Hybrid interpretable biophysical modeling of fluorescent protein biosensor function.

John N. Koberstein[1], Alison G. Tebo[1], Srinivas C. Turaga[1]

[1]HHMI Janelia Research Campus, Ashburn VA, USA

Abstract

Fluorescent protein biosensors are designed to sense physiological signals and report changes through altered fluorescence emission. These chimeric proteins combine ligand-binding and fluorescent protein domains such that binding allosterically regulates fluorescence. For optimal function, the reversible ligand-dependent transitions between conformational states must be tailored to the conditions of the intracellular environment. Biosensor function is not easily characterized by a single number, but instead depends on the kinetics of these transitions, which in turn are controlled by protein sequence and structure. However these parameters cannot be directly measured making biosensor optimization difficult. To address this challenge, we developed a hybrid model incorporating neural network layers and interpretable biophysics that infers the underlying dynamics. To assay biosensor function in the cellular context, we generated hundreds of sequence variants of the calcium biosensor GCaMP and expressed each in cultured primary neurons for testing. Given the amino acid sequence and the observed fluorescence, the network predicts the parameters for mechanistic models of biosensor function and cellular calcium handling and generates a reconstruction of the fluorescence trace. By modeling the response of biosensors in a physiologically relevant context, we identify how mutations alter the underlying biophysical properties. Similarly, by monitoring calcium transients using many biosensors with variable kinetic properties, the sources of heterogeneous dynamics across cells can be disentangled.

Introduction

Proteins are dynamic molecules that can adopt multiple conformations. Ligand-dependent transitions between conformational states enable these macromolecules to sense and respond to the environment. This phenomenon in which binding alters activity at a distal site, generally referred to as allostery, underlies crucial protein functions in gene regulation, signalling, and metabolism. While protein allostery is of fundamental importance to cellular physiology, it is currently difficult to design and optimize. Monitoring changes in protein conformational state and activity over time is challenging, resulting in relatively scarce data for training and evaluating models. Fluorescent biosensors provide a tractable system for measuring and modelling a dynamic protein function. These proteins are generated by combining existing ligand-binding and fluorescent protein domains such that their independent functions are allosterically coupled[1]. For example, high-performance biosensors are often dimly fluorescent in the apo-state and upon binding exhibit an increase in fluorescence emission. Crucially, these proteins produce a readily measurable spectroscopic signal that is directly linked to the state. Because this function can be recorded in live cells, these tools also provide a lens into the dynamics of cellular physiology.

The GCaMP family of calcium biosensors has become commonly used tools in biological research, most significantly in neuroscience (Fig. 1A). The widespread use in neurons derives from the observation of transient increases in intracellular calcium evoked by action potentials. These rapid calcium dynamics can then be converted into an optical signal using GCaMP. The high spatiotemporal resolution and non-invasive nature of fluorescence imaging enables neuroscientists to record the activity for thousands of neurons simultaneously[2]. The importance of this specific application has prompted extensive protein engineering efforts to develop new variants with improved sensitivity and faster kinetics[3–6]. State-of-the-art variants can achieve the detection of calcium transients evoked by a single action potential occurring on the order of 10 milliseconds[6]. However, the biophysical mechanisms underlying the function of these proteins remain poorly understood, impeding our ability to identify beneficial sequence changes in a principled fashion.

Methods and Results

Previously published datasets consisting of hundreds of GCaMP sequence variants tested in primary neuron culture were aggregated to use as training data[3–5]. GCaMP expressing neurons were imaged by fluorescence microscopy while electrically stimulating each well causing the neurons to fire a defined series of action potentials (Fig. 1B). A fluorescence trace, the mean intensity over time, was extracted from each individual cell soma. Metrics describing the normalized amplitude ($\Delta F/F_{max}$) and rate of decay ($t_{1/2}$) were extracted from fluorescence traces (Fig. 1C). Together these datasets comprise 345 unique sequences expressed in 67,963 neurons.

