What is Synthetic Biology?

• “... a maturing scientific discipline that combines science and engineering in order to design and build novel biological functions and systems” [SynBERC]

• Synthetic biologists are working on diverse applications:
  – New medical diagnostics and therapies
  – Extract harmful pollutants from the ground
  – Chemical production or detection

• Synthetic Biology is at a crossroads: AI can help!
The Vision – Programming Biology Like Computers

In the 1950s, computers were programmed by flipping switches and moving cables...

Synthetic Biology is at that point today...

• Manual creation of simple “parts”
• Tedious, ad hoc creation of simple systems

• Many opportunities in Design – Build – Test – Interpret loop to improve the process
• Goal: high-level program to functioning cells

If detect explosives: emit signal
If signal > threshold: glow red

Many iterations

>1 year/system

Simple systems
Complex Interactions in the Design-Build-Test Cycle

Design

Map behavior specification to nucleic acid sequence(s)

Test

Transfect/transform cells, assay behavior

Build

Synthesize/assemble nucleic acid sequence(s)

Collaborating laboratory

DNA/reagent supplier

Laboratory with new method
Challenges in Genetic Circuit Design

- Leveraging human expertise
  - Expert systems
- Parts can interfere, component availability varies
  - Constraint-based reasoning
- Large search space over possible designs; ‘impedance’ of parts must match; most of the constraint satisfaction problems are at least NP-Hard
  - Heuristic search
- Naive abstract networks and assembly steps are not viable
  - Optimization
- Large amounts of data with partially observable states
  - Machine learning
- Reason about and design colony behavior
  - Multi-agent systems
## Bioengineering Challenges & Potential AI Solutions

<table>
<thead>
<tr>
<th>Engineering Challenge</th>
<th>Key AI techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine-assisted gene circuit design</td>
<td>expert systems, constraint-based reasoning, heuristic search, optimization, machine learning, multi-agent systems</td>
</tr>
<tr>
<td>Flexible protocol automation</td>
<td>robotics, planning under uncertainty</td>
</tr>
<tr>
<td>Assay interpretation and modeling</td>
<td>machine learning, qualitative reasoning</td>
</tr>
<tr>
<td>Lab management and optimization</td>
<td>heuristic search, optimization, planning under uncertainty</td>
</tr>
<tr>
<td>Represent/exchange designs</td>
<td>semantic networks, ontologies</td>
</tr>
<tr>
<td>Represent/exchange protocols</td>
<td>semantic networks, schemas</td>
</tr>
</tbody>
</table>
Emerging Opportunities for AI-Enhanced Synthetic Biology

- Ability to produce large labeled experimental data at scale:
  - Foundries (e.g., MIT Broad)
  - Industry (e.g., Ginkgo Bioworks)
  - Cloud Labs (e.g., Transcriptic)

- Sophistication and efficiency of machine learning algorithms
  - AlphaFold: substantial improvements in protein structure prediction by Google’s DeepMind team\(^1\)
  - Experimental design via RL in quantum physics\(^2\), chemistry\(^3\) and molecule design\(^4\)

Access to design data at scale and maturing AI techniques can assist with accelerating bio engineering process

---

\(^1\) Evans R. et al. De novo structure prediction with deep-learning based scoring, 13 CASP, 2018
\(^2\) Alexey A. Melnikov et. al. Active learning machine learns to create new quantum experiments. Proc. of the National Academy of Sciences, 2018
\(^3\) Zhou et. al. Optimizing Chemical Reactions with Deep Reinforcement Learning, ACS Cent. Sci., 2017
\(^4\) You et. al. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation, 2018
Information Flow in Biological Systems and ML

- **Transcription**
  - Identify and characterize gene expression networks/modules
  - Predict transcription binding factors
  - Identify and characterize regulatory regions (promoters, enhancers)
  - Predict promoter/enhancer interactions

