Cell

EXTENDED EXPERIMENTAL PROCEDURES

Statistical Coupling Analysis

Statistical coupling analysis (SCA) was carried out on a multiple sequence alignment (MSA) of 418 members of the DHFR protein family using a process described in (Halabi et al., 2009) and implemented in version 4.0 of the SCA MATLAB toolbox (available from the Ranganathan laboratory website). Briefly, the basic analytic procedure involves three steps: (1) compute a conservation-weighted correlation matrix that estimates the degree of coevolution between each pair of sequence positions in the MSA, (2) carry out spectral (or eigenvalue) decomposition of this matrix and identify its statistically significant top eigenmodes, and (3) examine the pattern of positional contributions to these top eigenmodes to deduce the number and composition of groups of coevolving amino acid positions (sectors). In the case of DHFR, the spectral analysis indicates five significant top eigenmodes that represent a single protein sector. To systematically determine the composition of the sector, we fit histograms of the positional weights in these top five eigenmodes to the Students t-distribution and used probably density cutoffs in the tail of the distributions ranging from 0.005 to 0.015 to identify the sector residues. The t-distribution was selected empirically by quality of fit to the top eigenmodes and is simply meant to provide a well-defined basis for sector identification; no mechanistic basis for the use of this distribution is implied. Importantly, sectors identified by this method are consistent with previous definitions (Halabi et al., 2009; Lee et al., 2008). Consistent with the basic principle of sectors defined in (Halabi et al., 2009), the DHFR sector comprises a relatively small subset of total amino acid positions (13 to 31% based on the cutoffs indicated) and comprises a physically contiguous network of residues in the tertiary structure (Figure 1C). A MATLAB script comprising SCA for the DHFR family and the MSA will be made available for download on the Ranganathan lab website.

Chimera Construction

The DHFR/LOV2 fusions were constructed by PCR stitching performed in two consecutive rounds of PCR with overlapping oligonucleotides. In the first round of PCR, the portions of DHFR both N and C-terminal to the LOV2 insertion site were separately amplified from *E. coli* wild-type DHFR contained in the plasmid pHis8-3 (construct previously described in Lee et al. [2008]) using KOD polymerase (Novagen). These PCR products contained 15 or 16 base pair regions that overlap to the 5′ and 3′ regions of the LOV2 gene. The LOV2 domain was amplified from a previous construct (the A120 DHFR/LOV2 fusion in pHis8-3, also described in Lee et al. [2008]). The full-length gene product was then assembled from these three pieces (the 5′ end of DHFR, the LOV2 domain, and the 3′ end of DHFR) by amplification with flanking primers and cloned into pHis8-3 using the NcoI and XhoI restriction sites. To construct the barcoded constructs, the full-length DHFR/LOV2 fusions were PCR amplified from pHis8-3 using external flanking primers (the appropriate barcode was included in the 3′ antisense flanking primer, following the DHFR stop codon) and subcloned into the vector pACYC Duet-1 (Novagen) using either the BamHI/SalI or BamHI/NotI restriction sites. This construct contains the DHFR variant in multiple cloning site 1, and thymidylate synthase (a second enzyme needed for complementation of the *E. coli* folate auxotroph) in multiple cloning site 2, both under control of the T7 promoter. This barcoded construct co-expressing both DHFR and thymidylate synthase was used for all selection experiments in vivo.

DNA Barcode Strategy

The plasmid DNA for each of the individual DHFR variants (both point mutants and the library of DHFR/LOV2 fusions) and the PCR products generated for each experimental time point or condition were labeled using unique five nucleotide DNA barcodes. These barcodes were designed to specify the variant identity or experimental condition in four nucleotides (using base-four arithmetic), with the fifth as an error-detecting checksum. The code was specified as follows: A = 0; C = 1; G = 2; T = 3. By this rule, an identity of 255 (in base 10 arithmetic) would be encoded by 3,333 ((64*3) + (16*3) + (4*3) + (1*3)) = 255, or TTTT. The checksum (fifth base) was calculated as the total sum of the four base barcode, modulo 4. This strategy permits encoding of 4^4 or 256 unique sequences. The barcode collection was further filtered to remove strings of 3-base repeats, as well as barcodes beginning with "A" (to prevent a string of "A" s arising, as the barcode was following the stop codon "TAA"). This left a remainder of 152 barcodes, of which 87 were randomly chosen to label the set of 70 DHFR/LOV2 chimeras and 15 DHFR point mutants and control constructs. Each barcode was a minimum Hamming distance of two from any other to ensure that the barcodes would not be interchangeable by a single mutation or sequencing misread.

