Protein Paintbox™ - An Expansive Palette

Raul Duran[1], Drew Regitsky[1], Matthew Jones[3], Jerome Bonnet[2], Drew Endy[2], Sridhar Govindarajan[1], Claes Gustafsson[1], and Jeremy Minshull[1]

1. DNA2.0, Inc., 1140 O'Brien Drive, Menlo Park, CA 94025, 2. Stanford University, 473 Via Ortega, CA 94305,
3. University of Cambridge, Downing Street, Cambridge CB2 3DY, UK

Email: cgustafsson@dna20.com

Abstract

Fluorescent and chromogenic proteins are a valuable tool for research, with uses ranging from reporters in transgenic systems to biosensors. The tangle of intellectual property restrictions, as well as protein characteristics like brightness, spectral properties, oligomeric state and sensitivity to environmental conditions, limits the breadth of tasks to which fluorescent and chromogenic proteins can be applied.

We have created a large set of new colored and fluorescent proteins by backtranslating a set of sequences from Genbank, using GeneDesigner software to simultaneously minimize sequence differences and match GeneGPS derived *E. coli* codon bias for high protein expression. Gene fragments were designed to create chimeric genes when pooled. A subset of the resultant chimeras were selected as base genes from which multiple iterations of libraries with targeted substitutions were made. Each reiteration involved characterizing variants for spectral and sequence data which was analyzed to infer correlations between them. Correlations were then tested by creating specific individual sequences and determining whether their spectral properties were as predicted. Three amino acid positions were found to be individually responsible for cyan, green or yellow fluorescence emission. We also identified variants with non-overlapping spectra as candidates for applications using more than one fluorescent protein, for example microscopy and FRET (Fluorescence Resonance Energy Transfer) applications.

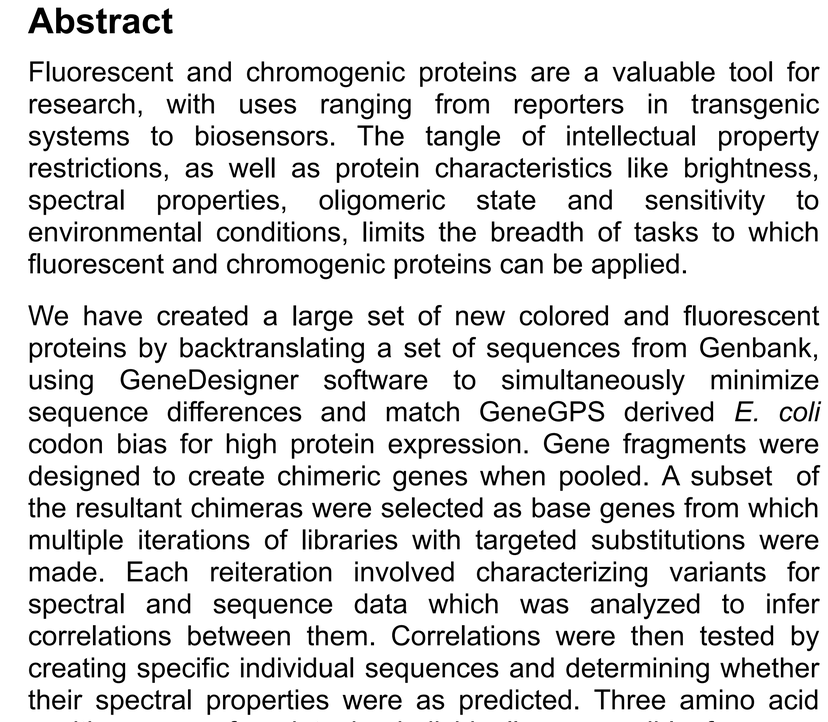

Figure 1. FPs library construction and screening.

Our Approach

Our basic approach is described in Figure 1. Genes encoding the starting proteins were backtranslated using GeneDesigner to simultaneously minimize sequence differences and match a GeneGPS derived *E. coli* codon bias. Oligonucleotides were designed to synthesize each gene in fragments, and subsets of these gene fragments were combined such that synthesis created chimeric genes. The resultant chimeras were cloned under control of the PTrc or PT5 promoters, and fluorescent colonies were picked under various illuminations. Excitation and emission spectra were recorded from pelleted cells resuspended in a transparent buffer using a fluorescence plate reader (Figs. 2 and 4).

