Functional Assays II: Computation

Mike White
Bio5488
4/3/24
Learning Goals

Be able to critically read reports of computational analysis and evaluate:

• **Aims of the method** (predict molecular or clinical impact of untested variants, quantify effects of assayed variants from complex data, predict functional enhancers, splice sites, etc.)

• **What is being quantified or modeled?** What is the output of the model?

• **What data or features are the model trained on?**
Not Today’s Learning Goals

• Details of machine learning/regression/computation

Check out Computational Molecular Biology, Machine Learning, Statistics/Bioinformatics courses
Why build computational models of massively parallel functional assay data?
The Current Vision
MAVE: Multiplexed Assay of Variant Effect
Rationales for modeling quantitative sequence-function relationships

Annual Review of Genomics and Human Genetics

Massively Parallel Assays and Quantitative Sequence–Function Relationships

Justin B. Kinney and David M. McCandlish

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; email: jkinney@csbl.edu, mccandlish@csbl.edu

Keywords
genotype–phenotype map, epistasis, variants of uncertain significance, biophysical modeling, cis-regulatory grammar, deep learning

Abstract
Over the last decade, a rich variety of massively parallel assays have revolutionized our understanding of how biological sequences encode quantitative molecular phenotypes. These assays include deep mutational scanning, high-throughput CRISPR–Cas9, single-cell experiments, and many others. Here we review the rational integration of these high-throughput assays and consider the development of quantitative sequence–function models.
Predicting which variants in the human genome are likely to be pathogenic requires a comprehensive and quantitative understanding of the molecular phenotypes produced by mutation (152).

Transcription factors (TFs) are proteins that regulate gene expression by binding to specific sites encoded in genomic DNA. Understanding the sequence specificities of TFs—that is, which sites a TF will bind and how strong this binding will be—requires quantitative models that integrate sequence information across the length of each candidate (77).
Big Themes of Functional Assay analysis

- Data is usually *sequencing data*, not a direct measure of the function of interest. Models connect data to the underlying function that we are interested in.

- To achieve practical goal of predicting variant effects, MAVE approaches aim to understand/model general, *quantitative* sequence-to-function relationships.
Why not just a lookup table?

Why do we need sophisticated models for MAVE data?

Don’t we just measure all the variants?

• Predict effect of variants in proteins that haven’t yet been assayed

• Assay itself is not a direct readout of phenotype: noise, non-linearities, etc.

• Epistasis! Even the biggest DMS only sparsely samples mutation space - you want to predict more combinations of variants than you can measure
Model Examples
MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect

Ammar Tareen, Mahdi Kooshkbaghi, Anna Posfai, William T. Ireland, David M. McCandlish & Justin B. Kinney

Genome Biology 23, Article number: 98 (2022)  
4504 Accesses  |  9 Citations  |  33 Altmetric  |  Metrics
Idea: Both sequence and assay effects need to be modeled to get accurate quantification of variant effect.
MAVE as Noisy Measurement of Genotype-Phenotype Map

Goal is to model both G-P map and measurement process
What Should a Genotype-Phenotype Map Look Like?

1. Additive (all positions in a sequence contribute independently)

2. Neighbor (interaction terms between neighboring positions)

3. Pairwise (each position interacts with every other position)

4. Black Box (roll your own relationship function)

5. Biophysical (model ΔΔGs) with neural network
Discrete vs Continuous Measurements

- Type of data influences how the model is fit.
- Barcode RNA-seq is continuous
- Sort-seq expression values are limited by the number of sorted bins.
Global epistasis:

“Mutations may act additively on some underlying, unobserved trait, and this trait is then transformed via a nonlinear function to the observed phenotype as a result of subsequent biophysical and cellular processes.”
Translation:

Global epistasis means mutations act *non-additively* on phenotype (epistasis), but this may be due to *additive* effects on some underlying trait.

Local epistasis means specific pairwise interactions between residues.
Phenotype

Unobserved molecular trait

Mutations move additively along x-axis, non-additive effects on y-axis
Predicting Splice Variants at *BRCA2* Exon 17
All 5’ SS PWM

BRCA2 site

MAVE-NN

Human (202,764 sites)

CAG/GC A A G U
Takeaways

- MAVE output is sequencing data, not a direct measure of the property of interest.

