Chapter 5

Ancestral Protein Reconstruction and Circular Permutation for Improving the Stability and Dynamic Range of FRET Sensors

Ben E. Clifton*, Jason H. Whitfield*, Inmaculada Sanchez-Romero, Michel K. Herde, Christian Henneberger, Harald Janovjak, and Colin J. Jackson

Abstract

Small molecule biosensors based on Förster resonance energy transfer (FRET) enable small molecule signaling to be monitored with high spatial and temporal resolution in complex cellular environments. FRET sensors can be constructed by fusing a pair of fluorescent proteins to a suitable recognition domain, such as a member of the solute-binding protein (SBP) superfamily. However, naturally occurring SBPs may be unsuitable for incorporation into FRET sensors due to their low thermostability, which may preclude imaging under physiological conditions, or because the positions of their N- and C-termini may be suboptimal for fusion of fluorescent proteins, which may limit the dynamic range of the resulting sensors. Here, we show how these problems can be overcome using ancestral protein reconstruction and circular permutation. Ancestral protein reconstruction, used as a protein engineering strategy, leverages phylogenetic information to improve the thermostability of proteins, while circular permutation enables the termini of an SBP to be repositioned to maximize the dynamic range of the resulting FRET sensor. We also provide a protocol for cloning the engineered SBPs into FRET sensor constructs using Golden Gate assembly and discuss considerations for in situ characterization of the FRET sensors.

Key words Ancestral protein reconstruction, Phylogenetic analysis, Protein engineering, Thermostability, Circular permutation, Förster resonance energy transfer, Fluorescence, Biosensor

1 Introduction

Optical biosensors, including those that rely on Förster resonance energy transfer (FRET sensors), allow robust, noninvasive quantification of small molecule dynamics in biological systems [1]. FRET is a physical phenomenon whereby excitation of a donor fluorophore results in nonradiative energy transfer to an acceptor

*These authors contributed equally.

Viktor Stein (ed.), *Synthetic Protein Switches: Methods and Protocols*, Methods in Molecular Biology, vol. 1596,
DOI 10.1007/978-1-4939-6940-1_5,

fluorophore in a distance-dependent manner, which results in fluorescence of the acceptor fluorophore. FRET sensors can be constructed by fusing a specific pair of fluorescent proteins to the termini of a suitable recognition domain, which undergoes a conformational change when it binds the small molecule of interest. This ligand-dependent conformational change alters the distance between the fluorescent proteins, causing an observable change in FRET efficiency, which can be observed as a change in the ratio of the fluorescence intensities of the donor and acceptor fluorophores. The dynamic range of the sensor is the maximum change in FRET efficiency between the unbound and ligand-bound states. In addition to having a large dynamic range, an ideal sensor would be specific for the target ligand and responsive over the physiological concentration range of the ligand; importantly, the sensor must also be stable for extended periods under the experimental conditions (e.g., temperature and pH) required for the biological system of interest. These properties of the sensor are largely determined by the choice of recognition domain.

The solute-binding protein (SBP) superfamily is one set of recognition domains that is commonly used in FRET sensors for small molecules [2–5]. SBPs exhibit high affinity and specificity toward a diverse array of ligands, and they undergo a large conformational change upon ligand binding, which can be transduced into an optical signal in a FRET sensor construct. However, the low thermostability of existing SBPs can limit the utility of the resulting sensors in biological environments, especially when destabilizing modifications to the recognition domain, such as circular permutation or specificity-switching mutations, are necessary to improve the dynamic range or specificity of the sensor [6].

Ancestral protein reconstruction (APR) is one method that has been shown to produce consistently thermostable proteins, with denaturation temperatures up to 40 °C greater than comparable mesophilic proteins [7–11]. We have previously used this method to reconstruct a thermostable SBP as a recognition domain for a robust FRET sensor for L-arginine [6]. The main steps involved in APR are: (1) collection and alignment of a sequence dataset representative of a protein family; (2) inference of a phylogenetic tree describing the evolutionary relationships between the protein sequences; (3) probabilistic reconstruction of the ancestral protein sequences; (4) synthesis and cloning of genes encoding the ancestral proteins; and (5) expression, purification, and biophysical analysis of the ancestral proteins.

It has been argued that the high thermostability of reconstructed ancestral proteins reflects the environment of the ancient, thermophilic organisms from which they originate [7, 8, 12]; thus, reconstruction of the ancestor of a sufficiently ancient protein family (for instance, one that predates the divergence of the major bacterial kingdoms) is a viable method for engineering a highly thermostable

protein [13]. Notably, computational evolutionary simulations [14] and the observation that ancestral proteins share similarities with consensus proteins in terms of both sequence and thermostability [8, 15, 16] suggest that the high thermostability of ancient proteins might also originate from bias in the maximum-likelihood method commonly used for APR, specifically, that more probable (often stabilizing) residues are chosen at every site, while less probable (often destabilizing) residues are neglected [14]. The implication of this potential bias for protein engineering is that APR may produce thermostable proteins regardless of their hypothetical age.

