A026 · Embedding kernels for sequence-function relationships

Presenter: Waverly Carabba — Tufts University

Authors: Waverly Carabba, Samantha Petti, Carlos Martí-Gómez, David McCandlish

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

A central challenge in biology is understanding how variation in biological sequences (DNA, RNA, and protein) impacts function. Although high-throughput experiments are now capable of measuring these sequence-function relationships at a massive scale, novel computational techniques for inference and interpretation are still needed. We address these problems using Gaussian process regression, a Bayesian method that imposes a Gaussian prior on sequence-function relationships via a kernel encoding how sequence similarity relates to covariance in measured functionality. Here, we explore kernels in which the covariance depends on the distance between sequences in an embedding space, making these kernels expressive yet interpretable. Embedding features can either be (i) precomputed, for example, from evolutionary sequences or large language models, or (ii) learned from experimental data via a factor analysis kernel. We describe the mathematical relationship between our embedding kernels and existing kernels for discrete sequence space as well as classical kernels for continuous spaces. Our embedding kernels improve predictive performance across various high-throughput sequence-function datasets, including protein datasets ranging from short motifs to full-length proteins. The inferred embeddings also capture known and novel features of the sequences' underlying functionality. Overall, these models provide new tools to understand and predict sequence-function relationships.

Embedding kernels for sequence-function relationships

Waverly Carabba[1], Carlos Martí-Gómez[2], Samantha Petti[1], David McCandlish[2]
[1] Department of Mathematics, Tufts University [2] Cold Spring Harbor Laboratory

Abstract

A central challenge in biology is understanding how variation in biological sequences (DNA, RNA, and protein) impacts function. Although high-throughput experiments are now capable of measuring these sequence-function relationships at a massive scale, novel computational techniques for inference and interpretation are still needed. We address these problems using Gaussian process regression, a Bayesian method that imposes a Gaussian prior on sequence-function relationships via a kernel encoding how sequence similarity relates to covariance in measured functionality. Here, we explore kernels in which the covariance depends on the distance between sequences in an embedding space, making these kernels expressive yet interpretable. Embedding features can either be (i) precomputed, for example, from evolutionary sequences or large language models, or (ii) learned from experimental data via a factor analysis kernel. We describe the mathematical relationship between our embedding kernels and existing kernels for discrete sequence space as well as classical kernels for continuous spaces. Our embedding kernels improve predictive performance across various high-throughput sequence-function datasets, including protein datasets ranging from short motifs to full-length proteins. The inferred embeddings also capture known and novel features of the sequences' underlying functionality. Overall, these models provide new tools to understand and predict sequence-function relationships.