Auxotroph Rescue Assay

All relative growth rate measurements were performed in the *E. coli* folate auxotroph strain ER2566 *ΔfolA ΔthyA*, previously developed in the Benkovic lab and described in (Saraf et al., 2004). To initiate the assay, 0.5 ml overnight cultures in rich media (LB with 50 mg/ml thymidine and 30 mg/ml chloramphenicol selection) were inoculated with streaks of ER2566 *ΔfolA ΔthyA* cells transformed with a given DHFR variant. Cultures were grown in 96-well deep well plates on a plate shaker at 37°C overnight. The cultures were harvested and washed two times with 0.5 ml minimal media as described previously in (Saraf et al., 2004), thymidine was omitted from the media as thymidylate synthase was coexpressed from the plasmid encoding DHFR) and resuspended in 0.5 ml MMA. The washed, resuspended cells were then used to inoculate a second round of 1 ml overnight cultures in MMA with chloramphenicol (30 mg/ml). This second round of cultures was grown for 18 hr at 30°C to allow the auxotroph to adapt to the new media and growth conditions, with the aim of preventing diauxic shift and/or variable lag phase times during the actual selection step. Following

adaptation, the culture densities (optical density at 600 nm) were measured in a plate reader (100 ul of culture in a 96-well flat bottom plate). The cultures for all 86 DHFR variants (mutants and DHFR/LOV2 fusions) were mixed in equal ratios by density and the mixed culture was used to inoculate two 50 ml cultures of MMA in baffled 250 ml flasks at a starting density of 0.04. The remainder of the mixed cells was taken as the t = 0 hr time point. All media for the dilution and growth steps was pre-warmed to 30°C to minimize temperature shock. These cultures were grown for 24 hr at 30°C, with one flask kept in a dark incubator and one under lit conditions (light provided by a 125 W daylight broad spectrum growlight [Hydrofarm, Petaluma, CA]). Samples were taken from each flask (10 mls each) at 4, 8, 16 and 24 hr. A single serial dilution into fresh media was made (to a starting density of 0.4) at t = 8 hr. The cells for each time point were harvested by centrifugation, resuspended in cell resuspension buffer (Wizard), and stored at −20°C for sequencing sample preparation.

Sample Preparation for Solexa Sequencing

Plasmid DNA for each time point of the selection experiment was prepared by miniprep (Wizard) and used as a template for generating the 250 base pair PCR product for Solexa sequencing. Two rounds of consecutive PCR with KOD polymerase (Novagen) were performed to add the adaptor regions necessary for the sequencing reaction as well as the barcodes specifying the time point and experimental condition. The first round of PCR used 1 μl of plasmid DNA (at roughly 20 ng/μl) as a template and added the barcode and a portion of the adaptor. The thermal cycling schedule used was: an initial 2 min at 95°C, followed by 25 rounds of amplification consisting of 20 s at 95°C, 10 s at 52°C, and 15 s at 70°C. This PCR product was then used as the template for a second round of PCR (performed as above, but with an annealing temperature of 55°C) that added the remainder of the adaptor region. The final PCR products were purified using the Zymo clean and concentrator kit and quantified by absorbance at 260 nm. The PCR products for each experimental time point and condition were then mixed in equal ratios, and this final mixture was gel purified. Concentration for the final sequencing reaction was measured by both band intensity on an agarose gel (in comparison to a standardized ladder) and by picogreen dsDNA quantitation reagent (Molecular Probes). The final PCR product was sequenced on a single lane of a version 2 (repeat 1) or version 4 (repeats 2 and 3) flowcell, using version 4 sequencing reagents on the Genome Analyzer IIx (Illumina) (see Table S6).

