Analysis of Variant Effect Data I

Mike White
Bio5488
4/11/22
MAVE: Multiplex Assay of Variant Effect
Learning Goals

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?**
Not Today’s Learning Goals

• Details of machine learning/regression/computation

Check out Computational Molecular Biology, Machine Learning, Statistics/Bioinformatics courses
Outline for the Week

1. Modeling protein MAVE assays
2. Modeling MPRA data
3. Variant effect predictions from evolution & population diversity
Why build computational models of variant effects?
The Current Vision
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
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@csil.edu, mccandlish@csil.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 and single-cell behaviors. How do we...
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).
Proteins typically have multiple molecular phenotypes, including folding energy, enzymatic activity, and expression level. Understanding how protein sequence governs these molecular phenotypes is complicated by the fact that the peptide chain of a protein typically folds into a specific three-dimensional structure, resulting in interactions between amino acids that are distant in the primary sequence. The study of protein sequence–function relationships (42) thus presents experimental and modeling challenges beyond those encountered in the study of TF–DNA binding.
Big Themes of MAVE analysis

- MAVE data is sequencing data, not a direct measure of the function of interest. Models connect data to function.

- To achieve practical goal of predicting variant effects, MAVE approaches aim to understand/model general, **quantitative** sequence-to-function relationships.
Example 1: Predicting Effects of Protein Variants

Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data

Vanessa E. Gray 1, Ronald J. Hause 1, Jens Luebeck 1, Jay Shendure 1,2, Douglas M. Fowler 1,3,4

https://doi.org/10.1016/j.cels.2017.11.003
Rationale for the approach

1. Existing predictors of variant effect are categorical, not quantitative, and are trained on imperfect clinical annotations.

2. Accurate clinical prediction should be founded on accurate molecular prediction (MAVE data).

3. Large-scale mutagenesis data may reveal general patterns that determine when variants are deleterious.


“Envision” procedure

1. Collect and normalize 21,026 variant measurements from 9 experiments (different proteins)

2. Annotate each residue with a set of features (chemical properties, conservation, etc.)

3. Train model (decision tree ensemble using supervised, stochastic gradient boosting learning algorithm) to predict variant effect from residue features.
Collect and annotate mutagenesis datasets

Train predictive regression models

Predict new mutations

Amino acid features used in the model

![Bar chart showing the feature importance](chart.png)
Models trained on individual proteins

Global predictor did less well than other models for predicting *clinical* effect, but did well on predicting MAVE results.
Takeaways

- Idea is to build good predictor of molecular effects as path to improve clinical prediction.

- Presumes that variant effects can be predicted from amino acid and protein ‘features’.

- Requires MAVE data plus curated features
Example 2: Modeling Assay Effects

bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer reviewed, practice/health-related behavior, or be reported in news media as established information.

New Results

MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect

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


This article is a preprint and has not been certified by peer review [what does this mean?].
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.
Contrast MAVE-NN with ENVISION

- Predict amino acid variant effects globally (ENVISION) vs model genotype-phenotype map of actual protein or cis-reg element sequences (MAVE-NN)

- Input data: Additional ‘features’ to describe amino acids (ENVISION) vs sequence and assay results only (MAVE-NN)
A simple test case: protein G binding to IgG
Genotype-Phenotype Map of GB1

GB1 = B1 domain of protein G, assayed for IgG binding
Modeling the Non-linear Relationship Between Assay Output and Modeled Phenotype
• 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.
Example 2B: Use the same approach to understand epistasis

Inferring the shape of global epistasis

Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin

PNAS August 7, 2018 115 (32) E7550-E7558; first published July 23, 2018; https://doi.org/10.1073/pnas.1804015115
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

Mutations move additively along x-axis, non-additive effects on y-axis

Unobserved molecular trait

Mutations move additively along x-axis, non-additive effects on y-axis
Model of global epistasis

\[ \phi = \beta_0 + \sum_{i}^{L} \beta_{i,a_i} \]

Unobserved additive trait

Sum of variant effects

Predicted phenotype

Unobserved non-linear map between additive trait and phenotype

\[ y = g(\phi) + \epsilon \]
GB1 binding to IgG
Biophysical models of variant effect

Mapping the energetic and allosteric landscapes of protein binding domains

Andre J. Faure, Júlia Domingo, Jörn M. Schmiedel, Cristina Hidalgo-Carcedo, Guillaume Diss & Ben Lehner

Nature 604, 175–183 (2022) | Cite this article

2182 Accesses | 237 Altmetric | Metrics

Abstract

Allosteric communication between distant sites in proteins is central to biological regulation but still poorly characterized, limiting understanding, engineering and drug development \textsuperscript{1,2,3,4,5,6}. An important reason for this is the lack of methods to
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’ 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 (for example, changes in stability reducing concentration or altered binding affinity).
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
The Data
“Biophysical Ambiguity”

Different $\Delta G$ values, same binding result

Ambiguity resolved!
Thermodynamic Model

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

\( p_f = \text{probability folded}, \ p_{fb} = \text{probability folded & bound} \)
Training thermodynamic parameters with machine learning
Raw data vs model results

Raw data

Modeled ΔG
Model identifies allosteric sites
(Mutations outside binding surface with binding effect)
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.
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
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
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?**
Wednesday

- Modeling MPRA data
- Predictions of variant effect from evolutionary/population variation/genomic annotation data