In silico cell working group

Stanford 2022

Who was here:

- Eran Agmon agmon.eran@gmail.com
- Jen Hofmann hofmannj@stanford.edu
- Alp Sunol asunol@stanford.edu
- Diego Alba dalba@uw.edu
- Ayush Pandey apandey@caltech.edu
- Ryan Cardiff cardiffr@uw.edu
- Richard M Murray murrayrm@cds.caltech.edu

Brainstorming hackathon

- Where/when?
  - Feb/Mar 2023: winter/warm standalone
  - May: EBRC 2023
  - May: Syncell/BuildAcell 2023
  - June: SEED 2023
  - Oct: IWBDA 2023
  - NIST-organized? (Elizabeth)
    - What measurements?
  - HARMONY/COMBINE. Propose a workshop.
  -
- Logistics
  - 3 days?
    - First day to arrive, outline, delegate tasks.
    - First evening to prepare data
    - 2nd day – hack it together
    - 3rd day – finalize and present?
  - Travel?
  - Experts
    - Present some data
    - Or integrate them with working groups.
- Funding
  - Build-a-cell (ask kate)
  - Center for Reproducible Biomodeling
- Bring prepared components. What components?
  - Pruned down E. coli model (Eran)
  - Simulation and parameter id (Ayush)
  - TXTL and/or PURE CRN (Ayush)
  - Biophysics with different granularity (Jen)
    - dynamics/spatial organization of plasmids (and nucleoid?) (Alp)
  - Plasmid model. Plasmid counts?

    - Gene circuits. How do proteins interact with metabolism? (Ryan)
        - RNA folding, multiple RNAs, competition (Diego)
        - Metabolic Dynamics in Escherichia coli-Based Cell-Free Systems
    - Regulation of ribosome counts. Growth rate control. (Leron)
    - Transport of waste products/nutrients to keep the lifetime going.
    - Multiple Liposomes merging
    - Environments, changing O2, how does this change the redox balance? (Leron)
- Outcomes/Application-focus (Ask integration group)
    - What environment do we put the synthetic cell in?
    - What does the synthetic cell DO in this environment?
        - Medical?
        - Sustainability?
        - Specific protein expression?
    - Can we extend the TXTL system's lifetime?
        - Balance in connection with metabolism
    - Self-replication
    - Sensors/ 2-component signaling
    - More general: **integrated model. Come in with a model component, come out with your component combined with other work.**
- Visualization for interrogation of model.
    - Reuse Causality Network
- Have a virtual meeting before the hackathon to prepare. Demo Vivarium and other modeling tools
- Invite experts
    - Liposomes
        - Experimentalist: K. Adamala
        - Programmable Fusion and Differentiation of Synthetic Minimal Cells
        - Confinement Modeling: R. Zia
        - Experimentalist: N Kamat
    - Spatial organization
        - Neal Devaraj – UCSD
    - Plasmids
        - How interaction between plasmids changes expression. (Mark Styczinski)
    - TXTL
        - Concentrations of DNA/polymerases (Vincent)
        - Quantitative modeling of transcription and translation of an all-E. coli cell-free system
        - Characterization of the all-E. coli transcription-translation system myTXTL by mass spectrometry
    - Metabolism
        - M. Covert
        - Paul Freemont (Imperial College)
    - Sensors and signaling
    - Environment. Gut microbiome. Epithelial models that couple to synthetic cells
- How to integrate w/ integration group
    - Align our goals.
    - How to integrate with experimental work?
    - What are their applications?

        - Metabolic engineering
        - characterizing components
    - How might models inform what they are doing?
    - What data do they have that could inform a model?

Goals(?):

1. Establish strong correspondence between model and experimental observations
2. Improve reproducibility of the individual modules
3. Integrate modules at both the experimental and modeling levels
4. *Improve lifetime, production capacity, productivity, etc. of system*

Caltech, 2022

Participant names (emails):

VivariumPURE repository: https://github.com/BuildACell/VivariumPURE

NIST, 2020

Participant names (emails): Akshay Maheshwari (akshaym@stanford.edu), William Poole (wpoole@caltech.edu), Miroslav Gasparek (miroslav.gasparek@eng.ox.ac.uk), Alysse DeFoe (alyssedefoe2021@u.northwestern.edu), Ankita Roychoudhury (aroychou@caltech.edu), Eran Agmon (agmon.eran@gmail.com), Ryan Spangler (ryan.spangler@gmail.com), Christian Zapata-Sanin, Jon Calles (jecalles@stanford.edu), Nadanai Laohakunakorn (nadanai.laohakunakorn@ed.ac.uk), Eric Wei (ewei2@stanford.edu), Sachi Laumas, Hannah Verdonk

Executive summary:

We propose developing a standard for computational models of biology that allow these modules to be composed in a systematic manner. We will use the PURE system as a motivating example of a biological system that may use multiple interconnected modules to fully model the system. We will start by attempting to integrate models representing two biological functions: translation elongation & energy regeneration. We will find and build simple models for both biological functions and integrate them with a model composer. We will simultaneously develop benchmark experimental datasets. From these experiences, we will propose a general schema for models built by the Build-a-Cell community such that any model following this schema will be composable.

This was a combined effort w/ PURE makes PURE & In silico cell subgroups.

Outline:

1. Build a list of (biological functional?) modules that can be implemented using PURE.
2. Describe the utility of each module, and how it can be modelled?
3. Integration platform -- Describe how modules could connect with each other? In other words, can the modules in (1) be classified into "super-modules" ie higher levels of abstraction. For example: "Metabolism", "Polymerization"
4. What questions about PURE can a modeling framework answer? For example, where does PURE break? What does this imply for modeling?
5. Think about a visualization framework for the resulting architecture

Let's work here!