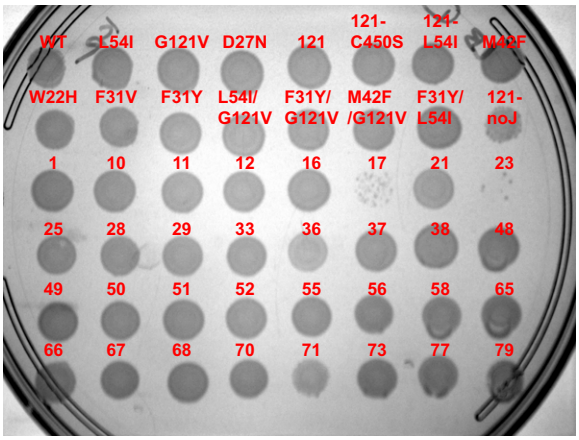
SUPPLEMENTAL REFERENCES

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774–786.

Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and Ranganathan, R. (2008). Surface sites for engineering allosteric control in proteins. Science *322*, 438–442.

Saraf, M.C., Horswill, A.R., Benkovic, S.J., and Maranas, C.D. (2004). FamClash: a method for ranking the activity of engineered enzymes. Proc. Natl. Acad. Sci. USA *101*, 4142–4147.

A



B

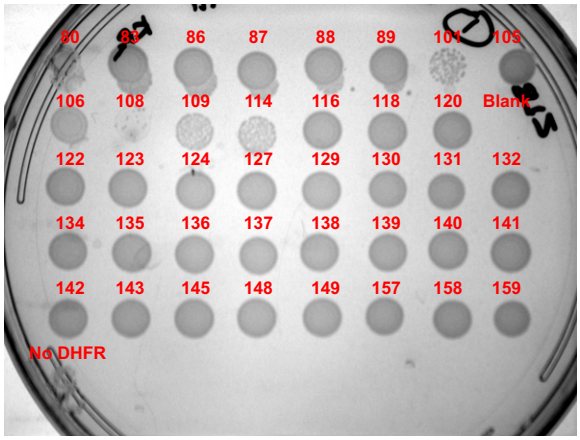


Figure S1. Auxotroph Rescue by the Library of DHFR Point Mutants and DHFR-LOV2 Chimeras, Related to Figure 3

(A and B) Under rich media conditions (LB+thymidine) the auxotroph cannot grow if DHFR activity is completely abrogated (No DHFR) but complementation is achieved for variants with substantially reduced activity (DHFR D27N). Here, each variant was grown overnight in liquid culture, diluted to a starting OD600 of 0.1, and spotted onto plates. The plates were photographed after 19 hr at 30°C. All variants are able to complement growth to some degree with the exception of DL17, DL23, and DL108. These data support Figure 3 in the main text.

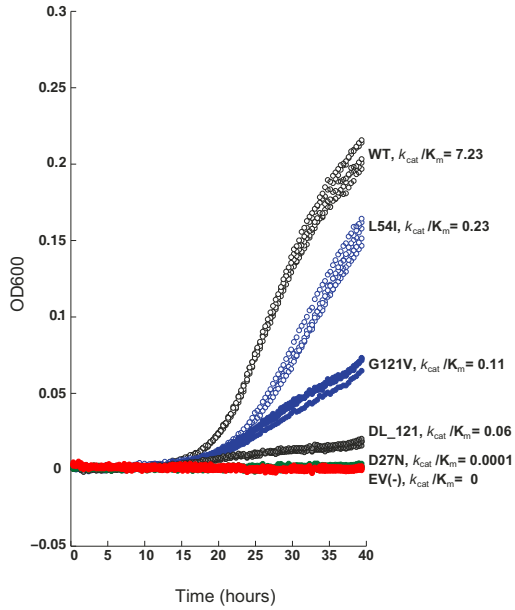


Figure S2. Auxotroph Growth Rate Varies with DHFR Catalytic Activity, Related to Figure 3

Growth rates for a set of five DHFR variants in ER2566 *ΔfolA ΔthyA* were monitored over a 40 hr period by optical density at 600 nm in a plate reader. These growth curves were conducted in minimal media at 30°C. EV(-) (shown in red) indicates empty vector, a negative control lacking both the DHFR (*folA*) and Thymidylate synthase (*thyA*) genes. Growth rate is seen to vary monotonically with catalytic activity.

