, but rather it provides a formal description of the species' chemical behavior according to a controlled vocabulary to support querying, synthesis, and verification.

Features are some of the most important species Act ascribes function to. Features are the DNA elements that

*Corresponding author.
This work was supported by the the DARPA Living Foundries initiative. SS is supported by NSF grant 1019343 to the Computing Research Association for the CIFellows Project. Additional support has been provided by the NSF Synthetic Biology Engineering Research Center (SynBERC).

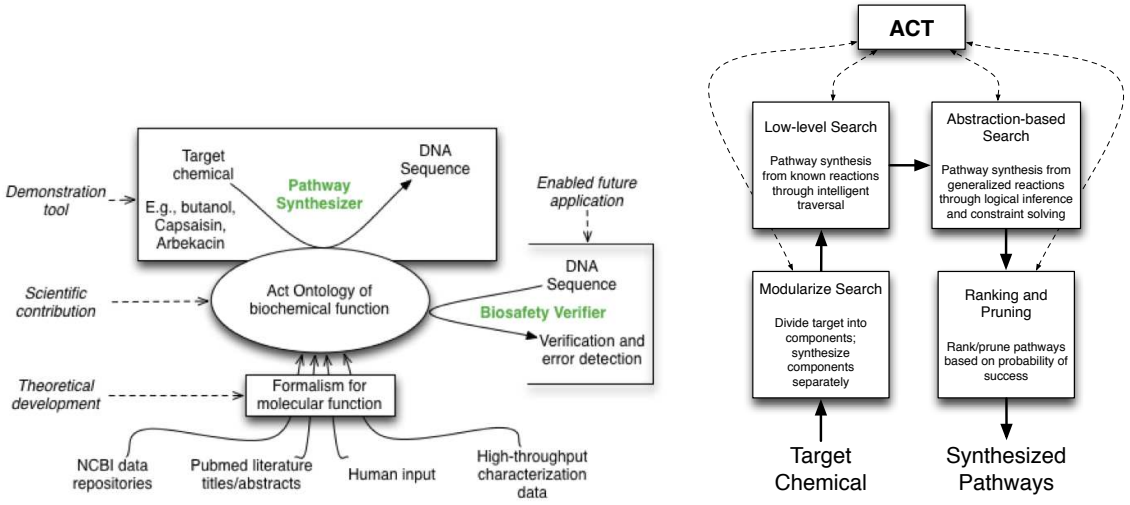

Figure 1: (a) Populating and using Act. (b) The architecture of the synthesizer.

directly encode a particular molecule, including functional DNAs, RNAs, and proteins. While a protein, RNA, DNA molecules may all come from the same sequence through translation, and transcription respectively, we distinguish them as different features because they have different functional characteristics in their various forms.

Every Act species is assigned a family, possibly more than one. An Act family encapsulates a unit of functional characteristic, e.g., the enzyme activity of a protein. The formal representation of a family is in terms of a finite state transducer, specifically a Mealy Machine [3]. The states correspond to states of existence of the species, e.g., native form, or bound to a small molecule etc. The input and output alphabet is the same and consist of the universe of species.

Formally, the Mealy machine for an Act family is the 6-tuple, (S, S0, Σ, $\wedge$, T, G), defined as: - A finite set of states S: the various states the molecule exists in.
- A start state S0 ($\in S$): The native state of the molecule.
- A finite input alphabet Σ, and output alphabet $\wedge$: The input and output are species and the empty symbol ϵ.
- A transition function ($T : S \times \Sigma \rightarrow S \times \wedge$) mapping pairs of a state and an input symbol to the corresponding next state and output symbol.

Currently, we populate Act by data repository and literature mining, but in the future Act will also get families from high-throughput characterization data, as shown in Figure 1.

3. ACT ENABLED PATHWAY SYNTHESIS

The architecture of the synthesizer and the role of the Act biochemical database is shown in Figure 1.

The pathway synthesizer has a very strict encoding of biosafety. We prohibit the synthesizer from exploring known harmful patterns in chemicals and known harmful families.

4. RELATED WORK

The GO and SO [2] ontologies are well known previous attempts at categorizing biochemical features into a functional hierarchy. The hierarchy of organization is the main contribution of these ontologies and provides the correlation between function. On the other hand, in Act, families are not defined by any hierarchy, but instead through their internal traits. A hierarchy can be inferred if so desired by checking a family pair whether one contains all the traits of the other.

The reaction operators that form the basis of our synthesizer are related to KEGG RCLASSes and BNICE operators. While those are manually authored and curated, in Act they are inferred based on chemical, biochemical, and chemoinformatics theory, and therefore scalable even with large amounts of fine-grained high-throughput data.

5. CONCLUSIONS

We have briefly described the Act Ontology, whose aim is to formally describe and encapsulate biochemical function of biomolecules. We also presented the encoding of enzyme function within Act, and its use in designing a synthesizer tool that automatically infers plausible metabolic pathways for unnatural target chemicals. These rationally designed new pathways consisting of natural and speculated enzymes, if inserted into *E. coli* or another chassis will allow the organism to produce the target chemical starting from its primary metabolites.

6. REFERENCES

[1] M. Ashburner, C. A. Ball, J. A. Blake, and et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1), May 2000.

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

[3] G. H. Mealy. A method for synthesizing sequential circuits. *Bell System Tech. Jour.*, 34(5):1045–1079, 1955.