Navigating the protein fitness landscape with Gaussian processes

Philip A. Romero¹, Andreas Krause² and Frances H. Arnold¹,*

1 Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA
2 Department of Computer Science, ETH Zurich, Switzerland
* E-mail: frances@cheme.caltech.edu

Manuscript No. 12-15251

Author Summary

As nature’s versatile molecular machines, proteins are an irresistible target for engineering functional nanostructures. To tailor-make proteins, however, we must know how changes in amino acid sequence impact function. Current models do not describe the relationship between the sequence and function of these complex natural polymers well enough to support engineering, particularly for functions such as catalysis.

The mapping of a protein’s sequence to its phenotype can be envisioned as a surface, or fitness landscape, over the high-dimensional space of possible sequences (1). Knowledge of this sequence-function mapping is important for understanding protein evolution as well as for engineering new proteins. Here, we introduce a class of models that learn the protein fitness landscape from experimental data using a machine-learning algorithm known as a Gaussian process. Gaussian processes take examples of points from a surface and use this information to infer the height of the surface at unobserved locations (Fig. P1A) (2). This inference assumes there is continuity in the surface (i.e., that close points are more likely to have similar values). Therefore, an important aspect of these models is the kernel function, which specifies the
distance, or similarity, between pairs of points. To describe the relationship between pairs of sequences in the protein fitness landscape, we use a kernel function that is inspired by the principle that sequences with similar structures are likely to have similar properties. Here, we represent the structure of a protein family using the amino acid residue pairs that are contacting each other in the 3D structures of its members. With this contact map, the structural distance between any two protein sequences is the number of contacting residue pairs that differ. This is similar to the mutational Hamming distance, the number of residue positions that differ, but includes important structural information as well.

We tested the Gaussian process landscape model using thermostability data from chimeric sequences that were generated by recombining three bacterial cytochrome P450 enzymes (3). These sequences can have dozens of mutations with respect to any of their parent proteins. The model displayed excellent predictive ability across this large and diverse sequence set ($r = 0.95$, Fig. P1B). Such predictive accuracy is not currently possible using physics-based models. Trained directly on experimental data, Gaussian process models implicitly capture all the factors which contribute to the measured property or properties, whether they are known or not.

Gaussian processes also provide an accurate estimate of model confidence: They are less certain about sequences that are distant from the observed locations. The model’s uncertainty can be used to identify sequences that, if characterized, would provide the most information about the landscape. We used the Gaussian process’s uncertainty to develop an experimental design algorithm that identifies small but highly informative sets of sequences for testing. This algorithm was used to design an informative set of 29 chimeric cytochrome P450s, which were constructed and tested for enzyme activity and affinity for binding dopamine and serotonin,
neurotransmitters of interest for functional MRI imaging in the brain. We used this data set to generate Gaussian process models that accurately predict the activity and ligand binding affinity of previously untested P450 sequences.

We also developed an algorithm that can be used to engineer optimized protein sequences. This algorithm works by iteratively evaluating regions of the landscape that are predicted to be both uncertain and have a high fitness value (4). In early iterations, this algorithm mostly explores uncertain regions of the landscape. But once it gains confidence, it begins to identify sequences that are optimized. We used this approach to create functional cytochrome P450s that are significantly more thermostable than any previously engineered using chimeragenesis, rational design, or directed evolution.

Gaussian process models can describe the protein sequence-function mapping by overlooking the physical details and instead learning from experimental data. These models may provide a useful alternative to physics-based models for protein design and optimization.

References


Figure Legend

Figure P1. Gaussian process landscapes. (A) Gaussian processes infer the mapping from protein
sequence to function using a small experimental sampling of the landscape. (B) The Gaussian process model shows excellent predictive ability ($r = 0.95$) for a set of 242 chimeric cytochrome P450s whose thermostabilities ($T_{50}$) have been measured. Shown are 10-fold cross-validated predictions versus measured $T_{50}$ values.

**Conflict of interest statement**
The authors declare no conflict of interest.

**Author contributions**
P.A.R., A.K., and F.H.A designed the research; P.A.R. performed the research; and P.A.R., A.K., and F.H.A. wrote the paper.