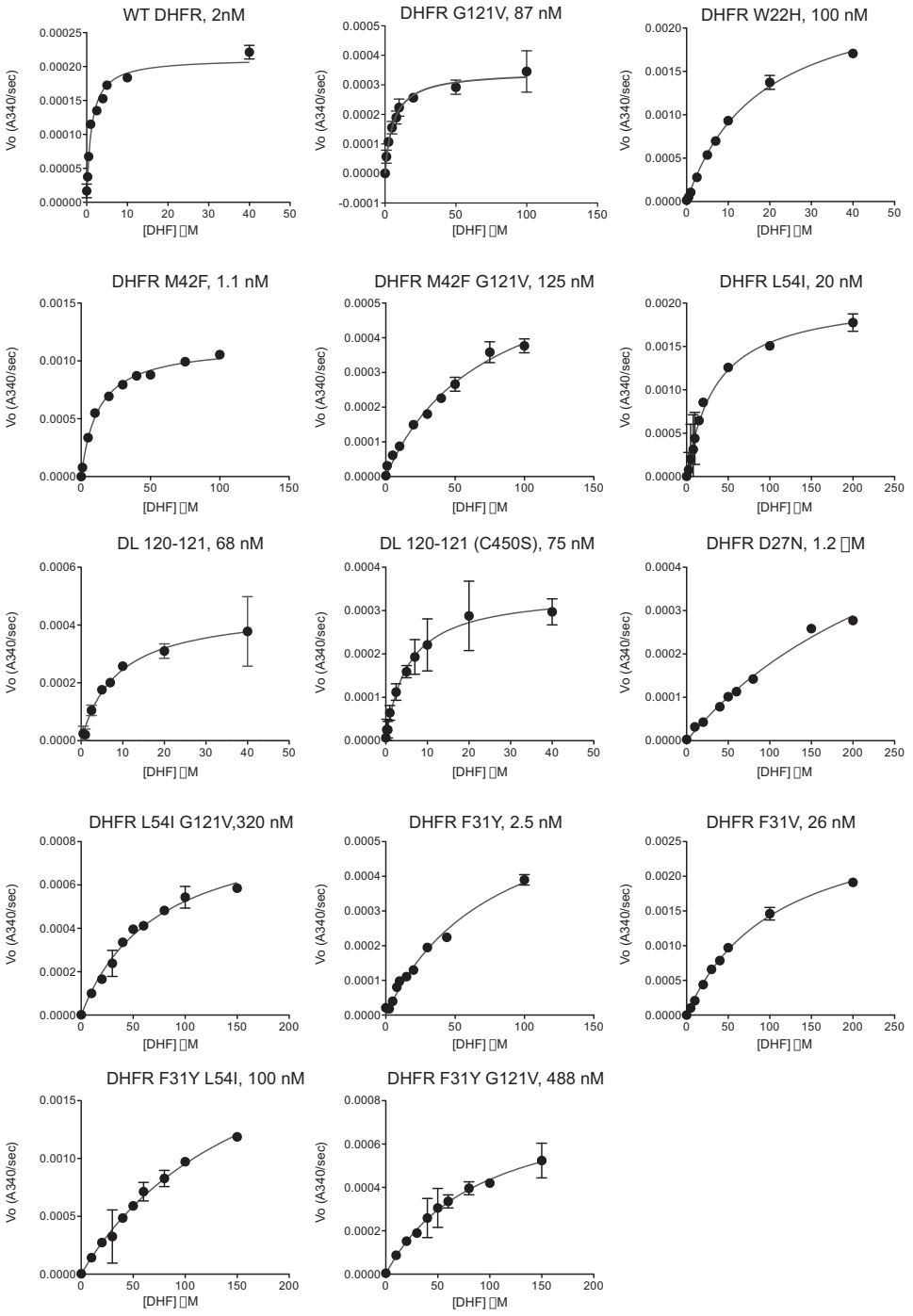


Figure S3. Michaelis Menten Curves for Steady-State Kinetics of the DHFR Point Mutants, Related to Figure 3 and Table S3

Initial velocity versus substrate (dihydrofolate) concentration is plotted for a set of 14 DHFR variants. Each experiment was performed in triplicate, and the final concentration of enzyme used in the reaction is indicated.

