Mathematical Representations
For Biological Systems

Sahil Loomba

17th August, 2018
Overview

- Brief Perspective on Modeling in Biology
Overview

● Brief Perspective on Modeling in Biology
● Systems Biology: Discovery of biological mechanisms
  ○ Markov Interaction Network: Uncertainty meets Knowledge
  ○ Hypothesis discovery as arbitrary probabilistic queries
Overview

- Brief Perspective on Modeling in Biology
- Systems Biology: Discovery of biological mechanisms
  - Markov Interaction Network: Uncertainty meets Knowledge
  - Hypothesis discovery as arbitrary probabilistic queries
- Synthetic Biology: Design of bioengineered parts
  - Language embedding models to represent biomolecules
  - Design challenges as arbitrary downstream machine learning
Overview

● Brief Perspective on Modeling in Biology

● Systems Biology: Discovery of biological mechanisms
  ○ Markov Interaction Network: Uncertainty meets Knowledge
  ○ Hypothesis discovery as arbitrary probabilistic queries

● Synthetic Biology: Design of bioengineered parts
  ○ Language embedding models to represent biomolecules
  ○ Design challenges as arbitrary downstream machine learning

● Closing the loop on iterative discovery and design
  ○ Organism-on-Chip for high-throughput drug discovery
Biology Exhibits Hierarchical Compositionality

Principle of Abstraction

ATOMS
Mass Spectrum, Molecular Properties

BIOMOLECULES
Structure, Transcriptomics, Proteomics, Metabolomics

CELLS
Cellular Phenotype

TISSUES
Images (Histology), EEG

ORGANS
Images (MRI), fMRI

ORGANISM
Organismal Phenotype
Biology Exhibits Hierarchical Compositionality
Biology Exhibits Hierarchical Compositionality

DESIGN SYNTHETIC BIOLOGY

TRANSCRIPTION TRANSLATION

INTERACTION

REACTION

REGULATION

GENOTYPE

DISCOVERY SYSTEMS BIOLOGY

PHENOTYPE
Problems of Discovery in Systems Biology
DeepMind’s AI can detect over 50 eye diseases as accurately as a doctor

The system analyzes 3D scans of the retina and could help speed up diagnoses in hospitals

By James Vincent | @jjvincent | Aug 13, 2018, 11:01am EDT
Biology: A Modeling Perspective

A Typical “Machine Learning” Problem

- Clearly defined **input** and **output**
- **Large** amount of data (usually cheap to acquire)
- Availability of labels: **supervised**
- Problems in medical biology at **phenotype level**

Atypical “Machine Learning” Problem

- **Arbitrary query** of interest
- **Small** amount of data (usually expensive to acquire)
- Few to no labels: **semi-supervised**
- Problems in systems and synthetic biology at **genotype level**
Work at Wyss: A Biological Perspective

- Discovery (Systems Biology)
  - Mechanism of tolerance to pathogens
  - Countermeasures to induce tolerance

- Design (Synthetic Biology)
  - Protein Stability Problem
  - Riboswitch Design Challenge

- Diagnostics (Medical Biology)
  - Predicting breathing severity in Asthma
  - Mass Spec for Pathogen Detection
A Typical “Machine Learning” Problem

Atypical “Machine Learning” Problem
Typical Machine Learning | case-in-point

*Predicting Breathing Severity in Asthma*

![Chart showing breathing severity levels]

**LOW SEVERITY**

**HIGH SEVERITY**
Typical Machine Learning | case-in-point
Predicting Breathing Severity in Asthma

LOW SEVERITY

HIGH SEVERITY
Typical Machine Learning | case-in-point
Predicting Breathing Severity in Asthma

LOW SEVERITY

HIGH SEVERITY
Typical Machine Learning | case-in-point
Predicting Breathing Severity in Asthma

Low Severity
Breathing is Monofractal

High Severity
Breathing is Multifractal

Simple ML model on this 2D space gives prediction accuracies of upto 97%
Atypical Machine Learning | case-in-point
MALDI-TOF MS for Pathogen Detection