Schematics

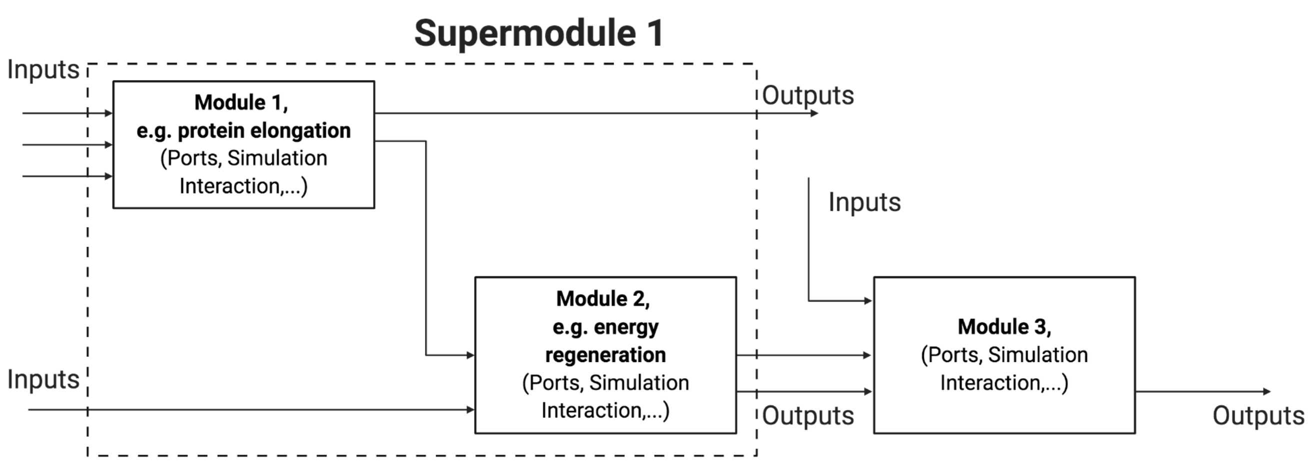

Figure 1. The idea of Cell Composer.

(1) Build a list of (biological functional?) modules that can be implemented using PURE.

Reference(s) for Models of PURE:

- Reaction dynamics analysis of a reconstituted Escherichia coli protein translation system by computational modeling - Matsuura et al. 2017
- Model-based inference of gene expression dynamics from sequence information - Arnold et al. 2005 (not PURE but has good descriptions of all TX-TL mechanisms)
- Experimentally validated model enables debottlenecking of in vitro protein synthesis and identifies a control shift under in vivo conditions - Niess et al. 2017 (updated version of Arnold model)
- A simple protein synthesis model for the PURE system operation - Mavelli et al. 2015
- Modelling cell-free RNA and protein synthesis with minimal systems - Doerr et al. 2019
- Experiment and mathematical modeling of gene expression dynamics in a cell-free system - Stögbauer et al. 2012.

List of Modules (and submodules) in PURE

1. Transcription
    a. Initiation, Elongation, Termination (trivial in PURE with T7)
2. Translation
    a. Initiation, Elongation, Termination (initiation and elongation are complex)
3. tRNA regeneration
    a. tRNA charging
    b. Ternary complex formation? (linked to TL elongation)
4. NTP Regeneration
5. Liposomes

List of Super Modules in PURE

1. Biophysics: [salt], pH, Temperature effects
2. ContactManager

List of PURE extension modules

1. Phosphate waste control
2. Encapsulation
3. Energy production (e.g., photosynthesis, catabolism)
4. Chemostat

See table below.

(2) Describe the utility of each module, and how it can be modeled?

-

(3) How do models of different processes talk to each other and integrate together?

- Specify the contact points (ports) between model modules

- Endow modules with the ability to “know” its state at each particular time and update itself
- General “structure” for the modules with:
    - Ports: default read + write, default to combinatorial datastream w/ description for folding into simple datastream
        - Inputs
        - Outputs
        - States
        - parameters,...
    - Simulation Interaction:
        - update rules
        - Interrupt port

Case study: protein elongation + energy regeneration modules

- Protein elongation
    - Ports:
        - mRNA
        - Ribosomes
        - Energy carrier
        - Ternary complexes
        - EF-G
        - EF-Ts?
- Energy regeneration
    - Ports:
        - Energy carrier (nGTP, nATP, nTTP, nUTP, nCTP, PPi, Pi)

Contact managers?

Contact modules?

Can’t fill out on a case by case basis. Abstract specification needed. Be able to do without (too much) coding. Wouldn’t have to code up every interaction between every module

Point of contact: RNA - these may be combinatorial objects e.g. RNA in many different states.

More generally: model modules are connected by contact modules. Model modules are like functions with inputs and outputs. A contact module maps each input and output of two model modules to one or more state-variables of the other module. In other words, an input “mRNA” into an mRNA degradation module can be mapped to many mRNA species/states that come out of a translation model (e.g. mRNAs with different ribosomal occupancy). This mapping forms a contact module specification.

(4) Describe how modules could connect with each other? In other words, can the modules in (1) be classified into "super-modules" ie higher levels of abstraction. For example: "Metabolism", "Polymerization"

Polysome that is being degraded by an endonuclease. How do we deal with this contact between translation and degradation?

(5) What questions about PURE can a modeling framework answer? For example, where does PURE break? What does this imply for modeling?

- How can PURE maximize GFP production?
- Sources of failure. When are substrates missing vs. molecules are degrading limiting? Depends on initial conditions and the proteins being produced.
- Long term: PURE makes PURE

(6) TODOs (next 6 months): We will be working off this new [CellComposer](https://github.com/BuildACell/CellComposer)repository!

- Case Study: protein elongation and energy regeneration in PURE GFP production.
- Build model registry: collect models of individual subsystems in a registry (which ones? The ones that are best described)
  - From papers! SBML models. Need references of published models -- ideally open source code & relevant to PURE.
  - Needs lots of people!
- Connect the models by building a contact module. This should result in combinatoric combinations of models.
  - Which combinations work? Which do not?
- MVP Benchmark data set
- Curate the required parameters as best we can (order of magnitude estimates are okay for now!).
- This can be used to assess the different mechanisms and aid in model selection. For example, predict variations in components of elongation or GFP production.
- Perform mock model selection to validate our ideas.
- Github repo and make pushes. cell-composer!