One problem that may be encountered in the construction of SBP-based FRET sensors is that the conformational change of the SBP may not translate into a change in distance between the N- and C-termini, which is often the case when the termini are located on the same domain of the bilobal SBP. A small change in the distance between the termini of the SBP results in a small change in FRET efficiency between the bound and unbound states of the sensor, which limits its dynamic range. In this circumstance, the dynamic range of the sensor can be improved by circular permutation, a method that allows the termini of a protein to be relocated [2, 6]. Circular permutation is achieved by rearranging the sequence of the SBP such that the original termini are connected by a flexible linker and new termini are created by disconnecting the sequence at a different location, which is chosen to maximize the distance change between the termini upon ligand binding. Since circular permutation is often destabilizing [17], the use of APR to obtain a thermostable SBP and the use of circular permutation to improve the dynamic range of the resulting FRET sensor are complementary.

Here, we show how APR can be used to create thermostable SBPs for the construction of robust FRET sensors. In this example, the maximum-likelihood statistical framework is used for phylogenetic analysis and reconstruction of ancestral sequences, although a variety of alternative methods are also available [18–20]. We focus on the reconstruction of ancestral SBPs, but similar approaches could be used to improve the thermostability of any protein, provided that a suitable sequence dataset can be obtained. We also show how the dynamic range of the sensor derived from an ancestral SBP can be improved by circular permutation prior to the insertion of the SBP into the FRET sensor construct.

2 Materials

2.1 Software

1. SeaView version 4.6 (http://doua.prabi.fr/software/seaview).
2. ProtTest version 3.4 (https://github.com/ddarriba/prottest3).
3. FigTree version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree).

4. PHYLIP version 3.695 (http://evolution.genetics.washington.edu/phylip.html).
5. PAML version 4.9 (http://abacus.gene.ucl.ac.uk/software/paml.html).
6. PyMOL version 1.8 (https://www.pymol.org).

2.2 Online Servers

1. NCBI BLAST and PSI-BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi).
2. UniProt-KB BLAST (http://www.uniprot.org/blast).
3. MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle).
4. PhyML (http://www.atgc-montpellier.fr/phyml).
5. Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2).
6. CPred (http://sarst.life.nthu.edu.tw/CPred).

2.3 Cloning

Prepare all solutions using ultrapure water.

1. Synthetic genes encoding the circularly permuted ancestral proteins (*see* Subheading 3.7 for design).
2. Primers to amplify the circularly permuted genes with flanking *Sap*I sites for Golden Gate assembly (*see* Subheading 3.7 for design).
3. High-fidelity DNA polymerase (e.g., Phusion Hot Start II Polymerase, Thermo Scientific).
4. dNTPs.
5. Thin-walled PCR tubes.
6. PCR thermocycler.
7. PCR purification kit (e.g., Wizard® SV Gel and Clean-Up System from Promega).
8. pDOTS4 or pDOTS10 plasmid [6] (*see* **Note 1**).
9. *Sap*I (10 U/μL) (e.g., from New England Biolabs).
10. T4 DNA ligase (400 U/μL) with 10× T4 DNA ligase buffer (e.g., from New England Biolabs).
11. Electrocompetent *E. coli* TOP10 cells (Invitrogen).
12. Electroporation cuvettes.
13. Electroporator.
14. YenB media: 7.5 g/L yeast extract, 8 g/L nutrient broth.
15. 5 mL culture tubes.
16. Luria-Bertani (LB) agar plates containing ampicillin (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, 15 g/L agar, 100 mg/L ampicillin).
17. Materials for colony PCR and analysis by gel electrophoresis.
18. PCR master mix with loading dye for gel electrophoresis (e.g., 2× PCR Super Master Mix from Biotool).

19. Primers for sequencing (*see* Subheading 3.8).
20. 1% agarose gel made with SB buffer (46 g/L boric acid, 8 g/L sodium hydroxide), stained with, e.g., GelRed™ (Biotium).
21. DNA ladder (e.g., GeneRuler 1 kb DNA ladder, Thermo Scientific).
22. Agarose gel electrophoresis apparatus.