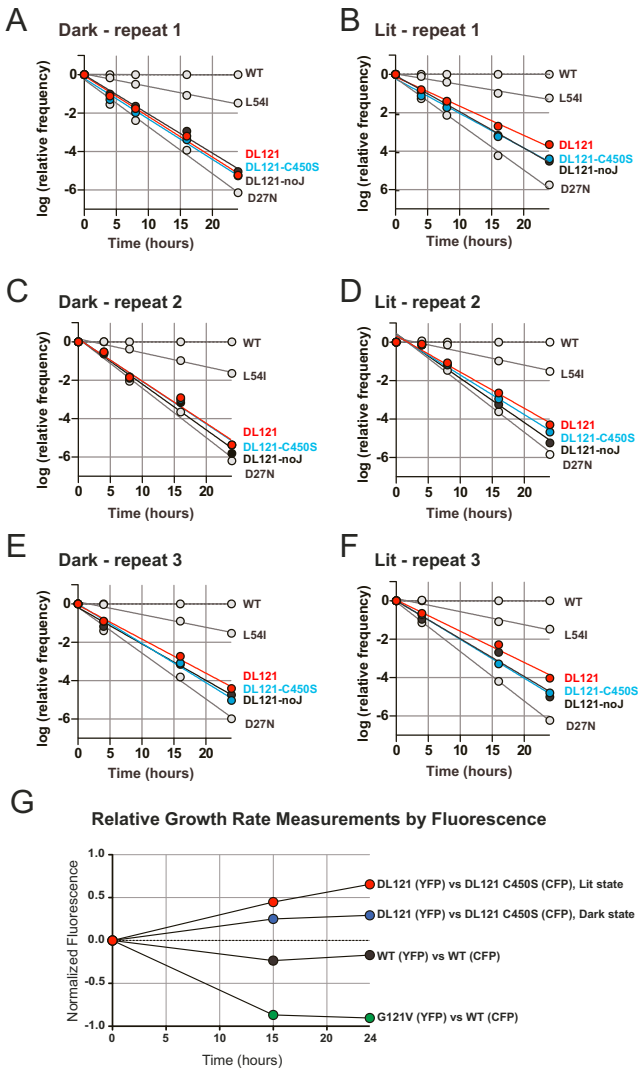


Figure S4. Comparison of Lit and Dark Growth Rates for the DHFR/LOV2 120-121 Chimera (DL121), Related to Figure 4

Relative growth curves reconstructed from sequencing data (A–F) are provided for a subset of DHFR variants grown under dark (A, C, E) and light (B, D, F) conditions. In panel G), relative growth rates measured by pairwise competition of fluorescently labeled strains are shown. For this experiment, one strain was labeled with yellow fluorescent protein (Venus) and the other with cyan fluorescent protein (Cerulean). The relative populations of cells were then quantified by flow cytometry at each time point shown, and normalized according to the equation:

$$f(t) = \log\left(\frac{N_t^{Mut}/N_t^{WT}}{N_{t=0}^{Mut}/N_{t=0}^{WT}}\right)$$

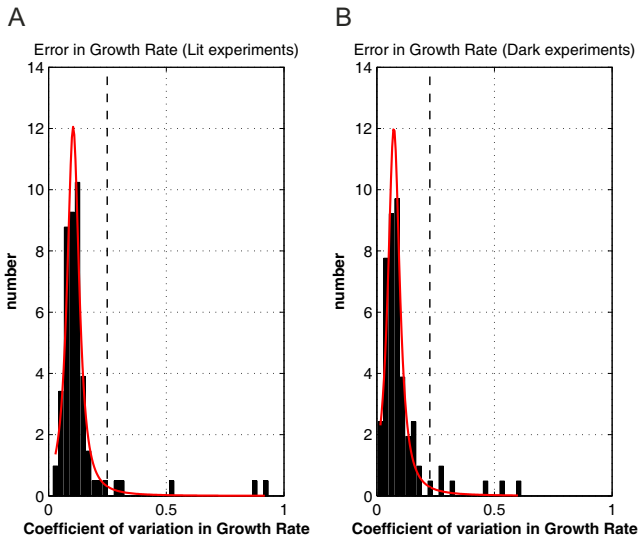


Figure S5. Elimination of Chimeras with Unreliable Growth Rates, Related to Figure 5

(A and B) Lit and dark growth rates were measured for all DHFR variants in each of the three independent experiments. Mean growth rates (μ) and standard deviations (σ) were determined, and used to calculate a coefficient of variation for each DHFR mutant (σ/μ, black bars of histogram). The histogram indicates that the majority of chimeras show coefficients of variation that are less than 0.2, while a few chimeras comprise a tail that indicates much large variation in growth rates. This histogram was fit to a t-location scale distribution (red line), and a p value of 0.07 (shown as dotted line) was used as an empirical cutoff to eliminate the obvious outliers in growth rate.