Shannon Duffy
Atypical Machine Learning | case-in-point
MALDI-TOF MS for Pathogen Detection

Shannon Duffy
Atypical Machine Learning | case-in-point
MALDI-TOF MS for Pathogen Detection

Cluster Frequency Weighting
Cluster Proportion Weighting

Accuracy of 100% on a Probabilistic Model with 12 blind samples

Shannon Duffy

BLIND LIBRARY
Core Tenets for Mathematical Representations in Biology

- Biology is uncertain and stochastic
- Biology is not discrete
- Biology is complex, but has some core underlying “latent” principles that govern “observed” complex phenomena
- Explicitly incorporate uncertainty through probabilistic modeling
- Develop “soft”, continuous and distributed representations
- Include domain knowledge to define structure of core variables that generate evidence
Uncertainty Meets Knowledge
With Great Knowledge Comes Great Power

drug2gene

gene2phene

drug

CTD, LINCS

gene

CTD, LINCS, KEGG,
Gene Ontology

phene

Gene Ontology

THERAPIES

MECHANISMS

OMICS

MORPHOMETRICS
Uncertainty Meets Knowledge: NeMoCAD
Network Model for Causality Aware Discovery

- **Network** of interactions between genes encode routes of causal mechanistic influence in a biological system
- **Probabilistic** weights obtain from gene knockout and/or overexpression data encode weight of causal interactions to form a Markov Network
- Arbitrary mechanistic queries can be turned into corresponding probabilistic inference queries for discovery
- Adding **drugs** as nodes of the network also allows us to discover new drug **therapies**
Uncertainty Meets Knowledge: NeMoCAD
Network Model for Causality Aware Discovery
Uncertainty Meets Knowledge: NeMoCAD
Drug Discovery | gene2drug

MARKOV INTERACTION NETWORK

LOOPY BELIEF PROPAGATION

{ 'hif1a': 1, 'epas1': 0 }
Uncertainty Meets Knowledge: NeMoCAD
Drug Discovery | gene2drug

MARKOV INTERACTION NETWORK

MARKOV INTERACTION NETWORK

LOOPY BELIEF PROPAGATION

{ 'hif1a': 1, 'epas1': 0 }
Uncertainty Meets Knowledge: NeMoCAD
Drug Combination Investigations | gene2drug

{\texttt{hif1a}:1, \texttt{epas1}:0}
Uncertainty Meets Knowledge: *NeMoCAD*

Drug Synergy Contextualized by Gene Space | drug2drug

![Heatmap of Drug Interaction Scores](image)

**Synergy**

**Orthogonality**

**Antagonism**
Uncertainty Meets Knowledge: NeMoCAD
Incorporating Gene Ontology | gene2phene
Uncertainty Meets Knowledge: NeMoCAD
Transcriptomics to Therapy and Phenotype | gene2drug+phene

Therapies to counter effects of Radiation

- Pick a condition to antagonize: 8 Gy → 16 Gy
- Run differential gene expression analysis → Input fold-changes + p-values into probabilistic model NeMoCAD to obtain drug therapies
- Enrich for phenotypes using Gene Ontology
Uncertainty Meets Knowledge: NeMoCAD
Transcriptomics to Therapy and Phenotype | gene2drug+phene

Top enriched phenotypes: gene2phene

1. Cholesterol monooxygenase activity
2. Helicase activity (important for DNA repair), regulation of response to DNA damage and integrity
3. Muscle cell fate specification, visceral muscle development

Top drugs selected: gene2drug

1. Dexrazoxane, a chemotherapy protective drug that is shown to reduce tissue damage.
2. Mercaptopurine, an immunosuppressive chemotherapy drug used to treat acute lymphatic leukemia
3. Atropine, an involuntary nervous system blocker
4. Ifosfamide, a chemotherapy drug used in treating multiple cancers
5. Pregnenolone, an endogenous steroid that is a precursor to most other steroid hormones
Uncertainty Meets Knowledge: NeMoCAD
drug+phene2gene

```python
hypoxia_drugphene2gene = nemo.query_phenet_given_drugs('hypoxia', ['bisphenol a', 'urethane'], combo_func=16)

found 18 GO terms of interest for query hypoxia
found 16 genes of interest for query hypoxia
Formed common network with 46 nodes
Formed common network with 48 nodes
gene network
```
Uncertainty Meets Knowledge: NeMoCAD