3 Methods

3.1 Sequence Collection

Given a reference SBP with the desired binding specificity, reconstruction of an ancestral SBP with improved thermostability and the same binding specificity requires a dataset of protein sequences that are homologous to the reference SBP (*see* **Note 2**). These sequences can be identified using a BLAST (Basic Local Alignment Search Tool) search of protein sequence databases, such as the NCBI database of nonredundant protein sequences or the UniProt-KB database. The protein sequences should be selected to maximize sequence diversity and phylogenetic diversity, but should have the same binding specificity as the reference SBP (*see* **Notes 3** and **4**). The sequences should be evenly distributed throughout sequence space; in other words, clusters of sequences with very high identity (>90%) and outlier sequences with low identity to any other sequence in the dataset should be avoided. A small set of outgroup sequences should also be selected for rooting the phylogenetic tree (*see* **Note 5**). In the case of SBPs, a suitable outgroup would be an SBP homologous to the reference SBP, but with a different binding specificity; for example, in our recent work, anionic amino acid-binding proteins served as an outgroup for cationic and neutral amino acid-binding proteins [6].

As a guideline, between 50 and 250 sequences should be selected for the phylogenetic analysis. If too many sequences are included, the phylogenetic analysis will be computationally intensive, while if too few sequences are included, there may be insufficient data to reconstruct the ancestral sequences accurately, or the sequence diversity within the dataset may be insufficient to make the ancestral proteins more thermostable than the reference SBP.

3.2 Multiple Sequence Alignment

Before the sequences can be used for phylogenetic analysis, they must be collated in a multiple sequence alignment. In the phylogenetic analysis and subsequent reconstruction of ancestral sequences, it is assumed that each column in the multiple sequence alignment corresponds to a set of residues that originated from a common ancestor. If the alignment is incorrect, this assumption is violated; hence, the quality of the multiple sequence alignment is a critical factor in the success of the project. The number of gaps should be minimized by careful editing of the alignment to remove poorly

conserved N- and C-terminal regions and insertions that are unique to individual protein sequences, since PAML (the program used for reconstruction of ancestral protein sequences) will attempt to reconstruct one ancestral residue for each column in the alignment.

1. Align the protein sequences from Subheading 3.1 using the MUSCLE server [21] (*see* **Note 6**).
2. Open the alignment in SeaView [22].
3. Inspect the alignment carefully. If necessary, make manual adjustments to improve the alignment, particularly in regions with many gaps.
4. Trim the alignment, removing insertions that would not be present in the ancestral proteins of interest (*see* **Note 7**). Remove poorly conserved N- and C- terminal extensions, including the N-terminal signal peptides in the case of SBPs.
5. Save the alignment in PHYLIP interleaved format.

3.3 Model Selection

Phylogenetic analysis and reconstruction of ancestral protein sequences using the maximum-likelihood method requires a probabilistic model of protein evolution, which is needed to evaluate the likelihood of different evolutionary scenarios. The main component of this evolutionary model is an empirical substitution matrix that specifies the frequency of each amino acid and the substitution rate between each pair of amino acids. Some modifications are also available to make the basic evolutionary model more realistic; for example, different sites in the protein are often allowed to evolve at different rates by modeling rate heterogeneity using the discrete-gamma model (denoted +G) [23].

The first step in maximum-likelihood phylogenetic analysis is to choose the evolutionary model most appropriate for the given sequence dataset, which can be achieved using programs such as ProtTest [24] (*see* **Note 8**). ProtTest evaluates different models using the Akaike information criterion (AIC), a metric that accounts for the likelihood of the tree inferred using the model (the goodness of fit) and the number of model parameters (to penalize overfitting).

1. Open the ProtTest graphical interface and load the alignment file.
2. Select "Compute likelihood scores" under the Analysis menu.
3. Select the substitution matrices to evaluate. At a minimum, select the general-purpose JTT, WAG, and LG matrices; most of the remaining matrices are based on specific types of proteins (for example, MtArt is based on sequence data from invertebrate mitochondrial proteins).
4. Ensure that the +I, +G, +I+G, and empirical amino acid frequency options are selected to evaluate these modifications.

5. Select "BIONJ tree" as the base tree for likelihood calculations (*see* **Note 9**).
6. After the computation is complete, select "Results" under the Selection menu. Identify the model with the lowest AIC, which will be used for the subsequent analysis.

3.4 Phylogenetic Tree Inference

Given an alignment of protein sequences and a probabilistic model of protein evolution, a phylogenetic tree can be inferred using the maximum-likelihood method. The goal of phylogenetic inference using ML is to identify the tree topology (branching pattern) and parameters (for example, branch lengths) that optimize the likelihood function. Likelihood is defined in this context as the probability of observing the data—that is, the probability that the present-day sequences would have evolved—given the tree topology, tree parameters, and an evolutionary model. In practice, the maximum-likelihood tree is inferred by first estimating the correct tree using a computationally simple method such as the BIONJ method; heuristic tree-searching algorithms are then applied to modify the topology of the tree and increase its likelihood until convergence on the maximum-likelihood tree is achieved.

Inference of a phylogenetic tree for the purpose of APR should be considered an iterative process; problematic sequences should be identified and removed, and alternative sequences should be added before repeating the phylogenetic analysis. The final tree should be biologically plausible, supported by high bootstrap values, and robust to variations in the evolutionary model used.