1. Make cell-composer a reality: https://github.com/BuildACell/CellComposer

   → William, Jon, Eran, Ryan, Christian Z.

2. Gather lots of models to put into cell-composer

   → Miro, Akshay, Nadanai, Eric, Ankita

3. Build MVP Benchmark dataset

   → Akshay, Nadanai, Eric

Motto: “Kittens running through a keyboard to generate fake experimental data” (W. Poole)

Figure 2. The proposed experimental method of collecting data.

List of molecules in 1fL of PURE

Molecule Species	Module(s)	Abundance	Radius	Volume fraction
Initiation factor 1	translation	7320.413019	1.32261563	7.09455E-05
Initiation factor 2	translation	3023.163764	2.81932376	0.000283782
Initiation factor 3	translation	2932.833767	1.7941126	7.09455E-05
Elongation factor G	translation	3884.435766	2.79356984	0.000354728
Elongation factor Tu	translation	13927.46618	2.29961328	0.000709455
Elongation factor Ts	translation	9909.244335	2.04450731	0.000354728
Release Factor 1	translation	1487.861622	2.24952763	7.09455E-05
Release factor 3	translation	1011.789535	2.55808783	7.09455E-05
Ribosome recycling factor	translation	2922.186962	1.79628887	7.09455E-05
Alanyl-tRNA synthetase	tRNA regeneration	4330.358397	2.99953497	0.000489524
Arginine-tRNA synthetase	tRNA regeneration	186.3709088	2.6292354	1.41891E-05

Asparginyl-tRNA synthetase	**tRNA regeneration**	**2522.571822**	**2.45360827**	**0.00015608**
Aspartate-tRNA synthetase	**tRNA regeneration**	**731.5618949**	**2.64580896**	**5.67564E-05**
Cysteinyl-tRNA synthetase	**tRNA regeneration**	**138.5678546**	**2.44785167**	**8.51347E-06**
Glutaminyl-tRNA synthetase	**tRNA regeneration**	**360.8274748**	**2.61280408**	**2.69593E-05**
Glutamyl-tRNA synthetase	**tRNA regeneration**	**1411.306766**	**2.47283546**	**8.93914E-05**
Glycl-tRNA synthetase	**tRNA regeneration**	**753.2701705**	**2.78431906**	**6.81077E-05**
Histidyl-tRNA synthetase	**tRNA regeneration**	**102.5413764**	**2.36415417**	**5.67564E-06**
Isoleucyl-tRNA synthetase	**tRNA regeneration**	**2311.430092**	**3.08322526**	**0.000283782**
Leucyl-tRNA synthetase	**tRNA regeneration**	**247.9380927**	**3.01197312**	**2.83782E-05**
Lysl-tRNA synthetase	**tRNA regeneration**	**669.7025874**	**2.52955918**	**4.54052E-05**
Methionyl-tRNA synthetase	**tRNA regeneration**	**165.9865508**	**2.77754419**	**1.48986E-05**
Phenylalanyl-tRNA synthetase	**tRNA regeneration**	**1172.596596**	**2.90655766**	**0.000120607**
Prolyl-tRNA synthetase	**tRNA regeneration**	**946.3392794**	**2.61575195**	**7.09455E-05**
Seryl-tRNA synthetase	**tRNA regeneration**	**236.569347**	**2.38713633**	**1.34797E-05**
Threonyl-tRNA synthetase	**tRNA regeneration**	**513.0387912**	**2.75006079**	**4.46957E-05**
Tryptophanyl-tRNA synthetase	**tRNA regeneration**	**177.1363025**	**2.19099078**	**7.80401E-06**
Tyrosyl-tRNA synthetase	**tRNA regeneration**	**76.10093164**	**2.37246204**	**4.25673E-06**
Valyl-tRNA synthetase	**tRNA regeneration**	**100.2685442**	**3.12115098**	**1.27702E-05**
Methionyl-tRNA formyltransferase	**tRNA regeneration**	**3528.97953**	**2.12522932**	**0.000141891**
Creatine Kinase	**NTP Regeneration**	**297.382716**	**2.83482945**	**2.83782E-05**
Myokinase	**NTP Regeneration**	**564.5625**	**2.08008902**	**2.12837E-05**
Nucleoside-diphosphate kinase	**NTP Regeneration**	**428.3903512**	**1.63228697**	**7.80401E-06**
pyrophosphatase	**Phosphate waste control**	**274.9771689**	**1.83307387**	**7.09455E-06**
T7 RNA polymerase	**transcription**	**60.82828283**	**3.03093759**	**7.09455E-06**
tRNAs	**tRNA regeneration, translation**	**90132.69382**	**3**	**0.010193768**
Ribosomes	**translation**	**7226.4**	**13**	**0.066502912**
NTPs	**Transcription,**	**2mM/4mM**		

	translation, NTP regeneration, tRNA regeneration	
Hepes-KOH, pH=7.6	**pH buffering**	**100 mM**
Creatine phosphate	**NTP regeneration**	**40mM**
Potassium glutamate	**electrostatics**	**200 mM**
Magnesium acetate	**electrostatics**	**26 mM**
Spermidine	**Ribosome stability**	**4 mM**
DTT	**Antioxidant**	**2 mM**
10-formyl-5,6,7,8-tetrahydrofolic acid (formyl donor)		**20µg/mL**
20x amino acids	**tRNA regeneration**	**0.2 mM**
Ternary Complexes (e.g. Bound Molecules)	**Electrostatics, everything**	

Nadanai: Increase phosphate in PURE → reduces magnesium concentration . When reduce Mg concentration -> ribosomes fall apart. How can we model pH/ion concentrations? One mode of failure for PURE.

