Empirical variance component regression for sequence-function relationships

Juannan Zhou David M. McCandlish
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory
email: jzhou@cshl.edu
preprint: https://doi.org/10.1101/2020.10.14.339804

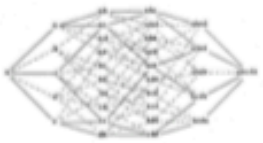
Summary

- High-throughput phenotypic assays reveal prevalent pairwise, as well as higher-order epistasis.
- Modeling and understanding these higher-order interactions remains challenging.
- We present a method for reconstructing sequence-to-function maps from partially observed data that can accommodate all orders of genetic interaction.
- The main idea is to make predictions for unobserved genotypes that match the type and extent of epistasis found in the observed data.
- This information can be extracted by estimating the fraction of phenotypic variance due to each order of genetic interaction (variance components)
- We model the sequence-function relationship using Gaussian process regression in which these variance components are used to define an empirical Bayes prior that in expectation matches the observed pattern of epistasis.

Background

- Modern high-throughput phenotyping assay can quantify the function 10^3 - 10^6 nucleotide or amino acid sequence simultaneously.
- This is allows us to study empirical fitness landscapes (sequence-function relationships) envisioned by Sewall Wright in 1932.
- Most data reveal complex pattern of genetic interactions (epistasis).

Fowler and Fields. 2014. *Nature Methods*



Wright.1932

Problems

A. How to understand the complex genetic interactions in the data?
B. How to use this information to make phenotypic predictions for novel sequences?

Our solution

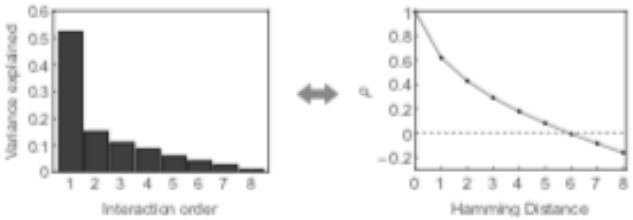
A. Study the type and extent of epistasis using variance component analysis
B. Reconstruct sequence-function maps using the infered variance components with Gaussian Processes regression

Variance component analysis for sequence-function relationships

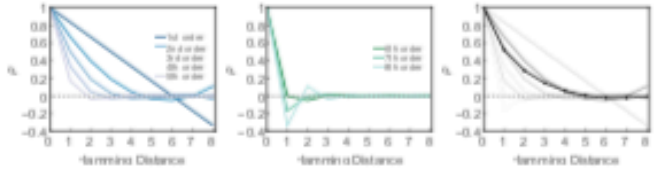
Ω_ℓ: fraction of variance explained by ℓ-th order interaction

$$\Omega_1 = \frac{\mathrm{Var}(A)}{\mathrm{Var}(G)} = \frac{h^2}{H^2}$$

Variance components determine correlation structure of the sequence-function map



Inference of variance components from partial data



- Distance correlation function for each order ℓ assumes a distinct shape

$$w_\ell^j(d) = \frac{1}{\alpha^j}\sum_{q=0}^{\ell}(-1)^q(\alpha-1)^{j-q}\binom{d}{q}\binom{l-d}{k-q}$$

- Estimate the variance components using the empirical phenotypic distance correlation

$$\rho(d) = \sum_{k=1}^{\ell}\lambda_k w_k^j(d)$$

Gaussian process regression

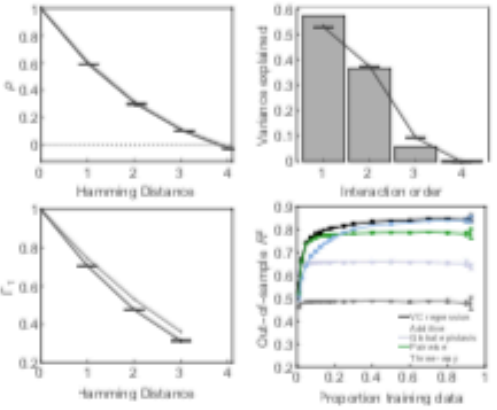
Prior distribution $\quad \mathbf{f} \sim \mathcal{N}(\mu, \mathbf{K}) \qquad K(d) = \sum_{k=0}^{\ell}\lambda_k w_k^j(d)$

Join Distribution $\quad \begin{bmatrix}\mathbf{f}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}_G\\ \boldsymbol{\theta}_m\end{bmatrix}, \begin{bmatrix}\mathbf{K} & \mathbf{K}_A\\ \mathbf{K}_A & \mathbf{K}_{AA}+\mathbf{E}\end{bmatrix}\right)$

Posterior Distribution $\quad \mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbf{K}_A(\mathbf{K}_{AA}+\mathbf{E})^{-1}\mathbf{y}, \mathbf{K}-\mathbf{K}_A(\mathbf{K}_{AA}+\mathbf{E})^{-1}\mathbf{K}_A)$

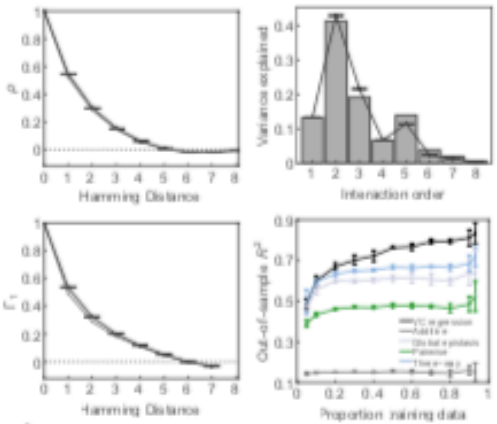
Example 1: Protein G

- Combinatorial mutagenesis: four sites on protein G domain B1 (Wu et al. 2016)
- Size of sequence space: 20^4 = 160,000



Example 2: *SMN1*

- Human 5' splice sites in gene *SMN1* (Wong et al. 2018)
- *In vivo* splicing assay
- Size of sequence space: $8^4 = 65536$



Conclusions

- Variance components analysis provide a high-level summary of the structure of the sequence-function map
- Incorporating the empirical variance components in the inference procedure greatly improves predictive accuracy

References

Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* 5:1-21.

Wong, MS., Kinney, JB, Krainer, AR. 2018. Quantitative activity profile and context dependence of all human 5' splice sites. *Molecular Cell*. 71(6): 1012-1026.