- **Splicing** → Prediction of Splicing Sites

- **Translation**
  - Predict mRNA binding targets

- **Folding**
  - Predict protein 3d structure from sequence

- **Metabolomics**
  - Identify biomarkers of phenotype
Current State of ML

- Very rare to have data that measures the process end-to-end
- ML are applied to only some of the parts of the process and for narrow instances
- Deep learning techniques have outperformed most of earlier ML techniques
- DL performance still overall poor
- Good progress made on promoter region prediction and gene expression for simple organisms
- Important advancement on predicting protein structure from DNA sequence (Deep Mind)
AI for SynBio Challenges and Opportunities

**Challenges**
- Overfitting/lack of robustness due to large size of design space
- Small changes in design space can lead to large changes in phenotypic properties
- Mismatch between distribution of designs with desired property and distribution of designs in training set
- Highly modular and hierarchical biological systems and contexts
- Expensive/Black Box Validation Process

**Opportunities**
- Generation of large datasets suitable for ML
- Identification of domain heuristics that restrict design space
- Incorporation of mechanistic models to aid data-driven learning
- Transfer learning
- Effective active exploration and sampling of design space
- Task-driven learning
- Interpretable ML
Research Opportunities: Expanding the System

• Can we improve design effectiveness predictions by expanding the system we measure and analyze?
Research Opportunities: Hierarchical Modeling of Biological Processes

- Reducing the genotype to phenotype divide
- Customize deep NN architectures by adding layer(s) that capture intermediate steps in the genotype-phenotype mapping

- Enforce intermediate latent spaces that capture cells interaction with the design
  - Various choices of latent space
- Use latent space to learn phenotypic classifier

Notional Architecture
Research Opportunities: Addressing Data Gaps and Need for System Modeling

• Multi-source, multi-experiment data fusion and enrichment
  – Leverage additional sensing and modalities to reduce data gaps
  – Task/context driven fusion

• Combine experiment-based data with synthetic data generated by mechanistic models
  – Mature mechanistic models that can be leveraged (e.g., Metabolic models, RBS models)
  – Trade-off between model realism and data completeness
Precision Medicine: Save Lives and Money

Medicine Today: Disease-based, “one-size fits all”

- 6% of baseline DoD Budget
- 10% of baseline DoD Budget

Medicine Tomorrow: Patient-based, customized
Eg. Transcriptional Variability in iPSC Motor Neurons

Increasing expression of a gene

Blue = SOA Informatics Methods
Orange = Personalization Thru Data Science

Increasing variation in a gene’s expression
Gene-Patient Relationships

Blue Nodes = SOA Informatics
Orange Nodes = Data Science
Added Personalization

This network visualizes the expression levels of purple gene nodes with relation to 3 ALS samples and the mean of 3 Control samples.

Genes expressed significantly in control patients but not in ALS patients

Genes expressed significantly in 3 ALS patients but not in control patients
Heterogenous Data Types

- **Life Sciences:** “cambrian explosion” of data **types** *(Volume/ Variety)*:
  - NGS: SNPs, CNV...
  - Microarray, RNASeq, ...
  - Western Blot, Mass Spec
  - Functional enrichment
  - Pathway Analysis
  - Experimental context

- **Data size:** orders of magnitude:
  - Genome -> Transcriptome -> Proteome [2]

- **Inconsistent and conflicting**
  - Variant, function annotations from literature disagree
  - Gene identifier formats vary
  - Protein -> gene correlation varies, post-transcriptional, translational changes [3]

- **Cybersecurity:** **Consistent**, well understood data types *(Volume/Velocity)*:
  - e.g. Netflow, Traceroute, BGP Routing
  - Unified across IP or MAC address
  - Logical model maps to physical
  - Strongly quantitative
  - Single data types sufficient for many problems
Tooling and Infrastructure Misaligned