		Red means that they have already been shown to work in vitro and within liposomes		
Main modules	Sub modules	Intended functionality	Gene Candidates	Specific function
Transcription	mRNA production	Generates mRNA from template DNA	T7 RNAP	Binds the T7 promoter and expresses mRNA based on template DNA
	tRNA production	Generates required tRNAs	T7 RNAP	Binds the T7 promoter and expresses mRNA based on template DNA
			tRNA maturation proteins?	Modify tRNAs to allow for proper maturation

	mRNA stability	Unwinds mRNA secondary structure	HrpA	ATP-Dependent helicase that unwinds mRNA secondary structure
Translation	Ribosome	Translates mRNA into proteins	Many proteins...	
	Translation factors	Aid the ribosome during translation	Initiation Factor 1	Not fully understood but it is suggested that it binds the 30S ribosome in the A site preventing aminoacyl-tRNA from entering
			Initiation Factor 2	IF2 is implicated in fMet-tRNA_fMet binding to 30S ribosomes and possesses a ribosome-dependent GTPase activity.
			Initiation Factor 3	IF3 discriminates against noninitiator tRNAs by selectively destabilizing other ternary 30S–tRNA–mRNA complexes
			Elongation Factor G	Facilitates the translocation of the ribosome by one codon along the mRNA molecule. The activity requires GTP hydrolysis
			Elongation Factor T (unstable)	In its GTP-bound (active) form, EF-Tu binds aminoacylated tRNAs. At the decoding site of the ribosome, this ternary complex is "tested" for a codon-anticodon match; if the proper aminoacyl-tRNA has been found, GTP is hydrolyzed and EF-Tu and GDP dissociate from the ribosome, while the aminoacyl-tRNA remains bound to the ribosome
			Elongation Factor T (stable)	EF-Ts interacts with aminoacyl-tRNA-bound EF-Tu and regulates its affinity for GTP and aa-tRNA ligands. EF-Ts may regulate the stability of the EF-Tu-GTP-aa-tRNA ternary complex by facilitating the conformational changes within EF-Tu that are responsible for binding and release of aa-tRNAs
			Release Factor 1	Release factor 1 (RF1) is one of two class 1 codon-specific factors in E. coli that facilitate the release of the growing polypeptide chain at stop codons. RF1 recognizes the termination codons UAG and UAA

			Release Factor 3	Release factor 3 (RF3) is a ribosome-dependent GTPase that stimulates the release of RF1 and RF2 from the ribosome after peptide chain termination. The action requires nucleotide exchange and hydrolysis of GTP
			Ribosome Recycling Factor	Ribosome recycling factor (RRF) recycles the ribosome upon translation termination. RRF, release factor RF3, and elongation factor EF-G are involved in this recycling process. At termination, RRF and EF-G catalyze release of the 50S ribosomal subunit from the 70S complex, a GTPase-dependent process; EF-G acts to release RRF from the ribosome. Release factor RF1 inhibits RRF activity. RRF with EF-2 and RF3 increase dissociation of peptidyl-tRNA from the ribosome.
	Protein folding	Allow for proper protein folding and localization	Trigger Factor	Associates with nascent polypeptide chains that have exposed hydorphobic regions to inhibit aggregation and misfolding
			dnaK?	Another protein chaperone that is typically used for longer proteins
			groEL?	Another chaperone that helps coordinating folds that sequester hydrophobic regions
	Membrane integration	Allows for proper membrane protein integration into the liposome membrane	SecYEG	Transmembrane complex known as a translocon that is responsible for the translocation of proteins targeted by the SR/SRP pathway
	tRNA loading	Charges tRNA with amino acids	All 20 amino-acyl tRNA synthetases	Attaches amino acids to their specific tRNA requiring the conversion of ATP to AMP
	tRNA Maturation	Matures the tRNA transcripts into functional tRNAs	RNase P (rnpB-Catalytic RNA, rnpA-protein component that	The RNase P protein-RNA complex is responsible for the 5' cleavage of the extended 5' end of tRNAs. If all of the tRNAs are monocistrons then RNase P is the only factor needed to mature the 5'

			stabilizes the RNA at physiological conditions)	end
Metabolism	ATP recharging	Replenish the ATP pool	Creatine Kinase	Uses creatine phosphate to recharge ADP, converting it to ATP
			alpha-Hemolysin	Generates pores that allow for nutirent transport into the cell, including ATP, Mg, etc.
			ATP/ADP translocase	Takes in extracellular ATP in exchange for intracelluar ADP
			Glycolysis	Includes all of the glycolytic proteins required to convert glucose to pyruvate or maybe have it undergo fermentation instead?
	NTP recycling	Convert ATP into all of the NTPs and recycle AMP	Myokinase (adenylate kinase)	Uses one ATP to and one AMP to create two ADP
			Nucleoside-diphosphate Kinase	Converts ATP into the other three NTPs
			Inorganic Pyrophosphotase	Breaks pyrophosphate into two inorganic phosphates
	AA replenishment	Replenish the pool of amino acids either through synthesis or import		
	Nucleotide replenishment	Replenish the pool of nucleotides either through synthesis or import		

Reagent Components			
Name	Function	Reference	Quantity (mM)
Magnesium acetate	Supplies magnesium for ppa, tRNA ligases, fmt, adk, ndk, etc.	ecocyc	9
Potassium phosphate	Supplies potassium for fmt	ecocyc	5
Potassium glutamate	Not sure but held at very intracellular concentrations		95
Ammonium chloride	Proper ion concentrations?		5
Calcium chloride	Proper ion concentrations?		0.5
Spermidine	Inhibits transient ribosomal subunit dissociation	Umekage and Ueda 2006	1
Putrescine	Polyamine, which is a cation and complexes with RNA, precursor to spermidine	Umekage and Ueda 2006	8
Dithiothreitol (DTT)	Reducing agent which helps for disulfide bonds	Wikipedia	1
ATP	Used in aminoacylation and incorporated in mRNA	Wikipedia	2
GTP	Used by EF-Tu to localize tRNA to ribosomes, RF-3 to dissociate the ribosomal subunits, and is incorporated into mRNA	Wikipedia	2
CTP	Incorporated into mRNA	Wikipedia	1
UTP	Incorporated into mRNA	Wikipedia	1
Creatine phosphate	Serves as the phosphate donor to replenish ATP	Wikipedia	10
tRNA mix	Provides all of the tRNA to allow for translation	Wikipedia	2.8 A260
10-formyl-5, 6, 7, 8-tetrahydrofolic acid	Substrate for methionyl-tRNA formyltransferase	Wikipedia	0.5 ug
Every amino acid	Used to generate polypeptides	Wikipedia	0.1