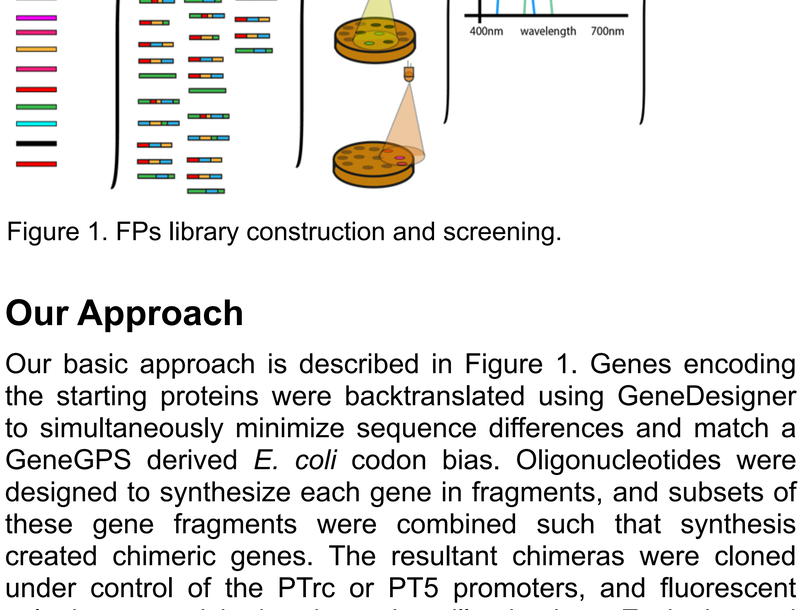

Figure 2. Phylogenetic analysis and example of spectral properties of 115 selected variants from the RFP family.

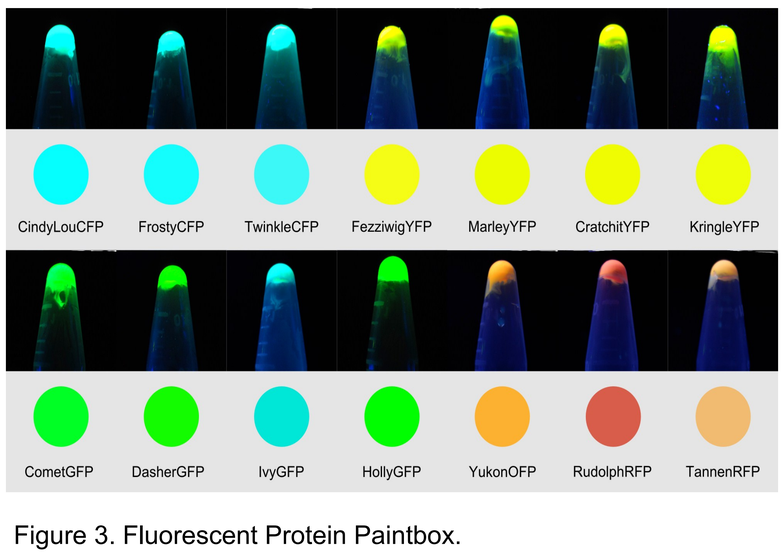

Figure 3. Fluorescent Protein Paintbox.

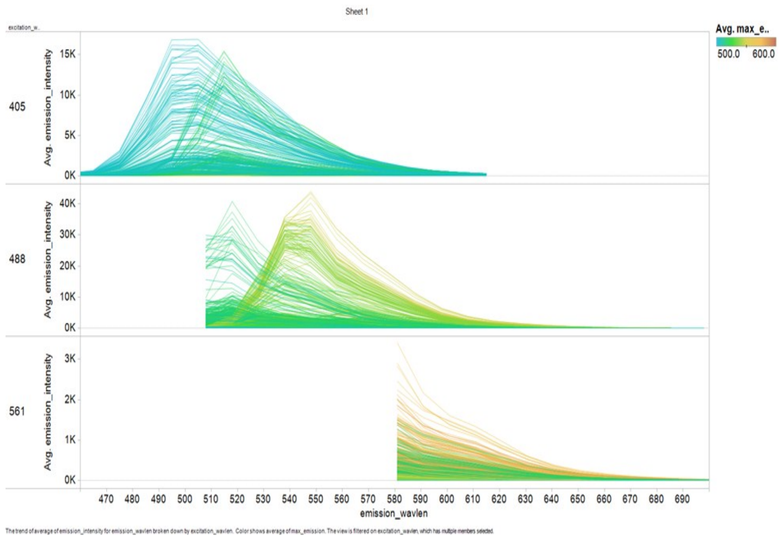

Figure 4. Overview of the spectral diversity in the starting library. Three scans at excitation wavelengths 405 nm, 488nm and 561nm over a broad range of emission wavelengths show the spectral diversity in the starting library.