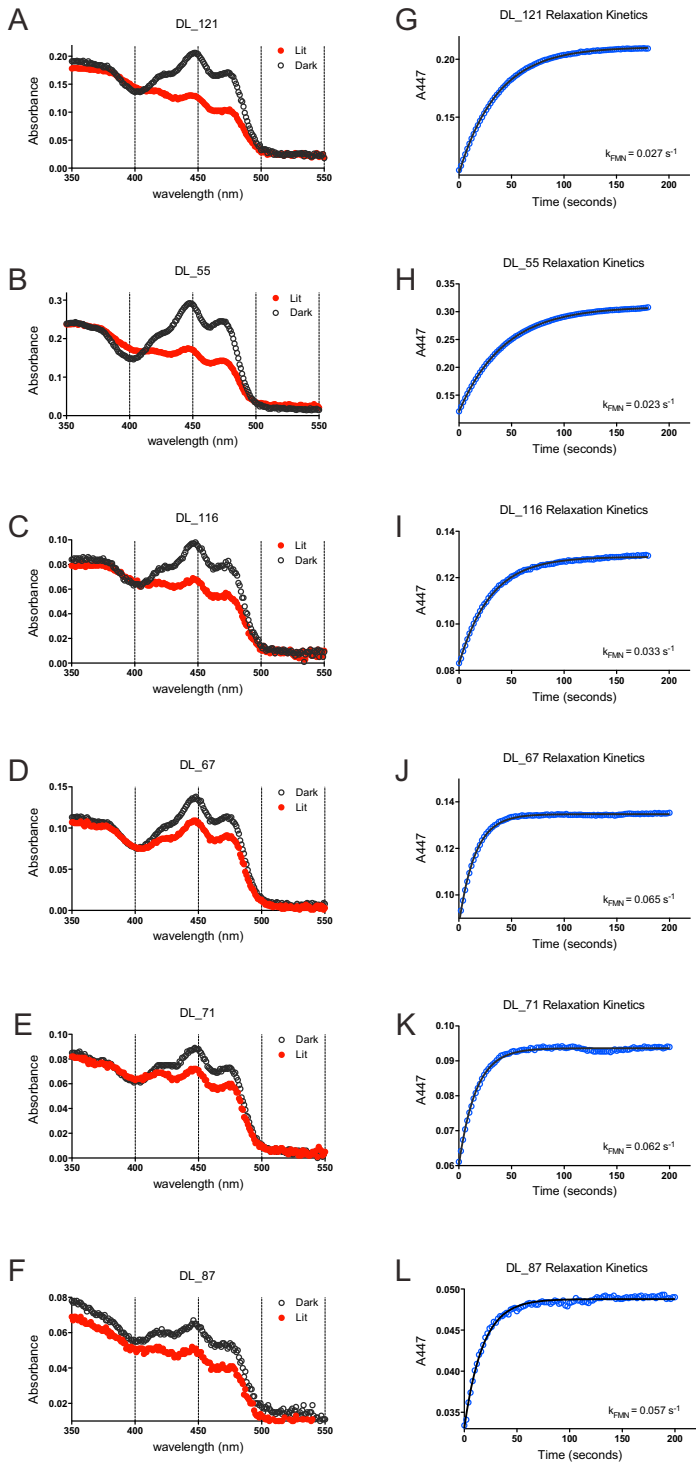


Figure S6. LOV2 Spectra and Relaxation Kinetics for a Sampling of DHFR-LOV2 Fusions, Related to Figure 5

These data include one light dependent DHFR-LOV2 fusion (DL121, A and G) as well as five randomly selected non-light dependent DHFR-LOV2 fusions (B–F, H–L). Spectra were collected from 350 nm to 550 nm under dark-equilibrated conditions (black circles) or following a 5 min excitation with blue light (480 nm, red circles) (A–F). In all cases, the LOV2 domains show a characteristic dark state spectra with a peak at 447 nm corresponding to the non-covalently bound FMN chromophore, and under lit conditions undergo a spectral shift to a 390 nm absorbing species due to formation of a covalent FMN-thiol adduct. Relaxation kinetics were monitored at 447 nm for 3 minutes following a 5 min excitation with blue light (G–L). All of the assayed LOV2 domains display exponential and reversible relaxation to the dark non-covalently bound state, with time constants near that of the isolated domain (.022 s^{-1}).