phene2drug

```
drugs_eye = nemo.query_drugs_given_phene('eye development', normalize='mean')
Found 15 relevant GO terms that map to query eye development

drugs_iron = nemo.query_drugs_given_phene('iron', normalize='mean')
Found 62 relevant GO terms that map to query iron
```
Uncertainty Meets Knowledge: NeMoCAD

drug2phene
Uncertainty Meets Knowledge: NeMoCAD

drug2phene

Gene Ontology Enrichment on urethane vs. bisphenol-a corrcorr 0.397

Drug Correlation at Gene Ontology Level

Drug Synergy at Gene Profile Level

Benzoic Acid
Bisphenol A
Emodin
Problems of Design in Synthetic Biology

DESIGN SYNTHETIC BIOLOGY

INTERACTION

REACTION

REGULATION

TRANSCRIPTION TRANSLATION

GENOTYPE
Representing Biomolecules as Symbolic Sequences

- Proteins are simply Amino Acid sequences: VPLLGLY...
- Genes/mRNAs are simply Nucleotide sequences: AATCGGTA...
Representing Biomolecules as Symbolic Sequences

- Proteins are simply Amino Acid sequences: VPLLGLY...
- Genes/mRNAs are simply Nucleotide sequences: AATCGGTA...
- Using language models for “representing” arbitrary biomolecules as mathematical entities
  - AA : Sequences = Words : Sentences

word2vec, Mikolov et al. (2013)
Representing Biomolecules as Symbolic Sequences

- Proteins are simply Amino Acid sequences: VPLLGLY...
- Genes/mRNAs are simply Nucleotide sequences: AATCGGTA...
- Using language models for “representing” arbitrary biomolecules as mathematical entities
  - AA : Sequences = Words : Sentences
- One simplifying assumption: local neighborhood decides global high-level properties
- Unsupervised: predict context words from current word

![Diagram showing word2vec model](word2vec_diagram.png)
Representing Biomolecules as Symbolic Sequences

prot2vec: a Learnt Space of Proteins

93,588 AA sequences from *Homo sapien* Proteome

prot2vec

100-d

VPLLGLY...

<0.21, -0.32, ..., 0.74>
Representing Biomolecules as Symbolic Sequences

prot2vec: a Learnt Space of Proteins

prot2vec
100-d

VPLLGLY...

<0.21, -0.32, ..., 0.74>

Protein Family
Protein Topology
Protein Stability

Any Machine Learning Model
Representing Biomolecules as Symbolic Sequences

prot2vec predicts Protein Families

**Insight:** proteins close in sequence space $\rightarrow$ close in prot2vec space $\rightarrow$ close in function space

- **prot2vec**
  - 100-d
  - $<-0.12, 0.78, \ldots, 0.56>$$

- 324,017 AA sequences from SwissProt across 7027 protein families

- Nearest Neighbor Classifier

- Protein Family
- Protein Topology
- Protein Stability

- **Nearest Neighbor Classifier**

- 73.2% accuracy
Representing Biomolecules as Symbolic Sequences
prot2vec estimates Protein Stability

prot2vec
100-d

<0.29, -0.10, ..., 0.98>

16,174 AA sequences from Rocklin et al. (2017)

Protein Family
Protein Topology
Protein Stability

Deep Neural Network Regressor

0.54 R² value
Representing Biomolecules as Symbolic Sequences
prot2vec captures Protein Topologies

2D tsne plot of prot2vec embeddings of Protein Stability Dataset

**Insight:** proteins close in sequence space $\rightarrow$ close in prot2vec space $\rightarrow$ close in topology space

Experimental Data from Rocklin et al. (2017)
Representing Biomolecules as Symbolic Sequences
prot2vec correlates to Biophysical Parameters