1. Open the PhyML [25] web interface and upload the sequence alignment.
2. Select "Amino-Acids" for data type.
3. Under "Substitution Model," specify the substitution model chosen in Subheading 3.3. +I denotes an estimated proportion of invariant sites; +G denotes an estimated gamma-shape parameter, and +F denotes empirical amino acid frequencies.
4. Under "Tree Searching," select "BIONJ" for the starting tree and "NNI + SPR" for the type of tree improvement (*see* **Note 10**).
5. Under "Branch Support," select the "aLRT SH-like" branch support test. Alternatively, for the final tree, select a bootstrap calculation with 100 replicates.
6. Once the analysis has completed, open the maximum-likelihood tree in FigTree. The tree can be rooted by selecting the branch connecting the ingroup and outgroup sequences and clicking the "Reroot" button.
7. Inspect the tree to identify problems such as long branches and low branch support values, and rectify these problems by repeating the multiple sequence alignment and phylogenetic analysis using a modified sequence dataset if necessary (*see* **Notes 11** and **12**).

8. Assess the robustness of the ML tree to variations in the substitution model by repeating the tree inference in PhyML using alternative substitution matrices that scored highly in Subheading 3.3. The resulting trees should be very similar to the maximum-likelihood tree.

3.5 Reconstruction of Ancestral Sequences

1. Ensure that the alignment file and tree file are properly formatted for input into PAML [26] (*see* **Note 13**).
2. Set up the program by modifying the control file ("codeml.ctl"), as shown in Table 1.
3. Run the program from the command line using the command codeml.

Table 1
Variables in the control file for ancestral protein reconstruction using PAML

seqfile	Filename of alignment
treefile	Filename of tree
outfile	Filename of output
noisy	9
verbose	**2**
runmode	0
seqtype	2
clock	0
aaRatefile	Filename of rate matrix (e.g., lg.dat for the LG matrix)
model	2
Mgene	0
fix_alpha	0 to use +G model, else 1
alpha	0.5 to use +G model, else 0
Malpha	0
ncatG	Number of rate categories to use for the +G model
getSE	0
RateAncestor	1
Small_Diff	.5e-6
cleandata	**0**
method	0

Deviations from the settings in the default control file are indicated in bold. The remaining variables in the control file are not applicable to amino acid-based analysis and can be removed

4. Open the result file ("rst") and scroll down to the text "tree with node labels." Copy the tree into a new file and open the file in FigTree. Identify the ancestral nodes of interest (*see* **Note 14**) and record the node labels that identify them.
5. Search the result file for the text "node #x" (where x is the number identifying the ancestral node) to find the maximum-likelihood ancestral sequence associated with that ancestral node.
6. Open the alignment file in SeaView and add the ancestral protein sequences to the alignment. Edit the ancestral sequences to remove any remaining insertions that are artifacts of the reconstruction process (*see* **Note 7**).

3.6 Characterization of Ancestral Proteins

Once the ancestral protein sequences have been obtained, they are back-translated into nucleotide sequences and codon-optimized for expression in *Escherichia coli*, and the genes are synthesized. The synthetic genes are cloned into expression vectors, and the ancestral proteins are expressed in *E. coli* and purified. The thermostability of the ancestral proteins can be assessed using methods such as circular dichroism spectroscopy, differential scanning fluorimetry (DSF), or differential scanning calorimetry. The binding specificity of the ancestral proteins can be assessed using methods such as isothermal titration calorimetry (ITC), DSF, or (in some cases) fluorescence spectroscopy. If the ancestral SBPs have high thermostability and the desired binding specificity, they are strong candidates for circular permutation and incorporation into FRET sensor constructs.

3.7 Design of Circularly Permuted SBPs

The circular permutation of SBPs has been described by Okada et al. [2]. It is critical to identify sites for the new N-and C- termini that will produce a sensor with high dynamic range without negatively affecting the binding affinity or stability of the protein.

1. Using a crystal structure or homology model of the ancestral SBP (created using the Phyre2 server [27], for example), select the positions of the new N- and C-termini. The new termini must be located on different lobes of the SBP to maximize the change in their relative positions due to the ligand-induced conformational change. This can be achieved by deleting a section of the protein that links the two lobes (i.e., a hinge strand) (*see* **Note 15**). The result is a theoretical protein with the original N- and C-termini, and additional N*- and C*-termini, i.e., two protein fragments.
2. Design a linker sequence to fuse the original N- and C-termini of the SBP. Each repeat of a flexible (GGS)*n* linker is approximately 11.4 Å in length. Measure the distance between the N- and C-termini of the SBP; the linker should have enough (GGS)*n* repeats to bridge this distance.