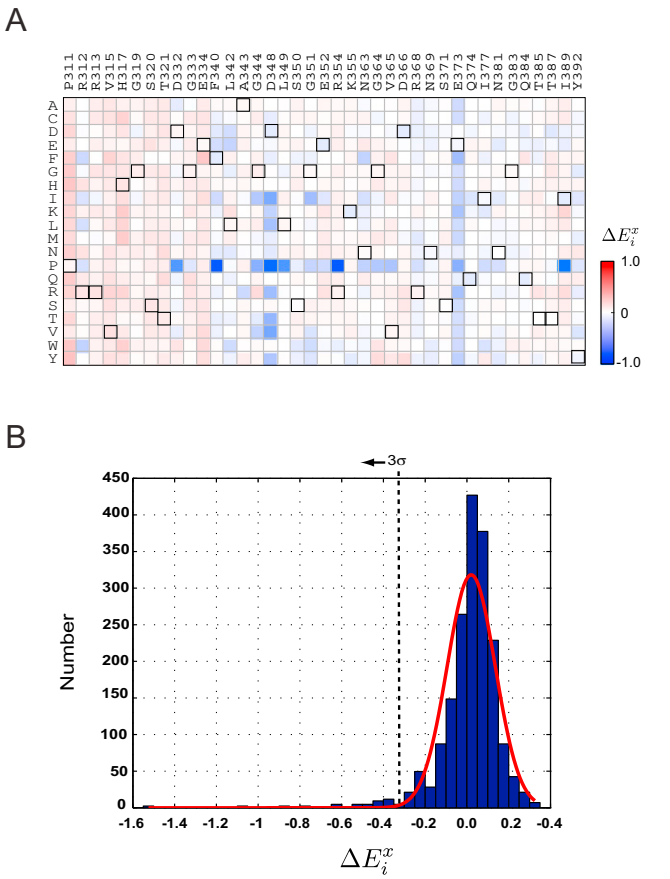


Figure S7. Saturation Mutagenesis of the PSD95pdz3 Domain Surface, Related to Figure 7

(A) Each surface accessible position *i* (columns) was mutated to all twenty amino acids *x* (rows) and the effect of mutation on function (ligand binding) relative to wild-type (ΔEix) was assessed by bacterial-two-hybrid assay. The data are shown colorimetrically, with blue representing loss of function and red representing gain of function, and the wild-type amino acid at each position is indicated by the bold outline. (B) A histogram of the data presented in the matrix in (A) finds that the mutations have a mean effect near zero, and that 11 positions have significant functional effects (>3σ).

p = .005 14% of DHFR	p=.008 23% of DHFR		p=.010 25% of DHFR		p=.015 31% of DHFR	
15	11	81	3	63	3	53
21	15	94	11	64	5	54
27	21	100	15	77	11	55
28	22	111	21	81	13	56
31	27	113	22	90	15	57
32	28	121	27	94	18	59
35	31	122	28	100	21	63
37	32	125	31	111	22	64
42	35	126	32	113	25	77
44	37	133	35	121	27	81
51	39	153	37	122	28	90
54	40		39	125	31	94
55	42		40	126	32	95
57	44		42	133	35	99
59	47		44	153	37	100
63	50		47		39	107
77	51		49		40	111
81	54		50		42	113
94	55		51		44	121
113	56		52		45	122
121	57		54		47	125
125	59		55		49	126
133	63		56		50	133
	64		57		51	153
	77		59		52	

Table S1. Related to Figures 1,2, 5, and 6

Sector composition at four significance cutoffs based on fitting the top significant eigenmodes of the SCA matrix to the Student's t-distribution (see Methods), and choosing probability density cutoffs as indicated. Residue numbering is from *E. coli* DHFR (PDB ID: 1RX2). The sector as defined at p=.015 is shown in Figures 1, 2, 5, and 6 in the main text.

	sector cutoff: p = .005	p=.008	p=.010	p=0.015
p-value	.0071	.0135	.0225	.0059

Table S2. Related to Figure 1

Fisher Exact Test p-values, testing the null hypothesis that those residues undergoing millisecond conformational exchange (as determined by NMR) and sector residues are uncorrelated. Over a range of significance cutoffs for sector definition (as described in table S1 and Methods), we find p-values < 0.023. Thus sector positions and the network of residues undergoing slow time scale dynamics are correlated.