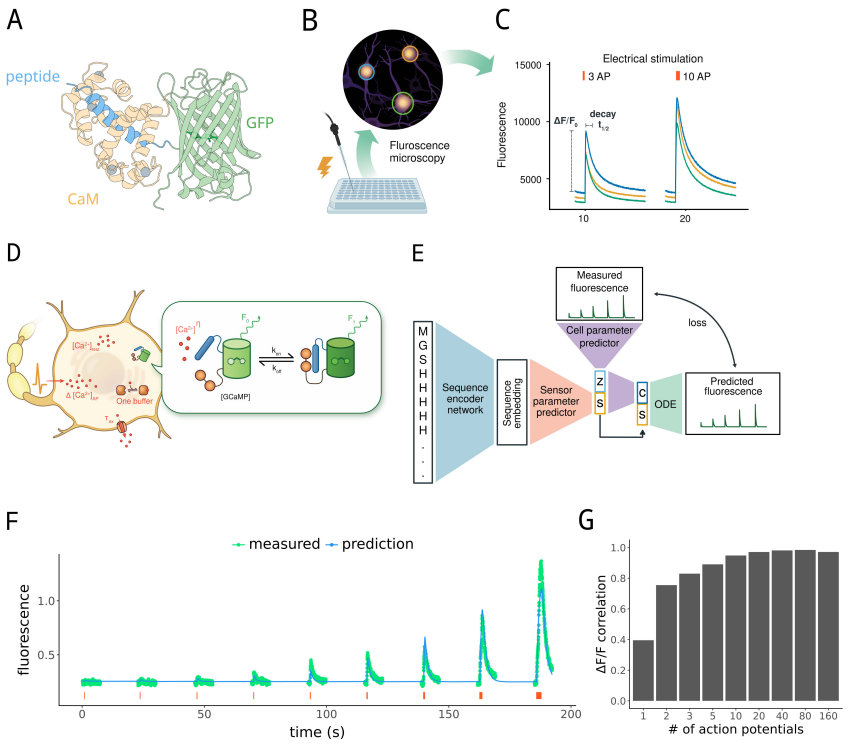

Figure 1: (A) GCaMP structure consists of a circularly permuted GFP attached to calmodulin (CaM) and CaM-binding peptide. **(B)** Fluorescence microscopy assay using GCaMP expressing neurons electrically stimulated with increasing number of action potentials (AP). **(C)** Example fluorescence traces with metrics corresponding to amplitude ($\Delta F/F_{max}$) and kinetics (decay $t_{1/2}$) labelled. **(D)** Graphical depiction of biophysical mechanisms used to simulate neuronal Ca^{2+} and GCaMP dynamics. **(E)** Hybrid model architecture using neural network layers to estimate sensor and neuron parameters of ODE model. **(F)** Model predicted trace from test set sequence. **(G)** Correlation between measured and predicted mean ΔF/F for test-set sequences.

Models were trained to generate a reconstruction of the fluorescence trace using protein sequence embeddings[7], electrical stimulus and fluorescence trace as inputs. To ensure interpretability, a mechanistic model was used that converts the electrical stimulus to changes in neuronal calcium concentration which drive transitions in protein conformation and the observed fluorescence emission (Fig. 1D). The resulting ordinary differential equations (ODE) describing this process can be solved to predict the fluorescence intensity over time for a given set of parameters. In our model, the parameters describing protein function and cellular physiology are the outputs of neural networks. The full model, combining a neural network and differentiable ODE solver, can then be optimized by gradient descent to minimize the error between predicted and measured fluorescence traces (Fig. 1E). The data was split into sets for training, validation and testing consisting of 80%, 10% and 10% respectively by randomly selecting sequences. Model performance was evaluated by comparing predicted metrics ($\Delta F/F_{max}$ and $t_{1/2}$) for held-out test set sequences. Models achieved a low MSE fit to test set data (Fig. 1F) and high correlation between measured and predicted metrics for novel sequences across the range of action potentials (Fig. 1G). Multiple mechanistic models of variable complexity describing GCaMP states and transitions, as well as calcium handling mechanisms were tested.

Discussion

Fluorescent biosensors are well positioned at the intersection of protein and cellular dynamics. Functional biosensors convert the concentration of intracellular signals into changes in protein state and altered fluorescence emission. We make use of this relationship to model the underlying dynamics of neuronal physiology and protein states. While disentangling the two sources of variability poses difficulty, it is also an opportunity to better constrain the estimated parameters for each system.

Incorporating biophysical mechanisms into the model, produces learned parameters that can be easily interpreted and transferred to new assays. The predicted effects of sequence mutations on kinetic rates and equilibrium constants from our model represent specific hypotheses to be experimentally tested. Model parameters equally apply to *in vitro* assays using purified protein in which calcium concentrations can be tightly controlled.

The mixture of mouse cortical neurons used in the current study represent a small fraction of the cell types present in the mammalian brain. A variant optimized in these assays might not be optimal for detecting action potentials in another type of neuron or monitoring calcium concentrations in cell types from different tissues. The flexibility of our model enables protein sequences to be tested *in silico* in response to novel electrical stimuli and in cell types with altered calcium dynamics. A sufficiently accurate model should eventually enable the generation of novel GCaMP variants with properties tailored to the physiology of specific neuronal cell-types.

References

1. Nasu, Y., Shen, Y., Kramer, L. & Campbell, R. E. Structure- and mechanism-guided design of single fluorescent protein-based biosensors. *Nat. Chem. Biol.* **17**, 509–518 (2021).
2. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, eaav7893 (2019).
3. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci.* **32**, 13819–13840 (2012).
4. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
5. Dana, H. *et al.* High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nat. Methods* **16**, 649–657 (2019).
6. Zhang, Y. *et al.* Fast and sensitive GCaMP calcium indicators for imaging neural populations. *Nature* 1–8 (2023) doi:10.1038/s41586-023-05828-9.
7. Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. 2022.07.20.500902 Preprint at https://doi.org/10.1101/2022.07.20.500902 (2022).