- Modeling gets you from sequencing data to the property of interest.
Biophysical models of variant effect

The energetic and allosteric landscape for KRAS inhibition

Chenchun Weng, Andre J. Faure, Albert Escobedo & Ben Lehner

Nature 626, 643–652 (2024) | Cite this article

18k Accesses | 1 Citations | 285 Altmetric | Metrics
Rationale:

Capturing biophysical effects underlying the phenotype will lead to better predictions of combinations of variants.

However, changes in affinity cannot be inferred simply by quantifying changes in binding to an interaction partner; even in the simplest genotype-to-phenotype (energy) landscapes, ‘biophysical ambiguities’\textsuperscript{18} exist, meaning that changes in a molecular phenotype (for example, binding to an interaction partner) can be caused by many different changes in the underlying biophysical properties\textsuperscript{18,19} (for example, changes in stability reducing concentration or altered binding affinity).
Energetic & Allosteric Landscape of KRAS

- Assay 26,000 KRAS mutations & assessed protein folding and binding affinity to six different partners.

- Use deep neural network implementation of a biophysical model to determine 22,000 free energy changes.

- Quantify how mutations tune binding affinity, map potentially druggable allosteric sites
Procedure:

- Assay variant effect in two ways (protein binding and protein stability/abundance)
- Include single mutations plus *many* double mutations
- Train a neural network model with biophysical ΔG values as trained parameters
- ΔG values produce biophysically-based, interpretable predictions
PCA: Protein Fragment Complementation

When blue and purple domains interact, yeast survive
Thermodynamic Model

\[ \Delta G_f = -RT \log(K_f) \]
\[ \Delta G_b = -RT \log(K_b c) \]

\[ p_f = f_f(\Delta G_f) = \frac{1}{1 + e^{\Delta G_f/RT}} \]
\[ p_{fb} = f_{fb}(\Delta G_f, \Delta G_b) = \frac{1}{1 + e^{\Delta G_b/RT}(1 + e^{\Delta G_f/RT})} \]

\( p_f = \) probability folded, \( p_{fb} = \) probability folded & bound
Training thermodynamic parameters with machine learning
Free Energy Table

### g) KRAS-RAF1 binding free energy change ($\Delta\Delta G$)

<table>
<thead>
<tr>
<th>Mutant AA</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### h) KRAS folding free energy change ($\Delta\Delta G$)

<table>
<thead>
<tr>
<th>Mutant AA</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### i) Binding interface

- **P-loop**
- **Switch-I**
- **Switch-II**
- **Core**
- **Surface**
- **HVR**
- **O-haired**
- **H-strand**

---

**Additional Information:**
- **GTP pocket**
- **Both**

---

**Legend:**
- **Black** indicates high energy change.
- **Color gradient** represents energy change from blue (low) to red (high).
KRAS mutations that affect RAF1 binding map to the binding interface.
Mapping Allosteric Sites that affect Binding
Takeaways

• Leverage 1) two types of functional measurements, 2) lots of double mutants, 3) machine learning

• Build an interpretable biophysical model that explains the effects of mutation on protein function

• Premise is that modeling molecular phenotypes correctly will produce better phenotypic predictions
Modeling Reporter Gene Assays
Problems with MPRA Measurements

- Measure is a ratio of counts (RNA/DNA) - can be unstable
- Multiple barcodes tagging same element: how should we account for them?
- Transfection of complex libraries across biological replicates can introduce significant variation
Problems with barcode RNA/DNA ratios

• Ratio of two noisy measurements

• When read counts are low, small differences in reads lead to very large changes in ratio

• Errors across measurements of multiple barcodes can be propagated to estimate variance, but common averaging approaches fail to take advantage of this.

If you want maximize the statistical power of your experiment, then you need a good estimate of variance!
MPRA Quantified as RNA/DNA

Array-synthesized library

Transfection into cells, Sequence BCs

Expression

BC RNA
BC DNA
BC RNA
BC DNA
BC RNA
BC DNA
...