- **Unreliable** database annotations
  - 90% of annotations are inferred electronically
  - Generic, not curated, e.g. ‘protein binding’ [5]
- Tooling environment lacks **standards**
  - 547 tools for pathway analysis at pathguide.org [4]
  - Links 404, SW is not updated
- Tools are not written for **distributed systems**
  - Unix piping and File I/O
  - “Basic” Taverna workflow
- Need **shared datasets** to combat growth
  - TCGA, ICGC (via AWS)
  - Some sending terabytes over ALS2 [6]
  - However, data gulf widening as $ / sequence drops [7]
Problems with *Shallow* Learning

*Looks for linear separability or direct relationships among variables in the data to extract patterns.*

<table>
<thead>
<tr>
<th>Problem</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformations and feature generation is complex for high-dimensional, sparse, non-linear omics data</td>
<td>Leads to research delays for timely diagnosis and treatment</td>
</tr>
<tr>
<td>Dimensionality reduction cannot capture both inter/intra omic relationships</td>
<td>Lose underlying relationships that model disease mechanisms</td>
</tr>
</tbody>
</table>
Move to *Deep Architectures*

- Well-equipped to handle high dimensional, sparse, noisy data with nonlinear relationships
- Provides high generalizability for *multiplatform* data common in the life sciences
- Abstracts data to learn complex features/patterns to identify the breadth vs depth trade-offs

![Multi-layer Multi Kernel Learning](image1)

![Deep Neural Network](image2)
Data $\rightarrow$ Deep Architecture

**Stacked Autoencoder**

Image $\rightarrow$ Features $\rightarrow$ Protein $\rightarrow$ Structure

**Convolutional Neural Network**

Image $\rightarrow$ Object $\rightarrow$ Gene Expression $\rightarrow$ Disease

**Recurrent Neural Network**

Sentence $\rightarrow$ Language $\rightarrow$ Cell Signalling $\rightarrow$ Signaling Cascades
How to Increase Collaboration and Reproducibility?

- Goals:
  - Capture details about system designs for reproducibility
  - Curate useful part databases
    - Qualitative and quantitative information enabling automation
  - Data exchange formats for collaboration

- Challenges
  - What is the minimum information required?
  - Detailed protocol descriptions are required to reproduce results
  - Comparable measurement units and techniques are just emerging
  - Results are stochastic and vary with time
  - Many biological processes poorly understood
Other Challenges

• Expert knowledge is not explicitly written down and can require “common sense” assumptions

• Many aspects of organism engineering may be proprietary or subject to IP
  – Can work to promote open exchange of knowledge like in CS

• Many unknowns in biological organisms
  – Some areas mature enough, AI can help reveal areas to study
Solutions: Knowledge Representation

- Synthetic Biology Open Language (SBOL) standard describing genetic parts, devices, modules, systems
  - Community effort much like W3C
  - Ontology based approach
  - Defines a standard visualization for designs
  - Slowly gaining recognition and acceptance
  - ACS Synthetic Biology adopted SBOL for depiction and representation of genetic construct for recording and sharing
How to Automate & Optimize Lab Equipment

• Goals:
  – Complex dependencies with temporal and resource constraints, non-deterministic outcomes during assembly and test
  – Effective use of shared wet-lab, shared supplies, shared equipment
  – Increasing precision of executed protocols

• Challenges:
  – Reliance on existing paper lab notebooks, software
  – Closed source lab management software
  – Complex scheduling including living cells that must be maintained and measured at certain times
Solutions

- Puppeteer (Vasilev et al. 2011)
  - Suite of tools for defining
    - robot specific resources
    - robot specific protocols
    - producing executable scripts.