3. Theoretically construct the circularly permuted SBP sequence by concatenating the three protein fragments: the N*-terminus to C-terminus fragment, the flexible linker, and the N-terminus to C*-terminus fragment.
4. Back-translate the circularly permuted SBP sequence to obtain a nucleotide sequence codon-optimized for expression in *E. coli*, and synthesize the gene (*see* **Note 16**).
5. Design primers to amplify the gene with flanking *Sap*I sites for insertion into pDOTS4 or pDOTS10. The forward primer contains the sequence 5′-GCTCTTCAATC-3′ followed by the first 10–15 nucleotides of the circularly permuted gene. The reverse primer contains the sequence 5′-GCTCTTCCGAG-3′ followed by the first 10–15 nucleotides of the reverse complement of the circularly permuted gene (*see* **Note 17**). The length of the primers should be chosen such that the melting temperatures of the primers are approximately 65 °C.

3.8 Cloning into a FRET Sensor Construct

Carry out all procedures at room temperature unless otherwise specified.

1. Using standard PCR with a high-fidelity DNA polymerase (e.g., Phusion Hot Start II), amplify the SBP gene using the cloning primers from Subheading 3.7. Purify the PCR product using a PCR purification kit, according to the manufacturer's instructions.
2. In a thin-walled PCR tube, chilled on ice, mix 2 μL of 10× T4 DNA Ligase buffer, 1 μL 400 U/μL T4 DNA ligase, 1 μL 10 U/μL *Sap*I, 100 ng pDOTS4 or pDOTS10, a fivefold molar excess of the purified PCR product from the previous step, and water to give a final volume of 20 μL.
3. Perform the Golden Gate assembly reaction using the following thermocycling protocol: 33 cycles of (37 °C for 2 min, then 16 °C for 3 min); 55 °C for 10 min; 80 °C for 5 min; then hold at 4 °C.
4. Transform one aliquot of *E. coli* TOP10 cells with 2.5 μL of the Golden Gate assembly reaction mixture by electroporation (*see* **Note 18**). After electroporation, resuspend the cells in 1 mL YenB and incubate with shaking at 37 °C for 1 h. Plate 100 μL of the culture on an LB/ampicillin agar plate and incubate at 37 °C overnight (*see* **Note 19**).
5. Use colony PCR and Sanger sequencing of the resulting PCR products to confirm correct insertion of the gene into the vector (*see* **Note 20**).

3.9 Characterization of the FRET Sensors In Vitro and In Situ

After expression of the FRET sensor in *E. coli* and subsequent purification (*see* **Notes 21** and **22**), the dynamic range of the sensor and the affinity of the sensor for its ligand should be measured

using fluorescence titrations. ITC can also be used to accurately measure the dissociation constant (K_d) of the sensor and the proportion of the sensor that is active.

The approach for in situ characterization of the FRET sensor will depend on where the sensor will be used to measure the resting level of its ligand, its dynamic changes or distribution. For instance, this could be intra- or extracellularly, in cell culture systems, or more complex tissue such as acute slices or in vivo. While intracellular ligand measurements typically require the sensor to be expressed by the cell of interest using an appropriate vector, extracellular sensor localization can be achieved by expression using plasma membrane targeting sequences and anchoring motifs [4]. However, sensor expressed endogenously and targeted to the membrane can be exposed to intracellular ligands during transit to the membrane. Both sensor populations, surface-presented and intracellular, will generate fluorescence that cannot be easily separated using diffraction-limited microscopy techniques, thus potentially limiting the interpretation of data. An alternative approach to target FRET exclusively to extracellular space relies on a biotin-streptavidin anchoring strategy [6, 28]. In our previous work, a biotin tag was incorporated into the FRET sensor using the pDOTS10 plasmid. Extracellular immobilization of the sensor was then achieved by linking the sensor to surface proteins, biotinylated using commercially available streptavidins and NHS-ester activated biotinylation reagents, via streptavidin [6].

Classically, FRET sensor imaging is performed using single photon excitation and quantification of donor/acceptor fluorescence intensity, donor lifetime, or emission spectra [4]. In thicker preparations like acute tissue slices and in vivo, two-photon excitation using near-infrared pulsed lasers has superior performance and provides good optical access to deeper structures [29] at the potential cost of reduced spectral separation of FRET donor/acceptor excitation (e.g., ECFP/EYFP) [30]. Using pulsed lasers, typically in the MHz frequency range (e.g., Ti:sapphire lasers around 80 MHz), also enables fluorescence lifetime imaging using time-correlated single photon counting (TCSPC-FLIM) of FRET donor fluorescence, a powerful tool to study binding of a ligand to a FRET sensor. Both modes of excitation allow monitoring of sensor fluorescence at high spatial and temporal, micrometer and millisecond, resolution.