Results

Analysis of sequence data correlated with spectral data suggested that specific amino acid residues were critical in determining the location of excitation and emission peaks in the spectra. There were 3 distinct sequence families of FPs within our group variants: (1) YFPs and UV-excitable GFPs, (2) Cyans and dual-excitable GFPs, and (3) a family including RFPs, OFPs and some GFPs. In sequence family (1) with GFPs and YFPs, two amino acid substitutions were identified that correlated with the observed spectral shifts. The amino acid C at position 151 was only found in UV-excitable GFPs, while 151Y was found in green-yellow FPs and YFPs (Fig. 5). Additionally, 206Y was only found in the more red-shifted YFPs, with 206V present in the GFPs and green-yellow FPs. To test whether these residues were causing spectral shifts, we introduced combinations of these two substitutions to our brightest GFP and YFP. The amino acid substitutions resulted in the expected changes in emission and excitation spectra. Similar analysis identified a single substitution, 146S/D, causing the major cyan to green shift in family (2).

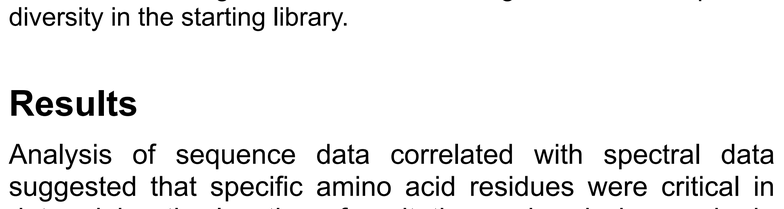

Figure 5. Identification of single amino acid substitutions that cause significant change in spectral properties. Analysis of a random library of substitutions, showing two single substitutions that cause spectral change.

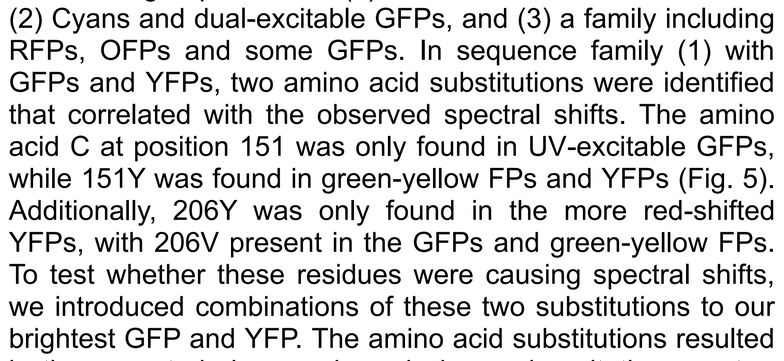

Figure 6. Overview of the spectra for three of the paintbox proteins used as compatible pairs.

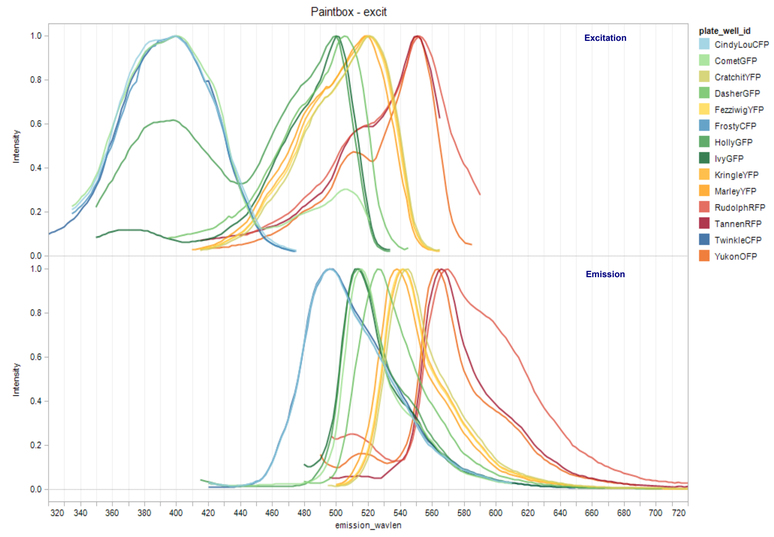

Figure 7. Excitation and emission spectra for Fluorescent Protein Paintbox.