NASA, 2020

MMMC: Maximally modellable minimal cell

What more needs to be done to construct a maximally modelable minimal cell? Modern knowledge of cell biology spans many scales, from the atomic structure of single molecules to the evolution and growth of populations. How much of this infinite knowledge space must we represent to enable engineers to design and operate synthetic cells? In our working group, we discussed **1) biological processes that must be represented to build a modelable cell, as well as the extent to which they must be modeled 2) The scales to which we must model these essential biological processes 3) applications of a modelable cell; 4) what more we need from you (and ourselves).**

1) **Essential biological process models:**

We've enumerated biological processes that we believe are general to all cells and that we must be able to model (e.g., genetic information storage and expression). These are listed in a set order, in which the first process will be modeled, tested, and refined, before integrating the next process model. We've also highlighted gaps in the modeling capacity of this group, as well as of the larger scientific community. The following examples represent our proposed "next step" models:

- Transcription/translation
- Metabolism, extra/intracellular transport (active and passive)
- Gene regulation, DNA replication
- Maintenance of cellular structures
- Spatial organization of macromolecules within the cell, cell shape & motility, cell division.

2) **Scales:**

We note that the models needed to understand a modelable cell span multiple scales, both in terms of physical space as well as biological complexity. In order to understand such a modelable cell, we need to integrate these models. Physical scales can be modeled in increasing levels of details, e.g. ranging from atomistic to colloidal dynamics all the way down to cell colonies. There are differing conditions where all-atom modeling may be required but it is generally accepted that 'legitimate' coarse-grain approaches should be used wherever possible - determining the difference is vital.

- **Subcellular dynamics** includes protein movement & interactions at both short and long time scales and at medium spatial resolution, with whole-cell properties such as cell shape, homeostasis, and robustness as *emergent properties*. The relevant modelling methods include all-atom or coarse-grained (e.g. molecular dynamics) and colloidal-scale models (e.g. Brownian & Stokesian dynamics).

- **Single cell network level** includes the kinetics of sensory, metabolic, signaling, regulatory, and protein-protein interaction networks. Relevant modeling methods include PDEs, FBA, Gillespie algorithm, graphical models, Markov processes
- **Population-level dynamics** describes the cell-cell interactions, population response to the external stress/stimuli, evolution, and processes at the longer time scale. Relevant methods include agent-based models, PDEs

3) Applications:

We anticipate that a "maximally modelable minimal cell" will give us insights into the following basic science and translational applications: genetic and metabolic circuit engineering, ecological modeling, cellular controls, novel (lineage-agnostic) organisms with new biochemistry and physical behaviors, possible insight into the origins of life, drug discovery (via "celluloid" models). We also anticipate that such models will enable outreach and education, and applying living organization to social or political systems, and bioethical issues such as disease-agent modeling.

4) What we need/are doing:

We've identified the following immediate action items towards building a maximally modelable minimal cell. What we need from everyone:

1. Benchmark datasets across scales, as well as associated "Turing tests" of model viability across scales (e.g. NIST for MNIST)
2. Development of the working models for cellular subsystems
3. Integration platform for subsystems across physical & time scales as described in Section 2).
4. Visualization schemes for developed minimum cell models.
5. Publicity and Outreach, including creating a build-a-cell competition to engage the general population & get feedback into what's actually needed (e.g., iGEM, DARPA Grand Challenge for self-driving cars)
6. Recruiting talent for the above (Experts needed for modeling different subsystems)
7. Resources & Finances for the above ($$)

What we're doing ourselves: Prototyping all of the above by next build a cell meeting. Meeting once a month to discuss progress.

Akshay Maheshwari: (1) Transcription/Translation experiments using PURE; come up with Turing test for transcription/translation at atomic, colloidal, whole-cell, and population scale. (2) Colloidal & whole-cell scale model of translation. (4) atomic-resolution + colloidal-resoluton snapshot of a minimal cell -- can be used for proposal (7) Submit this proposal.

Ryan Spangler: Integration platform: https://github.com/vivarium-collective/vivarium-core (3), Development of working models (2)

Eran Agmon: (3) Integration platform with documentation and prototype, (2) develop configurable network-level models of metabolism, transport, (2) a flagella-based model of motility.

Jen Hofmann: (1) Benchmark dataset (including particle tracking to correlate with dynamic simulation methods, coarse-grained macromolecular interactions for the colloidal level; come up with Turing tests for above)

Alp Sunol: Development of working models of cellular subsystems - nucleoid and/or active flows (2)

Hannah Verdonk: identifying benchmark datasets (literature search for ideal data) (1), preliminary model validation (make models from published datasets found in literature search/validate published models) (2)

Emma Gonzalez: Benchmark dataset: standardize measurements for intracellular transport relevant for colloidal dynamics (diffusivity, viscosity, …)

Miroslav Gasparek:

(2) Review the existing literature on the modeling of cell-free systems. Try to computationally find the optimal energy substrate/cell extract component inputs to cell free reactions that will allow for it to be sustained for extended period of time..

(5) Write a medium length article about synthetic cells for the general public (I have to do this anyways)
(5) Publish the interviews with Build-A-Cell steering committee on Medium. **Interviews are here**

Christian Clough: Support Eran Agmon on (2). Email already exchanged.:airhorn:

Andrew Cox - **visualization schemes (4)** - work with Kim group at UT Southwestern to develop a software to visualize whole-cell models in a modular fashion, **games (5)** - create a working group for next BaC to outline timeline and aspects of Life Design Game and communicate with in silico group to ensure integration between game and modeling is streamlined. **Funding (7)** - identify potential funding sources and construct narrative for why "maximal modeling of minimal cell" is an attractive funding opportunity. Tie in how timeline toward BaC1.0 (modulable protocell) opens up applications of model, use this to establish early deliverables for any industry funding.