1. Assignment to wells
2. Dilute with buffer
3. Combine assembly step
Solutions: Robotics

• Automated Assembly (Densmore Lab)

Assembly Planning Tools

Robot

Biomek 3000
• 1-way and 8-way pipette tools
• gripper to move plates to magnet
• heat block, shaker and -20C on deck
How to Handle Data Uncertainty

• Goals:
  – Many biological processes are inherently stochastic and not synchronized
  – Still many unknowns and noise in the data/models

• Challenges:
  – Some data collection methods destroy the cells
    • Can’t be used for data collection at later time points
  – Different trials of the same experiment may have different results
  – Modeling enough to guarantee predictability
Solutions

- Empirical Quantitative Incremental Prediction (EQuIP)
  - Accurate prediction of genetic regulatory network behavior from detailed characterizations of their components (Davidsohn 2015)
How to Predict Multi-Cellular Behavior

• Goals
  – Each cell can be viewed as an “agent”
    • Cells can communicate using small molecules
  – Complex applications rely on interaction between cells
    • Tissue differentiation, multi-cellular organisms

• Challenges
  – Communication rate and reaction time are slow
  – Need to model cell-to-cell communication for complex aggregate behavior
  – Test and validation of differentiated cells
Solutions: Multi-Agent Systems

• Spatial Computing / Aggregate Programming
  – Cells are seen as spatial computers: Executing the same program in reaction to the local sensor information

```lisp
(defun band-detector (signal lo hi)
  (and (> signal lo)
       (< signal hi)))

(let ((v (diffuse (aTc) 0.8 0.05)))
 (green (band-detect v 0.2 1)))
```

[Beal & Bachrach, ‘08]

[Weiss]
How to Design Complex Circuits

• Goals
  – Design is knowledge intensive
  – Understand what might be going wrong
  – What to do and how to correct it

• Challenges
  – Greater precision required for computation
  – Managing complexity
Solutions: Knowledge-Based Systems

- Formal grammars for verification (GenoCAD)
  - Is this a valid DNA sequence?
- Black-listed bad sequences (Eugene)
- BioCompiler: design motifs

GenoCAD CFG [Cai et al., ‘07]
Solutions: Bio Compiler

- Bio-Compiler: Motif-based compilation where operators are translated to motifs
Complex Example: 4-bit Counter

Optimized compiler outperforms human designers
How to Select Parts for a Circuit

• Goals
  – Transforming high-level organism descriptions to DNA sequences involves solving several constraint satisfaction problems
    • Which parts to use
      – Parts are not compatible with each other, interact and interfere
    • Which method to use
      – Better noise reduction at the cost of higher metabolosmic load

• Challenges
  – The data is missing or incomplete
  – Formulation of biological requirements as constraints
  – Domain is complex: identify necessary and sufficient conditions
Solutions: Constraints, Satisfiability, Search

How can we map the abstract parts in an AGRN to real parts? MatchMaker!

**Feature Mapping:** Satisfy the constraints on the edges of the AGRN

**Signal Matching:** Pick parts that are signal compatible accounting for noise and preserving digital behavior

---

<table>
<thead>
<tr>
<th>AGRN</th>
<th>Feature Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Factors</td>
<td>Transcription Factors</td>
</tr>
<tr>
<td>( \alpha_0 )</td>
<td>( q_0 )</td>
</tr>
<tr>
<td>( \alpha_1 )</td>
<td>( q_1 )</td>
</tr>
<tr>
<td>( x )</td>
<td>( q_x )</td>
</tr>
<tr>
<td>( y )</td>
<td>( q_y )</td>
</tr>
<tr>
<td>( P_0 )</td>
<td>( P_0 )</td>
</tr>
<tr>
<td>( P_1 )</td>
<td>( P_1 )</td>
</tr>
<tr>
<td>( P_2 )</td>
<td>( P_2 )</td>
</tr>
</tbody>
</table>

---

**Find Match**

- Signal compatible: \([Y] + [Z] = [Z] \) (Green)
- NOT signal compatible: \([Y] + [Z] \neq [Z] \) (Red)
Conclusion

• Synthetic biology has exciting applications
  – Curing cancer, diabetes, neglected diseases
  – Environmental remediation
  – Biofuels, nanofabrication

• Benefits to both fields
  – New exciting and complex application domain for AI
  – Enabling complex applications in biology through AI reasoning

• Tight collaboration is needed to realize the benefits to both AI and synthetic biology

• A goal is to build a set of community resources
Citations