4 Notes

1. pDOTS4 and pDOTS10 are mother plasmids for the cloning of SBPs to generate FRET sensors [6]. The pDOTS4 backbone is based on a pRSET plasmid [4] (pRSET FLIPE-600n, Addgene #13537, courtesy of Wolf Frommer) containing a 6× His-Tag at the N-terminus. For the construction of

pDOTS4, the native *Sap*I site was first removed by site-directed mutagenesis. To adapt this plasmid for Golden Gate assembly [31], the *ybeJ* coding sequence in pRSET FLIPE-600n was removed and replaced with *Sap*I-*Not*I-*Sap*I restriction sites. The SBP gene can be introduced into the pDOTS4 plasmid using Golden Gate assembly (*see* Subheadings 3.7 and 3.8); the SBP will be fused to a 6× His-Tag and ECFP at the N-terminus and to Venus fluorescent protein at the C-terminus. To generate pDOTS10, a biotin tag followed by a GSSG linker, synthesized de novo (Epoch Life Science) based on the sequence from the PinPoint™ Xa-1 plasmid, was amplified by PCR and cloned into pDOTS4 between the 6× His-Tag and ECFP using the *BamH*I restriction site. The SBP gene can be introduced into the pDOTS10 plasmid using Golden Gate assembly (*see* Subheadings 3.7 and 3.8); the SBP will be fused to a 6× His-Tag, biotin tag, GSSG linker, and ECFP at the N-terminus, and to Venus fluorescent protein at the C-terminus. pDOTS4 and pDOTS10 will be made available via Addgene.

2. Usually, protein sequences are used for ancestral reconstruction rather than nucleotide sequences because they are more highly conserved and therefore retain more phylogenetic signal over longer evolutionary distances; protein sequences are under stronger selection than nucleotide sequences because many nucleotide substitutions are synonymous and have low fitness costs. However, nucleotide- or codon-based phylogenetic analyses are also possible.
3. Information about sequence-function relationships from crystal structures of the reference SBP or other experimental data is useful for identifying protein sequences that are likely to have the same or similar binding specificity. If the structure of the reference SBP is unknown, it may be possible to predict binding site residues using the structures of homologous SBPs; these residues should be conserved when the binding specificity is conserved.
4. Several strategies can be used to increase the diversity of a sequence dataset: PSI-BLAST, which performs iterative BLAST searches, can be used to find sequences more distantly related to the reference SBP; multiple SBP sequences can be used as input for the BLAST search (provided that they are actually homologous); programs such as CD-HIT [32] can be used to filter redundant sequences out of the dataset; BLAST searches can be restricted to certain taxonomic groups to obtain sequences from phylogenetically diverse organisms. Phylogenetic diversity in the sequence dataset is desirable to enable reconstruction of more ancient (and more thermostable) ancestral proteins.

5. The root of a phylogenetic tree is the node corresponding to the ancestor of all sequences in the tree. Rooted trees show the position of the root and the direction of evolution, whereas the unrooted trees produced by the maximum-likelihood method contain no information about the direction of evolution. This problem can be circumvented by the inclusion of outgroup sequences, which are known to be more distantly related to the sequences of interest (ingroup sequences), in the phylogenetic analysis. The root of the tree will be located somewhere on the branch joining the ingroup and outgroup sequences.
6. Various alignment programs with different specialities are available. In addition to MUSCLE, these include MAFFT [33], T-COFFEE [34], and PRANK [35]. In particular, PRANK may be useful for APR, since its phylogeny-aware algorithm is able to distinguish independent insertion and deletion events that would result in erroneous inferences of homology in conventional multiple sequence alignment programs [36].
7. Programs for automatic curation of multiple sequence alignments are available (such as Gblocks [37]), but these programs may be too zealous in removing poorly aligned regions of the alignment for this particular application. Although insertions unique to individual sequences or a small cluster of sequences can be safely removed, care should be taken not to delete columns that might contain residues present in the ancestral proteins of interest. Since phylogenetic information is needed to make this judgement, the alignment should be edited conservatively before the reconstruction of ancestral sequences. Artifactual insertions in the ancestral sequences can be removed afterward, once the phylogeny is known.
8. More recent implementations of maximum-likelihood methods may give superior performance in terms of computational speed and/or accuracy. Alternative software options include IQ-TREE [38] for model selection, IQ-TREE or RAxML [39] for tree inference, and FastML [40] for reconstruction of ancestral sequences.
9. The assumption implicit in this choice is that the BIONJ tree is a good enough approximation to the maximum-likelihood tree that the substitution model of best fit will be identical for both trees. It is more accurate to use the maximum-likelihood tree as the base tree for the likelihood calculations, but the computation will be much slower.
10. If these settings are used, the tree search will be deterministic, that is, the same alignment will always produce the same tree. This can be problematic, since the tree search may stall at a locally optimal tree and fail to converge on globally optimal tree (the true maximum-likelihood tree). It is therefore good

practice to repeat the analysis in PhyML with randomized starting trees; if the same tree topology is recovered each time, it is more likely that the globally optimal tree has been recovered.