**Question:** What do these 100 dimensions mean? Are they arbitrary?
Representing Biomolecules as Symbolic Sequences

**ribo2vec**: a Learnt Space of Riboswitches

49,159 mRNA sequences of natural Riboswitches

```
AATCGGTA...
```

```
<0.09, -0.17, ..., 0.94>
```
Representing Biomolecules as Symbolic Sequences

\textit{ribo2vec}: a Learnt Space of Riboswitches

\textbf{ribo2vec}\hspace{1cm}10-d

- FMN riboswitch ([RFN element]
- TPP riboswitch ([THI element]
- \textit{yxyG} leader
- \textit{SAM} riboswitch ([\textit{S} box leader]
- Purine riboswitch
- Lysine riboswitch
- Cobalamine riboswitch
- \textit{glnN} glutamine-5-phosphate activated ribozyme
- \textit{yxyC} leader
- \textit{yN} leader
- Glycine riboswitch
- \textit{SAM} riboswitch ([\textit{alpha-proteobacteria]}
- \textit{PreQ1} riboswitch
- $S$-adenosyl methionine ([\textit{SAM}) riboswitch
- \textit{preQ1-HL} ([\textit{pre queuosine]) riboswitch
- Moco ([\textit{molybdenum cofactor}) riboswitch
- Magnesium sensor
- $S$-adenosyl-L-homocysteine riboswitch
- AdoCbl riboswitch
- M. fumarum riboswitch
- AdoCbl variants RNA
- SAM-AV variant riboswitch
- SAM-MNH riboswitch
- Fluoride riboswitch
- Glutaminine riboswitch
- ZMP/2ZTP riboswitch
- SMK box translational riboswitch
- Cyclic dGMP-II riboswitch
- \textit{SAM-V} riboswitch
- THF riboswitch
- \textit{PreQ1-III} riboswitch
- NiCo riboswitch

49,159 mRNA sequences of natural Riboswitches

\texttt{<0.09, -0.17, \ldots, 0.94>}

AATCGGTA...
Representing Biomolecules as Symbolic Sequences

ribo2vec correlates to Biophysical Parameters

**Question:** What do these 10 dimensions mean? Are they arbitrary?

Visualize 192 mRNA design sequences to detect DNT/TNT

**Insight:** some “arbitrary” dimensions of ribo2vec correlate with experimental activation ratio

*Experimental Data from Howard Salis*
Representing Molecules as Symbolic Sequences

- **Problem:** arbitrary metabolites, small molecules, drugs, ligands and carbohydrates are NOT simple linear sequences
- **Solution:** Simplified Molecular-Input Line-Entry System (SMILES) representation

**Bisphenol A**

CC(C)(C1=CC=C(C=C1)O)C2=CC=C(C=C2)O

**Ciprofloxacin**

N1CCN(C(C)(C(F)=C2)=CC=C(C2=CC=C(C=C2)=O)O)(C3CC3)(C(C)(C=O)=O)
Representing Molecules as Symbolic Sequences

*chem2vec*: a Learnt Space of Chemicals

First 1 million chemicals from PubChem

\[ C (C) C_1=C \ldots \]

\[ \text{chem2vec} \]

100-d

\[ <-0.66, -0.67, \ldots, 0.42> \]
Representing Molecules as Symbolic Sequences  
*chem2vec*: a Learnt Space of Chemicals

```plaintext
C(C)C1=C...
```

`chem2vec`  
100-d  

<-0.66, -0.67, ..., 0.42>

- Drug Family
- LINCS Gene Probabilities
- Protein-Binding Glycans

Any Machine Learning Model
Representing Molecules as Symbolic Sequences

*chem2vec* encodes Chemical Valencies

Visualizing just “atomized” chemical species
Representing Molecules as Symbolic Sequences

*chem2vec*

Visualizing just “atomized” chemical species

Encodes chemical valencies?
Representing Molecules as Symbolic Sequences

chem2vec correlates to Molecular Properties

**Question:** What do these 100 dimensions mean? Are they arbitrary?
Representing Molecules as Symbolic Sequences
chem2vec correlates to Molecular Properties

Linear Correlation between molecular properties and chem2vec dimensions
Representing Molecules as Symbolic Sequences

*chem2vec* serves as a drug query engine

Say we have drug candidates we find improving tolerance, and we wish to explore drugs “similar” to them in the functional space, but potentially better in PK/PD and toxicity.