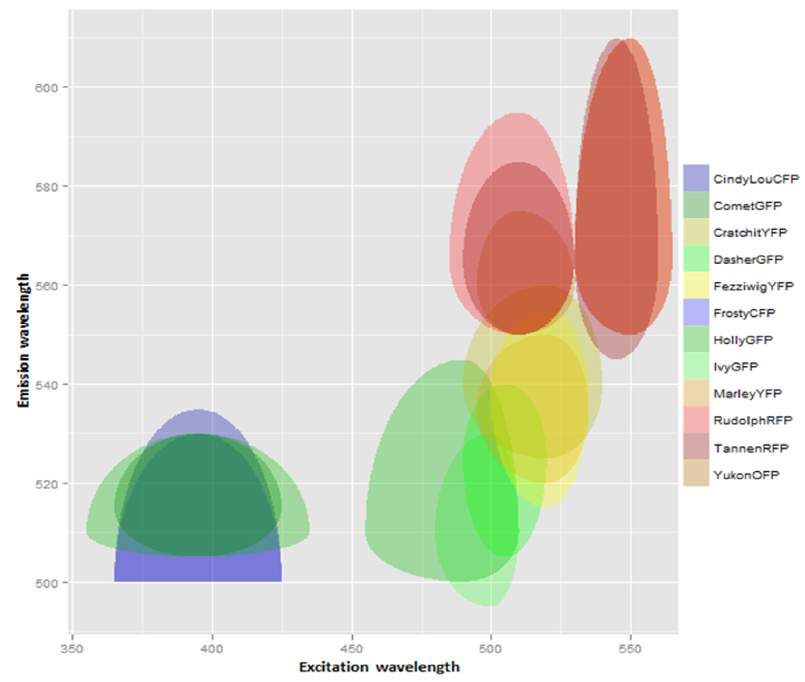

Figure 8. Refinement of a fluorescent protein "paintbox". Ovals show the 50% intensity contours for major excitation-emission peaks of all proteins in the paintbox, where color indicates the specific protein. Any FP with dual excit/emiss peaks has two ovals in this plot, one for each peak.

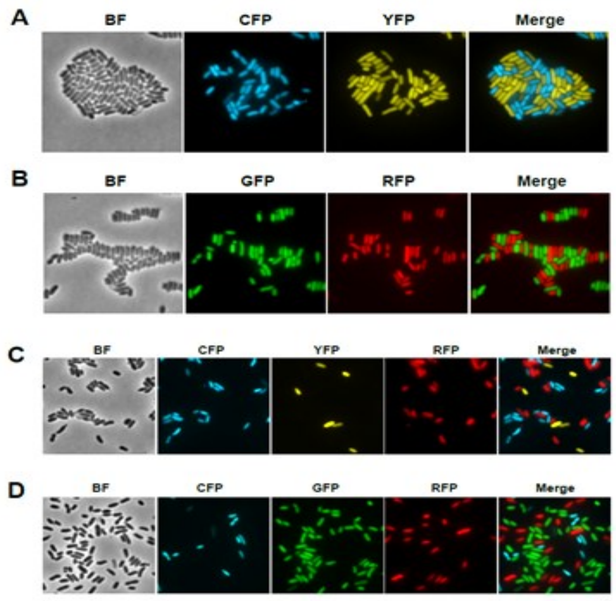

Figure 9. Multiple color imaging in live cells using FPs. Microscopy images of *E.coli* expressing FPs obtained under different excitation wavelengths demonstrate that pairs or triplets of our FPs can be used as distinguishable biosensors.

Conclusions

We have engineered a synthetic library encoding >1000 novel fluorescent and chromogenic proteins. Drawing from this library, we established a curated collection of fluorescent proteins encoding cross compatible optical properties and used machine learning and clustering algorithms to capture sequence-function correlations within the set.

DNA2.0's Protein Paintbox™ is available without intellectual property restrictions, for the Synthetic Biology research community to improve upon and to incorporate into innovative products. A subset of the Paintbox proteins reported here have been contributed to the public domain via the BioBrick Public Agreement. This open source protein library is a starting point for further engineering and refinement of free-to-use fluorescent proteins with useful properties.

DNA 2.0