11. Long branches on a phylogenetic tree arise from the inclusion of highly divergent sequences and cause a systematic error known as long branch attraction. If possible, long branches should be broken up by including additional sequences that have higher identity to the divergent sequences; otherwise, the divergent sequences may need to be removed.

12. Phylogenetic trees inferred using maximum-likelihood can be validated in several ways. First, the topology of the tree should be robust to variations in the evolutionary models used in the calculation; the phylogenetic analysis should be repeated using alternative substitution matrices that scored highly in Subheading 3.3, and the resulting tree topologies should be very similar to the maximum-likelihood tree. Second, statistical support for individual branches in the tree should be assessed using the bootstrap method, which involves repeating the phylogenetic analysis many times (usually at least 100) using multiple sequence alignments generated by random resampling of the complete alignment, to determine whether the same tree topology can be reproduced even when some of the data are removed or duplicated. The bootstrap value of a branch is the proportion of the bootstrap trees that contain the same bifurcation of sequences as denoted by that branch. Branches with bootstrap values <70% should be viewed with suspicion. Finally, if the tree describes the evolution of orthologous sequences, it should be consistent with established species-based trees (although this criterion is less applicable to SBPs due to the prevalence of horizontal gene transfer).

13. In our experience, the file formatting requirements of PAML are the most common cause of any errors encountered while running the program. The alignment must be in PHYLIP interleaved format. This format must be specified by adding the letter "I" (for "interleaved") to the end of the first line (after the number of sequences and number of columns). There must also be two spaces between each sequence name and the corresponding sequence. The tree must be in Newick format. PAML can only parse trees obtained from PhyML if they do not contain branch support values. Trees with branch support values can be reformatted using the program Retree in the PHYLIP package; open the tree in Retree and write the (unrooted) tree to a new file without further modification.

14. Ideally, multiple ancestral nodes between the reference SBP and the last common ancestor of the sequences of interest would be selected and characterized experimentally.

15. Previous examples of circularly permuted SBPs [2, 6] can be used as a guide to identify potential sites for circular permutation, provided that the proteins share sufficient homology. Otherwise, select sites in disordered loops of the protein (as indicated by high B-factors in the crystal structures of related proteins), and avoid the removal of any residues that contribute to secondary structure. Multiple sites may need to be tested. As a further check, the CPred server [41] can be used to evaluate the viability of potential sites for circular permutation.
16. Ensure that no *Sap*I sites (GCTCTTC) are present in the synthetic gene, as these will prevent successful cloning into pDOTS4 or pDOTS10.
17. Cloning into the pDOTS4 or pDOTS10 plasmids by Golden Gate assembly results in addition of an N-terminal isoleucine residue and a C-terminal leucine residue to the cloned gene, which should be taken into account if the linkers between the fluorescent proteins and recognition domain are optimized.
18. Competent cells with high transformation efficiency (10^7–10^8 cfu per μg DNA) are required for this step. If the transformation efficiency of the competent cells is low, plate the entire volume of electroporated cells by pelleting the cells ($18000 \times g$ on a benchtop centrifuge for 1 min), discarding the supernatant, resuspending the pellet in 100 μL YenB media, and plating the resulting suspension.
19. If no colonies are observed, the ligation by T4 DNA ligase may have been unsuccessful; if the colonies only contain empty vectors, digestion of the template gene by *Sap*I may have been unsuccessful.
20. We recommend sequencing in two reactions, first using a T7-specific primer (forward) in combination with a SBP-specific primer (reverse), then using an SBP-specific primer (forward) with a T7-specific primer (reverse), because the length of the full construct exceeds the length of most Sanger sequencing reads.
21. High levels of expression can be achieved using pDOTS4 or pDOTS10 by incubation of *E. coli* BL21(DE3) transformants in auto-induction media at 20 °C for 72 h. Expression can be monitored during growth by removing a 1 mL aliquot from the culture, centrifuging the sample to pellet the cells, resuspending the pellet in 1 mL water, and measuring the fluorescence spectrum of the resulting cell suspension. With excitation at 433 nm, an emission peak at 476 nm corresponding to ECFP and an emission peak at 525 nm corresponding to Venus should be apparent. Since maturation of the Venus fluorophore is slower than maturation of the ECFP fluorophore, the Venus peak may not be observed until after 48 h.

22. The FRET sensors are expressed with N-terminal hexahistidine tags, allowing purification by nickel affinity chromatography. In our experience, further purification by size-exclusion chromatography can also improve the dynamic range of the sensor. Finally, extensive dialysis may be required to remove any endogenously bound ligands, as sensor molecules that are already ligand-bound do not exhibit a ligand-dependent change in FRET, reducing the observed FRET efficiency.