CRE
DsRed
BC
CRE
DsRed
BC
CRE
DsRed
BC
MPRAnalyze Basic Idea:
Assume Linear Relationship

\[ RNA_{molecules} = \alpha \times DNA_{molecules} \]

Goal is to estimate \( \alpha \), the transcription rate.
Modeling Approach

External Covariates

- Plasmids
  - DNA counts
- Transcripts
  - RNA counts

Latent

Observed
DNA Sampling

Estimated

Plasmids

\[ \tilde{d} \sim \text{Gamma} (k, b) \]

Observed

DNA counts
RNA Estimate

\[ \hat{r} | \hat{d} \sim \text{Poisson} \left( \alpha \hat{d} \right) \]
Consider observed reads as sampling from a distribution

\[
\hat{r} \sim NB \left( \mu = \frac{\alpha \cdot k}{\beta}, \psi = k \right)
\]

1. Combine to get negative binomial for estimated RNA reads.
2. Replace equation parameters with DNA and RNA linear regression model.
How well does it work?

MPRA analyze produces lower variance among negative control sequences
Sensitive detection of active elements
DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers

Bernardo P. de Almeida, Franziska Reiter, Michaela Pagani & Alexander Stark

Nature Genetics 54, 613–624 (2022) | Cite this article

14k Accesses | 22 Citations | 104 Altmetric | Metrics
Training CNN on MPRA Data
How well does the model do?
Interpreting CNNs
Use the model to design sequences
Recommendations for using MAVEs for variant interpretation

- Assay should have sufficient dynamic range to robustly separate damaging variants from benign variants.

- Choose an assay that captures effects that are relevant, and note types of variants that would not be detected in the assay.

- Deposit data in repositories, disclose all experimental and statistical methods, make your code available

- Provide measures of reproducibility and error estimates

- Validate results with some single-variant tests and include known variants as controls
Clinical Relevance of Functional Data

Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation


On behalf of the Brotman Baty Institute Mutational Scanning Working Group

Genome Medicine 11, Article number: 85 (2019) | Cite this article

4316 Accesses | 18 Citations | 12 Altmetric | Metrics
Abstract

Variants of uncertain significance represent a massive challenge to medical genetics. Multiplexed functional assays, in which the functional effects of thousands of genomic variants are assessed simultaneously, are increasingly generating data that can be used as additional evidence for or against variant pathogenicity. Such assays have the potential to resolve variants of uncertain significance, thereby increasing the clinical utility of genomic testing. Existing standards from the American College of Medical Genetics and Genomics (ACMG)/Association for Molecular Pathology (AMP) and new guidelines from the Clinical Genome Resource (ClinGen) establish the role of functional data in variant interpretation, but do not address the specific challenges or advantages of using functional data derived from multiplexed assays. Here, we build on these existing guidelines to provide recommendations to experimentalists for the production and reporting of multiplexed functional data and to clinicians for the evaluation and use of such data. By following these recommendations, experimentalists can produce transparent, complete, and well-validated datasets that are primed for clinical uptake. Our recommendations to clinicians and diagnostic labs on how to evaluate the quality of multiplexed functional datasets, and how different datasets could be incorporated into the ACMG/AMP variant-interpretation framework, will hopefully clarify whether and how such data should be used. The recommendations that we provide are designed to enhance the quality and utility of multiplexed functional data, and to promote their judicious use.
Learning Goals Again

Be able to critically read reports of MAVE analysis and evaluate:

• **Aims of the method** (predict molecular or clinical impact of untested variants, quantify effects of assayed variants from complex data, etc.)

• **What is being quantified or modeled?** What is the output of the model?

• **What data or features are the model trained on?**
Cell-type-directed design of synthetic enhancers

Transcriptional enhancers act as docking stations for combinations of transcription factors and thereby regulate spatiotemporal activation of their target genes. It has been a long-standing goal in the field to decode the regulatory logic of an enhancer and to understand the details of how spatiotemporal gene expression is encoded in an enhancer sequence. Here we show that deep learning models, can be used to efficiently design synthetic, cell-type-specific enhancers, starting from random sequences, and that this optimization process allows detailed tracing of enhancer features at single-nucleotide resolution. We evaluate the function of fully synthetic enhancers to specifically target Kenyon cells or glial cells in the fruit fly brain using transgenic animals. We further exploit enhancer design to create ‘dual-code’ enhancers that target two cell types and minimal enhancers smaller than 50 base pairs that are still functional.
Reading Tips

• Don’t worry too much about biological details

• Focus on input data, modeling strategy, model output.

• How well did it work?