Jon Calles: identifying benchmark datasets (1), preliminary model validation (2). Lit search for dataset and share with all model makers. Or recieve model and help test. Help organize data.

Patrick Brennock: 4 work with Dr. Kim, 5 talk with Eran, Kim about creating work group for "cell simulation game".

You: ?

This is the landing page for this working group. **To participate in this group you do not need to edit this page.**

Please follow links to actual working group files. If you are at the workshop right now, follow the link to the workshop page.

Email group lead / contact below to learn more, get access to files and start contributing.

Workshop #6 file

https://docs.google.com/document/d/1AAN2R1Fe-NAPk2fvxZD40ol5Ja37IkYI5zjcXU_DMfg/edit?usp=sharing

Workshop #5 files

https://docs.google.com/document/d/1kHzrDTG21SEsIOw6FNdtrdpSWamxB0xmSzneFnmRxw8/edit?usp=sharing

Group contacts, name and email

Akshay Maheshwari <akshaym@stanford.edu>,

Ryan Spangler <ryan.spangler@gmail.com>,

Eran Agmon <agmon.eran@gmail.com>,

Jen Hofmann <hofmannj@stanford.edu>,

Alp Sunol <asunol@stanford.edu>,

Hannah Verdonk <hverdonk@stanford.edu>,

Emma Gonzalez <emmagg@stanford.edu>,

Miroslav Gasparek <miroslav.gasparek@eng.ox.ac.uk>, Christian Clough <christian.clough@gmail.com>,

Andrew Cox <andrew.cox@utsouthwestern.edu>,

Jon Calles <jecalles@stanford.edu>

Christian Zapata-Sanin <czapatas@caltech.edu>

Group tagline / short description:

What must we model to enable construction of a minimal cell? How long will it take & how much will it cost?

Group deliverables with approximate dates

1. 1/31 -- First draft of a proposal discussing the multiscale modeling needed to enable construction of a minimal cell. Should include modeling questions that need to be asked as well as time and cost.
2. 3/15 -- Prototype of first models, sufficient for proposal
3. 4/15 -- Complete & submit proposal

---

1) What must we model to enable the construction of a minimal cell?
2) What do we need to support the modeling?
3) How long will it take & how much will it cost?

What can we do vs. what should we do?

What more needs to be done to construct a modelable cell? Modern knowledge of cell biology spans many scales, from the atomic structure of single molecules to the evolution and growth of populations. How much of this infinite knowledge space must we represent to enable engineers to design and operate synthetic cells? Here, we discuss **1) biological processes that must be represented to build a modelable cell, as well as the extent to which they must be modeled 2) The scales to which we must model these essential biological processes 3) applications of a modelable cell; 4) what more we need from you.** We end with a proposed modeling plan that our sub-group will undertake.

Other routes could be taken to build a modelable cell that aren't constrained by existing life, and these "lineage-agnostic" routes could be fantastic; a good first step is to try to model existing life (and adjust as needed).

Biological processes that we need to model to build a first modelable cell

Our current understanding of cell biology is abstracted into biological processes. While much of the underlying chemistry and physics have been elucidated, much more remains to be understood.

All the things we know how to model that we'll probably need to represent to build a first modelable cell

Need:

1. Transcription/translation
    a. Required states: RNAP, RIB, amino acids, nucleotides, cell cycle checks, chromosome structure
    b. Initiation
    c. Elongation
    d. Termination
2. DNA replication
    a. Required states: chromosome, replosome, nucleotides
3. Gene regulation
    a. Required states: Transcription factors
4. Intra/intercellular Transport (active and passive)
    a. Required states: trans-membrane transporters
5. building/maintenance of cellular structures
    a. Required state: protein-degradation complexes, misfolded proteins, large multi-component complexes
6. Organization/compartmentalization of macromolecules within the cell
    a. Kinesin (or similar) transport, liposome (or similar) formation, Brownian motion, LLPS (distinguishing this from other similar types of molecular segregation, e.g. transient binding to other macromolecules)
7. Metabolism
    a. photosynthesis/chemosynthesis/cellular respiration
    b. Required states: metabolites, enzymes
8. Cell shape & motility
    a. Required states: flagella/cilia? membrane/cell wall? cytoskeleton?
9. Cell division
    a. Proteins responsible for membrane fission (can be coarse-grained)
    b. Orientation of mitotic spindle

All the things we haven't figured out how to model, but probably need to represent in order to build a first modelable cell.

1. How do all emergent properties of a cell arise?
    a. Need to fill out further
2. How macromolecules get distributed in cell, how do they know where to go? Which are the transport rates and mechanisms that keep the machinery so synchronized?
3. Cell death
4. Role of quantum mechanics
    a. e.g. molecular superposition/entanglement
    b. e.g. proposed involvement in photosynthesis

Scales of biological understanding and modeling

Physical scale?

1. Molecular dynamics (atomistic scale, high spatial resolution, short time scales)
2. Colloidal dynamics via either Brownian dynamics (colloidal scale) or even more advanced methods (Stokesian dynamics)
    a. Correlating properties of interest (e.g. diffusion, structure) to either experiments or all-atom modeling of *specific processes of interest* (e.g. translation, membrane transport, membrane fission)
    b. Key metric is understanding how to coarse-grain the interactions between molecules (e.g. proteins, membrane, nucleic acids) to this level - what can you ignore and what is vital?
    c. Collective motion of particles as an outcome of macromolecules sharing a common 'carrier' cytosol
3. Network/kinetics
    a. FBA level, regulatory networks, protein-protein interaction networks
    b. Networks?
4. Single-cell level dynamics (protein movement & interactions, short or long time scales, medium spatial resolution)
    a. Do we include cell shape? Can cell shape be emergent?
    b. Can we recapitulate a cell's shape?
        i. Pott's model
5. Population-level dynamics (cell-cell interactions, population response to stress/stimuli)

Complexity scale?

We test the emergent behavior from the level above, even though you're building one level lower.