Acknowledgments

Research was funded by Human Frontiers Science Program Young Investigator Award (HFSP to H.J., C.H., and C.J.J., grant number: RGY0084/2012), German Academic Exchange Service (DAAD-Go8) Travel Fellowship (to C.H. and C.J.J.), NRW-Rückkehrerprogramm (to C.H.), and German Research Foundation (DFG, SFB1089 B03, SPP1757 HE6949/1-1 and HE6949/3-1, all to C.H.).

References

1. Wang H, Nakata E, Hamachi I (2009) Recent progress in strategies for the creation of protein-based fluorescent biosensors. Chembiochem 10:2560–2577
2. Okada S, Ota K, Ito T (2009) Circular permutation of ligand-binding module improves dynamic range of genetically encoded FRET-based nanosensor. Protein Sci 18:2518–2527
3. Hires SA, Zhu Y, Tsien RY (2008) Optical measurement of synaptic glutamate spillover and reuptake by linker optimized glutamate-sensitive fluorescent reporters. Proc Natl Acad Sci 105:4411–4416
4. Okumoto S, Looger LL, Micheva KD et al (2005) Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. Proc Natl Acad Sci 102:8740–8745
5. Dwyer MA, Hellinga HW (2004) Periplasmic binding proteins: a versatile superfamily for protein engineering. Curr Opin Struct Biol 14:495–504
6. Whitfield JH, Zhang W, Herde MK et al (2015) Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. Protein Sci 24:1412–1422
7. Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704–707
8. Akanuma S, Nakajima Y, Yokobori S et al (2013) Experimental evidence for the thermophilicity of ancestral life. Proc Natl Acad Sci USA 110:11067–11072
9. Risso VA, Gavira JA, Mejia-Carmona DF et al (2013) Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. J Am Chem Soc 135:2899–2902
10. Perez-Jimenez R, Inglés-Prieto A, Zhao Z-M et al (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat Struct Mol Biol 18:592–596
11. Risso VA, Gavira JA, Sanchez-Ruiz JM (2013) Thermostable and promiscuous Precambrian proteins. Environ Microbiol 16:1485–1489
12. Gaucher EA, Thomson JM, Burgan MF et al (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285–288
13. Cole MF, Gaucher EA (2011) Utilizing natural diversity to evolve protein function: applications towards thermostability. Curr Opin Chem Biol 15:399–406
14. Williams PD, Pollock DD, Blackburne BP et al (2006) Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2:e69
15. Bershtein S, Goldin K, Tawfik DS (2008) Intense neutral drifts yield robust and evolvable consensus proteins. J Mol Biol 379:1029–1044

16. Risso VA, Gavira JA, Gaucher EA et al (2014) Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. Proteins 82:887–896
17. Heinemann U, Hahn M (1995) Circular permutation of polypeptide chains: implications for protein folding and stability. Prog Biophys Mol Biol 64:121–143
18. Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nat Rev Genet 13:303–314
19. Merkl R, Sterner R (2016) Ancestral protein reconstruction: techniques and applications. Biol Chem 397:1–21
20. Yang Z (2014) Molecular evolution: a statistical approach. Oxford University Press, Oxford
21. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797
22. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224
23. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314
24. Darriba D, Taboada GL, Doallo R et al (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165
25. Guindon S, Dufayard J, Lefort V et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321
26. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591
27. Kelley LA, Mezulis S, Yates CM et al (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10: 845–858
28. Okubo Y, Sekiya H, Namiki S et al (2010) Imaging extrasynaptic glutamate dynamics in the brain. Proc Natl Acad Sci USA 107: 6526–6531
29. Helmchen F, Denk W (2005) Deep tissue two-photon microscopy. Nat Methods 2: 932–940
30. Hashimoto H, Isobe K, Suda A et al (2010) Measurement of two-photon excitation spectra of fluorescent proteins with nonlinear Fourier-transform spectroscopy. Appl Optics 49:3323–3329
31. Engler C, Kandzia R, Marillonnet S (2008) A one pot, one step, precision cloning method with high throughput capability. PLoS One 3:e3647
32. Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152
33. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780
34. Di Tommaso P, Moretti S, Xenarios I et al (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 39:13–17
35. Löytynoja A, Goldman N (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 11:579
36. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632–1635
37. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577
38. Nguyen LT, Schmidt HA, Von Haeseler A et al (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32: 268–274
39. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313
40. Ashkenazy H, Penn O, Doron-Faigenboim A et al (2012) FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res 40:580–584
41. Lo WC, Wang LF, Liu YY et al (2012) CPred: a web server for predicting viable circular permutations in proteins. Nucleic Acids Res 40(W1):W232–W237