Mutant	k_{cat}	(+/-)	K_m	(+/-)	k_{cat}/K_m	(+/- error)
WT	7.95	0.38	1.1	0.2	7.231	1.423
DL 121	0.51	0.02	8.6	1.0	0.060	0.007
DL 121 C450S	0.34	0.01	5.3	0.5	0.065	0.006
G121V	0.30	0.01	6.1	0.6	0.050	0.005
L54I	7.88	0.28	35.0	3.4	0.225	0.023
D27N	0.05	0.01	330.0	78.9	0.0002	0.0004
W22H	1.89	0.06	18.0	1.2	0.105	0.008
M42F	79.20	1.72	13.0	1.0	6.092	0.478
F31Y	20.61	2.12	80.0	14.0	0.258	0.052
F31V	8.65	0.29	108.0	6.8	0.080	0.006
L54I/G121V	0.22	0.01	73.0	10.0	0.003	0.0004
F31Y/G121V	0.13	0.01	90.6	7.4	0.001	0.0001
F31Y/L54I	1.94	0.16	168.3	21.4	0.012	0.002
M42F/G121V	0.40	0.04	71.8	13.2	0.006	0.001

Table S3. Related to Figure 3

Steady state kinetic parameters for DHFR point mutants and two DHFR/LOV2 fusions (DHFR/LOV2 121 and DHFR/LOV2 121 C450S).

light dependent positions: 2σ (14 total)	2.2σ (11 total)	2.4σ (8 total)	2.6σ (6 total)	2.8σ (5 total)
16	16	33	48	48
33	33	48	73	73
36	48	73	80	121
48	73	80	121	127
50	80	83	127	134
73	83	121	134	
80	114	127		
83	121	134		
114	127			
121	132			
127	134			
131				
132				
134				

Table S4. Related to Figures 5 and 6

A list of light dependent positions at four significance cutoffs computed as shown in Fig. 5A. *E. coli* DHFR numbering is used (PDB 1RX2). The significance cutoffs derive from the distribution of lit – dark growth rate differences for non-light dependent DHFR mutants, and are chosen to span a range of cutoffs in the tail of the light-dependent distribution.

	Cutoff for sector definition:	0.005	0.008	0.010	0.015
	# of surface accessible (measured) positions contacting sector:	35	40	43	45
Cutoff for light dependency: 2σ (14 pos.)	# of light dependent positions contacting sector:	14	14	14	14
	p-value:	.0001	.001	.003	.007
2.2σ (11 pos.)	# contacting sector:	11	11	11	11
	p-value:	.001	.006	.013	.024
2.4σ (8 pos.)	# contacting sector:	8	8	8	8
	p-value:	.007	.026	.050	.073
2.6σ (6 pos.)	# contacting sector:	6	6	6	6
	p-value:	.030	.070	.100	.150
2.8σ (5 pos.)	# contacting sector:	5	5	5	5
	p-value:	.050	.110	.160	.200

Table S5. Related to Figure 5

Fisher Exact Test p-values for the null hypothesis that sector connection and light-dependence are independent properties for all 61 measured surface sites. Over a range of cutoffs for light-dependency and sector definition, the null hypothesis is rejected at a confidence level of 0.05 or better. Further, all light-dependent insertion sites remain sector connected at all cutoffs for sector definition, indicating that light dependence preferentially emerges at surface sites connected to sector positions with the strongest correlation signals. Thus, sector connection is significantly correlated to the emergence of light-dependence.

	Repeat 1	Repeat 2			Repeat 3	
	Lane 1	Lane 1	Lane 2	Lane 3	Lane 1	Lane 2
#of reads:	18,651,860	30,653,595	28,775,169	20,705,715	31,414,469	22,208,489
# of reads post filter:	9,963,009	28,263,744	18,774,987	16,756,941	14,112,059	15,436,165
fraction kept:	53%	92%	65%	81%	45%	69%
total # of reads for repeat:	9,963,009	63,795,672			29,548,224	

Table S6. Related to Experimental Procedures

Number of sequencing reads for individual experimental repeats.