FBA model that completely recapitulates life cycle but needs unrealistic tRNA. Bound your FBA w/ correct levels of tRNA. → You're model parameters should naturally result from or be bound by/constrained by one scale lower.

Unit testing via "knockouts"

<u>**No hard coding of the results that you want.**</u>

Integration?

1. What scales do the mentioned processes need to be in?
2. How do we integrate the different scales?

Applications of a modelable cell

1. Genetic/Metabolic circuit engineering
2. Modeling impacts of introducing non-natural nucleotides/amino acids
3. Ecological interactions
4. Predict cellular I/O

5. Identifying minimal sets of processes
6. Creating entirely novel organisms: new biochemistry and physical behaviors
7. Possible insight into the origin of life (not guaranteed)
    a. Underlying basic mechanisms emerging through changing interactions between subunits in vivo.
8. Drug discovery
    a. Drug off-target effects
    b. Drug/pathway interactions
    c. **Celluloids**
9. Teaching
    a. VR of inside cell
    b. Intra-Cellular LEGO (a.k.a. Build Your Cell!)
    c. Cell competition (Battle bots...for cells? Which can live the longest? Come up with a game in building cells. A lot of the parts are modular, so people not experienced in modeling can play. Drag and drop? Reflected in model what's useful to print. Can we have the models compete based on the input we put in. Have users asking for features, give us a reason to make a model useful. Hit more benchmarks as you go.
    d. Create a complex cell model that could serve as a benchmark for the Machine Learning tasks - could result in a whole bunch of ML engineers getting into this field
        i. Give people inputs and outputs to play with. Benchmark dataset
        ii. Benchmark dataset. What are the standardized experiments, etc. to make a benchmark.
            1. Standard for providing data in that's robust. Unit test for importing data into neuroinformatics -- get statistics around the data collected -- is it meaningful/are there anomalies? How can we do this for building a cell? Availability of data at every scale.
                a. Recruit teams to generate this data.
                b. Protocols for collecting data; normalized methods for doing experiments → Build a cell integration/NIST
                    i. We need a standard. We can use it ourselves. See what covers our own needs. Keep flexible until then.
10. Benchmarking *de novo* physical cells, with existing or different chemistry/physical characteristics
11. Applying living organization to social or political systems (economics, supply, collective decision making).
12. Disease-agent modeling
    a. Is this ethical? Are we making a cure or a bioweapon?

What do we need from you?

1. Benchmark datasets
    a. "Turing tests" across scales
2. Working models for subsystems
   a. Consult experts
3. Integration platform for subsystems

    a. Integration across scales (physical & timescales)
    b. Integration across biological subsystems
    c. PURE transcription/translation model across scales as an example
4. Visualization schemes
5. Recruiting talent
6. Publicity/Outreach [for maximal model of a minimal cell; in silico protocell]
7. Games
    a. http://herocoli.com
    b. Cell studio: A platform for interactive, 3D graphical simulation of immunological processes
8. Resources & Finances
    a. Compute (microcents/GigaJoule)
    b. Distributed compute (https://boinc.berkeley.edu/?)

1) Benchmark dataset definition

- What are the inputs/outputs?
- How complex?
- Training vs. testing vs. validation
- Automated evaluation?
- Datasets at different spatial scales
    - Validate models based on their predictions of emergent phenomena vs empirical results for those phenomena
    - E.g., use GFP diffusion data to validate predictions of cellular packing models
    - Cellular growth rate for the environmental stress experiments

Turing tests across scales.

1. Make an atomic-scale picture of a protocell?
2. Databank of empirical results of test cases, within 10% of a certain thing. Run the model, how does it line up.
    a. Empirical studies on E coli when drop glucose -> gene regulation. R^2 from data over time.
    b. Benchmark of diffusion/cytoplasmic streaming?
    c. Can we hit reaction rates, elongation rates, cell composition/mass fraction
    d. If we put early protocell, what are the protein content, etc. (internal composition of state)
    e. Some proteins are characterized w/ D constant, e.g., of ribosome between species.
    f. Fit model parameters to diffusion constants measured.
    g. Active/passive microrheology in cells; different types of microrheology, get agreement w/ viscosity, etc.. Better results. Microrheology
3. Multi-scale benchmarks → What are the multi-scale turing tests for building a cell

2) Working models for subsystems, Integration platform for submodels (also integrated into the visualization and video game platform?)

working models for… (should include parameters which can be changed for cell type, desired conditions, etc.)

- Cell membranes
- Nucleoid
- Cytoplasm
- Active flows/molecules (e.g. motor proteins)
- Metabolism
- Coarse-graining macromolecular structure
- Physical processes that can be readily applied to cellular environments

Test the emergent behavior from the level above, even though you're building one level lower.

FBA model that completely recapitulates life cycle but needs unrealistic tRNA. Bound your FBA w/ correct levels of tRNA. → You're model parameters should naturally result from or be bound by/constrained by one scale lower.

3) Integration platform for subsystems

- Existing successful multi-scale models, and how it / lessons learned can be applied to our system

4) Recruitment of talent from other fields:

- ML
- Video game design & marketing
- Computational science
- Minimal cell biologists
- DNA writing - both resources and talent
- Synthetic genome design
- Graphic designer for visualizations
- VR of inside cell
- Experts of different biological pathways

5) Publicizing potential utility of designing whole cell models, good for backing and building community

Come up with a good blurb. Identify needs (what are societal problems, critical knowledge gaps, etc?) and show how support of whole cell provides solution. A clear message of what we envision whole cell models enabling synthetic biology to accomplish.

6) Visualization schemes: different outputs depending on the scale or process of interest

→ Methods for showing what our model does.

→ Representation schemes

Dashboards: show how snapshot of moelcules in cells look; snapshot of cell game. Both directions of model.

7) Games

Make Spore but with real mechanisms

Flesh out cell game. Battle celluloids. Can battle the models.

8) Resources & Finance

- Computational resources
- Lab consumables
- Commercial Hardware & Software
- Money
- Travel
- DARPA -- benchmark dataset? Have them pay for benchmarking.
-

Model PURE transcription translation at multi-scale: atomic to multi-cell, as an example?