“Miracle” Drug for Xenopus

DFOA
Putting It Back Together

Drugs
Pathogen
Signal
Gene
Protein
Phene

NETWORKS
Omics

GENE ONTOLOGY

Morphometrics
Putting It Back Together

Drugs → Gene → Phene
Pathogen → Gene → Phene
Signal → Protein → Morphometrics

chem2vec → Gene
gene2vec → Gene
prot2vec → Protein

NETWORKS
GENE ONTOLOGY
Omics
Putting It Back Together

Drugs

Pathogen

Signal

Gene

Protein

Phene

Morphometrics

chem2vec

gene2vec

prot2vec

NETWORKS

GENE ONTOLOGY
Wyss Institute’s Organs-on-Chips
Xenopticon: Organism-on-Chip

High-Throughput Xenopus embryo analyzer

Capability to image >700 embryos every 15 min in 16 colors
Results in 100s – 1,000s of metrics per embryo per time-point

Youngjae Cho, Bret Nestor, Richard Novak
Xenopticon: A Cheaper Route to Drug “Design”

1. PHENOTYPE DATA (XENOPTICON)
   - $1-10 per image, 1 second per image

2. GENOTYPE DATA (TRANSCRIPTOMICS)
   - $1-10K per sample, 1 month per sample

3. DRUG DEVELOPMENT (R&D)
   - $1-10B per drug, 10 years per drug
Xenopticon: Acquire Phenotype Metrics
Estimate Embryo Viability

1-4-DPCA (Hif1a agonist) dose response

Survival rate (%)

Hours Post Infection (hr)

Vishal Keshari, Alex Dinis
Xenopticon: Acquire Phenotype Metrics
Estimate Embryo Viability

1-4-DPCA (Hif1a agonist) dose response

Classifier on Embryo’s Spectrum → Hidden Markov Model

93% accuracy on held-out test image sequences
**Xenopticon**: Acquire Phenotype Metrics

Obtain Survival Curves
Xenopticon: Acquire Phenotype Metrics
Track Tissue Development

DeepMind
Bret Nestor
**Xenopticon**: Acquire Phenotype Metrics

Making the Invisible, Visible | Visual Tracking of Immune Response

GFP expressing Macrophages to track spatiotemporal dynamics of “immune response”
Xenopticon: Acquire Phenotype Metrics
Making the Invisible, Visible | Visual Tracking of Pathogen Infection

mCherry expressing E. coli bacteria to track spatiotemporal dynamics of “pathogen infection”

Vishal Keshari, Alex Dinis
Xenopticon + NeMoCAD = XenoDoc

Drugs → Gene → Protein → Phene
Pathogen

Signal

NETWORKS

GENE ONTOLOGY

Morphometrics

Bret Nestor
Iterative Discovery and Design

Mechanism Discovery & Drug Design

- gene2drug
- Xenopticon (Xenopus + Pathogens)
- phene2gene
- GENOTYPE MECHANISMS
- PHENOTYPE METRICS

DRUG CANDIDATES
Iterative Discovery and Design

Gene2Drug

Mechanism Discovery & Drug Design

Phene2Gene

Gene2Drug

Mechanism Discovery & Drug Design

PHENOTYPE METRICS

PHENOTYPE

METRICS

Xenopticon

(Xenopus + Pathogens)

GENOTYPE MECHANISMS

Gene2Drug

DRUG CANDIDATES

PubChem

UniProt
Acknowledgements

Don Ingber, Jim Collins

Mike Super, Richard Novak

Bret Nestor, Vishal Keshari, Alex Dinis, Susan Clauson, Youngjae Cho

Shannon Duffy, Mark Cartwright

Joe Mooney, Ben Matthews, John Osborne

Nik Dimitriikakis, Shanda Lightbown, Kazuo Imaizumi, Dana Bolgen, Anna Waterhouse
Thank You, Questions?