Experimental plan

What are the key properties of the living “complete” cells that we aim to capture? What areas of expertise are represented in our group? What should we recruit for?

We need a strategy for incrementally introducing features to the model. Ideally, at each step we have what could be called a “working cell model” which progressively gains detail and specifics.

We need reliable ways to physically characterize cells, get inputs we need for our models.

Starting with transcription/translation, a la PURE system?

Processes (required mechanisms to target by model):

10. Transcription/translation
    a. Required states: RNAP, RIB, amino acids, nucleotides, cell cycle checks, chromosome structure
    b. Initiation
    c. Elongation
    d. Termination
11. Response to stimuli
    a. Toxins, signals, nutrients, temperature, pH
12. DNA replication
    a. Required states: chromosome, replosome, nucleotides
13. Gene regulation
    a. Required states: Transcription factors
14. Intra/intercellular Transport (active and passive)
    a. Required states: trans-membrane transporters
    b. Recover correct rates
15. building/maintenance of cellular structures
    a. Required state: protein-degradation complexes, misfolded proteins, large multi-component complexes
16. Proton-motive force
    a. Required states: internal/external pH, membrane potential, proton pumps
17. Cell shape & motility
    a. Required states: flagella/cilia? membrane/cell wall? cytoskeleton?
18. Complexation
    a. Proteins, transcripts
19. Organization/compartmentalization of macromolecules within the cell
    a. Kinesin (or similar) transport, liposome (or similar) formation, Brownian motion
20. Cell division
    a. Proteins responsible for membrane fission (can be coarse-grained)

    b. Orientation of mitotic spindle
21. Metabolism
    a. photosynthesis/chemosynthesis/cellular respiration
    b. Required states: metabolites, enzymes
22. Cellular immunity
23. Cell death
24. Cell life cycle stage progression/specialization (e.g. stem cell -> neuron)

Methods

Which processes can be integrated? How to integrate across time/space scales? What is the process for contributing to the model?

6. Whole-cell modeling
7. Deterministic modelling
8. Stochastic modeling (Gillespie)
9. Metabolic networks and FBA
10. Agent based modeling of cells in an environment
11. Molecular dynamics (atomistic scale, high spatial resolution, short time scales)
12. Colloidal dynamics via either Brownian dynamics (colloidal scale) or even more advanced methods (Stokesian dynamics)
    a. Correlating properties of interest (e.g. diffusion, structure) to either experiments or all-atom modeling of *specific processes of interest* (e.g. translation, membrane transport, membrane fission)
    b. Key metric is understanding how to coarse-grain the interactions between molecules (e.g. proteins, membrane, nucleic acids) to this level - what can you ignore and what is vital?
    c. Collective motion of particles as an outcome of macromolecules sharing a common 'carrier' cytosol
13. Ecological scale modelling
14. Visualization methods, 2d and 3d (audio?)
    a. GPU accelerated. 3D Game engine(Unreal, Unity)
15. Control theory (designing regulatory pathways)
16. Transcriptome modeling
17. Epigenome modeling
18. Machine Learning….
    i. Can be used to speed up physical models such as molecular and colloidal dynamics models
    ii. In areas where data is abundant can also predict properties
    iii. Unsupervised learning:
        1. Correlate things, i. e. protein-protein interactions, signalling
    iv. Supervised learning:
        1. Function approximation?
        2. But where do we get data from?
            a. Ties to the minimal cellular models
    v. Reinforcement Learning
        1. Maximize the reward ⇔ Maximize the fitness….

Emergent properties

While there are many detailed and low-scale processes going on in the living cells, the collections/tuples/supersets of these processes (pathways, networks, etc.) must have certain properties to be realistically utilized in the model of the cell working in the physical environment. The properties listed below are essentially “badges on the black boxes” that emerge at the top of the sets of more elemental interactions between systems and subsystems

1. Robustness
    a. The ability of the cell to function in the varying external/internal environment NB: this is different from evolution
        i. Variation in internal cellular resources
        ii. Variation in external environmental stress
        iii. Noise
        iv. Uncertainty in “parameters” (might not be the right “wording”)
        v. Disturbance rejection vs adaptation?
2. Homeostasis
    a. Maintain the internal state of the environment despite the external variation
    b. Metabolic production/degradation at (dynamic) equilibrium
    c. Cell structure building/maintenance/degradation
3. Evolution. growth/division/mutation
    a. General understanding of what is going on in this process?
    b. Utilization? (Can evolution engineer this for us?)
    c. How do multicellular organisms and communities evolve entirely novel functions?
4. Cell-cell interactions that arise from intrinsic cell dynamics in shared environments
    a. Collective behaviour? Synchronization?
    b. Organelle formation?
    c. Interspecies interaction

Applications of cell models:

What can the above processes and methods be used for. Which of these applications do you consider the 3 “simplest” to achieve?

13. Metabolic circuit engineering
14. Modeling impacts of introducing non-natural nucleotides/amino acids
15. Knockout predictions
16. Ecological interactions
17. Predict cellular I/O
18. Identifying minimal sets of processes

19. Possible insight into the origin of life (not guaranteed)
    a. Underlying mechanisms emerging through changing interactions between subunits in vivo
20. Drug discovery
    a. Drug off-target effects
    b. Drug/pathway interactions
21. Teaching
    a. VR of inside cell
    b. Intra-Cellular LEGO (a.k.a. Build Your Cell!)
    c. Cell competition
22. Benchmarking *de novo* physical cells, with existing or different chemistry/physical characteristics
23. Applying living organization to social or political systems (economics, supply, collective decision making).

Properties and integration based on chosen application:

What do each of the following require (minimal processes, methods, etc)?

1. Metabolic circuit engineering
2. Modeling impacts of introducing non-natural nucleotides/amino acids
3. Knockouts?
4. Ecological interactions
5. Predict internal response to external stimuli
6. Identifying minimal sets of processes
7. Possible insight into the origin of life (not guaranteed)